ABSTRACT

Title of Document: **MULTI-LEVEL AUDIO-VISUAL INTERACTIONS IN SPEECH AND LANGUAGE PERCEPTION**

Ariane E Rhone, Ph.D., 2011

Directed By: Professor William J. Idsardi, Linguistics

That we perceive our environment as a unified scene rather than individual streams of auditory, visual, and other sensory information has recently provided motivation to move past the long-held tradition of studying these systems separately. Although they are each unique in their transduction organs, neural pathways, and cortical primary areas, the senses are ultimately merged in a meaningful way which allows us to navigate the multisensory world. Investigating how the senses are merged has become an increasingly wide field of research in recent decades, with the introduction and increased availability of neuroimaging techniques. Areas of study range from multisensory object perception to cross-modal attention, multisensory interactions, and integration. This thesis focuses on audio-visual speech perception, with special focus on facilitatory effects of visual information on auditory processing. When visual information is concordant with auditory information, it provides an advantage that is measurable in behavioral response times and evoked auditory fields (Chapter

3) and in increased entrainment to multisensory periodic stimuli reflected by steady-state responses (Chapter 4). When the audio-visual information is incongruent, the combination can often, but not always, combine to form a third, non-physically present percept (known as the McGurk effect). This effect is investigated (Chapter 5) using real word stimuli. McGurk percepts were not robustly elicited for a majority of stimulus types, but patterns of responses suggest that the physical and lexical properties of the auditory and visual stimulus may affect the likelihood of obtaining the illusion. Together, these experiments add to the growing body of knowledge that suggests that audio-visual interactions occur at multiple stages of processing.

MULTI-LEVEL AUDIO-VISUAL INTERACTIONS IN SPEECH AND LANGUAGE PERCEPTION

By

Ariane E. Rhone

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor William J. Idsardi, Chair
Professor Naomi Feldman
Professor Norbert Hornstein
Professor David Poeppel
Professor Jonathan Z. Simon, Dean's Representative

Dedication

In memory of my grandparents

Dale & Judy Rhone, and Fred & Mary Sorensen

Acknowledgements

This dissertation was possible because of the support and encouragement I experienced as a graduate student in the Linguistics Department at the University of Maryland. My committee members gave me helpful feedback and new ideas, and I am glad that I was able to get such a smart bunch of people together in one room to ask me tough questions about my work. Bill Idsardi and David Poeppel have pushed me intellectually from the very first day of Sound++ lab meetings to the last minutes of my defense question period, and I am a better scientist because of them. Thanks to Jonathan Simon for serving as my dean's representative and for always finding time to answer signal processing questions over the past five years. Thanks to Naomi Feldman for offering an outside perspective and useful comments on the original draft, and for helping me clarify and strengthen some crucial sections of the thesis. Thanks to Norbert Hornstein for always being supportive, with amusing conversation, big-picture questions, and countless cookies.

Julian Jenkins was my AV buddy for most of my graduate school career, and contributed a great amount of time and energy to the analyses presented in Chapter 4. We started out trying to run what we thought would be a fun and simple audio-visual experiment, and ended up four years later with some interesting data and an actual publication. I don't know how many hours we spent trying to figure out the perfect stimuli and the most appropriate analysis techniques, but I do know that time was NOTHING compared to the days and days we yelled at each other and fought over

the laptop keyboard trying to figure out the best way of writing up our complicated design and even more complicated analysis so we could share our findings with the world. Together we learned that there is no such thing as low-hanging fruit when it comes to multisensory experiments. Thanks also to Diogo Almeida for giving me free rein over the McGurk experiment presented in Chapter 5.

Thanks to Phil Monahan for being my academic big brother, beating me at video games, and introducing me to a ton of good music (and some really really really bad music). Phil has been a great friend (despite all the teasing and the pranks!), and I look forward to collaborating with him on speech perception projects in the future.

Sharing an office (and teaching a class) with So-One Hwang was both hilarious and frustrating (in a good way) because she never let me get away with 'I don't know' as an answer. She encouraged me to think carefully about questions I might otherwise have shrugged off, and I appreciate her patience with me over the last five years.

Thanks also to other upstairs colleagues, especially Pedro Alcocer, Bridget Samuels, Mathias Scharinger, Max Ehrmann, and Dave Kush. It was great to have such a fun group of people to get together with every day (at exactly 12:20pm) to talk about classes, ideas, projects, life, and to produce elaborate pranks.

There are many other people I'd like to thank for being great friends and colleagues, especially Stacey Trock, Tim Hunter, Eri Takahashi, Akira Omaki, Clare Stroud,

Ellen Lau, Veronica Figueroa, Minna Lehtonen, Jon Sprouse, Matt Winn, Brian Dillon, Annie Gagliardi, Shannon Barrios, Wing Yee Chow, Terje Lohndal, Yakov Kronrod, Greg Cogan, Tim Hawes, Maki Kishida, Ilknur Oded, Johannes Jurka, Shayne Sloggett, Myles Dakan, and Nora Oppenheim. Thanks also to Peggy Antonisse, who was the best teaching mentor I could ever ask for. I would also like to thank Kathi Faulkingham and Kim Kwok for answering all my administrative questions with a friendly smile, and Jeff Walker for MEG training and support.

Thanks to Allard Jongman, Joan Sereno, and Ed Auer for giving me the opportunity to do speech perception research "full-time" as an undergraduate at the University of Kansas, and for encouraging me to go to graduate school in the first place.

Thanks to my family for understanding why I had to move so far away for school, and to Kenny's family for understanding why he wanted to come with me. And indescribable thanks to Kenny, for supporting me always.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The human brain possesses the remarkable ability to effortlessly integrate sensations from different modalities into unified percepts in space and time. Despite the ubiquity of multisensory experiences in our everyday life, the study of perception has largely focused on a single modality at a time. The historical bias toward studying individual sensory systems is a reasonable one for many reasons, primarily because these *are* very different systems. For example, the human auditory system detects air pressure fluctuations, uses a specialized transducer, the cochlea, to transduce this information, has a dedicated pathway from cochlea to cortex, and ultimately reaches its primary cortical destination in Heschl's gyrus of the temporal lobe. The visual system, on the other hand, is specialized for detection of photons, uses the retina for transduction, has its own dedicated pathway to cortex (including some structures distinct from the auditory pathway), and finally reaches its *own* primary cortical area– the calcarine fissure of the occipital lobe. These systems, each with its own distinct medium, organ of transduction, subcortical pathways, and primary cortical areas (spanning different lobes of the human brain) have been treated modularly in most studies of perception. The anatomical and physiological differences described briefly above seem to support the necessity of a modular approach to sensory systems. However, it is important to keep in mind that the overall objective of all of the sensory systems is the same: our senses are responsible for converting a distal stimulus into a coherent neural representation and ultimately to invoke interpretation and action. Furthermore, the

fact that these systems work together to provide us with an integrated percept of the world around us suggests that the study of these systems' interactions is also warranted.

Because the goal of the perceptual neuroscientist is to characterize the anatomy and physiology of the human sensory systems, it is important to consider them as near as possible to their actual, real-world roles. And, although these systems have historically been characterized separately, the ultimate task of the perceptual system is the same whether it is unimodally or multimodally considered. On the other hand, the linguist's goal is to characterize the mental representations used in language and also the processes that enable the language user to access and make full use of these representations. To have a full linguistic account of mental representations, a crucial piece of the story must be addressed: how does the external world interact with these representations? Presumably there is be some mapping from the physical world onto our mental representations, but the details of this mapping are not fully understood. This interface between multisensory perception and linguistic processing is the focus of this dissertation.

