2007

Audiovisual speech perception: A speech production approach

Michelle A. Jarick
Wilfrid Laurier University

Your file Votre référence
ISBN: 978-0-494-26586-4
Our file Notre référence
ISBN: 978-0-494-26586-4

AUDIOVISUAL SPEECH PERCEPTION: A SPEECH PRODUCTION APPROACH

by

Michelle A. Jarick

Honours Bachelor of Arts, Wilfrid Laurier University, 2005

THESIS

Submitted to the Department of Psychology

in partial fulfilment of the requirements for

Master of Science, Brain & Cognition

Wilfrid Laurier University

2007

Author: Michelle A. Jarick

Cognitive Neuroscience of Communication Laboratory

Department of Psychology

Wilfrid Laurier University

Waterloo, Ontario, Canada

Supervisor: Dr. Jeffery A. Jones

Cognitive Neuroscience of Communication Laboratory

Department of Psychology

Wilfrid Laurier University

Waterloo, Ontario, Canada

Advisory Committee: Dr. Todd Ferretti

Language and Cognition Laboratory

Department of Psychology

Wilfrid Laurier University

Waterloo, Ontario, Canada

Dr. Sukhvinder S. Obhi

Cognition in Action Laboratory

Department of Psychology

Wilfrid Laurier University

Waterloo, Ontario, Canada

Abstract

The purpose of these studies was to test the main assumptions outlined in the Motor Theory of speech perception that (1) speech perception is linked to speech production, (2) audiovisual integration of speech occurs *automatically* and after the motor commands are activated, and (3) we perceive the *intended* gestures, which are extracted by a specialized 'phonetic module' in the brain.

In Experiment 1, we used a Stroop-like paradigm, where participants viewed and listened to a speaker producing speech syllables (/aba/ or /aga/) in three conditions: audio-only, visual-only, and audiovisual. Participants were asked to ignore irrelevant speech stimuli, and to identify vocally or manually the target letters (BA or GA) that appeared over the speakers' face, as quickly and accurately as possible. If speech perception is closely tied to speech production, then we should find faster response times to vocally produce a syllable that matched the perceived syllable than mismatched. Indeed, we found that response times were quicker when the target and irrelevant speech were compatible, than when they were incompatible. This finding was consistent across all modality conditions for verbal, but not for manual responses, suggesting a close perception-production link for speech. The same stimulus-response interference was found when the irrelevant stimuli were static pictures portraying speech or non-speech gestures (Experiment 4), demonstrating that even implied speech gestures interfere with speech production.

Furthermore, these compatibility effects were seen in Experiment 2 when we used conflicting auditory and visual irrelevant speech information (e.g., auditory /aba/ dubbed over a visual /aga/). Our results showed both modalities interfered with speech

production to the same degree. However, once the modalities were integrated (i.e., eliciting the McGurk effect; Experiment 3), our data showed faster response times to produce the target that was compatible with the integrated percept, than those targets that were compatible with either modality alone. Although not statistically significant, the trends in Experiment 3 suggest that integration may occur before the response stage is reached, implying that another mechanism may be responsible for integration. Overall, our findings provide some support for the Motor Theory view of speech perception, demonstrating that speech perception effects speech production.

Keywords: Speech Perception, Speech Production, Motor Theory, Mirror Neurons, Stroop-Effect, Stimulus-Response Compatibility, Audiovisual Integration, McGurk Effect.

Acknowledgements

The research presented in this thesis was carried out in the Cognitive Neuroscience of Communication Laboratory (a subdivision of the Centre for Cognitive Neuroscience) at Wilfrid Laurier University. The work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

First and foremost, I would like to express my deepest gratitude to my advisor Dr. Jeff Jones for his patience, encouragement, and brilliant insight over the past three years. Thank you for believing in me enough to give me this opportunity. You are the best mentor, colleague, and friend one could ask for - I am forever indebted to you. Cheers to many more years of laughter and success!

I would also like to thank Sam Lee, Adrienne Steer, and Chris Schwint for their help in the creation of the stimuli and Farina Pinnock for her help with the data collection. Thank you very much to Dwayne Keough for his helpful comments on an earlier version of this paper. However, I could not have made it through this process without the assistance of all members of the CNC Lab, past and present. They have all made this a memorable, fun, and rewarding experience for me, and I sincerely thank each and every one of them – Adrienne Steer, Danielle Striemer, Jon Krohn, Dwayne Keough, Farina Pinnock, Lauren Eudoxie, and Darcy Maslen, as well as our honourary lab members, Jen Major, Tara Dumas, and Courtney Patterson. I wish you all the best in your endeavours.

Lastly, I would like to express my love and appreciation to my family and friends for their unconditional support and encouragement. I would not be where I am today without you.

Table of Contents

List of Tables

List of Figures

General Introduction

An overwhelming amount of evidence supports the multisensory nature of speech perception. Indeed, much research has focused on the importance of auditory and visual integration, because it occurs so frequently during face-to-face communication. For instance, in noisy environments where the acoustic signal is highly degraded, information from a speaker's face significantly enhances auditory intelligibility (Sumby & Pollack, 1957; Schwartz, Berthommier, & Savariaux, 2004). Still in optimal listening environments, a speech perception advantage is observed if accompanied by visual speech (Davis & Kim, 2004; Reisberg, McLean, & Goldfield, 1987). Even more compelling, phonetic information from the face can modify clearly audible speech (the "McGurk effect"). For example, a dubbed video of a face mouthing /ga/ with a voice saying /ba/, elicits the illusion of hearing the phoneme /da/ (McGurk & MacDonald, 1976). The McGurk effect is a striking revelation of the powerful role visual information can play during speech perception and is frequently used as a tool for investigating audiovisual integration. In order to have a complete account of speech perception, it is imperative for theories to include the weighted function of visual information and how it is used in combination with the auditory information. Yet, it is still not clear *how* these very different sensory experiences are integrated to form a unitary speech percept. In search of an explanation, prototypical theories of speech perception have diverged down two main pathways to understand audiovisual integration: perceptual and gestural.

Perceptual Theories of Speech Perception

One of the first proposals regarding how we integrate multisensory information was provided by Summerfield (1987), who proposed that we gather complementary information from each of the different sources and then integrate them together ('vision-place, audition-manner' hypothesis). For instance with speech, Summerfield suggested that the visual signal afforded the place of articulation (the obstruction of the vocal tract by the lips, tongue, and jaw) and the auditory signal supplied the manner information (the proximity of the speech organs to each other to create a sound). It was proposed that we not only perceive discrete information from the different modalities, but we ignore the unreliable information (auditory place of articulation and visual manner features) in each of the signals as well. This hypothesis assumed that the brain knew which modality was more reliable for certain features included in the speech signal. Yet, the theory did not survive long, as it could not explain why we perceived a combination of /bga/ when presented with an auditory /ga/ and visual /ba/ signal (McGurk & MacDonald, 1976). This combined percept demonstrated that the brain encodes place of articulation information from both auditory and visual modalities and questioned the idea that we ignore the information afforded by the unreliable modality (Fowler, 2004). Therefore, Summerfield himself (Summerfield, 1987) discarded this theory as a valid explanation for how we integrate audiovisual speech information.

Since then, more sophisticated perceptual theories have been developed. General perceptual accounts propose that we parse the acoustic (and presumably visual) signals into phonetic segments and then match these segments with phonetic templates (prototypes) stored through learned associations in our memory (e.g., Diehl & Kluender,

1989; Massaro, 1987, 1998). Proponents of these theories argue that speech should be treated as any prototypical event in the environment, and viewed simply as a form of pattern recognition in which stimuli are identified and categorized based on previous experience. The most widely recognized account of intersensory integration due to its quantitative modeling (using mathematical algorithms) is the Fuzzy Logical Model of Perception (FLMP) devised by Massaro (1987). To explain the McGurk effect for example (as well as all other instances of integration), the FLMP states that the prototype /da/ is selected based on the amount of phonetic features that the auditory /ba/ and visual /ga/ signals have in common. So, speech is perceived by choosing a prototype in memory that best matches the phonetic information afforded by the acoustic and/or visual speech signals. This best-match procedure operates in the following three stages: evaluation, integration, and decision. During the *evaluation* stage, both visual and auditory information are processed independently and continuously, where they are evaluated for the degree of support they lend to each alternative prototype in memory. The *integration* stage is similar, however now the auditory and visual information are combined to form one phonetic unit and the overall combined degree of support is calculated against various alternatives. And lastly, the *decision* stage operates by mapping the integrated output onto a response, which takes the form of an absolute decision or a rating indicating the extent of similarity with each alternative. Thus, phonetic perception is achieved once the decision has been made concerning the prototype that best matches the incoming visual and auditory information. Overall, the main assumptions of the model are that (1) the two sources are first evaluated independently and (2) that they are integrated to produce general measure of best-fit to a prototype available in memory.

Although the FLMP has been able to reliably model human data obtained in many studies of speech perception (see Massaro, 2004 for a review), it has also been criticized on the account that it can model random data as well (Cutting, Bruno, Brady, & Moore, 1992). For instance, Cutting et al. (1992) evaluated several perceptual models on how well they could fit data used for accurate depth perception provided by different sources, such as height, occlusion, relative size, etc. Each source was assigned a value indicating whether it was present or absent in the visual display presented. Observers were asked to rate the degree of depth perceived in the visual display and the data set obtained were modeled according to the FLMP (and others), for which the FLMP provided the most accurate predictions. However, the authors then tested the accuracy of the FLMP against 1,000 simulated data sets that were randomly created (random numbers were generated for each source), and the FLMP modeled the majority of data sets (608) with great accuracy as well. Thus, the FLMP proved to be a good model for demonstrating how we integrate multiple sources of information, but it was also a good model for capturing patterns that were of no interest to the researcher, and therefore cannot be used as a reliable theory of multisensory integration until it is able to factor out the random error available in the data.

Finally, other (less discussed) classes of perceptual theories believe that integration of the auditory and visual signals can occur because they share time-varying characteristics that are highly correlated (e.g., Davis & Kim, 2006; Kuratate, Munhall, Rubin, Vatikiotis-Bateson, & Yehia, 1999; Munhall & Buchan, 2004; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Yehia, Rubin, & Vatikiotis-Bateson, 1998). Indeed, research has demonstrated a linear relationship between the dynamic

movements of the face, the parameters of the acoustic signal, and the different shapes of the vocal tract (Yehia et al., 1998). Further, studies have shown that even having information from the movements of the head can improve the speech signal in noise (Davis & Kim, 2006; Munhall et al., 2004), and can help in deciphering different words from a continuous speech stream (for a review see Cutler, Dahan, & Donselaar, 1997). This correlated dynamic pattern of the face and voice characteristics also seems to be specific to individuals, in that subjects are able to match faces with voices of familiar people (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Munhall & Buchan, 2004). Thus, these theories hypothesize that integration of auditory and visual information may rely on a dynamic pattern of these signals which are highly correlated.

However, what all these theories are missing is an adequate description of *how* the different sensory experiences are integrated (Fowler, 2004). According to Prinz (1997), there is no 'common code', for which the auditory and visual signals are represented. In order for any theory to survive as a complete account of speech perception, an explanation is necessary of how auditory and visual signals are transformed into a 'common currency' for which integration can then occur (Fowler, 2004).

Gestural Theories of Speech Perception

Gestural accounts of speech perception do propose a 'common currency' in which auditory and visual information can be represented and integrated. Essentially, gestural theorists claim that speech is perceived by deciphering the articulatory information afforded by the speaker's vocal tract, which can be represented in both the auditory, visual, or tactile domains (e.g., Fowler, 1986; Fowler & Rosenblum, 1991; Liberman,

Cooper, Shankeiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985). One such perspective is proposed by the Direct-Realists (Fowler, 1986; Fowler & Rosenblum, 1991), who believe that we recover the phonetic features of speech directly from the signal that affords it. Similar to Gibson's theory of direct perception (Gibson, 1979), direct-realists state that both auditory and visual signals share a lawful relationship regarding a common linguistic event (i.e., gestures of the vocal tract), and it is this common information that is processed and integrated. In other words, we do not encode the raw auditory waveforms, visual wavelengths of light, or haptic mechanical receptor information that is projected directly onto our senses, but the representation of the gestural information that each of them affords (which is activated by our perceptual system) concerning the same event.

Another gestural account, the Motor Theory of Speech Perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985; Liberman & Whalen, 2000) goes beyond the Direct Realists to incorporate the involvement of the motor system for speech perception. Proponents of this theory postulate that the objects of speech perception are the *intended* vocal tract gestures used by the speaker during speech production. By 'vocal tract gestures' they meant the invariant configurations of the teeth, tongue, lips, jaw, etc. that make up an abstract phonetic segment. Thus, speech perception and production are intimately connected in such a way that we can extract the underlying intended gestures of the speaker through an "analysis-by-synthesis" process (Stevens & Halle, 1967). That is, we perceive speech through internal simulation, thereby activating the motor commands used to produce the speech. Accordingly, both auditory *and* visual sources are valuable because both

contribute information about the invariant motor commands used to create the speech signal, and are concurrently processed by a specialized 'phonetic module' in the brain where the perceptual-motor mapping takes place (Fodor, 1983; Liberman & Mattingly, 1985). This idea of an innate neural mechanism, where only speech information is processed, is a unique aspect of the Motor Theory, and is thought to be activated *automatically* when we perceive phonetic information, regardless of modality. Thus, it provides a common processing area where the signals can be integrated and offers a direct link between perception and production of speech.

Support of a Gestural Account

By focusing on abstract features (i.e., *gestures*) as the primary objects of speech perception, gestural accounts have been able to explain many perceptual conundrums of speech phenomena, such as speaker variability, co-articulation (e.g., Liberman, Delattre, Cooper, & Gerstman, 1954), and duplex perception (e.g., Mann & Liberman, 1983). For instance, Liberman et al. (1954) studied co-articulation (when a consonant and vowel overlap during the production of a syllable) using synthetic speech syllables, and found that the characteristics of the acoustic signal for the same consonant changed depending on the vowel that followed it, yet the perceptual experience of the consonant remained the same. A hallmark demonstration reported in their study was the difference in voice spectra between the phonemes /di/ and /du/ - the second formant transition frequency (commonly representative of the place-of-articulation feature in the speech signal) increased in /di/, but decreased in /du/. Even though the acoustic cues had changed, the percept of hearing a /d/ was consistent. This perceptual constancy of phoneme

categorization led Liberman et al. (1954) to conclude that we encode the invariant articulatory gestures used to produce the /d/, and not the variable raw auditory waveform, which changed depending on the neighbouring vowels.

For studies demonstrating duplex perception, researchers (e.g., Mann & Liberman, 1983; Whalen & Liberman, 1987) separated the acoustic signal and presented different parts of it to each ear. For instance, the information that makes up the "base" of the speech signal (usually the frequencies of the steady-state formants and the first and second formant transitions[1]) are presented to the left ear, and commonly perceived in isolation as an ambiguous stop-vowel syllable (e.g., /ba/ or /ga/). The rest of the signal (mainly the third formant transition[2] that separates /ba/ from /ga/) is presented to the right ear and frequently perceived in isolation as a 'chirp' sound. However, when presented together at the same time, two very different perceptions arise. Participants frequently report hearing a coherent phoneme (e.g., /ba/ or /da/) in the left ear and a 'chirp' sound in the right ear. Since the same sound (i.e., the third formant transition) can be perceived differently, as phonetic information in the left ear and a non-speech sound in the right ear, Mann and Liberman interpreted this as evidence that speech is processed differently from that of general auditory information (c.f. Fowler & Rosenblum, 1990). They claim there must be two perceptual systems that can operate simultaneously: the 'phonetic module' used to encode speech related information, and a general 'auditory module' that processes non-speech related information.

---

[1] The first formant transition is related to the voicing feature of stop-consonants in the English language and varies with the place of articulation during production. The second formant transition is more directly related to the place of articulation of the consonant that is produced (Liberman et al., 1967).

[2] The third formant transition is also dependent on the constriction of the vocal tract when a consonant is produced (Liberman et al., 1967).

Additional evidence supporting an innate 'phonetic module' in the brain comes from the observation that newborn babies can imitate adult facial gestures (i.e., tongue protrusion and mouth opening) without ever having seen their own faces (Meltzoff & Moore, 1997). The fact that the only sensory information available to the newborn is the proprioceptive and somatosensory feedback from their own articulators (i.e., tongue and lips), as well as the visual information provided by the adult facial gesture, suggests that there must be a common representation shared among the different senses for processing speech stimuli. Furthermore, since this is observable in newborns within their first few hours of life, it is believed to be an innate phenomenon and not learned through experience.

Also consistent with the proposal that the encoding of speech gestures does not require having previous experience with the stimuli, Folwer and Deckle (1991) observed that haptic information from touching someone's mouth producing a phoneme could alter a simultaneously presented acoustic phoneme - similar to that seen with the McGurk Effect (where visual information modifies auditory perception). For instance, participants heard a speech continuum ranging from the phoneme /ba/ to /ga/, while touching the face of a speaker mouthing either /ba/ or /ga/. They showed that there were more /ba/ percepts when the mouthed phoneme was /ba/, as opposed to /ga/. Since the majority of us do not have experience with feeling a speaker's face during speech comprehension, it is unlikely that we have learned associations between the haptic and auditory consequences of the phoneme produced. It is possible, however, for haptic and auditory information to interact if an innate mechanism existed that encoded the intended gestures available in each of the signals, and then mapped them onto a common representation (i.e., their

motor commands). In other words, these studies suggest that speech is not encoded through perceptual learning, but rather through a specialized domain representing phonetic information from all of the modalities. According to the Motor Theory (Liberman & Mattingly, 1985), this phonetic representation is gestural in nature and is stored in the form of motor commands.

This proposal that experience is not necessary to encode phonetic information is also supported by a recent study conducted by Fowler, Brown, Sabadini, and Weihing (2003). The authors used a speech shadowing task similar to that used by Porter and colleagues (Porter & Castellanos, 1980; Porter & Lubker, 1980), where a speaker produced the vowel /a/ for a variable amount of time and then switched to one of three syllables, either /pa/, /ka/, or /ta/. Participants were required to shadow the speaker saying /a/ and then to produce one of the syllables (/pa/, /ka/, or /ta/) once they noticed the shift by the speaker. There were two separate tasks used in the study - a choice response task and a simple response task. During the simple response task, participants were assigned one of the three syllables to produce, either /pa/, /ka/, or /ta/, and asked to produce this syllable as quickly as possible once they detected the shift by the speaker (regardless of the syllable produced). Thus, on some trials the produced syllable could match that of the speaker and on some it could not. However, for the choice response task, participants were asked to imitate the syllable that was produced by the speaker as quickly as they could. This required the participants to not only detect that there had been a shift in the speakers' production, but then to identify what the syllable was that the speaker produced and imitate that syllable as quickly as possible. Usually for manual responses, there is a substantial difference in response times between these two tasks, where the choice

response task is typically 100 to 150 ms longer on average than the simple response task (Luce, 1986). In this study however, Fowler et al. (2003) found this difference to be only 26 ms, suggesting that the perception of the speech gestures in the choice response task reduced the amount of time it would normally take a participant to make the correct choice and then to respond, making it more comparable to the simple response task where no choice needed to be made. The authors interpreted this as evidence to support the notion that we perceive speech in a gestural format, as the speaker producing the syllable in the choice response task provided further instruction to the participants as to which syllable they should produce, lessening the amount of time to make a choice. Furthermore, during the simple response task they found that response times to produce the assigned syllable was faster when it was congruent with the syllable spoken by the speaker, than when it was incongruent. This finding further corroborates the author's conclusion that the perception of speech gestures facilitates the production of the same gestures. It also provides support for the Motor Theory, in that the motor commands corresponding to the perceived gestures would have already become activated, resulting in quicker response time to verbally produce that gesture.

