Journal of Experimental Psychology:
Human Perception and Performance
2000, Vol. 26, No. 2, 634–647

Copyright 2000 by the American Psychological Association, Inc.
0096-1523/00/$5.00 DOI: 10.1037//0096-1523.26.2.634

Motor Activation From Visible Speech: Evidence From Stimulus Response Compatibility

Dirk Kerzel and Harold Bekkering
Max Planck Institute for Psychological Research

In speech perception, phonetic information can be acquired optically as well as acoustically. The motor theory of speech perception holds that motor control structures are involved in the processing of visible speech, whereas perceptual accounts do not make this assumption. Motor involvement in speech perception was examined by showing participants response-irrelevant movies of a mouth articulating /bʌ/ or /dʌ/ and asking them to verbally respond with either the same or a different syllable. The letters "Ba" and "Da" appeared on the speaker's mouth to indicate which response was to be performed. A reliable interference effect was observed. In subsequent experiments, perceptual interference was ruled out by using response-unrelated imperative stimuli and by preexposing the relevant stimulus information. Further, it was demonstrated that simple directional features (opening and closing) do not account for the effect. Rather, the present study provides evidence for the view that visible speech is processed up to a late, response-related processing stage, as predicted by the motor theory of speech perception.

Much research has demonstrated listeners' ability to recover phonetic information from visible speech. Listeners in a noisy environment have been shown to integrate visual and acoustic speech, resulting in improved intelligibility of the impoverished auditory signal (e.g., Sumby & Pollack, 1954). Perhaps the most compelling example of visual speech perception is the *McGurk effect* (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976), in which listeners integrate visual and acoustic cues in a common phonetic percept. McGurk and MacDonald showed participants a movie of a speaker's head in which repeated utterances of the syllables /ba/ or /ga/ had been dubbed onto lip movements for /ba/ or /ga/. When the lip movements did not correspond with the auditory syllable, the two sources of information were fused into the auditory percept /da/ for a visible /ga/ or /bga/ for a visible /ba/. However, if presented only in the auditory modality, the syllables were unambiguously perceived as /ba/ or /ga/. In another demonstration of the effect, the auditory syllable /ba/ was presented simultaneously with a speaker's lips producing the syllables /be/, /ve/, and /de/ (Liberman & Mattingly, 1985). Listeners' judgments of the speech event followed the place of articulation specified by the lip movements; that is, observers reported hearing the syllables /ba/, /va/, and /da/. Even when participants were told to selectively attend to the visual or auditory information, the "ignored" channel strongly affected judgments (Massaro, 1987).

Three theories attempt to explain visual speech perception: the motor theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985), the fuzzy logical model of perception (FLMP; Massaro, 1987), and the direct-realist theory of speech perception (Fowler, 1986). For our purposes, the important distinction among the three theories is that the former theory makes use of motor control structures to explain speech perception, whereas the latter two approaches are purely perceptual theories. In the following sections, we describe the three approaches in more detail. We then report four experiments in an interference paradigm that provided a direct test of the assumption made by the motor theory that a perception–production link is involved in the processing of (visible) speech.

The Motor Theory of Speech Perception

The revised version of the motor theory (Liberman & Mattingly, 1985) claims that the objects of speech perception are the speaker's intended phonetic gestures. Speech gestures are represented in the brain as motor control structures that do not have an invariant manifestation in the acoustic signal or in the observable articulatory movement. The motor control structure for a gesture configures the articulators to produce phonetic gestures such as "bilabial constriction" and "velum lowering." Thus, the same entities that are used in speech perception also command movements of the articulators in speech production. Liberman and Mattingly claimed that speech perception and production are realized by a specialized module, an "innate vocal tract synthesizer." After the listener has formed an initial hypothesis about what gestures may be contained in an acoustic speech signal, the module tests the hypothesis in a process of "analysis by

Dirk Kerzel and Harold Bekkering, Department of Cognition and Action, Max Planck Institute for Psychological Research, Munich, Germany.

We thank Bernhard Hommel, Dominic Massaro, Lawrence Rosenblum, and Douglas Whalen for valuable suggestions and Susanne Schorpp for helping to collect the data.

Correspondence concerning this article should be addressed to Dirk Kerzel, Department of Cognition and Action, Max Planck Institute for Psychological Research, Amalienstrasse 33, 80799 Munich, Germany. Electronic mail may be sent to kerzel@mpipf-muenchen.mpg.de.

synthesis." Here, the speech module translates the acoustic signal into its invariant gestural representation. Because some form of synthesis requiring motor commands is necessary to decipher the speech signal, speech perception and speech production are intimately linked in the motor theory of speech perception.

A primary objective in the development of the motor theory of speech perception was to explain invariance in the perception of stop consonants. For instance, formant transitions associated with the consonant /d/ are vastly different depending on which vowel follows the consonant. The formant transition rises if followed by an /i/ and falls if followed by a /u/ (Liberman et al., 1967). However, despite the contextual variation of cues, listeners perceive the same phoneme without difficulty. In the motor theory of speech perception, perceptual invariance in the face of context-dependent variation is possible because speech perception is based on invariant phonetic gestures, not on contextually varying acoustic cues. To the contrary, contextual variation provides information about articulation that is helpful in perceiving the gesture.

The perception of visible speech follows the general principle of the motor theory which states that speech perception is the extraction of information about intended vocal tract activity. According to the theory, both visual and acoustic sources of information may provide information about the speech gesture. Importantly, visual information is able to access the speech module where phonetic information is stored in a gestural format. Therefore, the same analysis-by-synthesis process that is used in dealing with acoustic input should be active when visual information about articulatory activity is supplied. In other words, visual information about vocal tract activity should lead to an activation of motor control structures that are used in speech production.

Perceptual Approaches

The Direct-Realist Theory

Similar to the motor theory of speech perception, the direct-realist theory (Fowler, 1986; Fowler & Rosenblum, 1991) assumes that the basic entities of speech perception are phonetically significant gestures of the vocal tract. That is, listeners to speech recover information about the articulatory activities of the vocal tract from various sources of information. However, unlike the motor theory, the direct-realist theory does not posit that perception of speech is achieved by the same units that also produce speech. Rather, recovery of the basic units of speech perception relies on purely perceptual processes. Consistent with Gibson's (1966, 1979) general theory of perception, the direct-realist approach states that activities of the vocal tract lawfully relate to patternings of the informational medium and thereby provide information about the distal event. Thus, when the ear of the listener is stimulated by the acoustic medium, the structure is imparted and the listener perceives the speaker's gestures, not the pattern of acoustic energy in the speech signal. Similarly, the structure of the optic medium may provide information about a distal event. In the case of the McGurk effect, for example, the influence of visual information on the perception of an auditory signal arises because visual and acoustic information apparently convincingly specify the same speech event. Because a speech event, not the patternings of the two media, is perceived, a unified percept results.

The Fuzzy Logical Model of Perception

Massaro's (1987) FLMP assumes that prototypes of syllables stored in memory are the basic units of spoken utterances. Both acoustic and visual cues may be used as specifications of prototypes. In the model, speech perception proceeds through three operations. In the first operation, features are evaluated in terms of prototypes of syllables in memory. This operation makes available the degree to which the features in the speech signal match the featural values associated with a prototype. In the second operation, the features corresponding to each prototype are integrated and the degree of correspondence to the prototype is determined. The third operation achieves pattern classification. In this stage, the relative goodness of match of each prototype is evaluated, and the prototype with the best match is selected. Importantly, the FLMP proposes that both visual and auditory cues are associated with prototypes. Presumably, repeated experience of a conjunction of visual and acoustic cues establishes the prototypes in memory.

When the observer is confronted with partially conflicting visual and auditory information, as in the McGurk effect, the prototype that best matches the collection of features is selected. Therefore, speech perception is explained by a best-match procedure that maps visual-acoustic input onto a memory representation that deviates the least from the sensory input.

In summary, the motor theory and the direct-realist theory share the assumption that speech gestures are the basic perceptual units in speech processing. The two theories differ with regard to the processes that are supposed to recover gestures from the speech signal. Whereas the motor theory attributes recovery to the analysis-by-synthesis process, which involves motor control structures used in speech production, the direct-realist theory assumes that lawful relations between the structure of the informational medium and gesture are being detected. Similar to the direct-realist theory, the FLMP does not invoke motor structures to explain speech perception. Rather, a process of matching perceptual input to units stored in memory accomplishes the recognition of syllables. No reference to speech production or a specialized speech module is made.