Chapter 2 contains a review of relevant literature on audio-visual perception, with an emphasis on the advantages that are observed when visible articulation is available in auditory speech perception tasks. I also briefly review neuroimaging literature that suggests that these interactions occur at early stages of processing. In Chapter 3, I show behavioral and neuromagnetic evidence for a flexible audio-visual advantage

using nonsense syllables in two different response set contexts. Chapter 4 contains two experiments that show neural entrainment to periodic audio-visual stimuli that share physical properties of the speech signal. The McGurk effect is explored in Chapter 5, and includes analysis of behavioral responses to large number of stimulus types that differ in phonological context, lexical status, and word position. Together, these experiments show influences of visual speech information on auditory perception at three levels: low-level perceptual processing, optical-phonetic prediction for auditory events, and in incongruent lexical items.

Chapter 2: Multisensory processing in speech perception

Introduction

Perhaps the most relevant multisensory stimulus, in terms of linguistics and cognitive science, is speech. Although typically discussed in terms of acoustic and auditory and properties, there is also a visual component that is inherently linked to the auditory speech signal. The articulation required to make the distinct sounds of a language often has visible consequences: the mandible raises and lowers, the tongue makes contact with articulatory landmarks in the oral cavity, and the lips open, close, make contact, protrude, spread, and round to various degrees. And, although speech *can* be perceived in the absence of these visual cues, they can provide disambiguating information that benefits the listener. For example, a conversation on the telephone lacks the visual information that is potentially utilized in face-to-face settings. As a result, the talker and listener must often make use of additional cues in order to complete the communicative function. When spelling out an unfamiliar name over the phone, one common compensatory strategy involves replacing difficult to understand letters with unambiguous words beginning with that letter, for example distinguishing "en" from "em" by saying "en as in November." In face-to-face situations, however, the visual cues provided by the talker's lips could provide disambiguation for these sounds; articulating the [m] of "em" requires full closure of the lips, while producing the [n] of "en" does not (the tongue makes contact behind the teeth, and the lips do

not close). This is only one example of the communicative advantage of face-to-face conversation in everyday speech perception situations, and while the intuition behind this advantage is straightforward, it is important to make every effort to assess and incorporate this advantage into our linguistically motivated and neurobiologically grounded theories of speech perception. And, although phonetics and phonology are commonly discussed solely in terms of their auditory properties, the study of audio-visual speech perception and its potential advantage in everyday communicative events has received increased attention over the past half century.

Visual speech contributions in suboptimal listening conditions

Trying to have a conversation in a noisy environment (near a busy street or in a loud party, for example) can be difficult, but if you are able to see the face of the person speaking, it seems easier to hear. Many studies have shown that visual speech information can be used to supplement auditory information, particularly in noisy situations or with stimuli that are easy to hear but hard to understand such as listening to your native language spoken by a person with a foreign accent, listening to a native speaker of a language that you are learning, or listening to complex sentences spoken in your native language by a native speaker (Reisberg, McLean, & Goldfield, 1987; Arnold & Hill, 2001).

The quantification of the advantage of audiovisual speech in degraded auditory conditions began with Sumby and Pollack's (1954) measurement of speech intelligibility at various signal-to-noise ratios (SNR) with and without visual speech

information to supplement the auditory signal. Although their study was designed to test possible communicative enhancements for noisy military or industrial workplace environments, this has proven to be the cornerstone of the vast body of work on the psychology of speechreading and the audiovisual advantage.

By evaluating the increase in intelligibility scores when auditory perception was supplemented with visual information in a variety of signal to noise ratios and several vocabulary sizes, Sumby and Pollack showed that the presence of visual speech information improved intelligibility scores, especially in very low SNR[1]. Their major claim was that the presence of bimodal (audio-visual) information resulted in higher resistance to noise or an increase in transmitted signal because allowing participants to see the face of the person speaking resulted in increased intelligibility. And although their major finding was that the visual signal was most helpful in low SNR conditions, this finding was for many years taken to suggest that the visual advantage is somehow more important or most relevant in seriously degraded conditions. However, Sumby and Pollack directly state that the audio-visual advantage is probably greater at poor SNR conditions simply because there is more room for improvement when auditory intelligibility is lower.

Sumby and Pollack (1954) also showed that by varying the vocabulary size that the participants were working within also affected the intelligibility scores. By

[1] A 0 dB SNR would indicate equal levels of target signal and masking noise, and a low SNR corresponds to the listening situation where the level of a masking noise exceeds the level of a target signal.

manipulating the size of the potential response list that was available to participants, they were able to show that the participants were able to obtain the greatest gain from bimodal signals when they had very limited response sets. Most notably, the effect of visual information in very limited (8 word) vocabularies was the greatest—increasing the percentage correct by 80 percent. Compared to the gain in the larger vocabularies (40 percent for the 256 response word list), this finding suggests that listeners performed best when the potential response set was limited. This finding is an important empirical demonstration of listeners taking advantage of reduced uncertainty in speech perception, using whatever information is available during a task (discussed further in Chapter 3).

Both of the findings of the Sumby and Pollack (1954) study are relevant motivators for the current set of experiments; the presence of additional visual speech information and the decreased uncertainty provided by a vocabulary set size both restrict the possible percepts that can result in increased intelligibility. The presence of visual information provides disambiguating information that is often difficult to recover from the auditory signal alone. By reducing the number of possible phonemes with this additional optical phonetic information, the perceiver has reduced uncertainty in the speech perception task. In the same vein, having a limited list of responses also aids the listener in reducing the number of lexical candidates they may have perceived. This decrease in uncertainty in both domains (visual speech and vocabulary size) is likely to facilitate the perception of speech and this facilitation is reflected in the increased intelligibility scores.

Many others (Erber, 1969; Middelweerd & Plomp, 1987; MacLeod & Q. Summerfield, 1990; Sommers, Tye-Murray, & Spehar, 2005) have confirmed that listeners with normal hearing benefit from having visual information available in a speech intelligibility task in degraded auditory environments. Traditionally, the advantage had been assumed to follow an inverse-effectiveness pattern, where visual input was assumed to have a larger impact on auditory speech perception in severely degraded listening conditions. However, Ma, Zou, Ross, Foxe, and Parra (2009) propose a Bayesian optimal model of cue integration for audio-visual speech perception and, via model fitting to a number of behavioral audiovisual speech-in-noise studies, found that the greatest contribution of visual information occurred at moderate SNRs, suggesting that the auditory signal does not have to be severely degraded for audio-visual interactions to occur.

Other studies of visual advantage have focused on listeners with impaired hearing who use visual cues to complement an intact distal stimulus that becomes degraded as a function of atypical auditory transduction caused by hearing loss. This is often not a complete replacement for auditory information, and the ability to lipread (and speechread[2]) is not an automatic consequence of having a hearing loss. Aural rehabilitation programs for people with decreased hearing acuity often incorporate training in lipreading, and although some debate exists regarding its effectiveness

[2] Following the convention of Summerfield (1992), I use the following terminology: "lipreading is the perception of speech purely visually by observing the talker's articulatory gestures. Audio-visual speech perception is the perception of speech by combining lipreading with audition. Speechreading [...] is the understanding of speech by observing the talker's articulation and facial and manual gestures, and may also include audition."