Neurological Evidence of Motor System Involvement

Current neurological support for a close relation between perception and production of actions has been provided by the recent discovery of a population of 'mirror neurons' in the premotor area of the monkey that respond to the sight/sound of an action performed by another monkey (or experimenter), as well as when the monkey executes the same action (e.g., Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti,

1992; Kohler et al., 2002; Rizzolatti et al., 1996). For example, the same neuron will fire when the monkey observes the experimenter grasp a banana *and* when the monkey herself makes a grasping motion towards the banana. The majority of studies have focused on hand or finger movements (i.e., grasps), however Ferrari, Gallese, Rizzolatti, and Fogassi (2003) found mirror neurons in the lower portion of area F5 in the monkey that respond specifically to mouth gestures - being ingestive and/or communicative. The potential involvement of mirror neurons for speech processing is still presently unknown, but has garnered recent attention since homologous areas in the human brain have been shown to produce similar effects - one region being Broca's area, which is known to be involved in speech processing and imitation (Calvert & Campbell, 2003; Leslie, Johnson-Frey & Grafton, 2004; Skipper, Nusbaum & Small, 2005).

For instance, using functional magnetic resonance imagining (fMRI), Skipper, Nusbaum, and Small (2005) recently identified a network of brain areas that are active bilaterally during audiovisual speech perception, including pars opercularis (Broca's area), premotor cortex, and adjacent primary motor cortex – all shown to be responsible for speech production planning and execution. Additional fMRI studies have shown an overlap in activation when participants passively listen to speech and when participants are asked to overtly produce speech (Pulvermuller et al., 2006; Wilson, Saygin, Sereno, & Iacoboni, 2004). Consistent with the fMRI results, studies using transcranial magnetic stimulation (TMS) have shown that stimulation over the left motor cortex produces a substantial amplitude increase in motor-evoked potentials (MEP's) recorded from the tongue (Fadiga, Craighero, Buccino & Rizzolatti, 2002) and lips (Watkins, Strafella, & Paus, 2003) of participants while listening to and viewing speech that required the use of

the tongue and lips, respectively. Furthermore, this perceptual-motor mechanism seems to be highly speech specific, in that the MEP's were more strongly elicited when the speech contained real words as opposed to pseudowords (Fadiga et al., 2002), and visible lip movements compared to eyebrow movements (Watkins et al., 2003). Another interesting possibility may be that the system is also sensitive to the different features provided by the visual and auditory modalities. Sundara, Namasivayam, and Chen (2001) found significant increases in MEP's recorded from the lip muscles when they saw a speaker producing /ba/ (a bilabial gesture), but not when they heard /ba/. This may be because the perception of place-of-articulation is more easily seen than heard (Summerfield, 1987). At present, the functional significance of mirror neurons is unknown, but one attractive possibility is that they are used to facilitate communication between sender and receiver by covertly simulating the observed gestures, leading to recognition of the intended action (Rizzolatti et al., 1996). Skipper et al. (2005) claims that this type of resonance mechanism may aid in speech recognition by matching the intended gestures of the speaker to the listeners' motor counterparts, thus narrowing the possible interpretations of the speech utterance for the listener. Clearly, this research is in its initial stages and more studies are needed to make stronger connections between mirror neurons and speech perception in humans.

Stimulus-Response Compatibility

In an attempt to provide empirical evidence of a common mechanism mediating perception and production of speech, recent behavioural studies have utilized stimulus-response compatibility paradigms (Prinz, 1997; De Jong, Liang, & Lauber, 1994;

Kornblum, Hasbroucq, & Osman, 1990; Hommel, 1993). According to the dimensional overlap model (Kornblum et al., 1990), if more than one stimulus and/or response have features in common, whether on a perceptual, structural, or conceptual dimension, the presentation of one may automatically activate the other. For instance, Kornblum et al. (1990) presented an irrelevant stimulus (one that was not required for a response) with a relevant stimulus (one required for a response), and found that which ever one correlated the most with the response in terms of common features automatically activated the response. So, when the relevant stimulus correlated with the response, facilitation of the response was observed. Yet, if the irrelevant stimulus was correlated to a greater degree with the response, then a delayed response was made.

The dimensional overlap model has been used to try and explain phenomena such as the Stroop-effect (Stroop, 1935). The Stroop-effect is a classic example for demonstrating how automatically processed stimuli can interfere with the production of a related response (Stroop, 1935). For example, Stroop found shorter response times for identifying the colour of text if it spelled the same colour-word (e.g., BLUE in blue ink) versus an incongruent colour-word (e.g., RED in blue ink). One explanation is that the two stimulus dimensions (colour of text and colour-word) are processed in parallel and compete at the response selection stage (naming the colour or reading the word; for a review see MacLeod, 1991). If both dimensions are congruent, then the same response will be activated and produce response facilitation. If they are incongruent, then the irrelevant dimension (colour-word) must be inhibited, causing response interference.

Similar paradigms have been used to address analogous questions regarding speech perception. For instance, Kerzel and Bekkering (2000) showed participants videos

of a model's mouth articulating either /ba/ or /da/, and had them vocally produce the same or different syllable (/ba/ or /da/) in response to the target letters 'Ba' or 'Da' that were briefly presented over the mouth. Based on the Motor Theory, they predicted that facilitative or interference effects should arise when the same or different speech gestures were viewed respectively (because seeing speech would activate motor areas involved in speech production). In their first experiment, they found that incongruent stimuli elicited slower responses than congruent stimuli. Furthermore, this effect carried over even when the target letters were replaced by arbitrary symbols (&& and ##) for which congruent or incongruent responses were assigned. For instance, participants were trained to say 'BA' when the target && appeared, and 'DA' when the target ## was presented. These findings led the authors to conclude that visual speech was processed up to a response-related stage, and not solely at a perceptual stage, since the arbitrary symbols were not speech related.

Results of similar studies have replicated and extended this perceptual-motor interference using auditory stimuli (e.g., Fowler et al., 2003; Gordon & Meyer, 1984; Porter & Castellanos, 1980). Interestingly, Gordon and Meyer (1984) found facilitation (~ 50 ms) when participants produced a syllable that contained the same voicing feature (vocal fold vibration) as a syllable heard, but no facilitation was observed when the syllable heard contained the same place of articulation information. However, this difference between voicing and place of articulation did not exist when the stimuli were presented visually. Again, this suggests the perceptual-motor system may be activated differently depending on the phonetic feature information available in each modality. In this case, the voicing feature of a speech stimulus is more easily heard than seen

(Summerfield, 1987), so the voicing information available in the auditory speech might have activated the motor commands for that gesture more strongly, than the place of articulation feature which is more readily perceived in visual speech. Although Gordon and Meyer found there to be no effect of place of articulation for the visually presented syllables, the results from Kerzel and Bekkering (2000) do show that the place of articulation feature can produce facilitation when presented visually. Thus, perhaps there are modality differences in motor activation for which future research should tease apart these differences.

However, Kerzel and Bekkering's (2000) results are not surprising when framed in terms of stimulus-response compatibility accounts (e.g., Hommel, 1993; Prinz, 1997; Kornblum et al., 1990), since the perceived and produced speech gestures, when compatible, had numerous features in common, leading to faster reaction times. So, in a following study Kerzel (2002) reduced the featural overlap between stimulus and response by using a manual instead of vocal response. By introducing a manual response, the stimulus and response no longer shared compatible features and response interference should disappear. Still, Kerzel (2002) again found compatibility effects that suggested that the interference observed was likely due to phonological correspondence at the perceptual stage. However, many details of this study bring uncertainty regarding these conclusions. One is that the comparison between studies (verbal vs. manual responses) should be made with caution, as the two responses vary remarkably in their complexity of execution (in our experience verbal responses tend to have increased variance and take longer to produce than manual responses). As well, no baseline measure of vocal or button-press response rates were included in the studies, making them hard to compare.

This also made Kezel (2002) unable to determine whether the response time differences were the product of response facilitation or response interference, thus, limiting the conclusions regarding the direction of influence. Finally, the participants were not required to ignore the irrelevant mouth movements when identifying the target, and they presented the response letters following the mouth movements for the majority of trials. Thus, participants probably processed the mouth movements and activated the phonetic gestures in memory long before the target letters even appeared. This could have allowed any type of response to be facilitated ('primed') provided it was compatible with the observed speech. Therefore, this comparison between response types needs to be accomplished under identical conditions and with a response baseline (to equate the variability differences) before strong conclusions regarding the level at which interference occurs can be made.

Purpose of the Current Studies

To our knowledge, researchers have yet to adequately investigate the degree to which interference effects differ across sensory modalities. Therefore, in the present studies we sought to identify the perceptual-motor interaction across different response types and modalities by including an essential baseline measure of each of the participants' response latency. We adopted a similar paradigm to that used by Kerzel and Bekkering (2000) and directly compared differences in response times caused by the compatibility between orthographic targets and irrelevant visual, auditory and audiovisual speech stimuli. If the perception of speech is closely linked to the production of speech, then there should be faster responses when the irrelevant speech and target are

compatible, and delayed responses when incompatible. We presented a speaker producing either /aba/ or /aga/ in three conditions: visual-only, audio-only, and audiovisual. Compatible or incompatible target letters ('BA', 'GA', or 'DA') were flashed over the speakers face, and participants were asked to identify the target as quickly and accurately as possible. An important addition to our work was the inclusion of a baseline control measure to which we were able to compare all other responses against. This allowed us to clearly identify whether facilitation or interference was occurring in each of the experimental conditions. According to the Motor Theory, if speech is perceived by encoding the articulatory gestures of the vocal tract, then the perception of speech (whether visually, auditorily, or audiovisually) should activate the motor commands needed to produce those gestures (leading to facilitation), while at the same time inhibiting production of any gestures not presented (leading to interference). Therefore, we predicted faster response times when participants had to produce the same gestures as those just observed as compared to producing conflicting speech gestures.

In a series of studies, we used this paradigm to test the assumptions made by the Motor Theory, that (1) speech perception is intimately tied to speech production ("analysis-by-synthesis"), (2) integration of speech gestures occurs *automatically* and after the appropriate motor commands are activated, and (3) it is the *intended* gestures that we perceive, not the raw sensory information, and that this is specific to speech. Experiment 1 was designed to establish whether the stimulus-response paradigm we adopted from Kerzel and Bekkering (2000) was a reliable measure of perceptual-motor interference for speech. Thus, we sought to determine whether any interference observed between the conditions was localized at the response-related stage or at a stimulus level.

We did this by including manual responses as well as verbal responses, expecting that if localized at the response stage, the manual responses will fail to show any differences in response times due to the lack of compatible features with the stimuli. If however, interference occurs at a stimulus level, then we should see the same effects for the manual responses as the verbal responses, suggesting that the effect is due to the compatible features inherent in the irrelevant and relevant stimuli.

In Experiment 1, we did find that the response times were faster to vocally (but not manually) produce the target when it was the same as the syllable seen or heard by a speaker, which provided a convincing demonstration of a relation between perception and production of speech gestures (satisfying the first assumption of the Motor Theory) for all three modality conditions. Thus, we devised Experiment 2 to investigate whether we could find any modality differences *within* the audiovisual condition when conflicting auditory and visual information were presented at the same time. For instance, if we presented a voice saying /aba/ and a face articulating /aga/ simultaneously (and both syllables were perceived, and not integrated as in the McGurk effect), then the degree of motor interference should depend on the modality in which the information was presented. According to the Motor Theory, motor activation of the speech gestures is independent of the modality in which gestures appear, so response interference should occur regardless of whether it was perceived through the visual or auditory modality. Therefore, response times should be similar when the target is compatible with the visual or auditory speech presented.

In Experiment 3, we took this a step further and investigated at what level the motor commands representing the auditory and visual speech stimuli were activated

during audiovisual integration of the signals, and whether integration was automatic. In this experiment, we utilized the McGurk effect, in which the irrelevant speech stimuli contained incongruent auditory /aba/ and visual /aga/ information that was integrated to form the perception of hearing /ada/, to see whether the individual auditory and/or visual gestures would lead to faster response times when the target was compatible (like Experiment 2) or whether only the target (DA) being compatible with the fused percept would be facilitated. If the results show that DA is faster to produce compared to BA or GA (auditory and visual percepts, respectively), then integration would be assumed to occur automatically and at a stage preceding motor activation. To our knowledge, this is the first behavioural experiment to examine at what stage audiovisual stimuli interact with the motor system during integration.

Lastly, Experiment 4 was designed to test whether static pictures of *intended* speech acts would prime the motor system to produce speech involving the intended gesture (the third assumption). The revised Motor Theory of speech perception as well as research pertaining to mirror neurons have suggested that motor areas are involved in the perception of the 'intended' actions of others, and not only the completed act (Liberman & Mattingly, 1985; Rizzolatti & Craighero, 2004). Thus, if people can perceive the intentions of others by viewing static pictures portraying the intended action (as reported by Nishitani & Hari, 2002), *and* if this is accomplished through internal stimulation of the perceived action (as stated in the Motor Theory), then we predicted that a static picture implying a speech gesture will affect the motor commands used to produce the same or different gesture, just as the dynamic gestures did in Experiment 1.

Ultimately, we hope to provide confirming or disconfirming evidence for the Motor Theory of speech perception and evaluate whether or not it can explain how we integrate auditory and visual speech information. In addition, we aim to provide a behavioural measure for possible mirror neuron activation during speech perception.

Experiment 1

As a preliminary investigation into the perception-production link for speech, we utilized a similar stimulus-response compatibility paradigm as Kerzel and Bekkering (2000). The aim of the proposed study was to establish the reliability of the paradigm by replicating the results found by Kerzel and Bekkering (2000), and to ascertain whether the interference observed was at a stimulus- or response-related stage during processing. Using a within-participants design with the same stimuli and task, we aimed to directly compare verbal responses to manual responses to see if the same pattern emerges when participants make a manual response compared to a verbal response. If this is the case, then the interference is likely to exist at the stimulus level. On the other hand, if verbal responses are differentially affected by the speech stimuli, compared to the manual responses, then interference may be localized at the response level. This finding would suggest that speech perception interacts with the motor system involved in speech production, and would provide additional behavioural evidence that there is an intimate perceptual-motor relationship for speech. In addition, we wished to examine whether the perceptual-motor effects differed across modalities by comparing response times during auditory-only, visual-only, and audiovisual conditions.

Method

Participants

Fourty-two university students (24 females, mean of 20.8 years) participated either for course credit or for an honourarium. All were native speakers of North American English (assessed using the language questionnaire in Appendix A), with normal or correct-to-normal vision, and reported no history of hearing or language impairments. All were right-handed according to the Dutch Handedness Questionnaire (mean score of 31.8, where a score of 34 indicates extreme right-handedness; Van Strien, 1988; see Appendix B). The Wilfrid Laurier University Research Ethics Board approved the procedures, and all participants gave written informed consent before participation.

Stimuli and Design

Stimuli were videos of a male speaker (from shoulders up) producing the vowel-consonant-vowel (VCV) syllables /aba/ and /aga/ (see Figure 1a for examples of the stimuli). An orthographic target (either 'BA' or 'GA' in black letters, Arial font) extending from the bottom lip to the top lip of the speaker when in a resting position (subtending approximately 2° of visual angle), appeared for three frames (~ 100 ms); one frame before, during, and after the consonant burst (see Figure 1b for a schematic timeline). Videos were 720 x 480 and viewed at an unrestrained distance of 80 cm. The videos were displayed on an IBM flatscreen LCD monitor (screen resolution of 1024 x 768 and refresh rate of 60 Hz) at 29.97 fps. The audio was heard through circumaural Sennheiser headphones (model HD 580 Precision) at an average of 60 dB (SPL). A DirectIN custom response box (Empirisoft Corp.) was used to collect manual responses and an AKG condenser microphone (model C 420 PP) was used to record the verbal

responses. The computer was equipped with a high quality sound card (Sound Blaster Audigy 2 ZS Platinum; Creative Technology Ltd.) to play the sound and record the voice responses accurately.

Three experimental conditions were presented randomly: an audio-only (AO) condition, in which the voice was heard but only a still-face was seen; a visual-only (VO) condition, in which the mouth movements of the speaker were seen but nothing heard; and an audiovisual condition, in which the mouth movements and the voice corresponding to the VCV syllable were presented. We also included a baseline control condition in which targets were presented over a still-face image of the speaker at rest, with his mouth closed. Note that the still-face was also presented during the AO condition. Nested within each condition, were compatible (speech stimuli and target letters matched) and incompatible (speech and target mismatched) trials. Each stimulus combination (see Table 1.1) was presented randomly five times for a total 70 trials per session, 140 trials in total.

Procedure

Participants sat in a dimly lit, sound attenuated chamber (Industrial Acoustics Company, Inc.) in front of a computer monitor. There were two experimental sessions: one requiring a voice response and another requiring a manual (button-press) response. The order of the sessions was counterbalanced across participants. Participants were instructed to ignore the irrelevant speech stimuli and to pronounce the relevant target (either 'BA' or 'GA') into a head-mounted microphone as soon as it appeared on the screen. During the manual sessions, they were instructed to indicate which target was presented by pressing the corresponding key labeled 'BA' and 'GA' on a button-box

(key order was counterbalanced). During both sessions, it was stressed that responses should be made as quickly and accurately as possible. Participants pressed the space bar to begin the next trial and each began with a prestimulus cue comprised of a row of x's presented for two seconds. Sessions took approximately 20 minutes with a five-minute break in between.

Results

Only correct responses were the focus of the analysis. See Table 1.2 for the percentage of incorrect responses for each condition across the manual and verbal response types (2.7% and 1.59% respectively). We classified trials with response times less than 200 ms and greater than 1000 ms, as anticipatory and neglected responses and removed them from the analysis (3.05% for manual and 1.03% for verbal). In total, very few observations were excluded (5.75% for manual and 2.62% for verbal), ruling out the possibility of a speed-accuracy tradeoff. On average, verbal responses were slower than manual responses (M = 560 ms, SD = 136 ms, 494.9 ms, SD = 128.65 ms, respectively). We therefore assessed facilitation and interference by subtracting the average response rate observed during baseline trials (for the respective response types) from the average responses during the experimental conditions for each participant.