Empirical Evidence

At least two different assumptions that are used to explain the perception of visible speech may be tested. First, a critical test between the FLMP on the one hand and the motor theory and the direct-realist theory on the other concerns the basic units of speech perception, that is, gestures versus syllables. Fowler and Dekle (1991) conducted experiments designed to distinguish the accounts offered by the FLMP and the direct-realist theory. According to the FLMP, the effect of visual influences on speech perception derives from a learned association between visual and acoustic cues for a particular syllable. In contrast, the direct-realist theory claims that the effect arises because observers perceive a unique distal speech event causing the structuring of light and sound. Fowler and Dekle examined the cross-modal influence on speech perception from two new sources, written letters and manually felt, mouthed syllables. Letters are thought to be associated in memory with the respective speech sound but have no lawful relation to it, whereas the opposite is true for haptically experienced syllables. Consistent with the direct-realist theory, pairing mouthed syllables with acoustic syllables influenced reports of the heard syllables. Only a small and unreliable effect was observed for pairings of letters and acoustic syllables (however, see Massaro, Cohen, & Thompson, 1988). Therefore, the authors suggested that the McGurk effect arises not from association in memory but from the bimodal specification of the same distal event, the phonetic gesture (see Massaro, 1998, for a different interpretation). Furthermore, Fowler and Dekle doubted that their results were compatible with the existence of a vocal-tract synthesizer, as hypothesized in the motor theory of speech perception. Because acoustic and visual information about vocal-tract gestures is readily available in the environment, exploitation of these sources might have adaptive significance. In contrast, haptic information is rarely available, so there is no reason to believe that the vocal-tract synthesizer should have anticipated its occurrence.

A second testable distinction among the three theories is the hypothesis of a perception–production link. Perception–action couplings are explicitly assumed by the motor theory but not by the direct-realist theory or the FLMP. If speech perception is mediated by the motor control structures used in speech production, then some evidence of their links should appear in reaction time (RT) data. In particular, perceiving auditory syllables should have an effect on the speed of producing vocal responses with similar features: The perception of an auditory syllable should activate specific motor control structures associated with the syllable's phonetic features. Because of the preactivation of motor commands, the production of vocal responses with the same features should be facilitated, the production of vocal responses with different features should be inhibited, or both. In other words, if perception and production of speech rely on the same structures, then perceiving certain features should prime production of those features.

Support for this idea comes from two studies. First, Porter and Lubker (1980) found that RTs were short when listeners were asked to repeat a syllable they heard. Although these shadowing responses constituted a choice task (i.e., different stimulus syllables required distinct responses), latencies were similar to those observed in a simple response task in which only one response was produced. Presumably, the priming of the vocal response via a direct linkage between speech analysis and speech production was responsible for the speed of shadowing responses.

Second, Gordon and Meyer (1984) demonstrated that the overlap between features of stimulus and response syllables produces faster RTs. Their stimulus material consisted of syllable pairs constructed from /pʌ/, /bʌ/, /tʌ/, and /dʌ/, which permitted orthogonal variation of place of articulation and voicing. The experimental task was to produce the second (response) syllable of a syllable pair as rapidly as possible in response to the first (stimulus) syllable. Consistent with the motor theory of speech perception, responses to auditory stimulus syllables were faster when the response syllable contained a consonant that shared the voicing feature of the consonant in the stimulus syllable. However, no effect of matched place of articulation was observed. Thus, response priming seemed to occur in the processing of the voicing feature but not in the processing of the place feature. This finding is surprising given that one of the major reasons for developing the motor theory of speech perception was to account for the invariant perception of place of articulation despite the context sensitivity of acoustic cues to it. Also, the lack of priming for place of articulation observed in the auditory modality favors a purely perceptual account of the processing of visible speech: Visible speech provides information about the place of articulation but not about voicing (Binnie, Montgomery, & Jackson, 1974; Green & Miller, 1985). Thus, given that motor priming was not found for the acoustic place dimension, the visible place feature may also fail to induce motor activation. Furthermore, the hypothesis of purely perceptual interactions in the processing of visual speech is corroborated by evidence for integral perceptual processing of the visual place and auditory voicing information in the McGurk effect (Green & Kuhl, 1991).

Purpose of the Study

The main motivation for the present study was to examine in more detail whether evidence for perception–action links can be obtained for visible speech. The question we tried to answer was whether watching a speaker's mouth movements would produce activity in response-related stages. Activation of motor codes during the perception of speech stimuli is predicted by the motor theory of speech perception but not by the direct-realist theory or the FLMP. The existing data question the account of visual speech perception offered by the motor theory in two ways: First, Fowler and Dekle (1991) suggested that the integration of haptic and acoustic information analogous to the integration of visual and acoustic information in the McGurk effect for visually presented speech gestures is unlikely to occur in a "vocal tract synthesizer." Therefore, the motor theory fails to explain the observed cross-modal influence of haptic information on the processing of acoustic information. Second, motor priming was not observed for place of articulation (Gordon & Meyer, 1984), which is the critical feature in visible speech perception.

However, both lines of evidence are inconclusive: Massaro (1998, pp. 352–355) argued against the direct-realist interpretation of Fowler and Dekle's haptic McGurk effect that listeners may simply relate information in the haptic modality to the visual or auditory modality. Also, Massaro showed that the FLMP may fit the data quite well. Furthermore, Gordon and Meyer's (1984) priming effect for voicing and the lack of a priming effect for place of articulation may result from the greater salience of the voicing feature than the place feature (Proctor, Dutta, Kelly, & Weeks, 1994). In addition, their interpretation of RT advantages in terms of response priming may be in agreement with models that assume a direct, automatic route from stimulus to response (dual-route models; see next section), but it is clearly at odds with translational models (Proctor et al., 1994; Proctor & Reeve, 1991). Translational models claim that when features of stimulus and response correspond, the stimulus–response (S-R) mapping can be characterized by rules that allow for easy S-R translation and fast RTs. Thus, faster RTs in conditions in which the place feature corresponded in stimulus and response syllables may be attributable to easy S-R translation, not to response priming. A similar objection may be raised against a priming interpretation of Porter and Lubker's (1980) shadowing responses: Repeating an acoustic syllable can occur as quickly as it does because the translation from stimulus to response is maximally easy, not because the motor response has been activated by the acoustic signal.

Therefore, in the present study we tested the hypothesis that motor activity is involved in the processing of (visible) speech. If the influence of visible speech on the perception of acoustic stimuli is attributable to the use of common motor commands, then seeing the articulatory movements of a speaker's lips should influence the speed of speech production. Because a visible speech gesture is supposed to be processed by activation of the corresponding motor commands ("analysis by synthesis"), production of the same gesture should be faster than production of a different gesture. However, it is necessary to rule out RT advantages arising from easy or difficult S-R translation and to show that differences in RT are not exclusively attributable to perceptual conflict between two stimulus dimensions; in other words, we need to ensure that the locus of the facilitation or interference occurs at a late, motor stage of processing. Research on S-R compatibility provides the required tools for distinguishing perceptual and response-related interference.

Stimulus–Response Compatibility

The most widely accepted models of stimulus–response compatibility (SRC) are dual-route models (e.g., De Jong, Liang, & Lauber, 1994; Hommel, 1993; Kornblum, Hasbroucq, & Osman, 1990). Generally, the idea is that SRC effects are attributable to a competition between automatic response activation and voluntary S-R translation. Perhaps the most general model is the Kornblum et al. (1990) dimensional overlap model, which classifies S-R sets according to their perceptual, structural, or conceptual similarity (i.e., their dimensional overlap). Two stimulus dimensions, one relevant for response execution and the other irrelevant, were considered by Kornblum et al. The relevant dimension correlates perfectly with the response and is translated into the response. The irrelevant dimension is completely unrelated to the response. According to the model, a stimulus dimension that shares features with the response may activate the corresponding response automatically. When the features of stimulus and response are congruent, performance benefits result because the correct response is activated. In incongruent trials, activation of the wrong response produces a cost.

For instance, Simon, Hinrichs, and Craft (1970) instructed participants to perform left- or right-hand keypresses in response to verbal instructions that were presented randomly to the left or right ear. When the stimulus occurred on the same side as the keypress, responses were faster than when it occurred on the opposite side; this advantage has come to be known as the *Simon effect* (for an overview, see Lu & Proctor, 1995). According to dual-route models of SRC, the verbal stimulus is voluntarily translated into the correct response. At the same time, the irrelevant location of the stimulation automatically activates the corresponding response, such that RTs are shorter when the irrelevant stimulus location and the response location are congruent. However, compatibility effects are not restricted to interference in response-related stages. In contrast to SRC, stimulus–stimulus compatibility (SSC) is due to interference at the level of stimulus identification; that is, the observed benefits and costs are generated at a stage prior to response selection (Kornblum, 1994). For instance, discrimination of dim or bright light is slowed when the visual stimulus is accompanied by irrelevant low- or high-pitched tones (Marks, 1987). The correspondence effect is assumed to be due to the dimensional overlap of the relevant and irrelevant stimulus sets and is therefore localized at the level of stimulus identification.