(Summerfield, 1992) supplementary training has been shown to increase in intelligibility scores in sentence recognition (Walden, Erdman, Montgomery, Schwartz, & Prosek, 1981; Richie & Kewley-Port, 2008), which likely translates to increased comprehension in day-to-day communication settings.

Likewise, individuals who receive cochlear implants are often also trained to use visible speech cues to facilitate comprehension of spoken speech (Lachs, Pisoni, & Kirk, 2001; Strelnikov, Rouger, Barone, & Deguine, 2009) in addition to auditory training techniques. Through a number of evaluation measures, it has been shown that speechreading ability in both hearing-impaired and normal hearing populations is highly variable (Bernstein, Demorest, & Tucker, 1998). However, these studies suggest that relative to auditory alone conditions, audio-visual speech perception is nearly always improved (Grant, Walden & Seitz, 1998). The neural mechanism underlying the perceptual advantage provided by visual information is thus the target of investigation in this thesis.

Although lipreading and speechreading can be important strategies for individuals with hearing loss, the more relevant case for the cognitive scientist relates to how visual speech information is utilized in speech perception for the typical listener. As mentioned above, visible articulators (such as the lips, the tip of the tongue, and the teeth) are responsible for creating the sounds of our languages. These articulators are part of the vocal tract filter that, once applied to the glottal source, modifies the acoustic output during speech. If there is an effect of seeing these movements (in

addition to hearing the acoustic consequences of the articulation for individuals without hearing loss), this deserves incorporation into models of speech perception. Furthermore, understanding the neural mechanisms underlying the integration of these two signals is a goal of neuroscience research. Put together, this raises the neurolinguistic question of how the brain integrates the auditory and visual speech information and maps this multisensory signal onto phonetic, phonological, and lexical representations.

For normal hearing listeners, it has been shown that the detection of speech in noise improves for audio-visual relative to auditory-alone stimuli (Grant & Seitz, 1998; Bernstein, Auer Jr, & Takayanagi, 2004), which will be discussed in detail in Chapter 4, and also phoneme detection—especially for real words—is improved (Fort, Spinelli, Savariaux, & Kandel, 2010).

Speech intelligibility scores are also bolstered by the addition of visual information MacLeod and Summerfield (1987) showed that the SNR at which keywords in sentences were identified correctly was significantly lower (i.e., identification was successful in conditions where a masking noise was greater) in audio-visual compared to auditory-alone conditions. Their quantification of the improvement in performance in audio-visual vs. auditory alone conditions offers further support for the Sumby and Pollack (1954) findings, with the additional methodological advantage of having an open response set (i.e. not limiting the vocabulary of possible response words and thus singling out the effect of visual contribution without additional top-

down biases) and also measuring the improvement as a function of threshold SNR for a set criterion level (such as 75%) rather than comparing percent correct in the two conditions (which potentially risks ceiling effects). Furthermore, their use of sentences (rather than isolated words) offers a more realistic evaluation of the use of audio-visual cues in speech perception.

In an additional demonstration of the contribution of visual information in speech perception, Rosenblum, Johnson, and Saldaña (1996) showed improved performance on speech perception tasks that included an impoverished visual input relative to performance without visual input (auditory-alone). They used a point-light display, rather than a natural face, as the visual input in a thresholding task similar to the MacLeod and Summerfield (1987) paradigm, and found that a coarse visual stimulus provided significant gains for audio-visual versus auditory-alone conditions. Point light displays use reflective dots placed on the articulators (usually lips, teeth, mouth, and chin) and special lighting to create video stimuli that contain only the kinematics of the reflective dots (see Fig 2.1). This provides articulatory information about the speech act to the perceiver without providing the extra facial identity information that is present in typical visual stimuli[3].

[3] Note that these "point-light" displays are unrecognizable when presented as a static image, unlike static images of fully illuminated faces, which can provide extralinguistic (affect, race, gender, age) as well as linguistic information (place of articulation, mouth diameter, etc.).

Figure 2.1 Schematic of point-light stimuli used by Rosenblum & Saldaña (1996)
Reflective dots are affixed to the talkers visible articulators (lips, teeth, tongue) and face (chin, cheeks, nose, etc.). When special lighting is used, only the illuminated dots are visible and provide kinematic information without facial detail.

The audio-visual speech identification improvement in the absence of fine spatial detail suggests that it is not necessary to view an actual *face* in order benefit from the information contained in the dynamic visual signal. However, Rosenblum et al. (1996) also found that speech comprehension thresholds improved as the number of reflective points adhered to the face increased. Although a coarse visual stimulus with as few as 14 points on the lip and mouth area was capable of improving thresholds relative to audition alone, increasing the visual resolution by increasing the number of illuminated points resulted in improved performance in the task. Furthermore, their "fully illuminated" condition resulted in the best threshold, so although impoverished stimuli *could* improve performance, the natural face video was the most beneficial relative to auditory alone stimuli. This suggests that the perceiver is able to use whatever information is present in the signal to help them perceive speech, and is consistent with a Bayesian-optimal view.

This is an important, yet often overlooked point; the audiovisual benefit is clearly not an all-or-nothing gain. Rather, this type of result suggests a flexible perceptual process where perceivers are able to take advantage of any and all cues available to them in a particular task. Further evidence for the flexibility and tolerance of visual degradation in the audiovisual speech perception can be found in the results of MacDonald, Andersen, and Bachmann (2000). They applied spatial degradation filters (mosaic transform) to the visual component of McGurk[4]-type audio-visual tokens. They presented dubbed stimuli at various spatial degradation levels and found that coarser visual input caused reduced number of illusory percepts. Interestingly, they also found that as spatial degradation increased, the clarity of the auditory stimuli was reported to increase as well; when the visual stream was more degraded, participants reported the auditory stream as being perceptually clearer. The participants were presumably able to modulate (or weight) their use of the auditory and visual information based on whatever modality was most beneficial to them at the time. This is further support for the flexibility of the perceptual system, and suggests that the audiovisual speech advantage reflects a complicated interplay of both auditory and visual sensory systems.

Responses to incongruent audio-visual stimuli

An additional paradigm for evaluating the contribution of visual information in speech perception has involved mismatched audio and visual signals. These

[4] McGurk-type stimuli generally consist of an auditory bilabial and a visual velar, which can result in a percept that corresponds to neither of the input modalities. This will be discussed further in this chapter, and is used in Chapter 5 as an experimental manipulation.

mismatches may be temporal (intentionally introducing temporal asynchrony) or they may mismatch in content. The most famous example of the latter type of mismatch is the McGurk effect (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978). In this compelling example of the potential effect that visual information can have on 'typical' speech perception, an audio track of a person speaking the syllable [ba] is dubbed on to a video of a speaker articulating the syllable <ga>[5]. A common result of this type of mismatch is the perception of a completely different syllable from what has been provided in either input modality: the listener perceives the alveolar consonant {da} (or, in some reports the labiodental {va} or interdental {ða})[6]. When the audio token [ga] is paired with the visual token <ba>, the resulting percept is often described as combination of the two input signals, such as {bga} or {gba}. The cue for the labial place of articulation is extremely salient (because the lips are highly visible articulators), and this cue seemingly cannot be overridden by discrepant auditory information (discussed further in Chapter 3 and Chapter 5).