A 2 (response type: verbal or manual) x 2 (compatibility: congruency between speech stimuli and target) x 2 (target: BA or GA) x 3 (modality: AO, VO, or AV) repeated measures analysis of variance (ANOVA) was conducted on the difference scores. The average response times relative to the baseline (i.e., 0) for each condition is shown in Figure 2.

The ANOVA revealed significant main effects of compatibility ($F(1, 41) = 44.69$, $p < .001$) and modality ($F(2, 82) = 21.62$, $p < .001$). As shown in Figure 2, response times were faster when the speech and target were compatible. As well, it appeared that responses were overall slower in the audiovisual condition compared to the audio-only and visual-only conditions. There was no main effect of response type ($F(1, 41) = 2.30$, $p = .137$). However, there was a reliable interaction between response type and compatibility ($F(1, 41) = 5.35$, $p = .026$), suggesting that the verbal responses were more affected by the compatibility between the target and speech stimuli, than were the manual responses. There was also a significant interaction between response type and modality ($F(2, 82) = 6.18$, $p = .003$). Post hoc comparisons (LSD) further indicated that response times in general were greater during the audiovisual condition when verbal responses were required ($p < .01$). Planned orthogonal contrasts revealed marginal compatibility differences between the response types for the visual-only condition, compared to the auditory-only and audiovisual conditions ($F(1, 41) = 4.03$, $p = .051$). It appeared that only the verbal responses were affected during the visual-only condition. In fact, paired t-tests (Bonferroni corrected to an alpha of .025 for multiple comparisons) showed that observing the same speech syllable facilitated pronunciation as compared to the baseline, $t(41) = 2.35$, $p = .024$ ($M = -14.31$ ms), whereas observing a conflicting speech syllable caused interference, $t(41) = -4.04$, $p < .001$ ($M = 20.55$ ms). Although, the manual response data appeared to follow a similar trend, these differences were not significant (Figure 2b).

Discussion

Our results suggest that observing a face and/or voice saying a speech syllable improves the speed of producing the same syllable, compared with producing a different syllable. For both the manual (excluding the visual-only condition) and verbal conditions, response times decreased when the irrelevant speech produced by the speaker in the video matched the target syllable that participants produced. For instance, participants produced the target 'BA' faster than the target 'GA' when they perceived the speaker utter /aba/. This effect was more prominent for the verbal responses, however the trend in the manual response data seemed to resemble that seen in the verbal response condition.

The most notable difference observed between the response types was that verbal responses showed a compatibility difference during the visual-only trials, whereas manual responses did not. In fact, we found response facilitation (compared to a baseline) when participants viewed the speaker in the video articulating the same target syllable, and response inhibition when the speaker articulated a different target syllable. Our results are consistent with Kerzel and Bekkering (2000), who also found a compatibility effect when visual speech was presented. However, because they did not include a baseline, the authors were unable to determine whether the reaction time differences were due to facilitation or differential interference. Our results now provide this essential information further supporting that the visual perception of speech gestures is processed at the response level, and either aids in the production of compatible speech gestures (facilitation), or inhibits the production of incompatible gestures (interference).

Interestingly, the only modality that failed to show any response interference for the manual response type was the visual-only condition. This result is at odds with Kerzel

(2002), who had participants make button-press responses and showed the same compatibility effects as previously observed for verbal responses. Based on this evidence, Kerzel concluded that interference was initiated at the stimulus level, in that visual speech interacted with processing the target letters. Our results on the other hand, suggest that interference may be localized at the response stage. We directly compared response times across the two response types using the same stimuli and task, and found significant compatibility effects when the responses were verbal, but minimal interference when the responses were manual. This difference in response types suggests the compatibility effects for the visual-only condition were indeed response related.

Surprisingly, we observed interference effects for both compatible and incompatible trials during the auditory-only and audiovisual conditions regardless of whether the response was verbal or manual. Thus, it seems that auditory stimuli disrupted responses differently than visual stimuli. One possibility is that only certain phonetic features that are available within the stimulus modality interact with the motor system. For instance, Gordon and Meyer (1984) found that only the voicing feature[3] presented in the auditory syllable affected vocal response times and not the place of articulation feature. Our findings, as well as those of Kerzel and Bekkering (2000), showed that the place of articulation presented in the visual stimulus affected vocal responses, however neither of these studies included stimuli that differed in voicing to test this hypothesis. Perhaps a future endeavour could compare across modalities and using syllables that contain different phonetic features to determine if this were true. For instance, examine whether the syllable /pa/ and /ga/ differ during the auditory-only conditions since they

differ in the voicing feature[3] (/pa/ is voiceless whereas /ga/ is voiced). Similarly, examine whether the two syllables would differ during the visual-only condition because they differ in place-of-articulation (/pa/ is bilabial and /ga/ is velar). If this were the case, it would speak against a Motor Theory view that the same abstract phonetic information is extracted equally from the auditory and visual modalities in the form of motor gestures (Liberman & Mattingly, 1985).

Another possibility could be that auditory speech in general does not generate motor activation (cf. Gordon & Meyer, 1984). Sundara, Namasivayam, and Chen (2001) used TMS to demonstrate modality differences when observing speech stimuli, and discovered that visual speech elicited strong motor-evoked potentials in the muscles used to produce the speech, but auditory speech did not (but also see Fadiga et al., 2002 and Watkins, et al., 2003). Because lipreading is inherently difficult, imagining the corresponding auditory speech may help to more accurately identify the message. This hypothesis is plausible considering some recent studies showing brain activation in auditory areas when participants silently lipread (e.g., Calvert et al., 1997; Mottonen, Krause, Tiippana, & Sams, 2002; Sams, et al., 1991).

Finally, perhaps the most parsimonious explanation for why the auditory modalities showed exaggerated effects might be that the voice simply attracted attention away from the visual target (letters BA and GA), resulting in an overall delayed response. In fact, some participants did report feeling distracted when they heard the voice but saw no facial movement. This attentional capture by the voice would also have increased the

[3] Voicing refers to the sound produced when the air passes through the vocal cords as they vibrate. For voiced consonants, the vocal cords vibrate and sound is heard, whereas during voiceless consonants the vocal cords do not vibrate.

level of processing of the irrelevant auditory stimulus, which in turn might cause increased interference at the response selection stage for the incompatible targets. Likewise, it is also possible that the increased interference during the audiovisual condition was caused by the irrelevant auditory and visual speech competing with the target for attention. Since speech forms a coherent audiovisual event, participants' visual and auditory attention might have been diverted from the target, resulting in delayed response times to the targets in general, for both verbal and manual responses Thus, Experiment 2 was devised to try to eliminate this distracting effect by examining any modality differences using only audiovisual stimuli.

Experiment 2

Even though the results of Experiment 1 showed that there was some effect at the stimulus-response level, it was not clear as to why more interference occurred when the auditory speech was present (particularly in the audiovisual condition), even when the target was compatible with the irrelevant speech. A possible explanation was that the audiovisual stimuli captured the participant's attention, causing a delay in their response. One way to exclude this possibility was to compare between modalities while keeping the amount of distraction constant. Therefore, Experiment 2 used solely audiovisual stimuli to investigate whether there were any stimulus-response differences across the visual and auditory modalities. This was achieved by creating videos that contained conflicting auditory and visual speech information (similar to that used for the McGurk effect). For example, one video contained an auditory /aba/ presented simultaneously with a visual /aga/ and the other video contained the reverse pairing. The effects of the incongruent

videos were compared to congruent videos, where the auditory and visual were the same (/aba/ and /aga/). The purpose of this experiment was to minimize any distracting factors that could explain the compatibility differences seen in Experiment 1, and examine if response differences would be found depending on the modality that contained the compatible information. In particular, whether there are differences in response time when the speech information is compatible with the visual versus the auditory signal.

According to the Motor Theory, both the visual and auditory modalities are thought to activate the motor commands to the same degree, implying that the response times to produce the targets should not depend on the modality in which the speech gestures were presented (assuming that both signals are processed at the same time). If this were the case, then response times should be similar in both the incongruent conditions for the targets BA and GA (compatible with auditory and visual signals), compared to DA (incompatible with both). The results from Experiment 1 however, suggest that there may be differences in motor activation depending on whether the information is available in visual or auditory modality, such that only the visual modality was shown to produce response facilitation. This suggests that response times might be faster when the target is compatible with the visual signal, and longer for the auditory signal. This would contrast with the Motor Theory by demonstrating that there are differences in motor activation depending on the modality that the information is available.

Method

Participants

Twenty-five Wilfrid Laurier University students (13 females, mean of 19 years) participated either for course credit or for a honourarium. All were native speakers of North American English, with normal or correct-to-normal vision, and no history of hearing or language impairments. All were right-handed (mean score of 31.4 on the Dutch Handedness Questionnaire; Van Strien, 1988). The Wilfrid Laurier University Research Ethics Board approved the procedures, and all participants gave written informed consent before participating.

Stimuli and Design

The stimuli were videos of a male (same as Experiment 1) uttering the nonsense bisyllables /aba/ and /aga/ and presented either congruently (visual and auditory /aba/ and /aga/) or incongruently (visual /aba/, auditory /aga/, or visual /aga/, auditory /aba/). The incongruent stimuli were created by aligning the acoustic burst of the consonant contained in the auditory syllable with that of the visual syllable. Typically when presented with conflicting auditory /aba/ and visual /aga/ stimuli, participants would report an integrated percept (/ada/), known as the McGurk effect. However, for this experiment, we wanted the visual and auditory information to be perceived simultaneously without being integrated. A perceptual experiment on a separate group of participants confirmed that the auditory and visual syllables were perceived individually and no combined percept was elicited. Since speech perception varies considerably across stimuli, it is not surprising that certain audiovisual stimuli will produce the McGurk

effect and some will not. Thus, the lack of integration just happened to be a product of the stimuli used.

The videos were the same as in Experiment 1 and displayed on a flatscreen monitor at 29.97 fps and the audio was provided through the same circumaural headphones. The computer was equipped with a high quality sound card (Sound Blaster Audigy 2 ZS) to play the sound and record the voice responses. During each video, one of three orthographic targets (either 'BA', 'GA', or 'DA') appeared for three frames (~ 100 ms); one frame before, during, and after the consonant burst. Due to the nature of the conflicting stimuli, the compatibility between target and irrelevant speech was now more complex. For the incongruent AV videos, the target 'DA' served as the incompatible target, while 'BA' and 'GA' were both compatible (either with the auditory or visual channel). Note, we chose DA as the incompatible target to be consistent with the stimuli in Experiment 3 where participants did perceive the McGurk effect, and DA served as the congruent target with the illusion /ada/. Like Experiment 1, a baseline control condition in which targets were presented with a still-face was also included. Refer to Table 2.1 for an outline of each stimulus type. Each trial was randomly presented five times, which amounted to 75 trials in total.

Procedure

Participants sat in a dimly lit, sound attenuated chamber, equipped with a computer monitor and were required to wear a head-mounted microphone and headphones. They were asked to ignore the irrelevant speech stimuli (i.e. the man in the video) and pronounce the relevant target (either 'BA', 'GA', or 'DA') into the microphone as soon as it appeared on the screen. It was stressed that responses be made

as quickly and accurately as possible. Each trial began with a prestimulus cue comprised of a row of x's presented for two seconds and the next trial began immediately after their response was recorded. Sessions took approximately 15 minutes to complete.

Results

The analysis was similar to Experiment 1, in that only correct responses were the focus of the analysis. Only 2.8% of responses were incorrect (see Table 2.2). Again, anticipatory and neglected trials were classified as response times less than 200 ms and greater than 1000 ms, respectively (0.93%). In total, this amounted to very few observations (3.73%). We assessed facilitation and interference by subtracting the average response rate during the baseline trials from the average responses during the experimental conditions for each participant. This resulted in a set of difference scores for each condition.

A 4 (condition: congruent /aba/ and /aga/, and incongruent V/aba/, A/aga/ and V/aga/, A/aba/) x 3 (target: BA, GA, DA) repeated measures ANOVA was conducted using the difference scores. An illustration of the results can be seen in Figure 3. The only significant finding was the two-way interaction between condition and target ($F(6, 144) = 2.537, p = .023$). Post hoc comparisons (LSD) revealed that response times were quicker when the target matched the auditory or visual gestures presented in the video. For instance, during the congruent /aba/ condition (for which the auditory and visual was /aba/) response times were significantly faster to produce the target BA than DA ($p < .05$), and faster for GA for the congruent /aga/ condition, compared to DA ($p < .001$) and BA ($p < .001$). This was also the case for both of the incongruent conditions, where

conflicting /aba/ and /aga/ syllables were presented simultaneously. We found quicker response times for the targets BA and GA, than for the target DA ($p < .05$). In fact, the response time to produce DA ($M = 43.69$ ms compared to baseline) was almost double that of BA or GA ($M = 23.46$ ms compared to baseline). No significant differences were found between the BA and GA productions for the incongruent conditions ($p > .05$). Despite this, visual inspection of the incongruent conditions suggests a subtle, but interesting pattern, where the response times seem to be a little faster when the target is compatible with the visual speech compared to the auditory speech.

Discussion

The results of Experiment 2 demonstrated that people were quicker to produce speech gestures when they matched with the observed speech gestures in the video. This was seen for both the congruent and incongruent conditions presented. For example, the congruent stimuli elicited faster vocal response times when the target (e.g., BA) was compatible with the irrelevant auditory and visual speech stimuli (e.g., audiovisual /aba/), than when it was incompatible (e.g., audiovisual /aga/). This finding replicated the pattern found in Experiment 1, where response time decreased when the target and irrelevant audiovisual stimuli were compatible.

However, the overall aim of this experiment was to examine whether conflicting information provided by the auditory and visual modalities would affect verbal responses to targets that were compatible with only one. Thus, we were mainly interested in the incongruent conditions, where participants either saw /aba/ and heard /aga/, or saw /aga/ and heard /aba/ simultaneously. Our results showed that participants' response times were

equally influenced by the auditory and visual information. For both of the incongruent conditions, where /aba/ and /aga/ signals were presented simultaneously, response times to produce the targets BA and GA were significantly faster than to produce an incompatible target DA (for which errors were also more prominent). This finding suggests that both the visual and auditory modalities are processed concurrently (even if not integrated) and can produce similar effects at the response stage, when the production of compatible and/or incompatible speech gestures are required. Thus, facilitation occurred for the compatible speech gestures and interference occurred for the incompatible speech gestures, regardless of whether the gestures were presented visually or aurally. Note also that the effects were comparable to the congruent stimuli, despite the distracting (and possibly masking) effect that the conflicting gesture might have had on the perception of the compatible modality. This suggests that perhaps each modality was being processed concurrently and independently.

Furthermore, the similarity in response times between the congruent and incongruent conditions suggests that the lack of difference between the modalities during the incongruent conditions is not likely caused by a differential weighting of concurrent facilitation and interference. Because the incongruent conditions contained speech information that was compatible *and* incompatible with the targets BA and GA, the similarities in response times to produce those targets may have been caused by an average of facilitation of the compatible target and interference of the incompatible target. However, both incongruent conditions showed comparable effects as the congruent conditions, where both modalities were compatible with the target and interference should not have occurred. Thus, the similarities in response times to produce

the targets when they were compatible with the speech perceived, suggests that both modalities can contribute equally at the response stage during speech perception, and can have a similar influence on speech production.

Finally, our findings from Experiment 2 are interesting in light of Experiment 1, where the audiovisual and auditory-only conditions produced more interference than the visual-only condition. This suggested that visual speech might have interacted with the motor system differently than that of auditory speech. Here, we found this not to be the case once the modalities were compared within the same conditions using all audiovisual stimuli. Therefore, it is plausible that the auditory stimuli used in Experiment 1 distracted participants and delayed their responses to the targets, whereas the visual stimuli did not. Perhaps another way to show this might be to compare the auditory-only and audiovisual conditions to an auditory baseline condition (for example a speaker saying /aaa/), as opposed to a visual still-face. This way, any extraneous effects caused by the auditory stimuli could be subtracted out. Thus, the modality differences we found in Experiment 1 for the auditory-only and audiovisual conditions compared to the visual-only condition might have been because we used an inappropriate baseline stimulus to compare the auditory stimuli to (where the still-face baseline was appropriate for the visual-only). Had we used a more appropriate baseline, we might have observed similar facilitation and interference effects for the auditory conditions as that observed for the visual-only condition. In sum, the results from this experiment using only audiovisual stimuli suggest that both the visual and auditory modalities can similarly affect speech production.

Overall, our findings are consistent with the Motor Theory (Liberman & Mattingly, 1985), which stated that the 'phonetic module' extracts all speech gestures

regardless of the modality in which they are presented in (visually, aurally, or haptically), and maps them onto their corresponding motor commands. However, the Motor Theory fails to explain how and when *integration* of the modalities occurs with respect to motor activation. This question could not be answered in the present experiment, since the audiovisual stimuli were created in such a way that participants could accurately identify the individual auditory and visual signals when presented simultaneously. In other words, integration of the auditory and visual information did not occur, for which any conclusions regarding integration could not be made. Therefore, Experiment 3 was designed to examine at what stage during audiovisual speech integration the motor commands would become activated by using stimuli that elicited the McGurk Effect (McGurk & MacDonald, 1976).

Experiment 3

We found from Experiment 2 that the perception of speech gestures interacted with the motor system in the same way regardless of whether the gestures were presented in the visual or auditory domain. This finding is consistent with the Motor Theory view that we perceive speech in a gestural code available through all of the senses to the same degree (Liberman & Mattingly, 1985). Yet, one crucial explanation that the Motor Theory does not discuss is *when* during audiovisual speech integration, are the auditory and visual speech gestures mapped onto their corresponding motor commands. Are the motor commands representing each of the modalities activated individually and then used to further activate the motor command representing the integrated gesture? Or does integration occur before motor activation, such that only the motor command

representing the integrated gesture would receive activation? As a result, Experiment 3 was designed to examine the question of whether integration occurs *prior* to motor mapping, such that the integrated percept activates the motor command representing the integrated percept, or whether integration occurs *after* the gestures presented in each modality are individually mapped onto their corresponding motor commands. In order to investigate this, we took advantage of stimuli that reliably elicited the McGurk effect (McGurk & Macdonald, 1976). For example, we created a video of a person saying /aba/ while their face articulated /aga/, for which the integrated percept of /ada/ was produced. It was predicted that if the integration of auditory and visual speech gestures preceded response activation (i.e., motor activation), then we should find faster response times when pronouncing the illusion percept (DA), than the gestures contained in the auditory and visual signals alone (BA and GA respectively). Yet, if response activation occurs before the integration stage, then faster response times should be seen for the auditory and visual gestures (BA and GA) over the combined percept (DA).