Predictions

Casting the above-mentioned accounts of speech perception into the framework of theories of stimulus and response coding leads to two conflicting sets of predictions. The perceptual accounts predict that visual speech is processed at the level of stimulus identification. Therefore, if one were to measure RTs to corresponding or noncorresponding stimulus dimensions in an appropriate experimental task, an SSC effect should occur. In contrast, motor accounts of speech perception localize the processing of speech at a response-related stage. In this view, motor activation is at the heart of speech perception. Therefore, an SRC effect for speech stimuli and verbal responses should obtain.

In the following experiments, we demonstrate that visually presented lip movements are processed up to a late, response-related stage. Our stimuli consisted of visually presented, response-irrelevant lip movements and response-relevant symbolic stimuli. In Experiment 1, we found an interference effect between lip movements and letters in a Stroop-like setup. This result supports both perceptual and response-related accounts because the two stimulus sets and the S-R sets show dimensional overlap. In Experiment 2, we eliminated the similarity between the two stimulus sets as in a Simon-type task but still obtained an interference effect. In

Experiment 3, the interference effect persisted even when response selection was completed before the irrelevant dimension was presented, so a perceptual conflict can be ruled out. Finally, in Experiment 4 we show that the effect cannot be reduced to simple spatial compatibility.

Experiment 1

The purpose of the first experiment was to establish a basic interference effect that could be localized at either a response-related or a perceptual processing stage. In subsequent experiments, we determined the exact locus of the interference. Participants were shown a speaker's mouth articulating either /bʌ/ or /dʌ/. While the mouth was moving, the written syllables "Ba" and "Da" were briefly presented on the mouth. Participants were instructed to respond to the letters by producing either /bʌ/ or /dʌ/ and to ignore the lip movements of the mouth. If—as claimed by the motor theory—the perception of articulatory movements and the production of speech pass through a shared processing stage, then both response-related and stimulus-related interference would be expected. In the motor theory, speech perception is motoric from the start, such that conflict may arise at a perceptual or response-related stage. In contrast, perceptual accounts predict that the simultaneous presentation of two kinds of linguistic input leads exclusively to perceptual conflict. Thus, both perceptual and motor accounts predict interference but diverge on where it is localized. In fact, the exact locus of interference may be hard to determine in this experimental setup. As in a Stroop experiment (Stroop, 1935), in which participants are required to either name the color of a color word or to read the color word, the response set overlaps with both the relevant and the irrelevant stimulus dimensions, which themselves overlap. Any interference might therefore be due either to SSC or to SRC. Consistent with the conceptual ambiguity of Stroop interference, evidence for response competition (Keele, 1972; Morton, 1969; Warren, 1972), perceptual conflict (Dunbar & MacLeod, 1984; Gumenik & Glass, 1970; Hock & Egeth, 1970; Melara & Mounts, 1993; Palef & Olson, 1975), and semantic interference (e.g., Dalrymple-Alford & Azkoul, 1972; Klein, 1964; Seymour, 1977; Stirling, 1979) has been obtained. Thus, there is no clarity about whether Stroop interference (i.e., difficulty in naming the color of a color word but no difficulty in reading a colored color word) is localized at a perceptual or a response-related stage. Therefore, both the perceptual and the motor accounts of visual speech perception predict that a reversed Stroop-like effect (i.e., difficulty in reading a syllable) can be obtained for visible speech gestures.

Method

Participants. Eight students at the Ludwig-Maximilians University of Munich were paid for their participation. All reported normal or corrected-to-normal vision and were naive as to the purpose of the experiment.

Apparatus and stimuli. The stimuli were created using a Matrox Millenium graphics adaptor controlled by a personal computer. The display had a resolution of 1,280 (H) × 1,024 (V) pixels on a 50-cm (diagonal) screen, and the refresh rate was 60 Hz. Viewing was unrestrained at a distance of 100 cm from the screen. Video recordings of a male speaker pronouncing /bʌ/ or /dʌ/ were digitized and converted into a standard picture-file format. Participants saw the lower portion of the speaker's face in a 6.62° × 4.95° window on an otherwise white screen. Only the mouth and parts of the chin and cheeks were visible. The mouth was approximately centered on the screen. In its resting configuration, it was closed and had a maximal extent of 3.26° × 1.09°. Visual presentation of speech gestures was accomplished by showing a sequence of 20 pictures at a rate of 30 Hz (667 ms). A gesture comprised the opening (10 pictures) and closing (10 pictures) of the mouth. In an animation showing a /bʌ/, the maximal vertical extent of the mouth decreased from its value at rest to a minimum of 0.23° after 133 ms (4 pictures) and then increased to a maximum of 2.58° after 333 ms (10 pictures). For sequences showing a /dʌ/, the vertical extent increased from the size at rest to a maximum of 2.86° after 333 ms (10 pictures). In both sequences, the closing movement from the maximal aperture to the rest configuration took 333 ms (10 pictures). The mouth at rest was visible during the intertrial interval.

As imperative stimuli, the printed syllables "Ba" and "Da" were presented for 100 ms in the center of the mouth. The capital letters measured 0.29° × 0.34°, and the "a" measured 0.29° × 0.29°. Participants responded by saying either /bʌ/ or /dʌ/. Responses were transmitted by a microphone attached to the participants' necks and to the computer's sound card (Soundblaster 16), which converted the analog signal into digital format at a rate of 12 kHz. The data were stored on hard disk for off-line analysis. To determine the onset of each utterance, we computed the moving average of the unsigned 16-bit samples for a 6-ms window. The values were adjusted for low background noise. To avoid selection of sounds resulting from lip, head, or hand movements as the speech onset, a search algorithm selected the largest window in which a certain threshold had been surpassed at least once and in which all sample values were higher than a second, somewhat lower, threshold. The start of this window was used to determine the speech onset. When the on-line analysis of the acoustic signal could not determine an onset, an error message appeared on the screen telling the participant that nothing had been said. In the off-line analysis, each onset was visually inspected and, if necessary, corrected. In addition, each utterance was listened to and classified as /bʌ/, /dʌ/, or unintelligible.

Design. There were 15 blocks composed of 16 trials each, which resulted from the factorial combination of visible speech gesture (/bʌ/, /dʌ/), spoken syllable (/bʌ/, /dʌ/), and stimulus onset asynchrony (SOA; 0, 167, 333, and 500 ms) between the first frame of the animated sequence (i.e., the last frame of the mouth in rest position) and the appearance of the imperative signal (i.e., the letters "Ba" or "Da"). The resulting 240 trials were administered in a single 30-min session. The first block served as practice and was not evaluated any further.

Procedure. The experiment took place in a dimly lit room. During the intertrial period, the mouth was visible in its resting position. To avoid anticipatory responses at the onset of the lip movement, we randomly varied the intertrial period between 1, 1.75, and 2.5 s. Lip movement thereby served as a warning signal for the imperative stimulus. After onset of the imperative stimulus, the sound was recorded for 2 s.

Results

Pronouncing the wrong syllable was counted as a choice error, and responses with RTs shorter than 100 ms or longer than 1,200 ms were considered anticipations and missing trials, respectively. Unintelligible utterances were considered missing. There were only a few anticipations and missing trials (0.8%). RT means are graphed in Figure 1. A three-way analysis of variance (ANOVA) (Visible Gesture × Spoken Syllable × SOA) was conducted on mean correct RTs. Responses decreased from 550 ms at the shortest SOA to 476 ms at the longest SOA, $F(3, 21) = 28.41$, $p < .0001$. Importantly, a highly significant interaction between visible gesture and spoken syllable emerged, $F(1, 7) = 45.35$, $p < .0003$, indicating that pronouncing /bʌ/ was faster when a /bʌ/ gesture was observed than when a /dʌ/ gesture was observed (491 vs. 519 ms). Conversely, producing /dʌ/ was faster when a /dʌ/ gesture was presented than when a /bʌ/ gesture was presented (502 vs. 543 ms). There was also a nonsignificant trend for spoken /bʌ/ responses, $F(1, 7) = 4.46$, $p < .0727$, and for responses in the presence of a visible /dʌ/ gesture, $F(1, 7) = 4.33$, $p < .0761$, to be faster than spoken /dʌ/ responses and responses to visible /bʌ/, respectively (504 vs. 522 ms and 510 vs. 516 ms, respectively). No other effect approached significance ($p > .13$). The proportion of errors (PE) was low (1.9%), so that many cells of the three-factorial design had values of zero. Therefore, PEs were collapsed across SOA and entered into a two-way ANOVA (Visible Gesture × Spoken Syllable). A significant interaction between visible gesture and spoken syllable emerged, $F(1, 7) = 10.29$, $p < .0149$, indicating that fewer errors were made when the visible gesture corresponded to the spoken syllable: 0.4% versus 3.5% for a spoken /bʌ/ and 0.6% versus 2.9% for a spoken /dʌ/. No other effects reached significance ($p > .6$).