Crucially, the perceptual effect goes away when the visual speech information is removed (i.e., the percept is not simply a case of mistaken auditory identity). This phenomenon has now been extensively studied, both for the sake of exploring such a robust effect of cross-modal discrepancy and also as a tool for understanding theoretical issues in audio-visual speech perception. Classic McGurk effect

[5] For audio-visually discrepant stimuli, the following conventions will be used. Items in square brackets [ ] denote the auditory stimulus; items in angled brackets < > denote visual stimulus; items in curly brackets { } denote percept.

[6] Typical McGurk fusion and combination dubs can be viewed at: http://www.files.ling.umd.edu/~arhone/Thesis/Chapter2

replications and expansions have been carried out for adult speakers of various languages including Japanese (Sekiyama & Tohkura, 1991; Sekiyama, 1997; Massaro, Cohen, Gesi, Heredia, & Tsuzaki, 1993), and the illusion persists even when auditory and visual stimuli come from mismatched genders (Green, Kuhl, Meltzoff, & Stevens, 1991). Regardless of the motivations for these McGurk studies, one underlying theme remains clear: visual input affects auditory speech perception, even in the absence of noise or other degradations.

The McGurk effect has also been exploited to test infants' ability to generalize over AV discrepant stimuli. Rosenblum et al. (1997) investigated the McGurk effect in prelingual infants. Five-month-old children were presented with synthetic audio stimuli dubbed onto a natural visual stimulus in a looking time habituation paradigm. After habituation to the congruent audio-visual stimulus /va/, looking time to audio [da] + visual <va> trials was significantly different from habituation, while audio [ba] + visual <va> was not, suggesting that 5-month olds can be influenced by discrepant audio-visual combinations. However, the visual and auditory features that overlap in the <va>+[ba] case (in particular, the shared labiality of these consonants) do not allow strong conclusions to be drawn about a "typical" McGurk effect for these children. Rosenblum et al. (1997) also present a series of follow up experiments to explore alternative accounts for the [ba]+<va> results, with the conclusion that infants can integrate audiovisual speech. The authors (rightfully) do not make strong commitments to issues of innateness or of statistical learning or experience in shaping the McGurk illusion, because within the first five months of life the infant has been

exposed to a great deal of multimodal input in his or her natural language environment.

Moreover, although studies of infants younger than five months may shed light on the developmental path of the McGurk illusion, failure to find expected results could be a result of insufficiently sensitive measures, task restrictions for extremely young infants, or simply physiological differences between adults and infants, since, visual and auditory development continues after birth. More recent electrophysiological studies using Electroencephalography (EEG) have shown effects in in event-related potentials (ERP) around 290 ms post-stimulus onset to McGurk "combination" stimuli ([ga]+<ba>={gba}) in five-month-old infants (Kushnerenko, Teinonen, Volein, & Csibra, 2008), but not for fusion responses ([ba]+<ga>={da}), suggesting that neural response profiles in the developing infant are indeed sensitive to (at least) the most salient audio-visual discrepancies (those involving visual bilabial and auditory non-labial input).

In addition to establishing that children, like adults, are susceptible to the McGurk illusion, the notion of a 'sensitive period' for multimodal integration has also been explored. Schorr et al. (2005) investigated the McGurk effect in 36 children with congenital deafness who had received cochlear implants (CI) and had used them for at least one year. They tested whether the drastically altered sensory experience of the deaf children who subsequently received CI would affect the magnitude of the McGurk effect. Compared to normal hearing controls, children with CI were less

consistent at fusing McGurk tokens. When fusion did not occur, the percept was generally dominated by the visual input, while the auditory signal tended to dominate for normal hearing controls. Importantly, the analysis included only those children with CI who accurately perceived the congruent control tokens, suggesting that a lack of fusion responses was not a byproduct of the child's general auditory perception. Furthermore, the age at which the child received the CI was related to the amount of consistent AV fusion while the effect of age at test and time using CI was not related to consistent bimodal fusion. They report that children who were implanted after about 2.5 years of age were less susceptible to the McGurk illusion, and interpret this finding as support for a sensitive period for developing typical bimodal fusion. The fact that the duration of CI use was not related to performance suggests that fusion ability is not acquired by purely statistical learning from the audiovisual input, although it does require early exposure[7].

Fowler and Deckle (1991) used haptic-acoustic and orthographic-acoustic in an attempt to tease apart theories of integration that rely on associations based on experience or convention from those theories that suggest the illusion is a function of the causal relationship that both modalities share. If two inputs that are related only by convention or association and are not related by the same causal source, such as orthographic-acoustic pairings, are susceptible to McGurk-like illusions, then that would offer support for theories that attribute the illusion/fusion/percept to perceptual integration of the two stimulus sources, such as the Fuzzy Logical Model of

[7] As with any sample taken from a special population, results for this study risk not being generalizable to the typically developing population

Perception (e.g. Massaro, 1987; Massaro & Cohen, 1983) Alternatively, a McGurk-type illusion that occurs for multimodal input that is not likely to result from typical experience or convention but which are causally related, such as haptic-acoustic pairs, suggests a model of perception in which representation of *events* is stored and thus susceptible to illusion, such as the Direct-Realist Theory (Fowler, 1986). To investigate these conflicting hypotheses, 3-formant synthetic audio stimuli were presented alone, in conjunction with independently paired orthographic syllables, and with independently paired "felt" syllables (audio + haptic) to determine which of the latter conditions would elicit McGurk-like illusions. Acoustic + orthographic trials required subjects to listen to an audio stimulus presented in synchrony with a visual display of an orthographic syllable, acoustic + haptic trials were mouthed by one of the authors in approximate synchrony with the auditory presentation of a syllable while participants felt the speaker's mouth with their hands. Cross-modal influences were found for the acoustic + haptic condition but not for the acoustic + orthographic condition, contrary to exemplar models of perception that would require some kind of experience to form the prototypical representation in memory, because participants had no previous experience perceiving spoken syllables haptically. Instead, they suggest that their results support the Direct Realist model (Fowler, 1986), which uses events as the basic unit of perception.

Brancazio (2004) explored the influence of lexical status on the magnitude of the McGurk effect in normal hearing adults in a series of three studies. The studies aimed to assess whether top-down effects of the lexicon would bias McGurk responses

toward lexical items by exploiting a modified version of the "Ganong effect" (Ganong, 1980) where listeners are biased toward perceiving actual lexical items rather than nonwords. For example, in the Ganong 1980 study an ambiguous alveolar segment with a voice onset time between prototypical /d/ and /t/ was more often perceived as /t/ when followed by "_ask" (preference for the real word percept "task" rather than nonword percept "dask") and as /d/ in the context "_ash" (preference for real-word "dash" rather than non-word "tash"). Brancazio (2004) used audiovisually discrepant stimuli that varied on lexical status of the physical and potentially illusory percept (e.g., [belt] [8] + <dealt> (both words), [beg] + <deg> (auditory word, McGurk nonword) [besk] + <desk> (auditory nonword, visual word) and [bedge] + <dedge> (both nonwords)). Brancazio expected more visually influenced responses and fewer auditory responses when the illusory[9] percept formed a word than when it formed a nonword, and found that participants did show fewer auditory-dominant responses for tokens created by dubbing an auditory nonword to a visual word. Conversely, more auditory-dominant responses were reported when a real auditory word was dubbed to a video nonword in both speeded identification and free response paradigms. These results were interpreted as support for models of lexical access that posit audio-visual integration occurring before, or possibly during, lexical access, which he used to question assumptions of modularity, because higher-level factors (in this case, lexical status) appear to have influenced lower level perception. However, in the free response task it is possible that lexical strategies were at play, and speeded detection

[8] In the case of actual lexical items (rather than nonsense syllables), orthographic representations will be used so that stimulus comparisons may be clearer to the reader.