Method

Participants

Fourteen Wilfrid Laurier University students (12 females, mean 18.9 years) participated either for course credit or for a honourarium. They were selected out of a group of 82 volunteers that participated in a perceptual experiment, where the stimuli used were identical to the stimuli used for the Stroop-like experiment, except that the task was to indicate on the keyboard what was said by the speaker in the video (either /aba/, /aga/, /ada/, or /abga/). Participants were chosen on the criteria that they perceived the

McGurk effect on 80% (4 out of 5) of the trials. For instance, they reported perceiving /ada/ for the fusion trials and /abga/ for the combination trials. Since the McGurk effect is a great tool for demonstrating audiovisual integration, this screening procedure made certain that the stimuli used in the following Stroop-like experiment were being integrated.

All participants were native speakers of North American English, with normal or correct-to-normal vision, and no history of hearing or language impairments. All were right-handed (mean score of 31.4 on the Dutch Handedness Questionnaire; Van Strien, 1988). The Wilfrid Laurier University Research Ethics Board approved the procedures, and all participants gave written informed consent before participating.

Stimuli and Design

In order to create stimuli that strongly produced the McGurk effect, we recorded a male and female speaker producing the nonsense bisyllables /aba/ and /aga/ using a Sony HD digital camcorder (model HDR-FX1). The stimuli were created in the same way as Experiment 2, either to be congruent (visual and auditory /aba/ and /aga/) or incongruent (visual /aba/, auditory /aga/, or visual /aga/, auditory /aba/) by aligning the acoustic burst of the consonant contained in the auditory syllable with that of the visual syllable (see Figure 4). However, the incongruent stimuli used in this experiment elicited two different perceptions. The video containing an auditory /aba/ paired with a visual /aga/ produced the perception of hearing /ada/ (a McGurk fusion), and the other video containing an auditory /aga/ paired with a visual /aba/ produced the perception of hearing /abga/ (a McGurk combination; McGurk & MacDonald, 1976).

The 720 x 480 videos were displayed on a Phillips 19" flatscreen monitor (screen resolution of 1024 x 768, refresh rate of 60 Hz) at 29.97 fps controlled by an LG computer with an Intel Pentium IV processor. The audio was provided through Sennheiser (model HD 580 Precision) circumaural headphones at an average of 60 dB (SPL). During each video, one of three orthographic targets (either 'BA', 'GA', or 'DA') appeared for three frames (~ 100 ms); one frame before, during, and after the consonant burst. The targets appeared in black, Arial font, measuring approximately 2^{o} of visual angle, from the bottom to the top lip of the speaker while in a resting position. The compatibility between the target and the irrelevant speech was different than Experiment 2, as the target could now be compatible with both the auditory and visual /ba/ or /ga/, or with the integrated percept /da/, depending on whether integration occurred before or after the audiovisual signals were processed at the response stage. Therefore, we left the compatibility of the target and speech stimuli undefined for the incongruent conditions. Consistent with the previous experiments, we included a baseline control condition where we presented each of the targets over a still-face of both speakers. DirectRT randomly presented each trial four times, for 120 trials in total.

Procedure

The procedure was the same as Experiment 2, for which participants sat in a dimly lit, sound attenuated room, in front of a computer monitor and were required to wear a head-mounted microphone and headphones. They were instructed to ignore the irrelevant speech stimuli (i.e. the man in the video) and pronounce the relevant target (either 'BA', 'GA', or 'DA') into the microphone as soon as it appeared. Since response times were being recorded, it was stressed that responses be made as quickly and

accurately as possible. Each trial proceeded with a prestimulus cue (a row of x's) presented for two seconds followed by the video. The next trial began immediately after their response was recorded and the sessions took approximately 25 minutes to complete.

Results

The analysis was also identical to that of Experiment 2, where we analyzed only correct responses. There were very few incorrect responses (0.53%) and were not examined further. Again, we classified response times less than 200 ms and greater than 1000 ms, as anticipatory and missed trials respectively (1.14%). Only 1.67% of the total data set was excluded from the analysis. Just like the previous experiments, we assessed facilitation and interference by subtracting the average response rate during the baseline trials from the average responses during the experimental conditions for each participant, resulting in a set of difference scores for each condition.

A 2 (speaker: male and female) x 4 (condition: congruent /aba/ and /aga/, and incongruent V/aba/, A/aga/ and V/aga/, A/aba/) x 3 (target: BA, GA, DA) repeated measures ANOVA was conducted on the difference scores. The ANOVA revealed a significant interaction between speaker and condition ($F(3, 45) = 7.57$, $p = .0003$), showing that the pattern of response times across each of the conditions was different depending on the speaker in the video. The data for each speaker can be seen in Figure 5 (5a for the male and 5b for the female).

Two separate 4 (condition) x 3 (target) ANOVA's were conducted for each speaker to investigate the individual patterns. For the male speaker (see Fig. 5a), no significant effects were found across the conditions and targets. However, for the female

speaker (see Fig. 5b) the ANOVA yielded a significant main effect of condition ($F(3, 45) = 12.037, p < .001$). The following post hoc tests (LSD) showed that response times were reliably faster during the fusion condition (V/aga/ A/aba/) and the congruent /aga/ condition ($p = .029$), than during the combination condition (V/aba/ A/aga/, $p = .001$). No other significant differences were found ($p > .05$).

Discussion

The findings from this experiment suggest that there appeared to be more interference (i.e., longer response times) to produce the targets overall when the female speaker was seen articulating the bisyllable /aba/ than when she articulated /aga/. However, response times to produce each of the targets were the same across all the conditions. This suggests that response times did not differ depending on the compatibility between the target and the irrelevant speech stimuli - a finding that is contradictory with the first two experiments. Due to the null results from Experiment 3, we were not able to identify the stage at which the motor commands for the auditory and visual signals were activated during audiovisual integration.

The lack of significant differences across the target and conditions could be attributed to the variability caused by the stimuli. Speech stimuli are very complex and the McGurk effect is known to vary across individuals (Massaro, 2004). This was even demonstrated in the present study by the fact that the McGurk effect was produced in only 14 out of the 82 participants tested in the perceptual experiment. Therefore, the stimuli we used were not very consistent, nor optimal for producing the McGurk effect. This is a problem for the majority of research conducted on speech perception, where the

stimuli vary considerably across experiments and laboratories. Perhaps using a simulation of speech stimuli (instead of a speaker in real-time) would reduce this variability in future research.

Another possible explanation for our null findings might be due to the extreme variation in response latencies within the participants and was most likely caused by the recording equipment used during the experiment. The computers used were not equipped with high quality sound cards like the computers used in Experiments 1 and 2 (Sound Blaster Audigy) and a lot of noise was created in the voice response data when response times were being collected through the microphone. This noise made it difficult to identify the signal (i.e., voice response) with great accuracy. Therefore, this experiment would benefit from using the same equipment as the previous experiments, whereby the signal can be accurately defined and clear differences (or lack of differences) across conditions can be observed. Future studies using more precise measurements and reliable stimuli need to be conducted to ascertain with confidence whether differences exist across conditions or not.

Yet, despite this lack of statistical significance, visual inspection of Figure 5b suggests that target differences could exist across the conditions. The most notable difference observed, was that the time to produce the target DA was consistently quicker than the time to produce BA and GA. This trend contrasted with the pattern observed in Experiment 2, where DA was shown to produce more interference than BA and GA. These observations are interesting given that audiovisual integration did not occur in Experiment 2 (i.e., no McGurk effect) but did occur in this experiment (i.e., elicited McGurk effects). Although these pattern differences could certainly be due to the

different stimuli used in the two experiments, another possibility could be that the fusion condition in the present experiment (where the perception of /ada/ was elicited) acted as a compatible stimulus with the target DA, and thus facilitated the response time. If the latter was true, it would suggest that integration of the auditory and visual speech signals occurred *before* the motor commands representing the integrated percept of /ada/ were activated. Even though not explicitly stated in the Motor Theory, it can be assumed that integration would occur after the motor commands for each of the signals have been activated. According to the Motor Theory, the 'phonetic module' *automatically* extracts the gestures from the speech stimuli and maps them onto their motor representations, so integration would have to occur after this motor-mapping process. Since the speech signals are represented in a gestural code, it provides a 'common currency' (Fowler, 2004) for which the auditory and visual signals can be integrated. However, the pattern observed in the present data contrast with the Motor Theory view by suggesting that audiovisual integration might occur prior to motor activation. Thus, integration might rely on a different mechanism before reaching the response stage during speech processing.

In sum, the results of Experiment 3 were unable to clearly answer any questions regarding the stage at which the motor commands were active during audiovisual speech integration, yet it was useful in illuminating an interesting relationship between audiovisual integration and speech production. In order to accurately investigate whether this is a valid and reliable relationship, future studies should replicate this experiment using the proper equipment for recording voice responses and more reliable speech

stimuli. Until then, we can only speculate the level at which motor activation occurs during audiovisual speech integration.

Experiment 4

Experiments 1, 2, and 3 tested two of the main assumptions offered in the revised version of the Motor Theory (Liberman & Mattingly, 1985): that speech perception is closely connected to speech production, and that audiovisual integration occurs automatically and following the activation of the speech motor commands. Finally, in this last experiment, we wished to examine a third assumption that we perceive speech by encoding the *intended* gestures of the speaker used during speech production (not the observable movements of the vocal tract). According to the Motor Theory, this is how we are able to overcome the problem of coarticulation in speech - where the movements of several phonetic segments overlap during speech production, yet we are still capable of maintaining perceptual constancy of the phonetic gesture. Thus, it is believed that the listeners' 'phonetic module' detects the abstract *intended* gestures produced by the speaker, which in turn activates the *invariant* motor commands representing those gestures in memory (Liberman & Mattingly, 1985).

This assumption made by the Motor Theory closely parallels some of the hypotheses surrounding the function of mirror neurons. Using single-cell recording of area F5 in monkeys, Umilta et al. (2001) demonstrated mirror neuron firing when an experimenter made a grasping motion to pick up an object, and when the monkey picked up the object itself. Interestingly, mirror neurons fired when the monkey observed an experimenter grasp behind a curtain (presumably to pick up the object hidden behind it),

but not when the experimenter was only seen making a grasping motion with no object present. The only difference between these two conditions was the perceived intention of the observed act – to pick up the object or not. This finding led the authors to propose that mirror neurons could function to encode the intentions of the actor observed. This is consistent with the views of Rizzolatti and Craighero (2004), who in a recent review suggested a possible role of mirror neurons may be to help us understand the underlying intentions of others behaviour by internally simulating them ourselves. Moreover, Skipper et al. (2005) claimed that this type of 'resonance' mechanism that mirror neurons afford could be used to aid in speech recognition, by matching the intended gestures of the speaker to the listeners' motor counterparts and narrowing the possible interpretations of the speech utterance for the listener. This would be very beneficial in a situation where the intentions of the speaker are not explicit and need to be inferred, like in photographs.

In a recent fMRI study by Calvert and Campbell (2003), similar cortical activation was found when participant's viewed static images of actors producing speech gestures compared to dynamic speech gestures, albeit the dynamic faces showed stronger activations. The areas found were predominantly more active in the left hemisphere and included the inferior frontal regions (Broca's area), superior temporal sulcus, and posterior areas (known for processing biological motion). However, the most intriguing finding was that the static speech images produced greater activation in the ventral premotor areas and the intraparietal sulcus, where mirror neurons are believed to be located (Gallese, Fadiga, Figassi, & Rizzolatti 1996; Gallese, Fogassi, Fadiga, & Rizzolatti, 2002). One possibility for the increased activation found in these areas may be that the static pictures were more difficult to interpret than dynamic speech gestures, and

therefore needed to rely more on a 'resonance' system to covertly imitate the implied gesture to infer the speakers' intentions. Consistent with this, a magnetoencephalographic (MEG) study conducted by Nishitani and Hari (2002) found similar cortical activation in Broca's area and primary motor areas when participants viewed static pictures that implied verbal or non-verbal mouth movements, and when participants imitated the mouth movements themselves. These findings suggest that not only dynamic speech gestures, but still pictures portraying gestures, can evoke the same pattern of cortical activation (presumably that of mirror neurons), and that this mechanism may be used to encode the underlying intentions of the speaker.

The purpose of this experiment was to investigate whether the perception of static images portraying verbal and non-verbal facial gestures would show similar response interference as the dynamic visual speech did in Experiment 1. To that end, we compared response times to produce speech gestures that were either congruent or incongruent with an *implied* gesture of a speaker presented in a photograph. Participants viewed still pictures of an actor producing either a speech gesture (consonant) or a non-speech gesture (orofacial gesture, such as tongue protrusion). The non-speech gestures were used to investigate whether the motor system was simply activated by the gesture observed (involving the use of the teeth, lips, and tongue), or whether the picture needed to imply speech in order to activate the speech motor system, as postulated by the Motor Theory ('phonetic module'). Hence, we compared pictures portraying non-verbal gestures to those depicting verbal gestures in order to examine whether or not response interference differed depending on the intentions of the actor (e.g., to speak or not to speak). Since static facial images have been shown to produce similar activation to dynamic faces

(Calvert & Campbell, 2003) - especially in motor areas responsible for speech production planning - we predicted that faster response times would be observed when participants produced the same gesture that was portrayed by the static image (compatible gesture), than when the gesture was not portrayed in the image (incompatible gesture). Moreover, if it is the case that we process the underlying intentions of the speaker during speech perception (according to the Motor Theory), then we should find significant response time differences between the verbal and non-verbal gestures. That is, only the images implying speech gestures should have a significant influence on speech production.

Method

Participants

Twenty-nine Wilfrid Laurier University students (three men, mean age of 19) participated for course credit or a honourarium. All were native speakers of North American English, with normal or correct-to-normal vision, and reported no history of hearing or language impairments. All were right-handed (mean score of 31.4 on the Dutch Handedness Questionnaire; Van Strien, 1988). The Wilfrid Laurier University Research Ethics Board approved the procedures, and all participants gave written informed consent before participation.

Stimuli and Design

A male and female actor were recorded using a Sony HD digital camcorder (model HDR-FX1) while producing phonemes and making facial gestures. The frame that signified the most robust facial gesture (characteristic of the *intended* action according to the experimenter) was chosen and used for the static pictures. The stimuli

included three pictures of speech-related gestures portraying the phonemes, [la], [pa], and [va], three pictures of non-speech gestures portraying the orofacial gestures, [licking lip], [protruding lips], and [biting lip]. We also included a neutral picture of each actor providing no gesture information, which served as a baseline control image (see Figure 6a for examples of the stimuli used). Non-speech gestures were used to examine whether any effects of the stimuli are due to processing the observed articulators used by the actor in the picture (i.e., lips, tongue, teeth, etc.) or whether the effects are due to processing of the underlying intentions of the actor (speech versus non-speech). Thus, the non-speech facial gestures were chosen based on the criteria that they used similar articulators as the speech gestures, yet depicted a non-speech act. For instance, the non-speech gesture [biting lip] was chosen because it incorporated similar articulators as the speech gesture [va] (where the teeth touch the bottom lip), yet did not imply a speech act. The neutral picture (baseline) was of the actor in a resting position with mouth closed, affording no gesture information. The baseline was used for the same reasons as the previous experiments, to get a measure of each of the participants' response rate to producing the targets. Thus, there were seven pictures from each actor, resulting in 14 pictures in total.

The pictures were displayed on a Philips 19" flatscreen monitor (screen resolution of 1024 x 768, refresh rate of 60 Hz) controlled by an LG computer housing an Intel Pentium IV processor. The pictures were 600 x 800 pixels and the orthographic target letters (LA, PA, or VA) were in presented in black, Arial font, measuring approximately 2° of visual angle. Using DirectRT (Empirisoft Corp.), we displayed the static pictures for 500 ms and had the target letters flash over the actors' mouth for 100 ms in the middle of the presentation (see Figure 6b for a schematic timeline). An AKG condenser

microphone was used to record the verbal responses, which were collected using DirectRT and stored for later offline analysis. Response times were determined by hand using MatLab (The Mathworks Inc.).

Procedure

The procedure was the same as in the previous experiments, in which participants were seated in a sound attenuated room in front of a computer monitor unrestrained at a distance of approximately 80 cm. They were equipped with a head-mounted microphone and headphones (to keep the microphone in place and reduce any noise). Although, for this experiment they were asked to observe the pictures of the man and woman making facial gestures (in Experiments 1, 2, and 3 they were asked to ignore the speech stimuli) and pronounce the target that appeared over the mouth (LA, PA, or VA) into the microphone, as quickly and accurately as possible. Since we were interested in the underlying intention of the image presented, we wanted to make certain that the images were being processed (i.e., not ignored) and the intention perceived. Thus, any differences in response time between the gestures portrayed could not be attributed to one gesture capturing more attention than another because all gestures were equally attended to.

Each trial began with a row of crosses (+++++) in the center of the screen for one second as a fixation, followed by a static picture for 200 ms, then the same picture *with the target* for 100 ms, and then the picture again for 200 ms (see Figure 6b for a schematic timeline). Once the participant responded, the next trial began immediately. Each trial was randomly presented four times for 168 trials in total. Each session lasted about 30 minutes

Results

Similar to the previous experiments, only correct responses were the focus of the analysis. Very few responses were incorrect (0.31%) and were not examined further. We excluded response times that were less than 200 ms and greater than 1000 ms, as they represented anticipatory and missed responses, respectively. These errors were minimal (0.86%). In order to accurately compare across the different types of responses (i.e., targets), we measured each participants' average response rate to the different targets in the neutral condition (baseline) and subtracted that average from the experimental trials, in which we obtained a set of difference scores. The average response time differences with respect to the baseline (i.e., 0 ms) are illustrated in Figure 7a for the verbal stimuli and Figure 7b for the non-verbal stimuli.

The difference scores were submitted to a 2 (actor: female or male) x 2 (intention: verbal or non-verbal) x 3 (gesture: congruent with either la, pa, or va) x 2 (target: LA, PA, or VA) repeated measures analysis of variance (ANOVA). The ANOVA revealed a significant main effect of intention ($F(1, 28) = 4.496$, $p = .043$), showing response times to be greater overall for verbal compared to non-verbal stimuli. Likewise, there was a reliable main effect of gesture ($F(2, 56) = 4.085$, $p = .022$), where the stimuli congruent with the /pa/ gesture elicited quicker responses than the /va/ or /la/ gestures. Most importantly, the ANOVA yielded a significant three-way interaction involving intention, gesture, and target ($F(4, 112) = 3.276$, $p = .014$), for which post hoc comparisons were conducted to investigate this further.