Discussion

It is clear from the results that visible lip movements influenced the pronunciation of either the corresponding or the noncorresponding syllable. RTs and PEs were lower (35 ms and 2.7%, respectively) when visible speech gesture and spoken syllable corresponded than when they did not. Because the response set shows dimensional overlap with both the relevant and the irrelevant stimulus sets, which overlapped themselves, the observed interference might have been due to either SSC or SRC. If the interference effect was due to SSC, perception of the visible gestures /bʌ/ or /dʌ/ interfered with perception of the printed "Ba" or "Da." In contrast, if the interference was due to SRC, the perception of the visible gesture interfered with the pronunciation of /bʌ/ or /dʌ/. Therefore, this result would be expected from both the perceptual and motor accounts of visual speech perception. Although no test of the predictions from the two conflicting accounts can be achieved at this point, the experiment is useful in establishing that reading responses may be influenced by a visible speech gesture. The influence of irrelevant information on reading responses is referred to as the "reversed Stroop effect," and it has been notoriously difficult to obtain (M. O. Glaser & Glaser, 1982; W. R. Glaser & Düngelhoff, 1984). For instance, in a Stroop experiment, reading color words suffers little from incongruent color information unless the discrimination of the color word is decreased (Melara & Mounts, 1993). Therefore, the interference from mouth movements may result from privileged access to the response programming stage. This is what is predicted by the motor theory of speech perception; however, alternative explanations in terms of SSC cannot be ruled out.

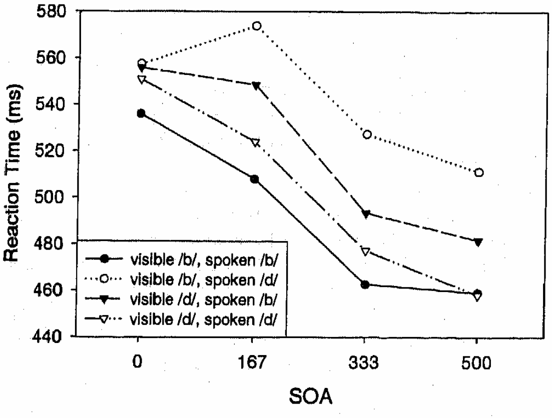

Figure 1. Mean reaction times as a function of visible gesture, spoken syllable, and stimulus onset asynchrony (SOA) in Experiment 1.

Experiment 2

Experiment 2 was designed to evaluate whether the interference effect obtained in Experiment 1 resulted from motor or perceptual conflict. To exclude the possibility of SSC, we eliminated dimensional overlap between the relevant and irrelevant dimensions by using symbols unrelated to /bʌ/ or /dʌ/ as imperative stimuli. As such, only the irrelevant stimulus set (i.e., the visible mouth movement) was similar to the response set. The relevant stimulus set, on the other hand, was dissimilar to both the irrelevant stimulus set and the response set. Consequently, any interference effects have to be attributed to S-R interactions.

The ensemble of S-R dimensions in the present experiment is akin to that in the Simon effect. In the Simon effect, spatially defined responses are faster when the irrelevant stimulus position and response location are congruent than when they are not (e.g., Simon & Small, 1969). Dual-route models of SRC (e.g., Kornblum et al., 1990) hold that the activation of the correct or incorrect response by the irrelevant stimulus dimension leads to performance benefits or costs. Consistent with this interpretation, electrophysiological studies have shown that irrelevant spatial information automatically induces response-related activation. For instance, the effects of irrelevant spatial information on lateralized readiness potentials (LRPs) were investigated in tasks that required participants to perform left- or right-hand keypresses. The LRP waveforms revealed activation of the response congruent with the irrelevant spatial position of the imperative stimulus (De Jong et al., 1994) or to the spatial information provided by a noninformative symbolic cue (Eimer, 1995). Thus, there is reason to believe that an interference effect obtained with nonoverlapping stimulus sets and overlapping S-R sets arises from direct activation of response codes by the irrelevant stimulus dimension. Consequently, such a paradigm offers an excellent opportunity for testing the predictions of perceptual and motor accounts of visual speech perception. If an interference effect is found with visible speech gestures as the irrelevant dimension and arbitrary symbols as imperative stimuli, strong evidence for a motor explanation is obtained: Any observed interference can be attributed only to a direct activation of the phonological codes generating the response. A purely perceptual explanation, however, predicts the absence of Simon-like interference because the stimuli are separate nonoverlapping dimensions that do not permit perceptual conflict or fusion.

Method

Participants. Eight students at the Ludwig-Maximilians University of Munich were paid for their participation. All reported normal or corrected-to-normal vision and were naive as to the purpose of the experiment.

Apparatus and stimuli. Apparatus and stimuli were the same as those used in Experiment 1 with the exception that the printed symbols ## and && were used as imperative stimuli. The size of the letters was 0.29° × 0.34°.

Design. The design was the same as that used in Experiment 1.

Procedure. The same procedure was used as in Experiment 1 with the following exception. Half the participants were instructed to pronounce /bʌ/ in response to ## and /dʌ/ in response to &&. The other half received the reverse mapping.

Results

Data treatment was the same as in Experiment 1. There were only a few anticipations and missing trials (0.4%). RT means are graphed in Figure 2. A three-way within-subject ANOVA (Visible Gesture × Spoken Syllable × SOA) on RTs showed that RTs decreased from 643 ms at the shortest SOA to 537 ms at the longest SOA, $F(3, 21) = 34.38$, $p < .0001$. Importantly, a significant interaction between visible gesture and spoken syllable was confirmed, $F(1, 7) = 45.16$, $p < .0003$. Pronouncing /bʌ/ was faster when a visible /bʌ/ gesture was presented than when a /dʌ/ gesture was presented (559 vs. 605 ms). Conversely, producing /dʌ/ was faster when a visible /bʌ/ gesture was presented than when a visible /dʌ/ gesture was presented (578 vs. 615 ms). No other effects reached significance ($p > .12$). Again, the PEs were low (1.3%); therefore, we collapsed them across SOA and entered them into a two-way ANOVA (Visible Gesture × Spoken Syllable). No significant effects emerged ($p > .27$). Inspection of the distribution of errors did not reveal any signs of a speed–accuracy trade-off. For a spoken /bʌ/, the PEs were approximately equal in corresponding and noncorresponding trials (1% vs. 0.8%); for a spoken /dʌ/, the PEs were somewhat lower when the corresponding syllable had to be uttered (1% vs. 2.3%).

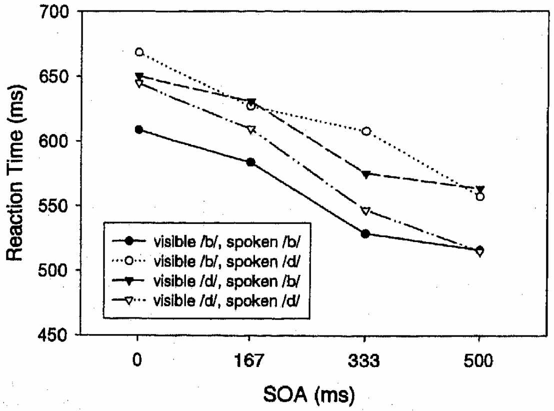

Figure 2. Mean reaction times as a function of visible gesture, spoken syllable, and stimulus onset asynchrony (SOA) in Experiment 2.

Discussion

The purpose of Experiment 2 was to determine whether the interference effect observed in Experiment 1 was due to perceptual or response-related conflict. We used arbitrary imperative (relevant) stimuli to eliminate stimulus–stimulus overlap. Consistent with a motor account of visual speech perception, responses were faster by 42 ms when irrelevant mouth movements and responses were congruent. Presumably, looking at the visual articulation of /bʌ/ and /dʌ/ caused an activation of the response codes associated with these consonant–vowel (CV) syllables. That is exactly what would be expected if speech perception relies on perceptual–motor structures, as claimed by the motor theory of speech perception: the perception of speech stimuli should lead to the activation of structures also used in speech production. In contrast, our findings are at odds with perceptual approaches to visual speech perception. If visible mouth movements are processed only up to a perceptual level, no Simon-like interference effect should be observed. Rather, interference should be restricted to the level of stimulus identification. Given that irrelevant and relevant stimulus dimensions did not overlap in the present experiment, the RT benefits and costs could not arise at a perceptual level.