[9] Note that Brancazio (2004) does not use typical McGurk [ba] + <ga> = {da}, but instead considers any variation from the auditory input as an illusory response (see Chapter 5, and *Discussion*)

tasks may also tap into "higher" processing stages despite efforts to minimize these effects.

Windmann (2004) examined the effect of sentence context on the McGurk illusion in adult speakers of German by manipulating expectations: holding the physical input constant, the subjects were presented with real word McGurk stimuli that occurred in either expected or unexpected sentence context. An effect of sentential context would suggest that the listener treats the McGurk illusion no differently than ambiguous or noisy phonemes. An increased number of McGurk illusions were reported and were given higher goodness ratings when the sentential/semantic context was biased toward the illusory percept compared to environments in which the McGurk illusory percept was unexpected, which suggests that previous descriptions of illusion that emphasized the "autonomy and cognitive inaccessibility" were inaccurate, and that the discrepant audio-visual stimuli that form the McGurk illusion are in fact no different from noisy/ambiguous phoneme stimuli, and no different from the rational/optimal perceiver described above. Windmann also suggests that the illusion is probabilistic and experience dependent rather than a hardwired, automatic, innate process. However, the question of whether the illusion affects primary perception or is a result of a post-perceptual artifact could not be directly answered from this study. Further investigation in the time course of the McGurk effect using electrophysiological methods may elucidate the issue, although Windmann (2005) lacks clear definition of what 'post perceptual' is and how the perceptual/post perceptual distinction could be reliably assessed. In contrast, Sams et al. (1998) found

no effect of sentential context using McGurk-type stimuli with Finnish speakers. The complicated interplay between the perception of an incongruent audio-visual item and the processing involved in making an overt response in the experimental task is likely responsible for these—and potentially other—seemingly conflicting results, and have continued to be explored (see Chapter 5).

Despite the huge number of replications, expansions, and variations, the full range of uses for McGurk-type stimuli has yet to be fully explored. Although the sample population and unique input pairings have been manipulated in numerous ways, the actual structure of stimuli has remained remarkably consistent—most typically employ individual nonsense syllables (usually CV or V.CV) and a lesser number of studies using real words (see the discussion of Brancazio (2004), this chapter; Easton and Basala (1982) and Dekle et al (1992), Chapter 5). The vowel contexts /i/, /a/, and /u/ have been generally accepted as a representative vowel inventory for these studies, but differences have been found even within this narrow phonetic context (Hampson et al, 2003), and the stimulus set size can also have an effect on reported percepts (Amano & Sekiyama, 1998).

Responses to temporally mismatched audio-visual stimuli

In addition to mismatches in content, as in the McGurk-type stimuli, temporal mismatches have also been utilized in the study of audio-visual speech perception. Although one might expect listeners to be most sensitive to temporally synchronous audio-visual events, van Wassenhove, Grant, and Poeppel (2007) showed that

temporal asynchronies of up to 80 ms are tolerated when the auditory signal precedes the visual signal, and up to 131 ms in the reverse case (visual preceding auditory information) in a simultaneity judgment task. In fact, the plateau point for the judgments was not at 0 ms (absolute synchronous), but was centered at 23-29ms of auditory lag (depending on syllable type). This suggests that an absolute zero time point for synchronization is not necessary for the perception of simultaneity, at least for audio-visual speech.[10] Predicting the onset of sound, then, is unlikely to be the major driving force behind the audiovisual speech advantage, because these results suggest that even out-of-synch audio and video signals can be tolerated and perceived as congruent.

Soto-Faraco & Alsius (Soto-Faraco & Alsius, 2009) tested whether fused McGurk-type percepts would differ in simultenaity judgments from non-fused responses that were temporally offset would affect the responses. They found that even when percepts are fused at a categorical level, participants can still be aware of temporal mismatches of the stimuli. Because the participants perceived the fusion, yet still were able to detect temporal mismatches, Soto-Faraco and Alsius (2009) conclude that multisensory integration is a non-homogenous process where different attributes of physical stimuli are bound at one processing stage, or that perceivers recover multiple representations of the same physical event. This view challenges some previous assumptions about the modularity of the integration system, but makes

[10]Caution should be taken to avoid over interpretation of this result—the 23-29ms lag reflects only the center of the model that was fit to their data. It is important to note that performance at the 0ms lag was near that of the centered plateau point.

progress toward a more realistic model of audio-visual speech perception. Given that unisensory object perception occurs at different stages, it is not unreasonable to expect multisensory interactions to occur at various stages as well.

With the exception of McGurk-type and asynchronous audiovisual perception studies, studies of the contribution of visual information in speech perception in undegraded listening conditions have been restricted by the overall success of auditory speech perception without vision. The level of accuracy at which normal hearing individuals perform in non-degraded auditory intelligibility studies is so high that it is difficult to evaluate the contribution of an additional visual signal. This has resulted in a large number of studies that have utilized degraded listening conditions to evaluate whether performance on intelligibility tasks improves with the presence of visual speech information (Erber, 1975). These studies are often thought to be more ecologically valid than McGurk-type studies (because the likelihood of encountering speech in noise far outweighs the likelihood of encountering discordant audio-visual information); however, in order to incorporate visual features into models of speech perception it is important to understand how these features contribute to speech more broadly construed. In other words, it is important to evaluate whether a listener uses visual cues and features in a scenario where he has access to undegraded auditory information. Otherwise, visual speech is relegated to its previous status as a compensatory strategy rather than a valid source of speech information.

Visual speech contributions in intact listening conditions

The finding that visual speech increases intelligibility scores or decreases detection thresholds in noise confirms the intuition that seeing a talker is beneficial in degraded auditory situations. However, recent work has suggested that visual speech is also utilized in clear auditory listening conditions. By showing that visual information is not simply a backup or compensatory strategy in speech perception, these studies provide even stronger support for the claim that visual speech information should be taken into account when constructing theories of speech perception in general.

One approach that has been utilized to avoid the ceiling effect seen in many auditory-visual experiments (where auditory perception is so good that it is difficult to find improvement with the addition of visual information) takes advantage of situations where speech is easy to *hear* but difficult to *understand*. For example, Arnold and Hill (2001) showed that listener comprehension improves with the presence of visual speech information in difficult to understand passages from the Neale reading assessment (Neale, 1999). When participants were able to hear and see as a person speaking a conceptually difficult passage of text (e.g. 'Other knowledge is accumulated from international co-operative endeavors to create sanctuaries for vulnerable species of bird, and animals and plants') they performed better at comprehension tasks following the passage than when presented with audio alone. Although this result does not address the specific question of how auditory and visual information is perceived and integrated, it does suggest that the presence of visual information may reduce the burden of the auditory speech perception system and

potentially "frees up" mental resources that are then presumably available for higher level comprehension.