The Least Significance Difference (LSD) test was performed to examine the factors included in the three-way interaction between intention, gesture, and target. For

the verbal stimuli (Figure 7a), we found significantly faster response times when the implied and produced gestures were compatible (see Table 3.1). For instance, when the stimulus presented was an intended [la] gesture (see Figure 7b), participants were significantly faster to produce the target LA, than PA or VA ($p < .001$). Similarly, when the stimulus was an intended [pa] gesture, participants were faster to pronounce the target PA, than LA or VA ($p < .001$). Lastly, when presented with the intended [va] gesture, response times were significantly faster to produce the target VA, compared to PA ($p = .05$) or LA ($p < .001$). However, this same compatibility trend between implied and produced speech did not reach significance for the non-verbal stimuli (Fig. 7b). Post hoc tests showed no significant differences in response times for each of the targets ($p > .05$) when presented with the [licking lips] gesture. For the other two gestures, LA was the only target that showed reliable interference, for instance the [protruding lips] gesture showed significantly slower response times to produce the target LA compared to PA ($p = .002$) or VA ($p = .04$) and participants were reliably slower to produce LA compared to VA ($p = .004$), during the observation of the [biting lip] gesture.

When comparing between the speech and non-speech gestures that we matched for incorporating similar articulatory configurations, there were no significant differences found in response times to produce the compatible target. For example, there was no difference in response time to produce the target LA when participants viewed either a picture implying a [la] gesture or the [licking lips] gesture. There was also no difference in response time to produce the target PA, whether the implied gesture was [pa] or [protruding lips]. Lastly, response times were the same to produce the target VA when participants observed the implied gesture [va] and when they observed the non-speech

gesture [biting lip]. There were however, some significant differences in response interference for the incompatible targets. We found that response times to produce PA and VA were significantly longer when the picture implied the incompatible speech gesture [la] than the [licking lip] non-speech gesture ($p < .05$). As well, the response time to produce LA decreased when the picture implied a non-speech [biting lip] than the speech gesture [va] ($p < .05$). No other comparisons were significant.

Discussion

Our results provide a behavioural compliment to the neurological findings of Nishitani and Hari (2002) and Calvert and Campbell (2003), showing that indeed, static images implying mouth gestures can access the motor system and affect speech production. Since the majority of research investigating motor activation from visible speech gestures have used neuroimaging techniques (such as fMRI and TMS) without any behavioural correlate, providing this behavioural data is critical in order to link the neurological data to an observable behaviour. Although our results cannot show that response interference is directly related to mirror neuron activation, we do provide behavioural evidence suggesting a close relation between the perception and production of speech gestures. This evidence is demonstrated in the verbal condition, where the observation of congruent speech gestures facilitated responses (i.e., faster than baseline) when participants were required to produce those gestures, as opposed to an incongruent gesture, which produced interference (i.e., slower than baseline). This pattern was not as obvious during the non-verbal condition, where the pattern did not depend on the compatibility of the observed and produced gestures. This is surprising considering we

chose the non-verbal gestures on the bases that they closely resembled that of the verbal gestures (Figure 6). We did, however, find that the target LA showed response interference during the non-verbal conditions [protruding lips] and [biting lips]. This finding may be due to the similarities in producing PA and VA (both involve the use of the lips) compared to LA (mostly involves the use of the tongue), suggesting that response interference was not specific to the images implying speech.

We see at least two possible explanations for our results: that (1) the still images of the gestures caused the participants to covertly imitate the gesture observed (possibly through the activation of 'mirror neurons') or that, (2) we extract the underlying *intentions* of the actor from the static images (either to produce speech or not to produce speech), through an 'analysis by synthesis' process and those gestures are mapped onto our motor commands to produce those gestures.

This latter explanation is in line with the Motor Theory of speech perception (Liberman & Mattingly, 1985), which proposed that we extract the *intended* gestures from the speaker and not the actual movement of the articulators. These intended gestures provide the basis for the invariant phonetic categories stored in our memory. In other words, they are the primary 'objects' of speech perception. In our case, the verbal images used in this experiment that *implied* a speech gesture were able to activate the motor commands corresponding to that gesture, and 'prime' the production of it. The non-verbal stimuli on the other hand, could not have accessed the speech module (because it is not speech), and therefore would not have interfered with speech production. One possible exception to this rule would be if the non-verbal gestures were perceived as 'noisy' speech gestures. If this were the case, then the non-verbal gestures would have

access to the speech motor system. Although, because these gestures are degraded, they might not activate the motor commands corresponding to that gesture to the same degree as a speech gesture would. Therefore, our findings provide partial support for the Motor Theory such that we found the observation of speech gestures affected speech production more than the observation of non-speech gestures. However, our results do not strongly support the assumption that the perceptual-motor interference is special to speech, since the non-verbal stimuli did elicit response facilitation similar to the verbal gestures as well as some response interference (i.e., to produce LA). In fact, response times to produce the compatible targets were not significantly different across the speech and non-speech gestures, yet there seemed to be greater response interference to produce the incompatible targets when presented with the speech gestures as opposed to the non-speech gestures. Nevertheless, these findings suggest that observing mouth gestures, whether they imply speech or not, can affect the speech motor system. Strong support for the Motor Theory would have been provided if the non-verbal stimuli produced little or no interference or facilitation with the production of speech. However, if it is the case that the non-speech gestures were not perceived as non-speech to the participants, rather they were perceived as 'noisy' speech gestures, then a Motor Theory explanation for our findings would still be appropriate.

We believe a more plausible explanation for our results is that participants were covertly imitating the gestures portrayed in the images, regardless of whether they were verbal or non-verbal. This internal simulation mechanism would have already prepared the motor system for action execution, causing faster response times to produce that perceived gesture, as opposed to producing a different gesture. Furthermore, obligatory

imitation of the gestures could also explain why interference was observed for the non-verbal stimuli, where no connection to speech production would exist. If participants were inclined to internally imitate the mouth gestures portrayed in the images, then those gestures primed during imitation would be faster to produce regardless of whether the intended act was speech related or not. For instance, we found that when participants were presented with an image of the actor biting their bottom lip (congruent with a /va/) or protruding their lips (congruent with /pa/), they were faster to pronounce the gesture /va/ and /pa/ (which both involve the use of the lips – labiodental and bilabial gestures, respectively). However, participants were significantly longer to produce /la/ (which involves the tongue). Since [protruding lips] and [biting lip] portray common gestures, it makes sense that PA and VA resulted in similar response times, whereas LA shares no similar features with the observed gestures and took significantly longer. It would be interesting to see in a future experiment if participants were required to produce the non-verbal gesture (i.e., touch your alveolar ridge), whether response facilitation would occur for compatible gestures in the same way as the verbal condition. However, because we are limited to using a voice key to collect response times, we were unable to explore the perceptual-motor interference for the non-verbal condition in this way.

Overall, the findings from Experiment 4 suggest that implied speech gestures (and to some degree non-speech gestures) are processed up to a late response-related stage analogous to the dynamic speech gestures presented in Experiment 1. Thus, we demonstrated that the perception of static and dynamic speech gestures can affect subsequent production of those gestures. This perceptual-motor effect may occur via activation of the mirror neuron system, which might function as an 'action resonance'

mechanism used for imitation, or to encode the phonetic gestures produced by a speaker during speech perception (Motor Theory; Liberman & Mattingly, 1985). Future research investigating this perception-production relationship will likely uncover how these two processes are related.

General Discussion

This research provided more insight into the complex relationship between the perception and production of speech, as well as added information to the small body of evidence demonstrating the behavioural significance of mirror neuron activation. Using a stimulus-response paradigm similar to Kerzel and Bekkering (2000), we set out to test three of the main assumptions proposed by the Motor Theory of speech perception (Liberman & Mattingly, 1985). In Experiments 1 and 2, we investigated whether there was a direct relationship between the perception and production of speech, and whether this relationship was the same across modalities. Experiment 3 was designed to determine at which stage during audiovisual speech processing the speech motor system becomes involved. And lastly, we examined in Experiment 4 whether photographs of implied speech or non-speech gestures would interfere with speech production.

Generally, we found in Experiment 1 that participants were quicker to produce the speech gestures vocally when they were compatible with the observed irrelevant speech gestures, than when they were incompatible. This occurred regardless of the modality in which the observed speech was available. For example, we showed that participants were significantly faster to say BA when they heard or saw a speaker produce the utterance /aba/, than when they heard or saw the utterance /aga/. Interestingly, we found that the

visual-only trials facilitated verbal response times (faster than baseline) when presented with compatible irrelevant speech stimuli, and delayed response times (slower than baseline) when presented with the incompatible stimuli. This was however, the only modality to show response facilitation, whereas the audiovisual condition produced the greatest amount of interference (i.e., longest response times). We thought one possibility for the modality differences could be due to the amount of attention captured by the auditory versus visual stimuli. This factor was kept constant in Experiment 2, where we presented only audiovisual stimuli for which modality differences were not observed.

In Experiment 2, we found response times to produce the targets were significantly faster and more accurate when they matched the visual or auditory speech gesture presented (BA or GA), than when it was incompatible with both (DA). However, our findings from Experiment 2 failed to show response time differences across modalities. Whereas the auditory-only condition in Experiment 1 produced greater interference compared to the visual-only condition, the auditory and visual modalities in Experiment 2 showed similar response interference. Although, this discrepancy in modality differences could be attributed to many factors, the results of Experiment 2 do support the possibility that the increase in response interference found during the audiovisual condition in Experiment 1 may have been due to the amount of attention the stimuli captured. Therefore, our findings from Experiment 2 seem to suggest that the auditory and visual modalities are processed in the same way at the response stage during speech perception.

A possible future study could be to degrade the amount of information available in each of the modalities, and observe whether response times changed in a linear fashion

depending on the degree of degradation. If this were to occur, then additional evidence would be provided to support the notion that both auditory and visual modalities are processed similarly during speech perception (consistent with Experiment 2), and that both have access to the speech motor commands used during speech production.

In Experiment 3, we explored the previously unexplored domain of identifying the level at which motor commands for speech gestures are activated during audiovisual integration. We were interesting in whether activation occurred (1) before integration, whereby the motor commands corresponding to each modality would be activated first and integrated later on; (2) during integration, whereby motor activation would be involved in the integration process, such that the motor commands may become active for the gestures perceived in both modalities, and those involved in the integrated percept at the same time; or (3) after integration, whereby integration would occur first and then the motor commands representing the integrated percept would be activated later on. Unfortunately, the results failed to indicate any significant differences across the conditions and targets, so we could not make any strong conclusions regarding the level at which motor activation occurred. Despite our null findings, a trend in the data showed that participants were the fastest at producing DA when the incongruent stimulus elicited a McGurk fusion (perception of /ada/), suggesting that perhaps the motor commands for DA were activated *after* integration of the speech signals occurred. Since we did not find statistical evidence for this, this proposal is only speculative and further research is needed to support this.

Lastly, Experiment 4 investigated whether static images only implying speech gestures could interfere with the speech motor system, and whether this interference was

speech-specific. We found that response times were fastest to produce the targets that were the same gesture as that portrayed (e.g., viewing [va] and producing the target VA), compared to producing a target not portrayed (e.g., viewing [va] and producing the targets PA or LA). We also found that the non-speech facial gestures caused response interference as well (especially for LA), suggesting that the perceptual-motor interference observed may not only pertain to speech stimuli.

Overall, the experiments presented in this thesis provided evidence supporting a close relationship between the perception and production of gestures during speech processing. In general, the data (1) provided support for the Motor Theory of speech perception, (2) contributed information relating to stimulus-response compatibility priming, and (3) offered insight concerning the behavioural consequences 'mirror neuron' activation.

Support for the Motor Theory

These series of studies were motivated to test some of the central assumptions outlined in the Motor Theory of speech perception proposed by Liberman and colleagues (Liberman et al., 1967; Liberman & Mattingly, 1985; Liberman & Whalen, 2000). First, we set out to show behaviourally whether speech perception was intimately tied to speech production. According to the Motor Theory, the primary objects of speech perception are the abstract vocal gestures used during speech production. It is believed that a 'phonetic module' in our brain encodes these phonetic gestures and maps them onto their corresponding motor commands in our motor repertoire. Therefore, since we perceive speech gestures in terms of motor acts, then observing speech should interfere with

producing speech. Moreover, this perception-production interference should be observed in response time differences.

In Experiment 1, we found that response times to produce a target syllable were faster if the same syllable was observed simultaneously, whereas production took longer if a different syllable was observed. This finding is consistent with the Motor Theory, such that when the speech gestures were the same, the perceived gestures would have already activated the motor commands needed to produce the target gestures, and would result in quicker pronunciation times for that target. However, if the target gestures were different, then there would have been no activation 'priming' offered by the perceived gestures, and production times for that target would take longer. Further support for a Motor Theory explanation is provided by our findings that there was no response 'priming' of speech gestures when the target was identified by a button-press response. Thus, these findings suggest that the perception of speech gestures interacts with the production of speech gestures at a late response stage during speech processing, and is consistent with the Motor Theory view that speech perception and production are closely connected.

Furthermore, we were also interested in whether this perception-production link was dependent on the modality in which the phonetic gestures were presented. According to the Motor Theory, the 'phonetic module' extracts the phonetic gestures equally from both the visual and auditory channels. Our results of Experiment 1 showed that the gestures presented in the visual-only condition influenced speech production differently than those presented in the auditory-only condition. The visual-only data appeared to show response facilitation when the observed speech gestures were compatible with the

target gestures, and response inhibition when the observed speech gestures were incompatible. A different pattern was observed when the speech gestures were presented aurally, where both compatible and incompatible speech gestures produced delayed response latencies (i.e., interference). However, we investigated these differences further using only audiovisual stimuli in the second experiment by presenting conflicting gestures from each modality simultaneously. We found that both modalities effected response times to a similar degree as that found when both modalities contained congruent speech gestures. Our findings from Experiment 2 agree with the Motor Theory, showing that both auditory and visual speech gestures are processed to the same degree at the response stage, where motor activation is produced. Perhaps the reason why response times in Experiment 1 showed more interference for the auditory-only and audiovisual conditions were because we compared them to a visual baseline (still-face) as opposed to a more appropriate auditory baseline to calculate the difference scores. Had we compared them to their appropriate baseline (auditory /aaa/ and visual still-face), we might have seen similar response facilitation and interference as the visual-only condition, showing that both modalities can elicit similar effects on speech production – consistent with Experiment 2.

Secondly, we sought to test whether integration of auditory and visual speech gestures happened *automatically*, before the appropriate motor commands were activated, or whether each activated their own motor commands and then integration occurred afterward. The level at which motor activation relates to audiovisual integration is not directly addressed in the Motor Theory. However, it can be assumed that since we perceive speech in the form of abstract motor commands, then integration must occur

after the motor commands corresponding to the auditory and visual speech gestures are activated separately. The level at which activation occurs was investigated in Experiment 3, where we took advantage of the McGurk effect (McGurk & MacDonald, 1976), and created stimuli with conflicting auditory and visual speech information. When the information is integrated, participants perceive either a novel percept or a combination of the conflicting signals. The findings of Experiment 3 were inconclusive, and so the question concerning the level of activation during audiovisual integration could not be answered. However, this remains to be an important question and future research should investigate this using more accurate measurements and reliable speech stimuli.

Even though not statistically significant, the data revealed an interesting pattern, whereby participants response times were observed to be the fastest when they produced the target DA while viewing the incongruent stimuli that elicited the perception of /ada/ (McGurk fusion). Indeed, the response times were almost double to produce the targets GA or BA, which were compatible with the individual auditory and visual signals during the incongruent conditions, but compatible with both auditory and visual signal during the congruent condition (see Figure 5b). Since the response times were the fastest when the target was compatible with the integrated percept during the incongruent fusion condition, we hypothesized that integration might automatically occur *before* motor activation is reached for the individual auditory and visual speech components. If this hypothesis were true, then it would argue against a Motor Theory account for audiovisual speech integration, seeing as though proponents of the theory believe that the 'phonetic module' would automatically translate the auditory and visual signals into the motor commands. Thus, the 'phonetic module provides a 'common currency' (i.e., motor

representation) for which the gestures could be integrated. If the signals are being integrated before the 'phonetic module' is reached, then this suggests that there must be some other mechanism in place to integrate them. A recent fMRI study conducted by Jones and Callan (2003) investigated brain activations when the McGurk effect was elicited and found the superior temporal sulcus and posterior parietal regions to be the main activation sites (in addition to inferior frontal and premotor areas). These findings suggested that those regions are important for audiovisual integration of speech signals. It is possible that the auditory and visual speech signals are integrated in those regions before premotor and motor areas are active, however it is still unknown where and when in the brain audiovisual integration takes place. Since the results of Experiment 3 could not provide strong evidence for the level at which integration occurred with respect to motor activation, we leave this as an important question to be examined in future studies.

Lastly, we aimed to test a third assumption of the Motor Theory that we perceive the abstract *intended* gestures of the speaker and not the actual movements of the articulators during speech production. We aimed to test this in Experiment 4 using static photographs of actors that portrayed speech gestures or non-speech facial gestures, and found that response times were faster when the implied gestures were compatible with the target gestures. For instance, participants were quicker to produce the target VA if they saw a picture portraying the gesture [va], as opposed to viewing a picture portraying [pa] or [la]. This compatibility effect was true for the other targets PA and LA as well (see Fig. 7a). These findings suggested that even pictures only implying speech gestures can activate the motor commands for producing those gestures and can affect subsequent speech production, consistent with the Motor Theory.

However, even though the same interference effects were not seen when the non-speech facial gestures were incompatible with the targets (with the exception of LA), the response times were the same to produce the targets that were compatible with the speech gestures and their non-speech equivalents. For example, there were no differences in response times to produce the target VA whether it was presented with the implied speech gesture [va] or with the non-speech equivalent [biting lip]. Thus, it seems that the facilitation found to produce the compatible speech gestures was not specific to speech and applies to general facial gestures as well. This finding goes against the Motor Theory claim that only the *intended* phonetic gestures are detected by a specialized 'phonetic module', which is where motor mapping occurs. Instead, our findings indicate that there might be a general perception-action mechanism responsible for processing implied facial gestures. This proposal is consistent with Nishitani and Hari (2002) who showed similar cortical activations using MEG when participants viewed static images of a person making speech and non-speech lip movements, as well as when the participants imitated the lip movements themselves. In sum, the results from Experiment 4 support the Motor Theory view that the perception of implied speech gestures affect speech production, however our findings suggest that this perceptual-motor effect is not speech-specific and likely involves a general mechanism responsible for action imitation.