Although the SRC account of the Simon effect has received much support (see Hommel, 1995, for an overview), doubts about the response-related nature of the interference have been put forth by Hasbroucq and Guiard (1991; Guiard, Hasbroucq, & Possamai, 1994; see Hommel, 1995, and O'Leary, Barber, & Simon, 1994, for rejoinders) and Stoffels, Van der Molen, and Keuss (1989). Hasbroucq and Guiard argued that in the Simon paradigm there is an inevitable confounding of the critical S-R relationship with stimulus–stimulus (S-S) congruity. Each time the values on the irrelevant spatial dimension and the response location correspond (e.g., if color is relevant and horizontal position irrelevant, a color patch presented on the left corresponds with the left response key), so do the two values of the irrelevant and relevant dimensions (e.g., the left stimulus position and green signifying left). Similarly, if values on the irrelevant stimulus dimension and response location conflict (e.g., a color patch presented on the right and the left response key), so do the two stimulus values (e.g., the right stimulus position and green signifying left responses). In other words, S-S congruity correlates perfectly with S-R congruity.

The question of how a nonspatial stimulus dimension such as pitch or color and a left–right spatial dimension may be linked in a correspondence relation was addressed by Hasbroucq and Guiard (1991) in the following way: The instruction to respond to color with a spatial response gives color a new spatial significance (e.g., green comes to signify left and red signifies right). That is, because a nonspatial response-relevant dimension is correlated with a spatial response, it acquires spatial significance. Thus, presentation of the two stimulus dimensions amounts to two simultaneous left–right messages, one stemming from the irrelevant spatial position and the other from the color.

This perceptual account of the Simon effect was disproved by Hommel (1995). The logic of his most compelling experiment was to present the irrelevant spatial stimulus information after identification of the relevant stimulus feature was completed but still before the response was actually executed. If the Simon effect is due to perceptual interference, the effect should disappear because the irrelevant location information cannot impair identification of the relevant dimension. SRC accounts of the Simon effect, however, predict a Simon effect because spatial information is assumed to automatically activate response codes. The latter result was obtained, so a perceptual account was refuted.

Experiment 3

In Experiment 3 we sought to apply the logic of Hommel's (1995) Experiment 1 to our present paradigm. Following Hasbroucq and Guiard's (1991) argument, the arbitrary symbols && and ## used in Experiment 2 may have acquired the meaning of /bʌ/ and /dʌ/ because the instruction tied them to these syllables. Therefore, in each trial the stimulus contained two perceptual signifiers for a CV syllable, allowing for perceptual conflict. To rule out perceptual interference between response-relevant and response-irrelevant stimulus dimensions, we had the presentation of the response-relevant dimension precede the presentation of the irrelevant information. Participants saw a response cue ("Ba" or "Da") indicating which response was to be performed at least 1 s before the irrelevant mouth movements were presented. Thus, they had enough time to complete identification of the response-relevant stimulus. Then, after a randomly determined interval, the mouth started moving while pronouncing either /bʌ/ or /dʌ/, which indicated that the previously cued response should be emitted. Participants were instructed to initiate the prespecified response irrespective of what the mouth seemed to pronounce. Note that the relevant feature for response initiation, motion onset, was not correlated with the response, so it should not have acquired any response-related significance in the sense described by Hasbroucq and Guiard (1991). Because identification of the response cue should have been completed before the mouth started moving, any effects of the irrelevant mouth movement can be attributed to activation of response codes independently of S-R translations. That is, no perceptual conflict is possible.

Method

Participants. Eight students at the Ludwig-Maximilians University of Munich were paid for their participation. All reported normal or corrected-to-normal vision and were naive as to the purpose of the experiment.

Apparatus and stimuli. The apparatus and stimuli were the same as those used in Experiment 1.

Design. The same design used in Experiment 1 was used here except that the SOA indicated the time between the onset of the cue indicating the response and the onset of the motion of the mouth. It was varied between 1,033, 1,767, 2,467, and 3,200 ms.

Procedure. The same procedure was used as in Experiment 1 with the following exception: After an intertrial interval of 1.5 s, the cue (syllables "Ba" or "Da") was presented for 100 ms. Participants were instructed to delay the response until the mouth started to move. In other words, lip movement was used as a "go" signal for the verbal response.

Results

Data treatment was the same as in Experiment 1. RT means are graphed in Figure 3. There were only a few anticipations and missing trials (0.5%). A three-way within-subject ANOVA (Visible Gesture × Spoken Syllable × SOA) on RTs showed that RTs decreased from 443 ms at the shortest SOA to 365 ms at the longest SOA, $F(3, 21) = 41.16, p < .0001$. Responses with a visible /bʌ/ as the go signal were faster than with a visible /dʌ/ (380 vs. 403 ms), $F(1, 7) = 44.71, p < .0003$. Importantly, a significant interaction between visible gesture and spoken syllable emerged, $F(1, 7) = 22.05, p < .0022$. When a visible /bʌ/

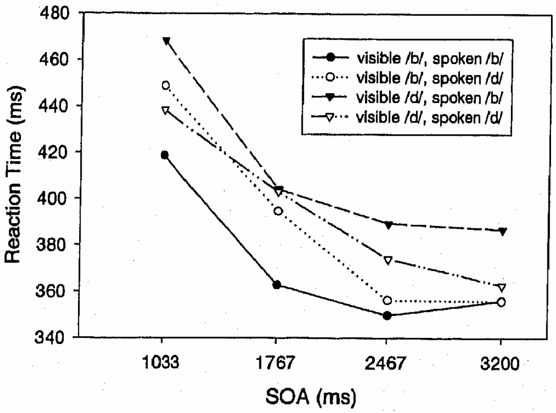

Figure 3. Mean reaction times as a function of visible gesture, spoken syllable, and stimulus onset asynchrony (SOA) in Experiment 3.

gesture was presented as the go signal, /bʌ/ responses were faster than /dʌ/ responses (372 vs. 388 ms). Conversely, producing /dʌ/ was faster than producing /bʌ/ with a /dʌ/ gesture as the go signal (394 vs. 411 ms). No other effects reached significance ($p > .3$). Again, the PEs were low (1.0%), so they were collapsed across SOA and entered into a two-way ANOVA (Visible Gesture × Spoken Syllable). No significant effects emerged ($p > .27$). Inspection of the distribution of errors did not reveal any signs of a speed–accuracy trade-off. For visible /bʌ/ and /dʌ/ gestures, PEs were approximately equal with corresponding and noncorresponding spoken syllables (1.2% vs. 0.6% and 0.8% vs. 1.2%, respectively).

Discussion

The results are clear-cut in showing that interference from visual mouth movements persisted even if perceptual conflict was ruled out. The congruency of the irrelevant speech gesture and spoken syllable resulted in a 17-ms RT advantage. Perceptual conflict was rendered unlikely because identification of the response-relevant stimulus preceded presentation of the irrelevant stimulus. This result is unexpected from a perceptual approach to the effect of irrelevant mouth movements on verbal responses and provides strong support for a response-related interpretation. Furthermore, the interference effect was not modified by the SOA, which provides further evidence against the perceptual account. If one were to argue that stimulus identification was not complete by the time the irrelevant stimulus was presented, then an interference effect should be observed at the smallest SOA and should disappear at large SOAs. The reason is that the likelihood of stimulus identification being complete is expected to increase with the time between onset of the relevant stimulus and onset of the irrelevant stimulus. Thus, an interaction of SOA with interference, that is, a three-way interaction, was expected and clearly did not occur.

In the present experiment, the size of the effect (17 ms) was smaller than in Experiments 1 (35 ms) and 2 (43 ms). Hommel (1995), however, found a Simon effect comparable in size to those in previous studies with simultaneous presentation of irrelevant and relevant stimulus dimensions. The explanation for this discrepancy is twofold. First, we used the onset of the mouth movement as a go signal for the initiation of the response (i.e., there was no response uncertainty). Such a setup resembles simple RT tasks, which are known to yield only small compatibility effects (see Hommel, 1996, for an overview). Second, the irrelevant mouth movement extended over time, such that motion detection might have preceded identification of the mouth gesture. Thus, the response might have been initiated before the irrelevant information could exert its influence. In contrast, the previous experiments required participants to watch the animation until the response-relevant stimulus appeared. Therefore, identification of the mouth movement at the time of response initiation was ensured.