The presence of visual cues has also been shown to facilitate perception of non-native phonemic contrasts in a second language. Navarra and Soto-Faraco (2007) showed that native Spanish-dominant bilingual speakers of Spanish and Catalan were unable to distinguish the Catalan phonemes /e/ and /ɛ/ in a unimodal auditory task, but with the addition of visual information the listeners did show discrimination ability. In a speeded syllable classification task, disyllabic nonword lists were presented that either held the second-syllable vowel constant (always /e/ or always /ɛ/ within a given list) or had both types of vowels in the second-syllable vowel (50% /e/ and 50% /ɛ/ within the list). Participants have to categorize the first syllable, and reaction times are collected. In this paradigm, increased reaction times are expected for the mixed "orthogonal" condition relative to the constant "homogenous" condition only if the listener can discriminate the two sounds.

When only auditory information was presented, Catalan dominant bilinguals showed discrimination (increased reaction time in the orthogonal relative to the homogenous), and Spanish dominant bilinguals did not differ between the two lists. However, with the addition of visual information that facilitated discrimination of these vowels (see Figure 2.2) Spanish dominant bilinguals did show increased reaction times to the orthogonal condition.

Figure 2.2 Frames from the /ɛ/ (left) and /e/ (right) stimuli from Navarra and Soto-Faraco (2007). With the addition of visual information, Spanish-dominant bilinguals were able to discriminate sounds that they were not able to discriminate auditorily.

Furthermore, Wang, Behne, and Jiang (2008) found that native Mandarin speakers who spoke Canadian English as a second language showed improved performance on identification of interdental consonants not present in the Mandarin speech sound inventory when the speech was presented audiovisually rather than only auditorily. These results suggest that listeners are able to take advantage of visual cues even when dealing with contrasts (either auditory or visual) that are not present in their native language. This addresses the role of specific experience with auditory-visual combinations. The observation that visual information is helpful even with contrasts that are not regularly encountered by a listener suggests that the advantage is not due to experience with regularly encountered auditory-visual pairings from their own language. Furthermore, evidence from infant studies has shown that within a few months of birth, human babies are sensitive to the congruence of audiovisual pairings outside of the McGurk paradigm (Kuhl & Meltzoff, 1982; Baier, Idsardi, & Lidz, 2007).

Although the individual studies described above have addressed very specific questions and manipulated different parameters of audio-visual speech stimuli, they all converge on one important point: auditory perception is flexibly influenced by visual information. The visual information that is available in conjunction with auditory speech is utilized by listeners in a variety of situations, and listeners are able to take advantage of whatever cues are available to them in order to perform in these experimental tasks as well as to perceive audiovisual speech events in the real world. Furthermore, the use of audiovisual speech cues does not seem to be dependent on experience with particular articulator-acoustic pairings, suggesting that perception of audio-visual speech events is generalizable to new multisensory combinations and the audio-visual advantage probably reflects a general cognitive process rather than an effect of experience.

Neural correlates of audio-visual (speech) perception

Recently, the application of neuroimaging methods such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (EEG), and magnetoencephalography (MEG) has provided an additional tool for understanding audio-visual speech perception. These electrophysiological (MEG, EEG) and hemodynamic (fMRI, PET) studies tend to investigate when and where AV signals are integrated in the brain and to test hypotheses of "early" versus "late" integration. While the stage at which integration takes place is still a topic of much debate (Campbell, Dodd, & Burnham, 1998) a

pattern of results has begun to emerge which suggests that multisensory interactions, if not integration per se, occur early in the processing stream.

Neuroanatomy of multisensory processing

Lesion studies and anatomical tracings from nonhuman primates were the initial source of information about sensory processing before noninvasive functional imaging became widely available. These studies suggested that multisensory binding took place after being processed extensively as unisensory streams.

One of the most notable models for sensory processing was proposed by Mesulam (1998). This mechanism included multisensory integration stages that incorporated feedback from higher synaptic levels (the "heteromodal" areas) to modulate and influence synaptic activity in unimodal or sensory areas. Mesulam's (1998) model includes both serial and parallel processing streams, and is anatomically precise. However, this mechanism is perhaps too broad, as conclusions are often interpreted as support for the model. In particular, the "transmodal areas" that Mesulam discusses have been used as a catch-all for any brain area that does not have strict specificity. The applications for human behaviors such as working memory, language, and object recognition is clear from Mesulam (1998), but the presence of interactions at earlier stages has called for more attention to the processing of multisensory stimuli at the level of unimodal sensory areas.

Classic neuroanatomical structures that have been considered potential sites for multisensory convergence include the anterior superior temporal gyrus (STS), posterior STS (including temporal-parietal association cortex), ventral and lateral intraparietal areas, premotor cortex, and prefrontal cortex. Subcortically, the superior colliculus, claustrum, thalamus (including supragenicular and medial pulvinar nuclei), and the amygdaloid complex have also been implicated in multisensory perception (see Figure 2.3 from Calvert & Thesen (2004) presented below, and Campbell (2008) for reviews).

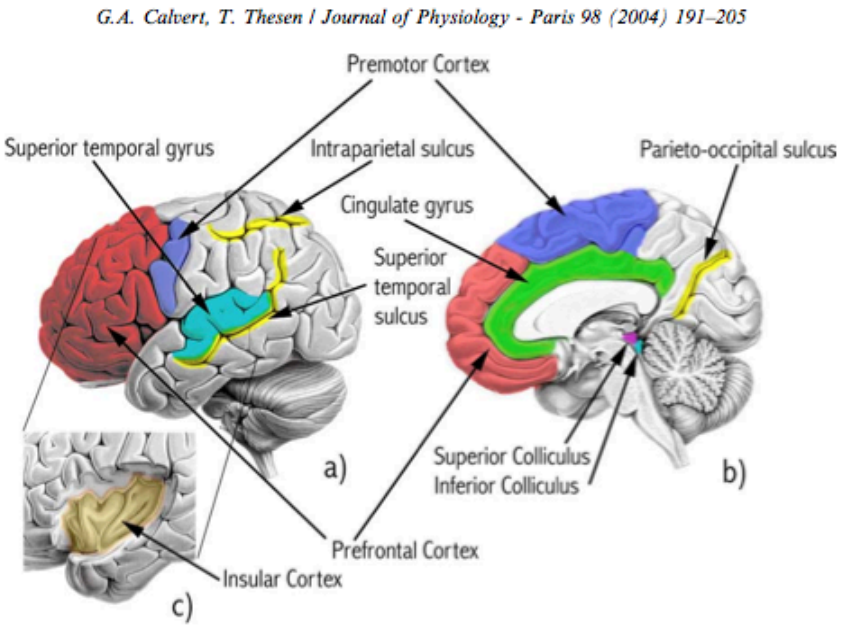


Figure 2.3 Potential sites of audio-visual integration in humans (from Calvert and Thesen, 2004)
a) lateral and b) mid-sagittal view of left hemisphere integration sites; c) shows insular cortex (portion of temporal lobe removed)

However, more recent studies have suggested that multisensory processing is not limited to "association" areas, but that areas once considered unisensory are also involved in multisensory processing. Converging evidence from human and

nonhuman primate studies of the anatomical pathways and response patterns over now suggest that neocortex is fundamentally multisensory, and that preferential responses for one modality does not preclude any area from having interactions with other sensory systems (Ghazanfar & Schroeder, 2006).