Support for Stimulus-Response Compatibility

According the dimensional overlap model (Kornblum et al., 1990), if two dimensions overlap in structural, perceptual, or conceptual features, then activation of the one dimension will automatically activate the other dimension and result in compatibility

priming. This is frequently demonstrated by the Stroop-effect (Stroop, 1935), where reading the colour-word (irrelevant stimulus) and vocally identifying the colour of the word (relevant response) overlap structurally producing faster response times, as opposed to identifying the colour of the colour-word (relevant stimulus), which does not share any features with the vocal response. This compatibility effect can also be seen in the present experiments, where response times were faster when the speech and target were compatible, than when they were incompatible. It is possible that the paradigm used in this series of studies produced stimulus-stimulus interference (interaction of speech stimulus and target) and not stimulus-response interference (interaction of speech stimulus and vocal production), where stimulus-response interference is what will demonstrate a link between perception and production of speech gestures. However, the paradigm we used had been previously shown by Kerzel and Bekkering (2000) to produce interference at a response-related stage. By training participants to verbally respond /ba/ or /da/ to arbitrary symbols like && and ## (Experiment 2), they created a situation were the irrelevant stimuli and targets no longer shared features with each other, which minimized any perceptual overlap between the dimensions. Using this design, they still found the same compatibility advantage as that observed when the letter targets 'ba' and 'da' were displayed (Experiment 1). This result allowed the authors to conclude that the facilitation of response times to produce the target letters must have been localized at the response level. However, one could still argue that by training the participants to respond BA or GA to the target symbols ## and &&, that those symbols are no longer arbitrary and now represent their corresponding responses - just the same as the letters

'ba' and 'ga' did. Thus, stimulus-stimulus interference might not have been eliminated in this design.

In a following study, Kezel (2002) conducted a similar experiment to examine whether their previous results (Kerzel & Bekkering, 2000) were due to stimulus-stimulus or stimulus-response compatibility by reducing the similarity between the stimulus and response. To accomplish this they had participants press a button corresponding to the target letters, instead of identifying the target vocally. Since both of the stimuli dimensions no longer had features in common with the response, response priming should have been eliminated. This was the same logic we used in Experiment 1, where both verbal and manual response types were measured. We predicted that stimulus-response interference would occur for the verbal condition and not for the manual condition. Our findings confirmed this prediction, where we showed response times to vocally produce the target to be significantly quicker when the speech and target matched than mismatched. Since this compatibility priming was not replicated in the manual response data, our results suggested that the effects observed for the verbal responses were due to stimulus-response interference. The failure to observe response time differences in the manual response data contrast with the findings of Kerzel (2002), who were able to find similar compatibility effects for manual and verbal responses. Kerzel therefore claimed that stimulus-stimulus interference could explain the results of Kerzel and Bekkering (2000), and not necessarily stimulus-response effects. Although there are many differences between Kerzel (2002) and our studies, we believe that our greater sample size (42 versus 18), within-participants design (directly comparing manual to verbal responses), and use of a critical baseline measure (to correct for response time

differences between targets and across response types), provides us with the power to confidently conclude that there were indeed stimulus-response interference when participants viewed videos of a speaker producing speech syllables. Nevertheless, because the manual response data in Experiment 1 did produce a trend that seemed to follow that of the verbal responses, showing faster response times to compatible stimuli than incompatible stimuli, we cannot completely rule out the presence of stimulus-stimulus interference within this paradigm. It seems as though the dimensional overlap model is able to account for the conflicting findings in such paradigms, by making the assumptions that there were features shared amoung the two stimuli dimensions as well as with the response dimension, and that this featural overlap produced compatibility priming effects.

This conflict in deciphering the locus of interference is also true for paradigms investigating the Stroop-effect. Numerous studies have shown interference at the level of stimulus encoding where the stimuli possess perceptual and conceptual similarities (for an extensive review see MacLeod, 1991). A good demonstration showing stimulus-stimulus interference can be seen in a study by Zhang and Kornblum (1998). They presented participants with three colour-words (e.g., *blue-green-blue*) and asked them to respond by saying a digit that corresponded to the middle word, for instance they were trained to say "two" whenever they saw the word *blue*, or "three" when they saw the word *yellow*, and so on. Using this design, they removed any structural overlap (Kornblum et al., 1990) between the stimulus and response, since the stimulus was a colour-word and the response was a digit. Even though they removed response interference, they still found that participants were quicker to say the digit corresponding

to the colour-word in the middle if the other two words were congruent (e.g., *green-green-green*), than if they were incongruent (e.g., *blue-green-blue*). This suggested to the authors that the difference in performance was due to interference at the stimulus level, where the two words that were irrelevant for the task interfered with the word that was relevant for the response when they were incongruent. Thus, they provided clear evidence that stimulus-stimulus effects can occur using the Stroop-effect.

On the other hand, removal of structural overlap between stimulus and response features has also been shown to reduce performance differences in the Stoop-effect. For instance, Durgin (2000) had participants use a mouse to point to a corner of the screen representing the colour of the colour-word, for which now the word and the manual response shared no features. They used the classic Stroop paradigm, but instead of having the participants verbally pronounce the colour of the colour-word, they asked them to use a mouse to indicate the colour of text the word was printed in. In this experiment, they found that the colour-word had no effect on response latencies to move the mouse to the correct colour corner, and thus stimulus-response interference was eliminated. According to Kornblum et al.'s dimensional overlap model, no compatibility interference occurred because there was no overlap between the stimulus and response dimensions.

In sum, the dimensional overlap model seems to be able to accommodate many different findings showing compatibility effects between stimuli and responses in Stroop-like paradigms. However, it is difficult to dissociate between stimulus-stimulus and stimulus-response compatibility with the paradigm we used because both the irrelevant (speech syllable) and relevant (target syllable) dimensions overlapped in features, and they also overlapped with the response (syllable production). Therefore, the compatibility

interference we found could be localized at the stimulus or response level. Although our findings contrast with those of Kerzel (2002) who was able to find compatibility effects using a manual response, we believe (according to the conclusions based on Kerzel's design) that the lack of compatibility priming observed in Experiment 1 for the manual condition compared to the verbal condition is the result of interference at the response level more so than the stimulus level.

Support for 'Mirror Neuron' Activation

Finally, each of the experiments presented in this paper support the existence of an observation-execution matching system for speech (Sundara et al., 2001) that can provide a link between actor and observer during online communication. Our findings repeatedly demonstrated that when participants responded verbally to a target syllable, their response times decreased if they simultaneously observed a speaker producing or implying the same syllable. It has previously been proposed that this system may function to understand others intentions by internally generating the actions we observe and comparing them with our own motor repertoire for producing specific behaviours (Rizzolatti & Craighero, 2004). In fact, a prerequisite for 'mirror neuron' activation in monkeys seems to be that the action must have a clear intention and directed at a goal. Umilta et al. (1996) demonstrated in monkeys that mirror neurons are responsive only when an object-directed action was observed or produced, and not when the object was observed alone. However, they demonstrated that mirror neuron activation could be elicited when a goal-directed action is observed or executed in isolation, without the presence of an object. This finding that only actions be directed at a goal produce mirror

neurons activity suggested that a potential function of the mirror neuron system may be to encode the intentions of others to help us understand the reasons behind their actions (Rizzolatti, Fogassi, & Gallese, 2001).

Numerous studies have shown similar activation patterns in humans using EEG recordings (e.g., Cochin, Barthelemy, Roux, & Martineau, 1999), MEG recordings (e.g., Hari et al., 1998; Nishitani & Hari, 2000, 2002), TMS techniques (e.g., Fadiga et al., 1995, 2002, 2005; Sundara et al., 2001; Watkins et al., 2003; Watkins & Paus, 2004) and fMRI (e.g., Pulvermuller et al., 2006; Skipper et al., 2005). Within these, the most compelling evidence for an observation-execution system in humans can be demonstrated in studies using TMS. Fadiga, Fogassi, Pavesi, and Rizzolatti (1995) used TMS to send magnetic pulses to the left motor cortex representing the hand and arm, and recorded motor-evoked potentials (MEP) from the arm and hand muscles of the participants. They found enhanced MEP's when the participants viewed an experimenter making hand and arm actions, and specifically from the muscles that participants would use to make those actions themselves. In the realm of speech perception, TMS studies have shown that MEP's are enhanced in the lip and tongue muscles when participants observe or hear a speaker producing speech using the lips or tongue (Fadiga et al., 2002; Watkins et al., 2003). For instance, Watkins et al. (2003) recorded from the lips muscles of participants, while stimulating the face area of the primary motor cortex. They measured MEP's from the lips during four experimental conditions: (1) a speech condition, where they only listened to speech, (2) non-verbal condition, where they listened to non-speech sounds (like bells ringing), (3) lips condition, where they viewed dynamic movements of the lips, and (4) eyes condition, where the viewed dynamic movements of the eyes and eyebrows.

The authors found that the MEP's were significantly greater when the participants listened to speech or viewed lip movements, compared to listening to arbitrary sounds and viewing eye and eyebrow movements. Thus, these findings provide evidence for a 'motor resonance' mechanism in humans that includes speech related actions, in addition to hand and arm movements.

This activation of the 'mirror neuron' system when speech stimuli are observed could possibly account for the stimulus-response compatibility found in the present experiments. If the visual and auditory speech syllables presented in Experiment 1, 2 and 3 activated a 'resonance' mechanism that internally generated the observed speech, then this would decrease the amount of time to overtly produce that same speech act (in a sense, 'priming' the execution of it). This explains the faster response times that we found when participants produced the same target syllable as that presented visually or aurally. Furthermore, this not only occurred for the dynamic speech syllables, but also for static images that implied speech syllables (Experiment 4). These findings are consistent with a MEG study conducted by Nishitani and Hari (2002), where they showed participants still photographs of an actor making lips movements that implied either a verbal gesture or a non-verbal gesture, while also asking them to either imitate the observed lip movement or to produce a spontaneous lip movement themselves. The authors found similar areas to be active when participants observed the verbal or non-verbal lip forms, and when they imitated them. Broca's area, commonly used for overt and covert speech production, (Grafton, Fadiga, Arbib, & Rizzolatti, 1997) and imitation of meaningful goal-directed actions (e.g., Iacoboni, Woods, Brass, Bekkering, Mazziotta, & Rizzolatti, 1999) was one of the areas that were activated. Thus, mirror neurons might

provide the link between sender and receiver during online communication and constitute a common mechanism for which the perception and production of speech gestures are processed.

Conclusions and Future Directions

The experiments presented in this paper support the notion of tight perceptual-motor relationship between speech perception and production, however without supplementary neuroimaging data, we cannot ascertain whether the response time differences found were indeed modulated by the mirror neuron system. Thus, additional research is needed before a strong connection between speech perception and mirror neurons can be asserted.

Future research could identify this perception-production link using a variety of neuroimaging techniques to uncover the mechanisms responsible for the behavioural results found in our studies. One idea would be to replicate these experiments in an fMRI scanner to elucidate the cortical areas that are active during both speech perception and production. For example, three scanning conditions would be needed: a speech perception condition, a button-press condition, and speech production condition. Since speech perception and production are believed to involve similar brain areas, making a vocal response to the speech stimuli could disguise any activity produced by the perceived stimuli. Thus, each condition would have to be scanned separately and then compared post hoc to reveal the overlapping cortical areas. During the speech perception condition, participants could be shown the same videos as the behavioural experiments presented in this paper, however they would press a button identifying the target letters

instead of a vocal response. During the button-press condition, for which only the target letters would be presented on a blank screen, participants would be required to press a button identifying the target. This condition would be used as a control for the activation elicited from pressing the button and seeing the target letters in the speech perception condition. Finally, for the speech production condition, participants could be asked to identify the targets vocally. It should be stressed that all responses in the three conditions be made as quickly and accurately as possible to resemble the behavioural experiments as presented here.

For this proposed fMRI experiment, one would expect to find activity in areas responsible for speech perception (e.g., auditory cortical areas, superior temporal sulcus/gyrus, Broca's area; Skipper et al., 2005), as well as a network of areas for speech motor planning and production (e.g., Broca's area, premotor and primary motor areas; Bohland & Guenther, 2006; Soros, et al., 2006). However, the critical prediction would be to find common activation in Broca's area, potentially supporting the existence of a mirror neuron system responsive to both speech perception and production. Furthermore, Broca's area might also be important for processing images implying speech or non-speech gestures, as suggested by Experiment 4. Therefore, if we could combine the behavioural response time data with the neuroimaging data provided by fMRI, we would be able to more clearly identify the mechanisms that underlie the behavioural results presented. Until then, we can only assume that our behavioural data closely parallel the neuroimaging studies demonstrating a close perceptual-motor mechanism for speech.

Table 1.1

Each trial type presented in Experiment 1. The [ ] represent the visual stimulus and / / represent the auditory stimulus. The capital letters signify the targets.

Condition	Compatible	Incompatible
Visual-Only	[aba] + BA [aga] + GA	[aba] + GA [aga] + BA
Audio-Only	[still]+/aba/ + BA [still]+/aga/ + GA	[still]+/aba/ + GA [still]+/aga/ + BA
Audiovisual	[aba]+/aba/ + BA [aga]+/aga/ + GA	[aba]+/aba/ + GA [aga]+/aga/ + BA
Baseline Control	[still] + BA [still] + GA	

Table 2.1

Percentage of incorrect responses for each of the conditions in Experiment 1.

Manual Condition	Compatible	Incompatible
Visual-Only	2.14	4.05
Audio-Only	1.19	2.38
Audiovisual	2.62	3.81
Verbal Condition	**Compatible**	**Incompatible**
Visual-Only	0.48	2.62
Audio-Only	1.19	2.38
Audiovisual	0.95	1.90

Table 2.1

Each trial type presented in Experiment 2 and 3. The [ ] represents the visual stimulus and / / represents the auditory stimulus. The capital letters signify the targets.

Condition	Compatible	Incompatible
Congruent	[aba]+/aba/ + BA	[aba]+/aba/ + GA
		[aba]+/aba/ + DA
	[aga]+/aga/ + GA	[aga]+/aga/ + BA
		[aga]+/aga/ + DA
Incongruent	[aba]+/aga/ + BA	[aba]+/aga/ + DA
	[aba]+/aga/ + GA	
	[aga]+/aba/ + BA	[aga]+/aba/ + DA
	[aga]+/aba/ + GA	
Baseline Control	[still] + BA	
	[still] + GA	
	[still] + DA	

Table 2.2

Percentage of incorrect responses for each of the conditions in Experiment 2.

Condition	BA	GA	DA
/aba/	0	0.2	0.67
/aga/	0.2	0	0.73
Fusion V/aga/ A/aba/	0.13	0	0.27
Combination V/aba/ A/aga/	0	0.07	0.53
Total	0.33	0.27	2.20

Table 3.1

Each trial type presented in Experiment 4. The [ ] represents the visual stimulus and the capital letters signify the targets. Each combination below was presented using the female and male actor.

Condition	Compatible	Incompatible
Speech Gesture	[pa] + PA	[pa] + LA
		[pa] + VA
	[la] + LA	[la] + PA
		[la] + VA
	[va] + VA	[va] + PA
		[va] + LA
Non-speech Gesture	[protruding lips] + BA	[protruding lips] +LA
		[protruding lips] +VA
	[licking lips] + LA	[licking lips] + PA
		[licking lips] + VA
	[biting lip] + VA	[biting lip] + PA
		[biting lip] + LA
Baseline Control	[still] + PA	
	[still] + LA	
	[still] + VA	

Figure Captions

Figure 1: (a) Examples of the 'BA' and 'GA' target positions on the face of the speaker during the production of the irrelevant speech syllable (/aba/ on the left and /aga/ on the right). (b) A schematic timeline of the video presentation in Experiment 1. The middle frame displaying the target was presented for three frames (approx. 100 ms at a 29.97 fps).

Figure 2: Response time differences across all modalities relative to each individuals' baseline response rate (i.e., 0 ms) for (*a*) the verbal response and (*b*) manual response conditions in Experiment 1. For both, response times were quicker in all three modality conditions when the speech and target were compatible (*C*) than when they were incompatible (*IC*). The visual-only condition showed a verbal response facilitation when compatible and response inhibition when incompatible. There were no significant compatibility differences within the visual-only condition for the manual responses. The error bars represent the 95% confidence intervals of the means.

Figure 3: Response time differences compared to the baseline response rate (i.e., 0 ms) across the different audiovisual conditions in Experiment 2. The incongruent conditions contained a visual /aga/ and an auditory /aba/ (V/aga/ A/aba/) and a visual /aba/ presented with an auditory /aga/ (V/aba/ A/aga/). The congruent conditions contained both visual and auditory /aba/ and /aga/. The error bars represent the 95% confidence intervals.

Figure 4: A schematic example of how we created the incongruent audiovisual stimuli used in Experiments 2 and 3. As shown, we aligned the acoustic bursts of the consonants with that of the visual stimuli.

Figure 5: Response time differences across the congruent and incongruent audiovisual conditions relative to each individuals' baseline response rate (i.e., 0 ms) for the male speaker (*a*) and female speaker (*b*) in Experiment 3. There were no significant response time differences to produce the targets across all the conditions when the male speaker was presented. However, there were differences observed when the female speaker was presented, such that the fusion condition seemed to produce the fastest response times relative to the other conditions. The error bars represent the 95% confidence intervals of the means.

Figure 6: (a) Examples of the seven different types of visual gestures used in Experiment 4. The verbal gestures are along the top and non-verbal on the bottom. (b) A schematic timeline of when the target appeared during the stimulus presentation in Experiment 4.

Figure 7: Mean response times to produce the different targets (pa, la, and va) when presented with congruent and incongruent visual images for the verbal condition (a) and the non-verbal condition (b). The error bars represent the 95% confidence intervals.