In summary, Experiments 1–3 show that visually presented mouth movements are automatically processed up to a late, response-related stage. Even when identification of the response-relevant stimulus dimension preceded presentation of the irrelevant mouth movements, the visible gesture still influenced responses. The results are fully consistent with the account of visual speech perception furnished by the motor theory of speech perception.

Experiment 4

The purpose of Experiment 4 was to examine whether the interference from speech gestures could be accounted for by directional coding of the mouth movements. It is known that the McGurk effect obtains even if only isolated kinematic properties of visible speech are presented (Rosenblum & Saldaña, 1996). This leads to the question of whether visible speech that is reduced to simple directional features suffices to produce the present SRC effect. In a related study by Langton, O'Malley, and Bruce (1996), investigating deictic gestures, weak interference effects were observed when complex biological stimuli were replaced by abstract stimuli: Langton et al. presented pointing gestures of a human model paired with congruent or incongruent verbal equivalents. Participants' responses to either gestures or words were symmetrically influenced by irrelevant information from the other dimension. The interference from the nonverbal material persisted when the deictic gestures were replaced by arrows, but the size of the effect was reduced. Thus, it is possible that gestures convey largely the same information as spatial symbols. Informal inspection of the present stimulus material suggests that the predominant spatial feature of the displayed biological motion is an opening–closing movement of the speaker's mouth. Thus, the hypothesis may be entertained that the interference effect observed in Experiments 1–3 was caused by an overlap of direction of displayed mouth movement (opening or closing) and response direction. That is, the influence of the visible speech gestures on verbal responses can be accounted for by spatial compatibility, not by gestural compatibility. If the former alternative is accurate, then presentation of lines moving up and down in the same way as the speaker's mouth should have the same effects as presentation of the speaker's mouth. However, if the nature of the effect is not spatial but gestural, as a strong version of the motor theory of speech perception would predict, then no interference from line movements should be expected.

Method

Participants. Eight students at the Ludwig-Maximilians University of Munich were paid for their participation. All reported normal or corrected-to-normal vision and were naive as to the purpose of the experiment.

Apparatus and stimuli. The apparatus and stimuli were the same as those used in Experiment 1 with the following exceptions. At the upper and the lower lips of the speaker's mouth, horizontal white lines were attached. The lines extended 3.15° from the horizontal center of the mouth to the left and right. The vertical positions of the lines were aligned with the outer edges of the lips, encompassing the maximal vertical extent of the mouth. The center portion of the lines (0.86°) was deleted so that the imperative stimuli were not covered by the moving lines. During the experiment, the picture of the mouth was rendered black, such that only lines and imperative stimuli were visible. For a "/bʌ/" gesture, the lines moved closer together before moving apart. For a visible "/dʌ/" gesture, the lines only opened up. In both cases, the lines returned to their starting position after 333 ms (10 pictures; see Experiment 1 for a more detailed description of the lips' movement).

Design and procedure. The design and procedure were the same as those used in Experiment 1.

Results

Data treatment was the same as in Experiment 1. There were only a few anticipations and missing trials (0.7%). RT means are graphed in Figure 4. A three-way within-subject ANOVA (Visible Gesture × Spoken Syllable × SOA) on RTs revealed a significant main effect of SOA, $F(3, 21) = 15.75, p < .0001$. RTs decreased from 533 ms at the shortest SOA to 468 ms at the longest SOA. Responses with line movement for a visible /dʌ/ gesture tended to be nonsignificantly faster than responses to line movements for a visible /bʌ/ (492 vs. 499 ms), $F(1, 7) = 3.94, p < .0876$. No other effects reached significance ($p > .22$). Again, because the PEs were low (1.9%), we collapsed them across SOA and entered them into a two-way ANOVA (Visible Gesture × Spoken Syllable). No significant effects emerged ($p > .14$).

Discussion

The results show unambiguously that the observed SRC effect from mouth movements was not spatial. Lines mimicking the vertical opening–closing motion of the speaker's mouth while producing /bʌ/ or /dʌ/ failed to interfere with the production of CV syllables. Thus, the current effect of irrelevant mouth movements on the production of CV syllables cannot be classified as a variant of the spatial interference observed in Simon- and Stroop-like settings (e.g., Kornblum, 1994; O'Leary & Barber, 1993). Rather, the interference from mouth movements is related to gestural information conveyed by the display. Thus, our results support the view that gestures are important units in the process of speech processing.

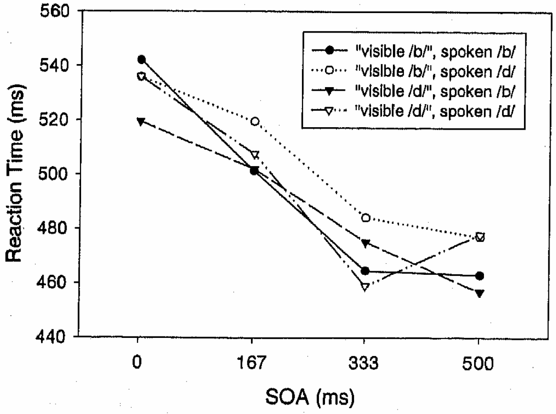

Figure 4. Mean reaction times as a function of visible gesture, spoken syllable, and stimulus onset asynchrony (SOA) in Experiment 4.

Gestural information may be contained in the rich kinematic pattern of mouth movements (Rosenblum & Saldaña, 1996) but cannot be reduced to simple opening–closing features. Hence, we conclude that the directional coding of the stimulus features does not account for the interference observed in Experiments 1–3.

General Discussion

Perceptual and motor accounts of visual speech perception were tested in an RT paradigm. The motor account of visual speech processing, as derived from the motor theory of speech perception (Liberman & Mattingly, 1985), holds that motor control structures are involved in the processing of visible speech. If visible speech is indeed analyzed by perceptual–motor structures, as the motor theory holds, looking at a speaker's mouth should lead to activation of these structures. Consequently, pronouncing the same utterance as the speaker should be easier than pronouncing a different utterance; that is, S-R compatibility effects for visually presented mouth movements and verbal responses would be expected. Perceptual explanations do not share the assumption of motor involvement in the perception of visible speech but claim that visual speech input is treated as a cue (Massaro, 1987) or as lawful information about a distal event (Fowler, 1986). Thus, no perceptual–motor interference effects should emerge. Rather, interference from mouth movements should be restricted to perceptual interference.

We put these two conflicting views to a test by showing participants response-irrelevant movies of a mouth articulating /bʌ/ or /dʌ/ and by asking them to verbally respond with either the same or a different syllable. In Experiment 1, we used the letters "Ba" and "Da" that appeared on the speaker's mouth to indicate which response was to be performed. We observed an interference effect that was attributable either to perceptual fusion or conflict or to perceptual–motor facilitation or interference because both stimulus dimensions showed dimensional overlap with the response and overlapped themselves. The purpose of Experiment 2 was to render perceptual interference unlikely by assigning arbitrary symbols to the responses. In this Simon-like setup, only the irrelevant stimulus dimension and the response showed overlap, so SRC accounted for the observed interference from mouth movements. In Experiment 3, any perceptual interference was ruled out by preexposing the relevant stimulus dimension. Finally, Experiment 4 showed that the stimuli were not coded as simple directional features (opening and closing), so the nature of the interference effect may be hypothesized to be gestural and not spatial. In summary, the present experiments provide compelling evidence for the view that visible speech is processed up to a late, response-related processing stage and that gestures constitute important units in speech perception. These results are fully consistent with the motor theory of speech perception.

Gestures as Basic Units?

In Experiment 1, we found interference from visible mouth movements in a reading task. This result was surprising because it has proved difficult to find interference from a picture in a word-reading task. W. R. Glaser and Düngelhoff (1984) showed participants words printed inside an outline drawing of concrete objects. When participants were asked to name the picture, responses were strongly affected by the accompanying word. However, no such interference was observed when participants were asked to read the word, even if the picture was preexposed. The pattern of results was similar to that found in a related study with Stroop dimensions (M. O. Glaser & Glaser, 1982). In Experiment 1, a situation similar to Glaser and Düngelhoff's experiments was created. A (moving) picture of a mouth was presented and a syllable was to be read. However, the apparent conflict between the strong interference effect observed in our setup and the lack of interference in the experiments of W. R. Glaser and Düngelhoff may be resolved by looking at the stimulus material that was used. The main difference between the picture–word paradigm used by W. R. Glaser and Düngelhoff and ours is the linguistic unit under consideration. We used simple CV syllables, whereas words were used in W. R. Glaser and Düngelhoff's experiments. Unlike syllables, words have semantic and lexical representations that come into play in picture–word interference: Picture naming requires processing at a semantic level before the word form can be retrieved from the lexicon (W. R. Glaser & Glaser, 1989). Therefore, interference from a word in a picture-naming task may arise at either the lexical (Starreveld & La Heij, 1995) or the semantic (W. R. Glaser & Düngelhoff, 1984; Seymour, 1977) level. Presumably, word naming remains unaffected by the presentation of a picture because words have privileged access to the lexical level and are therefore resistant to interference from a picture via the semantic route (W. R. Glaser & Glaser, 1989).