In an fMRI study of normal hearing listeners, Calvert et al. (1997) found increased BOLD response in left lateral temporal cortex (including part of Heschl's gyrus) for silent speech relative to non-linguistic facial movements. This activity in auditory cortex—despite the lack of auditory stimulation—suggested that the cortical network for auditory speech perception was also sensitive to *visual* speech (with a replication without scanner noise (MacSweeney et al., 2000)). MacSweeney et al. (2001) tested the effect of audio-visual experience on the activation of auditory cortex by visual speech by comparing normal hearing participants with deaf individuals (all profound hearing loss from birth who used speechreading as their primary form of communication) and found significantly less temporal activation for speechread silent numbers (relative to a still face) for the deaf group than the normal hearing group, suggesting that the development of the network involved in this response is affected by experience.

Bernstein et al. (2002) questioned the findings of Calvert et al. (1997) on the grounds that cross-subject averaging obscured the site of true activation on an individual-by-individual basis. They attempted to replicate the findings of Calvert et al. (1997) and did find significant areas of activation *around* Heschl's gyrus, but not in primary

auditory cortex proper. However, Pekkola et al. (2005) –using a stronger magnet (3T compared to the 1.5T magnet used by Bernstein et al., 2002)—localized the BOLD signal based on individual anatomical landmarks and found support for the original Calvert et al. (1997) finding that primary auditory cortex, in addition to surrounding areas, was activated by silent speechreading in normal hearing individuals.

It should be noted that activation in primary sensory areas does not preclude synaptically higher areas from functioning during multisensory processing. Furthermore, the poor temporal resolution of hemodynamic methods does not provide direct evidence for or against "early" or "late" effects, except in terms of the neuroanatomical pathways (i.e., the number of synaptic junctions between the sensory organ and a particular cortical area). In addition to the inherently slow temporal resolution of the BOLD response (on the order of 6 seconds), the particular design (block, event-related, etc.) and analysis methods (control or subtraction condition, voxel size, contrast level, etc.) used in each of these studies can also influence the outcomes that are reported and should be carefully considered. These discrepancies could be true non-replications or could be attributed to differences in technique and paradigm. If anything, the fact that several studies show "early" multisensory effects and several others support "late" integration can offer support for a multi-stage audio-visual interaction and integration model, because, as discussed by Soto-Faraco and Alsius (2009), it is likely overly simplistic to think that multisensory convergence, interaction, binding and integration can be described as a monolithic process. Rather, the processing of multisensory stimuli likely unfolds over brain-space and time, just

as other processing is now accepted to do. With that in mind, seemingly conflicting accounts of multisensory processing can be reframed as evidence for a vast network of interactions occurring in sensory cortex as well as areas that have been traditionally considered heteromodal.

Electrophysiology of audio-visual speech

Although hemodynamic studies have shown that cortical areas long presumed to be unimodal can be affected by multimodal stimuli, the poor temporal resolution of these methods does not allow for a detailed understanding of the time course of audio-visual speech perception. A more reliable way to address *when* multisensory perception is occurring is through the use of electrophysiological methods, which offer a more direct measure of the electrical activity of (and the corresponding magnetic field generated by) populations of neurons, at a temporal resolution of about 1 ms.

Sams et al. (1991) investigated the effect of conflicting visual information on an auditory MEG response using McGurk-type stimuli. They showed that the neuromagnetic "change detection" response known as the mismatch magnetic field (MMF - generated in auditory cortex and generally considered indicative of pre-attentive changes in auditory properties (Sams, Paavilainen, Alho, & Näätänen, 1985)) could be elicited by a change in audio-visual percept without change in the acoustic stimulus. As in a standard MMF oddball paradigm, participants were presented with a large proportion of one type of token (standard), with infrequent

tokens of a different type interspersed randomly (deviant). Crucially, the only property that differed between the standard and deviant types was the *visual* stimulus; the auditory stimulus remained the same. However, because the McGurk-type audiovisual dubbing results in the perception of a "fusion" consonant that is not present in either the auditory or visual physical stimulus as described above, there is a perceived change in the deviant stimulus.

For example, participants were presented with:

Auditory Stimulus:	[pa]	[pa]	[pa]	[pa]	[pa]	[pa]	[pa]	[pa]
Visual Stimulus:	<pa>	<pa>	<pa>	<pa>	**<ka>**	<pa>	<pa>	<pa>
Perceived Stimulus:	{pa}	{pa}	{pa}	{pa}	**{ta}**	{pa}	{pa}	{pa}

to see if an MMF was elicited to the deviant stimulus (here in **bold**). They found that an MMF response was elicited to perceptually deviant items, despite the fact that the auditory component of the "deviant" was physically identical to "standard." They contrast this with several control conditions (visual speech alone, or a change from standard red to deviant green lights) where the visual input changed and auditory input stayed the same, but in these cases—where no fusion of the auditory and visual signals occurred—no MMF was elicited (see Figure 2.4).

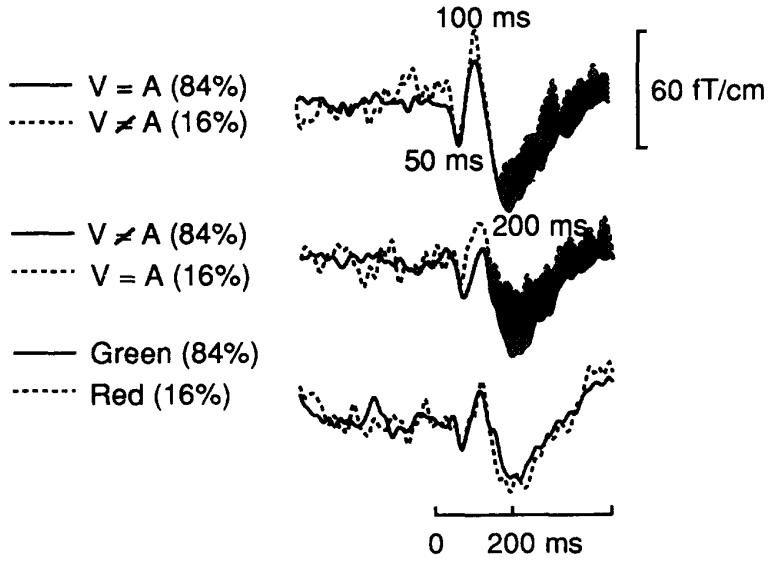


Figure 2.4 MMF responses to McGurk stimuli found by Sams et al. (1991)

Based on the localization of this MEG response, they concluded that visual input is able to affect responses generated in auditory cortex. Kislyuk, Möttönen, and Sams (2008) followed up on this study by examining whether an MMN would be elicited to a change in auditory stimulus in the absence of a change in percept. Unlike the Sams et al (1991) paradigm, the auditory standard matched the oddball in *percept* rather than in acoustic properties (the standard was auditory [va] + visual <va> = perceptual {va} and the deviant was auditory [ba] + visual <va> = perceptual {va}), and no MMN was elicited (see Figure 2.4), confirming the prediction of Sams et al. (1991).

These studies do offer support for a relatively early audiovisual integration process, because the MMF response that they observed occurs as early as 140-180ms after the onset of the auditory stimulus. Because the MMF is traditionally localized to the auditory cortex, and support the idea that primary sensory areas are at least interacting

with—and possibly integrating with—other sensory systems prior to that point of processing. This offers support for an early (pre-categorical) model of AV interaction and also demonstrates that sensory-specific "auditory" areas can be sensitive to extra-auditory influences.