Fig. 1a

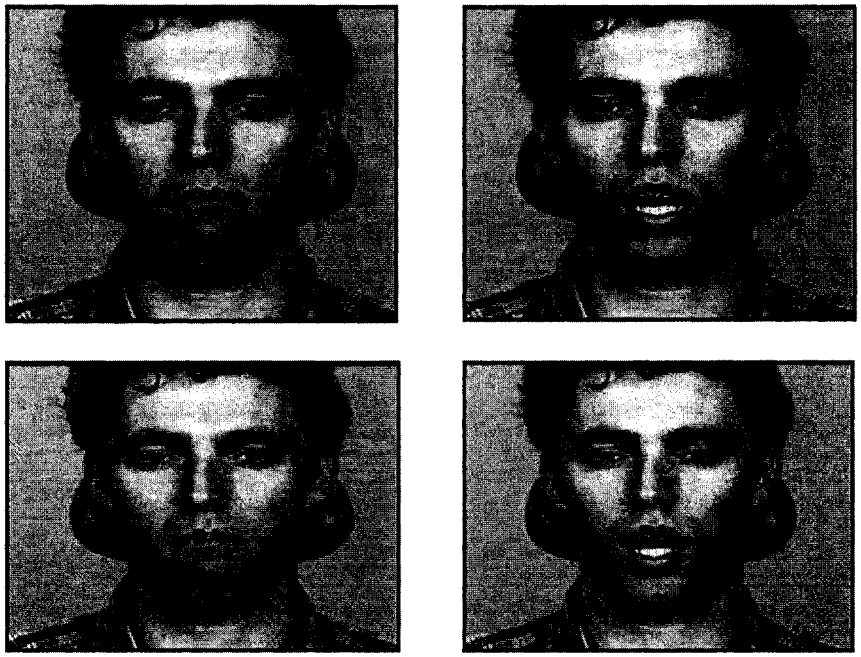


Fig. 1b

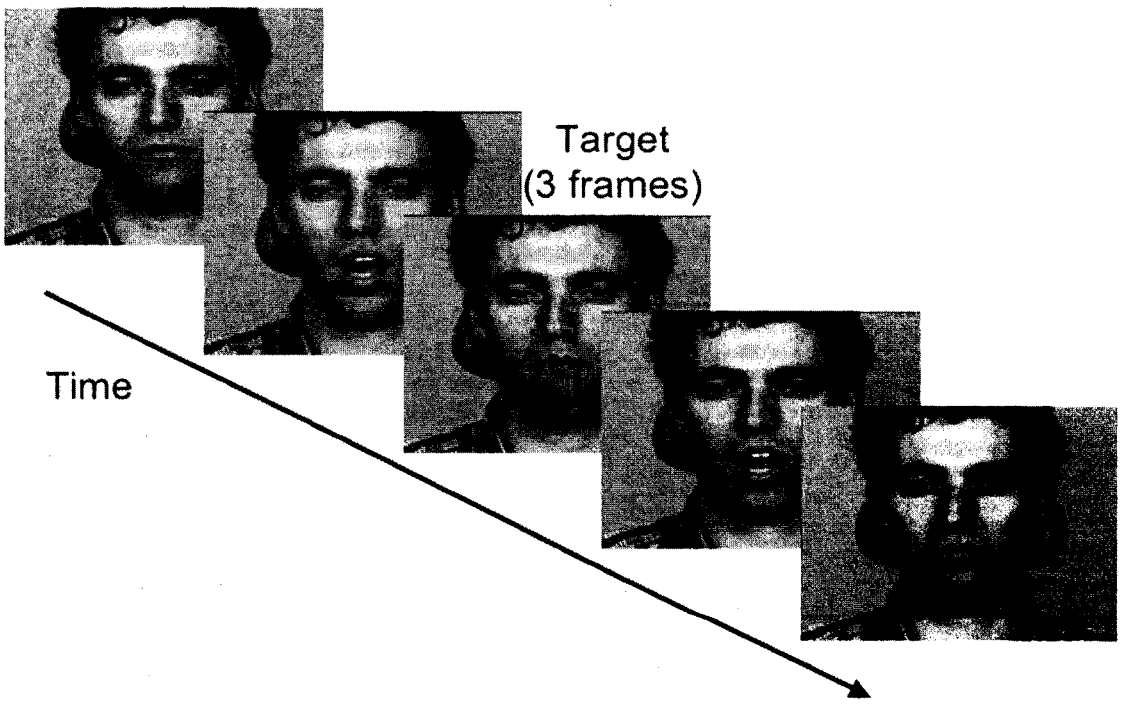


Fig. 2a

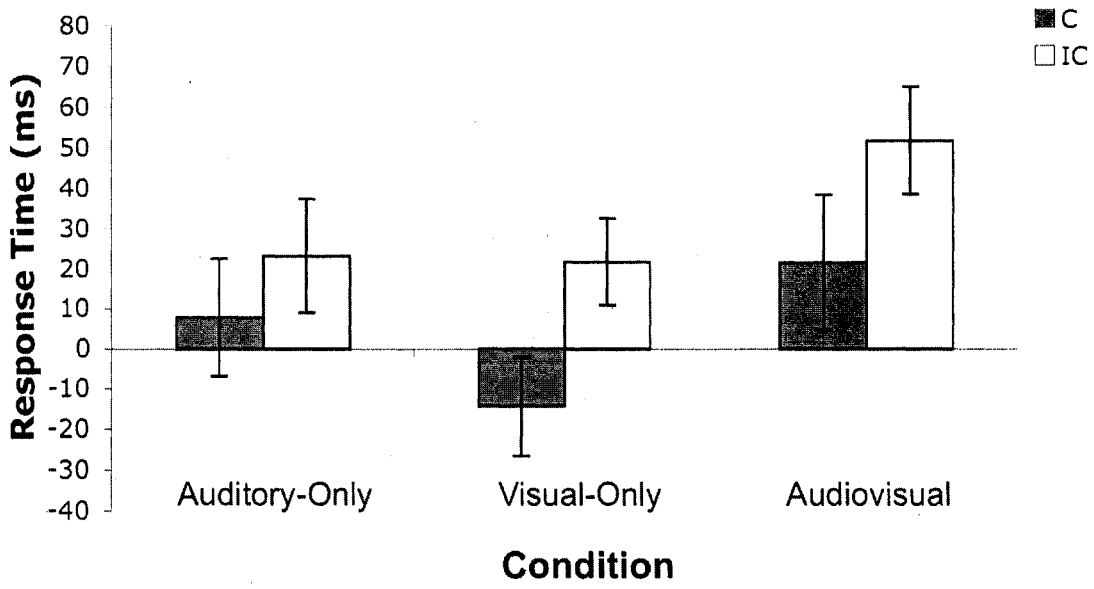

C
IC
Response Time (ms)
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
Auditory-Only
Visual-Only
Audiovisual
Condition

Fig. 2b

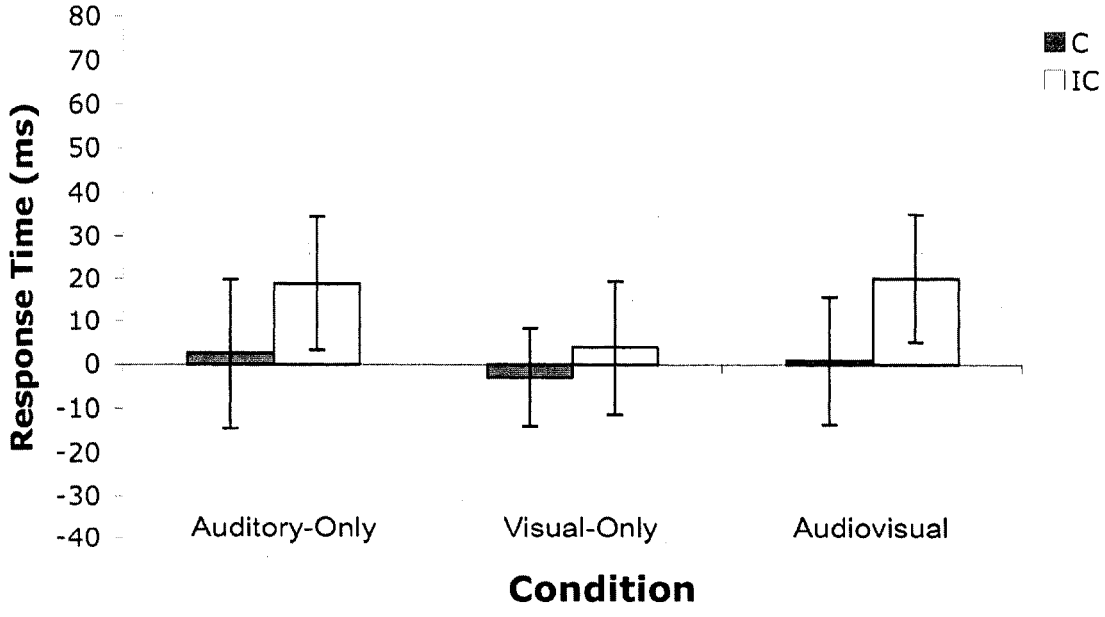

C
IC
Response Time (ms)
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
Auditory-Only
Visual-Only
Audiovisual
Condition

Fig. 3

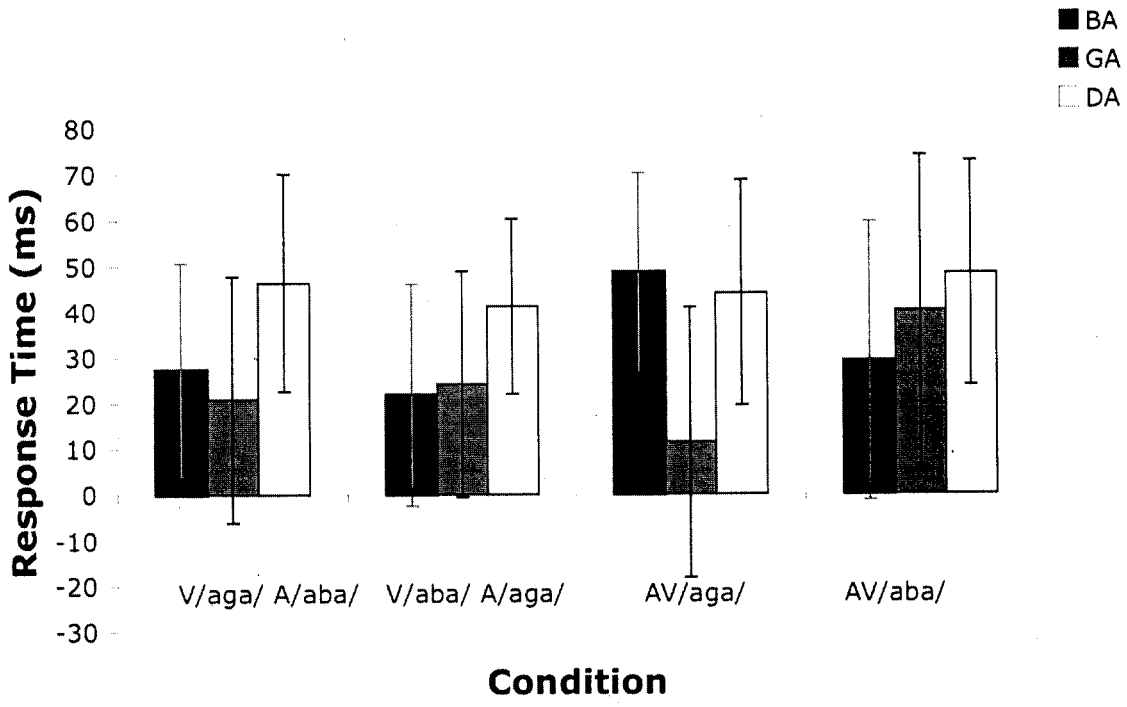

BA
GA
DA
80
70
60
50
40
30
20
10
0
-10
-20
-30
Response Time (ms)
V/aga/ A/aba/
V/aba/ A/aga/
AV/aga/
AV/aba/
Condition

Fig. 4

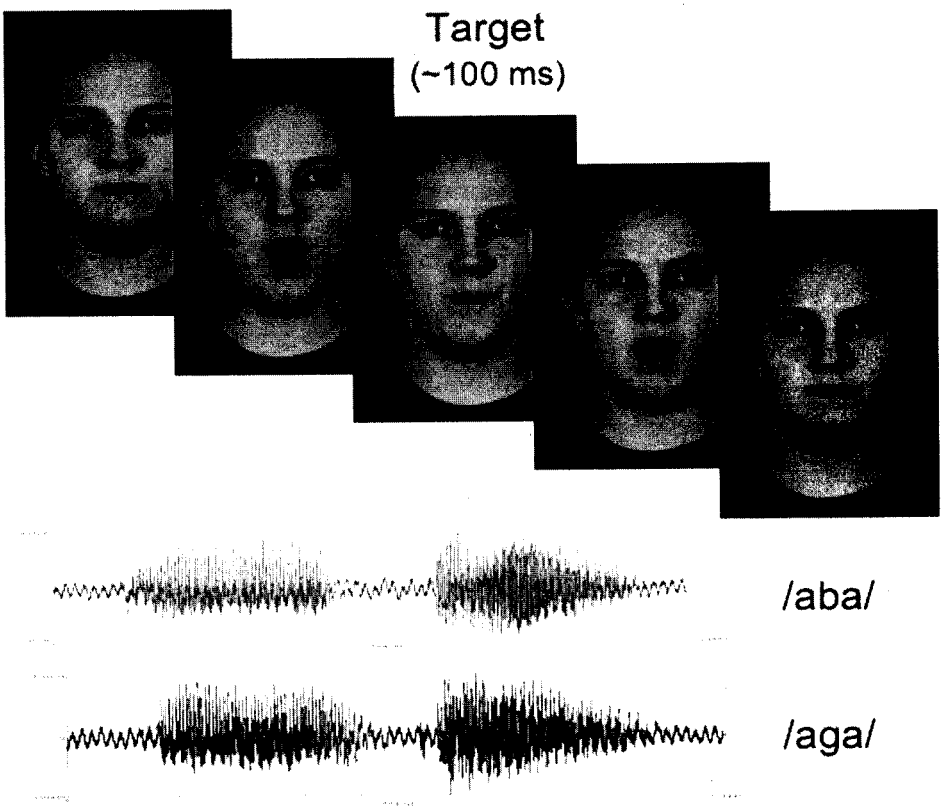


Fig. 5a

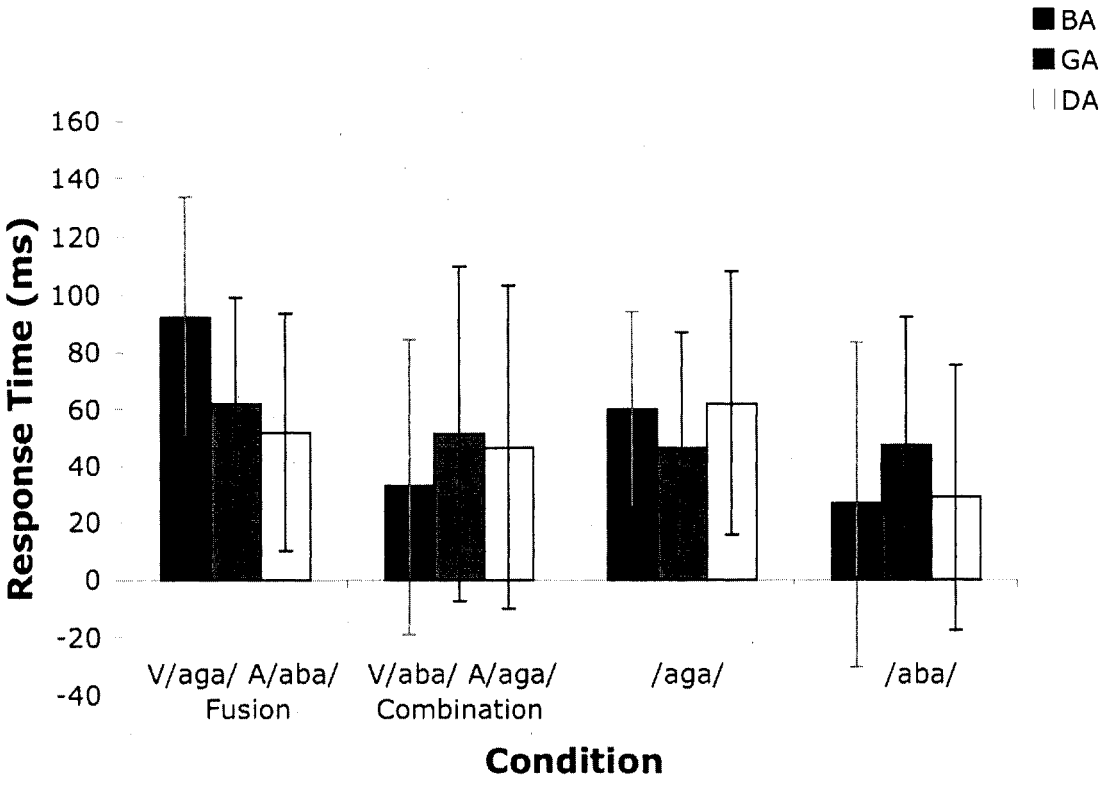

BA
GA
DA
160
140
120
100
80
60
40
20
0
-20
-40
Response Time (ms)
V/aga/ A/aba/ Fusion
V/aba/ A/aga/ Combination
/aga/
/aba/
Condition

Fig. 5b

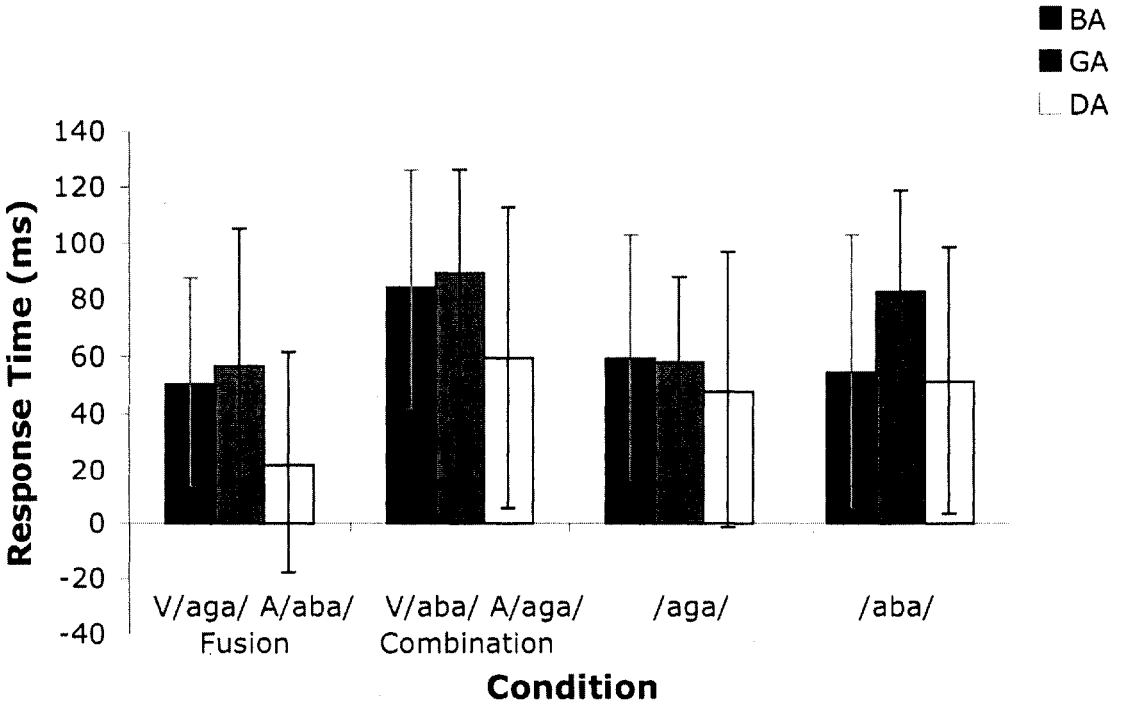

BA
GA
DA
140
120
100
80
60
40
20
0
-20
-40
Response Time (ms)
V/aga/ A/aba/ Fusion
V/aba/ A/aga/ Combination
/aga/
/aba/
Condition