The situation is vastly different for syllables. The appropriate level of description for CV syllables is phonological. Traditionally, phonemes were considered to be units representing bundles of phonological features (e.g., Kenstowicz & Kisseberth, 1979). However, there is also reason to believe that phonemes are represented as gestures (Browman & Goldstein, 1992). The observed interference from visual articulatory gestures provides evidence for the latter view. If phonological information were exclusively acoustic, the transformation from a symbolic representation (letters) to a phonological representation should not be perturbed by simultaneously presented visual mouth movements, even more so considering the fact that letters may be recoded phonologically (e.g., Lupker, 1982). Therefore, at least some information about articulatory gestures has to be represented at the phonological level for gestures to interfere in a reading task. Thus, our results support the claim that gestures constitute the basic units of speech.

Deictic Gestures

Our experiments clearly demonstrate that visible gestures are analyzed up to a late, response-related stage of processing. This interpretation is at odds with a recent claim that gestures are analyzed in a specialized system operating only at the perceptual level. To determine the relative contributions of gestures and verbal information in the comprehension process, Langton et al. (1996) investigated mutual interference effects of static deictic gestures and verbal information. In a Stroop-type paradigm, static pictures of a model whose arms pointed into one of four different directions (up, down, left, or right) were paired with congruent or incongruent verbal equivalents. Participants were instructed to respond to either a gesture or a verbal stimulus. Consistent with the idea that gestures and speech are complementary in comprehension (McNeill, 1985), symmetrical interference effects emerged: Irrelevant verbal material influenced the processing of the gestural information by the same amount as irrelevant deictic gestures influenced processing of verbal material. This was the case regardless of whether the verbal material was auditory or visual (written). When the deictic gestures were replaced by arrows, the pattern of interference remained symmetrical. However, the size of the interference effect was smaller, indicating that arrows cause much less interference than gestures. Although alternative explanations for the difference between interference from arrows and gestures could not be ruled out, Langton et al. suggested that a specialized system concerned with the identification of gestural material accounted for the observed interference. Importantly, the locus of interference was placed at a perceptual level because interference was correlated with perceptual discriminability (Melara & Mounts, 1993).

From the viewpoint of a dimensional overlap model (Kornblum et al., 1990), the Langton et al. results are far from conclusive. As mentioned earlier, Stroop-type paradigms are ambiguous about the locus of interference. Because both stimulus dimensions overlap with the response and among themselves, interference may be due to SSC, SRC, or both. As demonstrated in Experiment 1, Stroop interference may be obtained for stimuli that are processed up to a late, response-related stage (Experiments 2 and 3). These results are hard to reconcile with exclusively perceptual processing of gestures. To salvage the concept of a perceptual gesture analyzer, one might argue that deictic gestures and phonetic gestures are part of two different modules. In fact, this would lead to the hypothesis shared by motor theorists that speech perception is accomplished by a specialized module different from modules dealing with deictic gestures.

Is Processing of Speech Gestures Special?

The primary aim of the present study was to test a specific hypothesis derived from the motor theory of speech perception (Liberman & Mattingly, 1985). Fully consistent with the theory was our finding that looking at visible speech gestures induced activity in a response-related stage. A further claim of the motor theory is that speech processing takes place in a specialized module that has to be distinguished from other modules such as the auditory. One line of evidence supporting this view consists of phenomena demonstrating that the same cues are perceived differently depending on whether they are part of a speech or a nonspeech sound. For instance, formant transitions that cue place of articulation in CV syllables are perceived differently in a CV context and in isolation. In a CV syllable, the transition may be rising or falling if paired with different vowels, but it still cues the same consonant (i.e., the consonant sounds the same). When presented in isolation, however, rising and falling transitions sound like two distinctly different glissandi or chirps (Mattingly, Liberman, Syrdal, & Halwes, 1971; Xu, Liberman, & Whalen, 1997). The logic underlying these experiments is that differences in processing characteristics for speech and nonspeech sounds are explained by two different modes of acoustic sound perception, one dealing with speech and the other dealing with the remaining auditory input (see Liberman & Mattingly, 1985, for a review). Consequently, if the same processing characteristics are observed for speech and nonspeech stimuli, the claim of a specialized speech module loses force and a "speech-is-not-special" view is favored. For instance, evidence for the integration of visual and auditory nonspeech stimuli analogous to the McGurk effect has been obtained (Saldaña & Rosenblum, 1993). In a similar vein, there is evidence that the processing of visual speech and nonspeech gestures takes place in the same cortical areas. Smeele, Massaro, Cohen, and Sittig (1998) found a left-hemisphere advantage both for the discrimination of visible speech and for the discrimination of nonlinguistic facial movement. Therefore, Smeele et al. suggested that the laterality effect was not due to activity in a specialized module dealing with speech but rather to a left-hemisphere advantage in the processing of dynamic visual information, a function that is not specific to gestural stimuli. This conclusion, however, may be premature in light of research by Rizzolatti and colleagues (see Gallese, Fadiga, Fogassi, & Rizzolatti, 1996, for an overview).

Rizzolatti and colleagues (e.g., Rizzolatti et al., 1988) discovered neurons in area F5 of the macaque monkey that discharged both when the monkey performed a given action and when it observed a similar action performed by the experimenter. These so-called "mirror neurons" showed a clear relation between the visual and motor actions they responded to and are thought to form a system for matching observation and execution of motor actions. The observation and execution system in monkeys may serve to generate an internal representation of movements that is involved in the understanding of motor events. A link between observer and actor is thereby formed that is interpreted to be a precursor of human speech (Rizzolatti & Arbib, 1998). Empirical support for the existence of an observation and execution system in humans was provided by a transcranial magnetic stimulation study (Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995) and a positron emission tomography study (Rizzolatti et al., 1996). Fadiga et al. (1995) showed that cortical motor excitability was increased when participants observed grasping movements. Motor-evoked potentials recorded from the hand–arm muscles during observation of 3-D objects being grasped were higher than in the two control conditions in which participants merely observed the objects or detected the dimming of a light. Heightened motor excitability was restricted to those muscles necessary to actively perform the action. In addition, Rizzolatti et al. (1996) found that, during observation of grasping, regional blood flow increased in the region of the superior temporal sulcus and in the posterior part of the inferior frontal gyrus (Broca's area).

Although traditionally considered to be a speech area, Broca's area may contain representations not only of effectors related to language production but also to hand movements (see Rizzolatti et al., 1996). Thus, the existence of an observation and execution matching system seems well established, and stunningly, the cerebral location of this system points to a close connection between speech and hand movements. This fact may reflect the development of speech from hand gestures (Rizzolatti & Arbib, 1998) and the existence of a human communicative system that embraces both gestures of the vocal tract, which are predominantly linguistic, and hand gestures.

We interpret the action recognition systems in humans to be evidence for the view that a perceptual–motor structure such as the "vocal tract synthesizer" in Liberman and Mattingly's (1985) motor theory is not special to speech. Rather, we posit that manual and articulatory gestures, both of which serve communicative functions, are perceived by accessing their motor representations. In line with this view, both phonetic and nonlinguistic mouth gestures appear to be processed more efficiently by the left hemisphere (Smeele et al., 1998), indicating that there is no left-hemispheric specialization for speech gestures. However, hemispheric asymmetries may not emerge in studies opposing linguistic and nonlinguistic stimuli. Rather, the crucial difference in left–right hemispheric processing of mouth movements may concern their meaningfulness. Differences in the meaning of hand actions have been shown to lead to different patterns of brain activity and clear left–right asymmetries (Decety et al., 1997; Grèzes, Costes, & Decety, 1998). Meaningful actions strongly engaged the left hemisphere, whereas meaningless actions involved mainly the right hemisphere. If the left-hemispheric specialization for meaningful actions included mouth movements, the left-hemispheric advantage for nonlinguistic mouth movements in the study by Smeele et al. may be attributed to the meaningfulness of their stimuli (e.g., kissing, tongue protrusion). Further research will have to show whether a processing pattern that parallels the meaningful–meaningless distinction for hand gestures also exists for mouth gestures and whether meaningful gestures are further differentiated according to their linguistic content.

Thus, we share the motor theoretical assumption that speech perception relies on perceptual–motor structures, but we reject the claim that speech perception is special. Rather, it appears more plausible to assume that perceptual codes are represented in a format commensurate with motor functions (Hommel, Müsseler, Aschersleben, & Prinz, 1999; MacKay, 1987; Prinz, 1990).

To conclude, we conducted four experiments that provide strong support for a perception–action link in the processing of visible speech that was postulated in the motor theory of speech perception (Liberman & Mattingly, 1985). We found that visible speech gestures strongly interfered with the production of CV syllables. Because the influence of the speech gestures persisted even after any possibility of perceptual conflict was eliminated, the locus of the interference was supposed to be a response-related stage. An additional experiment showed that directional coding of the mouth movements did not take place; rather, information about the gesture was important. We argue that perceptual-motor structures handling perception of speech gestures are not special but constitute a general processing mechanism in human communication.

References

Binnie, C. A., Montgomery, A. A., & Jackson, P. M. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research, 17,* 619–630.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49,* 155–180.

Dalrymple-Alford, E. C., & Azkoul, J. (1972). The locus of interference in the Stroop and related tasks. *Perception & Psychophysics, 11,* 385–388.

Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., & Fazio, F. (1997). Brain activity during observation of actions: Influence of action content and subject's strategy. *Brain, 120,* 1763–1777.

De Jong, R., Liang, C. C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus–response correspondence. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 731–750.

Dunbar, K., & MacLeod, C. M. (1984). A horse race of different color: Stroop interference patterns with transformed words. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 622–639.

Eimer, M. (1995). Stimulus–response compatibility and automatic response activation: Evidence from psychophysiological studies. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 837–854.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology, 73,* 2608–2611.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3–28.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 816–828.

Fowler, C. A., & Rosenblum, L. D. (1991). The perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 33–59). Hillsdale, NJ: Erlbaum.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119,* 593–609.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 875–894.

Glaser, W. R., & Düngelhoff, F.-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 640–654.

Glaser, W. R., & Glaser, M. O. (1989). Context effects on Stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118,* 13–42.

Gordon, P. C., & Meyer, D. E. (1984). Perceptual-motor processing of phonetic features in speech. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 153–178.

Green, K. P., & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 278–288.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics, 38,* 269–276.

Grèzes, J., Costes, N., & Decety, J. (1998). Top-down effect of strategy on the perception of human biological motion: A PET investigation. *Cognitive Neuropsychology, 15,* 553–582.

Guiard, Y., Hasbroucq, T., & Possamai, C.-A. (1994). Stimulus congruity, irrelevant spatial SR correspondence, and display-control arrangement correspondence: A reply to O'Leary, Barber, and Simon (1994). *Psychological Research, 56,* 210–212.

Gumenik, W. E., & Glass, R. (1970). Effects of reducing the readability of the words in the Stroop Color-Word Test. *Psychonomic Science, 20,* 247–248.

Hasbroucq, T., & Guiard, Y. (1991). Stimulus-response compatibility and the Simon effect: Toward a conceptual clarification. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 246–266.

Hock, H. S., & Egeth, H. (1970). Verbal interference with encoding in a perceptual classification task. *Journal of Experimental Psychology, 83,* 299–303.

Hommel, B. (1993). The relationship between stimulus processing and response selection in the Simon task: Evidence for a temporal overlap. *Psychological Research, 55,* 280–290.

Hommel, B. (1995). Stimulus-response compatibility and the Simon effect: Toward an empirical clarification. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 764–775.

Hommel, B. (1996). S-R compatibility effects without response uncertainty. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 49A,* 546–571.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (1999). *The theory of event coding (TEC): A framework for perception and action.* Manuscript submitted for publication.

Keele, S. W. (1972). Attention demands of memory retrieval. *Journal of Experimental Psychology, 93,* 245–248.

Kenstowicz, M., & Kisseberth, C. (1979). *Generative phonology: Description and theory.* Boston: Academic Press.

Klein, G. S. (1964). Semantic power measured through the interference of words with color naming. *American Journal of Psychology, 77,* 576–588.

Kornblum, S. (1994). The way irrelevant dimensions are processed depends on what they overlap with: The case of Stroop- and Simon-like stimuli. *Psychological Research, 56,* 130–135.

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

Langton, S. R. H., O'Malley, C., & Bruce, V. (1996). Actions speak no louder than words: Symmetrical cross-modal interference effects in the processing of verbal and gestural information. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1357–1375.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Lu, C.-H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin & Review, 2,* 174–207.

Lupker, S. J. (1982). The role of phonetic and orthographic similarity in picture word interference. *Canadian Journal of Psychology, 36,* 349–367.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24,* 253–257.

MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills.* Berlin: Springer-Verlag.

Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 384–394.

Massaro, D. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, MA: MIT Press.

Massaro, D. W., Cohen, M. M., & Thompson, L. A. (1988). Visible language in speech perception: Lipreading and reading. *Visible Language, 1,* 8–31.

Mattingly, I. G., Liberman, A. M., Syrdal, A. M., & Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology, 2,* 131–157.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

McNeill, D. (1985). So you think gestures are non-verbal? *Psychological Review, 92,* 350–371.

Melara, R. D., & Mounts, J. R. W. (1993). Selective attention to Stroop dimensions: Effects of baseline discriminability, response mode and practice. *Memory & Cognition, 21,* 627–645.

Morton, J. (1969). Categories of interference: Verbal mediation and conflict in card sorting. *British Journal of Psychology, 60,* 329–346.

O'Leary, M. J., & Barber, P. J. (1993). Interference effects in the Stroop and Simon paradigms. *Journal of Experimental Psychology: Human Perception and Performance, 19*(4), 830–844.

O'Leary, M. J., Barber, P. J., & Simon, J. R. (1994). Does stimulus correspondence account for the Simon effect? Comments on Hasbroucq and Guiard (1991). *Psychological Research, 56,* 203–209.

Palef, S. R., & Olson, D. R. (1975). Spatial and verbal rivalry in a Stroop-like task. *Canadian Journal of Psychology, 29,* 201–209.

Porter, R., & Lubker, J. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research, 23,* 593–602.

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.), *Relationships between perception and action* (pp. 167–201). Berlin: Springer-Verlag.

Proctor, R. W., Dutta, A., Kelly, P. L., & Weeks, D. J. (1994). Cross-modal compatibility effects with visual-spatial and auditory-verbal stimulus and response sets. *Perception & Psychophysics, 55,* 42–47.

Proctor, R. W., & Reeve, T. G. (1991). The prevalence of salient-features coding in choice reaction tasks. In J. Requin & G. E. Stelmach (Eds.), *Tutorials in motor neuroscience* (pp. 17–26). Dordrecht, the Netherlands: Kluwer Academic.

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neuroscience, 21,* 188–194.

Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Mattelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements. *Experimental Brain Research, 71,* 491–507.

Rizzolatti, G., Fidiga, L., Mattelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996). Localization of grasp representations in humans by PET: 1. Observation versus execution. *Experimental Brain Research, 111,* 246–252.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 318–331.

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics, 54,* 406–416.

Seymour, P. H. K. (1977). Stroop interference with response, comparison, and encoding stages in a sentence picture comparison task. *Memory & Cognition, 2,* 19–26.

Simon, J. R., Hinrichs, J. V., & Craft, J. L. (1970). Auditory S-R compatibility: Reaction time as a function of ear-hand correspondence and ear-response-location correspondence. *Journal of Experimental Psychology, 86,* 97–102.

Simon, J. R., & Small, A. M. (1969). Processing auditory irrelevant information: Interference from an irrelevant cue. *Journal of Applied Psychology, 53,* 433–435.

Smeele, P. M. T., Massaro, D. W., Cohen, M. M., & Sittig, A. C. (1998). Laterality in visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1232–1242.

Starreveld, P. A., & La Heij, W. (1995). Semantic interference, orthographic facilitation, and their interaction in naming tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 686–698.

Stirling, N. (1979). Stroop interference: An input and an output phenomenon. *Quarterly Journal of Experimental Psychology, 31,* 121–132.

Stoffels, E. J., Van der Molen, M. W., & Keuss, P. G. J. (1989). An additive factor analysis of the effect(s) of location cues associated with auditory stimuli on stages of information processing. *Acta Psychologica, 70,* 161–197.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Warren, R. E. (1972). Stimulus encoding and memory. *Journal of Experimental Psychology, 94,* 90–100.

Xu, Y., Liberman, A. M., & Whalen, D. H. (1997). On the immediacy of phonetic perception. *Psychological Science, 8,* 358–362.

Received November 10, 1998
Revision received March 25, 1999
Accepted April 28, 1999 ■