Fig. 6a

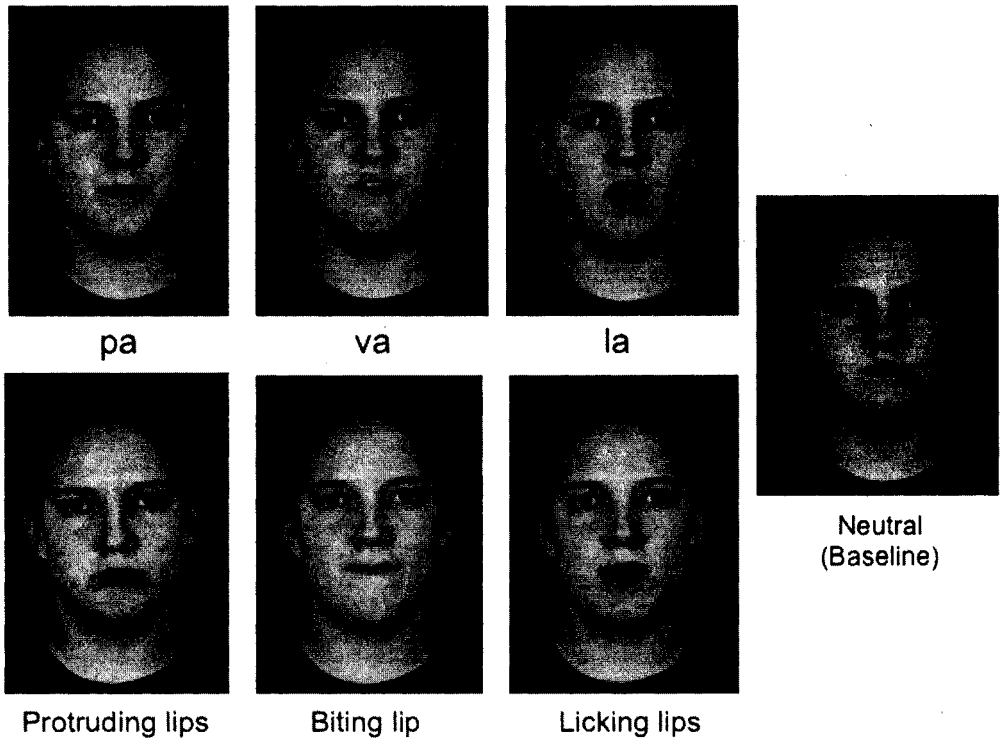

pa
va
la
Neutral
(Baseline)
Protruding lips
Biting lip
Licking lips

Fig. 6b

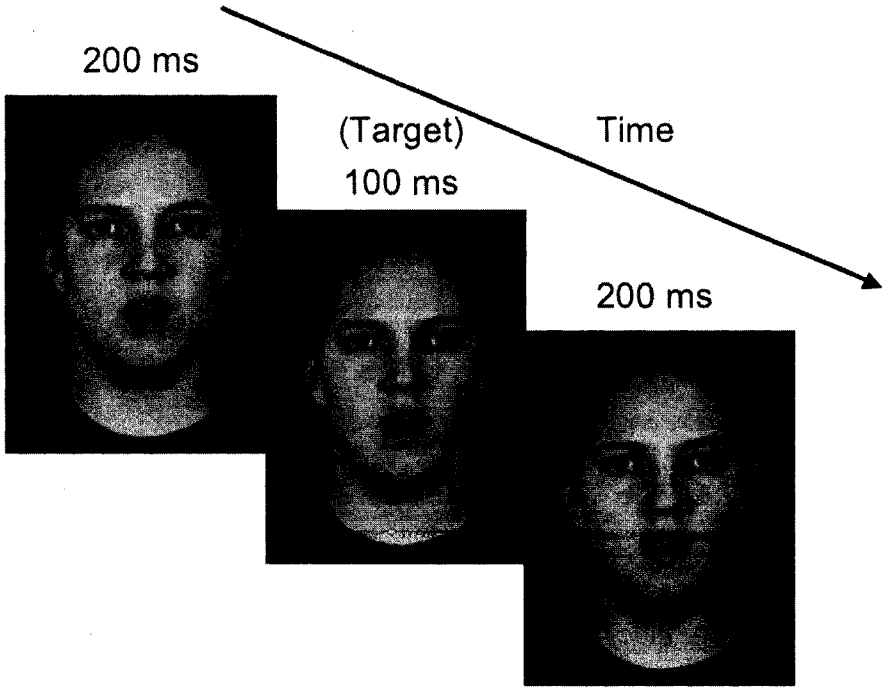

200 ms
(Target)
100 ms
Time
200 ms

Fig. 7a

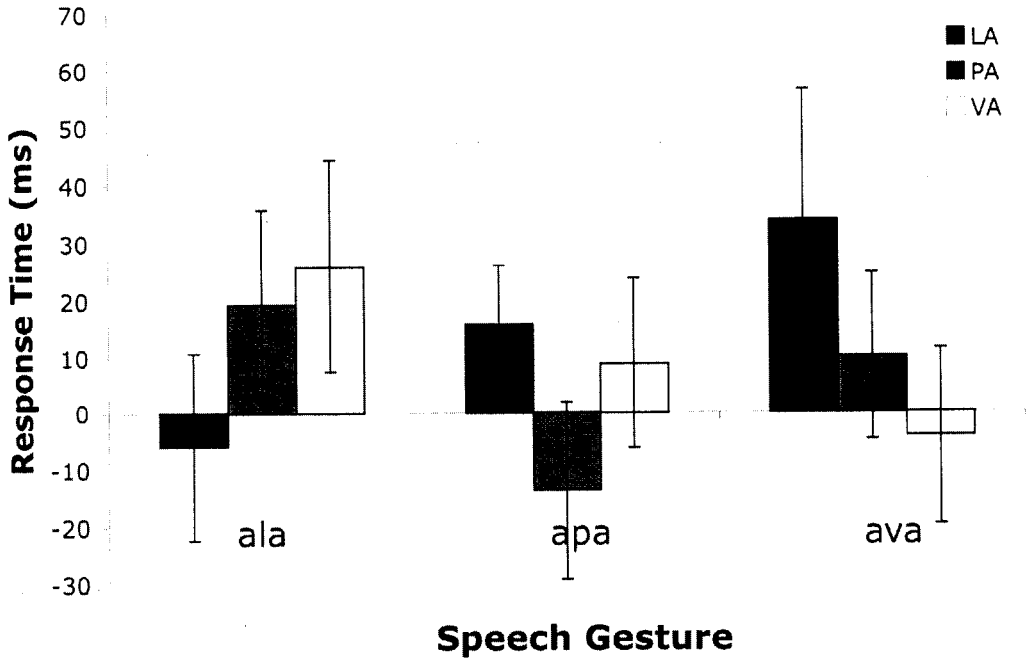
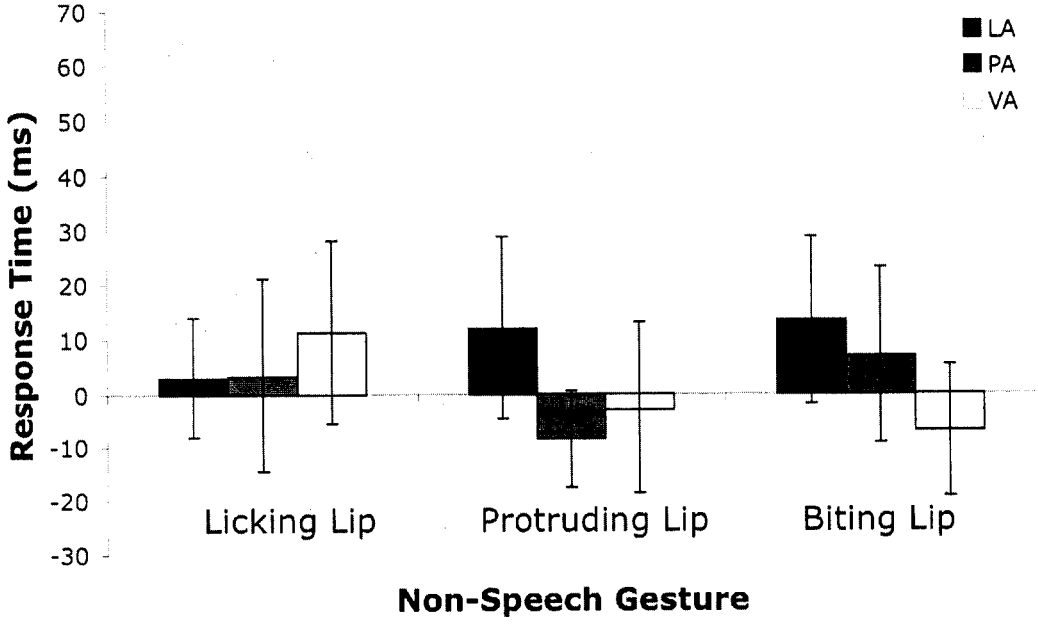

70
60
50
40
30
20
10
0
-10
-20
-30
Response Time (ms)
LA
PA
VA
ala
apa
ava
Speech Gesture

Fig. 7b


70
60
50
40
30
20
10
0
-10
-20
-30
Response Time (ms)
LA
PA
VA
Licking Lip
Protruding Lip
Biting Lip
Non-Speech Gesture

Appendix A

Language Questionnaire

1. What is your mother tongue (the first language you learned)?
2. What other languages do you know?
3. What is your best language for speaking?
4. What is your best language for writing?
5. What language(s) did your family speak at home?
6. In what city (and country) were you born?
7. How long did you live in the city that you were born?
8. In what city did you go to elementary school?
9. In what city did you go to high school?
10. How many years have you lived in Canada?

Appendix B

Handedness Questionnaire

Instructions: Think carefully about each of the following tasks and indicate by circling, whether you use your left hand, right hand or either hand.

1. Which hand do you use to hold scissors?

   Left Either Right

2. With which hand do you draw?

   Left Either Right

3. With which hand do you screw the top off a bottle?

   Left Either Right

4. With which hand do you deal cards?

   Left Either Right

5. Which hand do you use to hold a toothbrush when cleaning teeth?

   Left Either Right

6. With which hand do you use a bottle opener?

   Left Either Right

7. With which hand do you throw a ball away?

   Left Either Right

8. Which hand do you use to hold a hammer?

   Left Either Right

9. With which hand do you thread a needle?

   Left Either Right

10. With which hand do you hold a racket when playing tennis?

    Left Either Right

11. With which hand do you open the lid of a small box?

    Left Either Right

12. With which hand do you turn a key?

    Left Either Right

13. With which hand do you cut a cord with a knife?

Left Either Right

14. With which hand do you stir with a spoon?

Left Either Right

15. With which hand do you use an eraser on paper?

Left Either Right

16. With which hand do you strike a match?

Left Either Right

17. With which hand do you write?

Left Either Right

References

Bohland, J. W., & Guenther, F. H. (2006). An fMRI investigation of syllable sequence production. *NeuroImage, 32,* 821 – 841.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iverson, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*, 593-596.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15,* 57-70.

Cathiard, M. A., Lallouache, M. T., & Abry, C. (1996). Does movement on the lips mean movement in the mind? In D. Stork & M. E. Hennecke (Eds.) *Speechreading By Humans and Machines*. Berlin, Germany: Springer-Verlag.

Cochin, S., Barthelemy, C., Roux, S., & Martineau, J. (1999). Observation and execution of movement: similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience, 11*, 1839 – 1842.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech, 41*, 141-201.

Cutting, J. E., Bruno, M., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: a lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121,* 364-381.

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition, 100,* B21-B31.

Davis, C., & Kim, J. (2004). Audio-visual interactions with clearly audible speech. *Quarterly Journal of Experimental Psychology Section A, 57,* 1103-1121.

De Jong, R., Liang, C. C., & Lauber, E. (1997). 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.

Diehl, R., & Kluender, K. (1989). On the objects of speech perception. *Ecological Psychology, 1*, 121-144.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91,* 176 – 180.

Durgin, F. H. (2000). The reverse Stroop effect. *Psychonmic Bulletin & Review, 7*, 121 – 125.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience, 15,* 399-402.

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

Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience, 17*, 1703 – 1714.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fowler, C. A. (2004). Speech as a supramodal or amodal phenomenon. In G. A. Calvert, C. Spense, & B. E. Stein (Eds.). *The Handbook of Multisensory Processes.* Cambridge, MA: MIT Press.

Fowler, C. (1994). Speech perception: Direct realist theory. In *Encyclopedia of Language and Linguistics* (vol. 8, pp. 4199 – 4203). Oxford, England: Pergamon Press.

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., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: evidence from choice and simple response time tasks. *Journal of Memory and Language, 49,* 396-413.

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

Fowler, C. A., & Rosenblum, L. D. (1991). Perception of the phonetic gesture. In I. G. Mattingly & M. Studdert-Kennedy (Eds.). *Modeularity and the Motor Theory.* Hillside, N.J.: Lawrence Earlbaum.

Fowler, C. A., & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 742 – 754.

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

Gallese, V., Fogassi, L., Fadiga, L., & Rizzolatti, G. (2002). Action representation and the inferior parietal lobule. In W. Prinz and B. Hommel (Eds.). *Attention &*

Performance XIX. Common Mechanisms in Perception and Action (pp. 247 – 266). Oxford, UK: Oxford University Press.

Gentilucci, M., & Bernardis, P. (in press). Imitation during phoneme production. *Neuropsychologia.*

Gentilucci, M., & Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. *Experimental Brain Research, 167*, 66-75.

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

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.

Grafton, S. T., Fadiga, L., Arbib, M. A., & Rizzolatti, G. (1997). Premotor cortex activation during observation and naming of familiar tools. *NeuroImage, 6*, 231 – 236.

Hari, R., Forss, N., Avikainen, S., Kirveskari, S., Salenius, S., & Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proceedings of the National Academy of Sciences of the United States of America, 95*, 15061 – 15065.

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.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science, 286*, 2526 - 2528.

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *NeuroReport, 14,* 1129-1133.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson (2003). "Putting the face to the voice": matching identity across modality. *Current Biology, 13*, 1709-1714.

Kerzel, D. (2002). Evidence for effects of phonological correspondence between visible speech and written syllables. *Psychological Research, 66*, 195-200.

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 634-647.

Kim, J., Davis, C., & Krins. (2004). Amodal processing of visual speech as revealed by priming. Cognition, 93, B39-B47.

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297*, 846 – 848.

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

Kuratate, T., Munhall, K. G., Rubin, P. E., Vatikiotis-Bateson, E., & Yehia, H. (1999). Audio-visual synthesis of talking faces from speech production correlates. *Proceedings of EuroSpeech'99, ESCA.*

Leslie, K. R., Johnson-Frey, S. H., & Grafton, S. (2004). Functional imaging of face and hand imitation: Towards a Motor Theory of empathy. *Neuroimage, 21,* 601-607.

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., Delattre, P., Cooper, F. S., & Gerstman, L. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied, 68,* 1-13.

Liberman, A. M. & Mattingly, I. (1985). The Motor Theory revised. *Cognition, 21,* 1–36.

Liberman, A. M. & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Science, 3,* 254 – 264.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Mann, V., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition, 14*, 211-235.

Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. A. Calvert, C. Spense, & B. E. Stein (Eds.). *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.

Massaro, D. W. (1998). *Perceiving talking faces*. Cambridge, MA: MIT Press.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163-203.

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

Meltzoff, A., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting, 6,* 179-192.

Mottonen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research, 13*, 417-425.

Munhall, K. G., & Buchan, J. (2004). Something in the way she moves. *Trends in Cognitive Sciences, 8*, 51-53.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2003). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological Science, 15*, 133-137.

Munhall, K. G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America, 104*, 530-539.

Nishitani, N., & Hari, R. (2002). Viewing lips forms: cortical dynamics. *Neuron, 36*, 1211-1220.

Nishitani, N. & Hari, R. (2000). Temporal dynamics of cortical representation for action. *Proceedings of the National Academy of Sciences of the United States of America, 97*, 913 – 918.

Porter, R., & Castellanos, F. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America, 67*, 1349-1356.

Porter, R., & Lubker, J. F. (1980). Rapid production 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. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9,* 129-154.

Pulvermuller, F., Huss, M., Kherif, F., Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 7865 - 7870.

Reisberg, D., McLean, J. & Golfield, A. (1987). Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In Campbell, R. & Dodd, B. (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 97 – 113). Lawrence Erlbaum Associates, London: UK.

Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Reviews in Neurosciences, 27*, 169-192.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3,* 131-141.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of actions. *Nature Reviews Neuroscience, 2*, 661 – 670.

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S .T., & Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127*, 141-145.

Schwartz, J., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition, 93,* B69-B78.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech production. *Neuroimage, 25,* 76-89.

Smeele, P. M. T. (1994). Perceiving speech: integrating audio and visual speech. *Unpublished Doctoral Dissertation,* Delft University of Technology.

Soros, P., Sokoloff, L. G., Bose, A., McIntosh, A. R., Graham, S. J., & Stuss, D. T. (2006). Clustered functional MRI of overt speech production. *NeuroImage, 32,* 376 – 397.

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition, 92,* B13-B23.

Stevens, K. N. & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Wathen - Dunn (Ed.). *Models for the Perception of Speech and Visual Form.* Cambridge, MA: MIT Press.

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

Sumby, W. H. & Pollack, I. (1987). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212 – 215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.). *Hearing by eye:The psychology of lip-reading* (pp. 3-52). Lawrence Erlbaum Associates, London: UK.

Sundara, M., Namasivayam, A. K., & Chen, R. (2001). Observation-execution matching system for speech: a magnetic stimulation study. *Neuroreport, 12,* 1341-1344.

Umilta, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing: a neurophysiological study. *Neuron, 32*, 91 – 101.

Van Strien, J. W. (1988). Handedness and hemispheric laterality. Dissertation: Vrije Universiteit Amsterdam.

Watkins, K. E., & Paus, T. (2004). Modulation of motor excitability during speech perception: the role of Broca's area. *Journal of Cognitive Neuroscience, 16,* 978-987.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excited the motor system involved in speech production. *Neuropsychologia, 41*, 989 – 994.

Whalen, D. H., & Liberman, A. M. (1987). Speech-perception takes precedence over nonspeech perception. *Science, 237*, 169 – 171.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*, 701 – 702.

Yehia, H., Rubin, P., & Vatikiotis-Bateson (1998). Quantitative association of vocal tract and facial behaviour. *Speech Communication, 26*, 23-14.

Zhang, H., & Kornblum, S. (1998). The effects of stimulus-response mapping and irrelevant stimulus-response and stimulus-stimulus overlap in four-choice Stroop tasks with single-carrier stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 3 -19.