Although Sams et al. (1991) and Kislyuk et al. (2008) provide a useful discussion of visual effects in auditory cortex, these studies provide only an indirect upper bound on multisensory integration effects in these particular neurophysiological responses. Investigations into earlier auditory responses have been explored more recently in an effort to establish more precisely when and where audio-visual integration occurs.

Much of the neuroimaging literature on audio-visual interaction has focused on simple audiovisual stimuli; pairings of simple multisensory objects (such as a light flash and a pure tone) provide the opportunity to test neural correlates of multisensory integration in highly controlled experiments but risk not scaling up to more complex (yet more ecologically valid) stimuli such as speech. For example, Shams, Kamitani, Thompson, and Shimojo (2001) presented subjects with paired flashes and tones and found that the auditory signal affects sensory evoked responses (the visual C1 and N1 in ERP). Behaviorally, this is tied to an illusory phenomenon first described by Shams, Kamitani, and Shimojo (2000) where listeners were presented with a single brief flash of light combined with either a single beep or multiple beeps and reported seeing multiple lights in the condition where they heard multiple tones. The ERP study demonstrated neurophysiological correlates of the illusory flash by testing a

sensory evoked component, the visually evoked potential (VEP) that traditionally had been thought to reflect only unimodal visual processing. This behavioral finding and the subsequent ERP results offer further evidence for an early interaction of auditory and visual processing in time windows and cortical regions previously thought to reflect dedicated unimodal sensory processing.

Studies of multisensory object recognition have provided further support for audiovisual interactions at an "early" stage. Giard and Peronnet (1999) collected simultaneous ERP and behavioral categorization data to test the time course of audio-visual integration and to localize brain regions active in multisensory object recognition. They introduced participants to two objects that could be categorized by their auditory or visual properties (by pairing one tone frequency with a circle that deformed into an ellipse vertically and a different tone frequency with a circle that deformed horizontally) and investigated the reaction time (RT) and evoked responses to the objects in audio-alone, visual-alone, or audio-visual presentations. They tested whether the reaction time to objects in audiovisual presentations would have shorter reaction times than those that were presented unimodally. Because both sensory streams provided unambiguous cues to the "identity" of the object, the audiovisual condition provided redundant information for the categorization task. The facilitation provided by this redundancy is reflected behaviorally in reduced reaction times and increased accuracy in the audio-visual stimuli condition. The simultaneous ERP experiment showed that the multisensory presentation/redundant cues condition

showed effects in the first 200ms after stimulus onset, a time window that had often been regarded as reflecting sensory-specific processes.

They also showed that the sum of the ERP waveform to auditory-alone and visual-alone stimuli did not equal the ERP waveform to combined audio-visual information. The difference between the sum of auditory (A) and visual (V) and the audiovisual (AV) wave was taken to reflect neural processes involved specifically in integrating the two modalities. The logic behind this is as follows:

$$ERP\ (AV) = ERP\ (A) + ERP\ (V) + ERP\ (A \times V\ interactions)$$

If the A and V signals have been processed separately up to the level of the sensory A or V ERP generators, the A x V interactions should be zero and the sum of ERP(A) + ERP(V) should equal ERP (AV). If, however, there is A and V integration at or before the level of processing reflected in the ERP component, any A x V interactions will be reflected in the difference between the right and left sides of the equation.

Giard and Peronnet (1999) found an increase in ERP amplitude in the auditory N1[11] wave for AV versus A-alone object recognition, along with a decrease in the amplitude of the visual ERP wave N185[12]. They also found behavioral facilitation reflected in reduced reaction times for categorization of combined AV stimuli relative

[11] Auditory N1: auditory evoked response, generated in auditory cortex and sensitive to physical properties of the stimulus
[12] N185: visual evoked response, generated in visual cortex

to either of the unimodal stimuli. This finding violated the *race* model of redundant information processing, which would have predicted the AV reaction time to be equal to the fastest of the unimodal conditions, and led Giard and Peronnet (1999) to conclude that "multisensory integration is mediated by flexible, highly adaptive physiological processes that can take place very early in the sensory processing chain and operate in both sensory-specific and nonspecific cortical structures in different ways." However, it is unclear whether the arbitrary audio-visual pairing used in this study reflects different cognitive demands and potentially different processing strategies than intrinsically linked audio-visual stimuli such as speech.

Klucharev, Möttönen, and Sams (2003) tested audiovisual facilitation effects to congruent ([a]+<a>) and incongruent ([a]+ <y>) vowel pairings and found two distinct ERP correlates of audiovisual integration and processing. They found an early (roughly 85ms post auditory onset) audio-visual integration effect that was not affected by congruence (when audio and visual were mismatched in content, the ERP to AV was still greater than that of A + V alone) and also a later component, peaking around 155ms post auditory onset, which was sensitive to the congruence of the audio-visual stimuli. They interpret this result as a demonstration of one early, pre-phonetic effect of having any multisensory speech stimulus and a separate later, post-phonetic process that is sensitive to the content of the audio-visual stimuli. Their finding offers further support for a model where audio-visual interactions do not occur only once in the processing stream, but unfold over time. This viewpoint has

since gained increased attention with the addition of similar studies that have shown variation in sensitivity to different manipulations of multi-sensory signals.

In an effort to bridge the gap between highly-controlled, yet arbitrary, pairings of audiovisual stimuli and the more complex, yet more ecologically valid, stimuli such as speech, Besle, Fort, Delpuech, and Giard (2004) used a similar simultaneous behavioral/EEG study to explore the neural mechanisms underlying these audio-visual speech integration and behavioral multisensory facilitation effects. They investigated whether the auditory ERP responses to combined audio-visual speech stimuli differed from the summation of responses to auditory and visual speech stimuli presented alone. Based on the result of the findings of Giard and Peronnet (1999), they expected to see nonadditive responses reflecting the interaction or integration of auditory and visual information in the speech perception process. Of the ERP components that they analyzed, the auditory N1 (with generators in auditory cortex (Picton, Woods, Baribeau-Braun, & Healey, 1976; Näätänen & Picton, 1987) showed the greatest A x V interaction effects. This offers further support for relatively early effects of visual information on auditory processing. However, rather than the *increased* nonadditive amplitude for auditory responses seen in Giard and Peronnet (1999), Besle et al. (2004) found a *decrease* in the amplitude of the auditory N1. Although previous findings had shown amplitude enhancement for audio-visual

stimuli, they suggested that this difference reflects different processes in speech vs. nonspeech multisensory processing[13].

Besle et al. (2004) also collected reaction time data for their stimuli and found that the reaction time for audiovisual stimuli was faster than reaction times to either AO or VO stimuli. Crucially, their results supported Giard and Peronnet (1999) by also falsifying *race* models where the reaction times for multimodal signals would have been determined by the first of the unimodal processes that was completed. Their finding that reaction times to audio-visual stimuli were faster than either stimuli presented either auditorily or visually alone offers support for a model of audiovisual speech integration that has the multisensory stimuli interacting at a predecisional stage of processing to facilitate speech recognition as reflected by response times. Furthermore, the demonstration of the influence of visual information on auditory cortical responses demonstrates and supports a role for crossmodal computations in areas of the brain that were traditionally considered to be unisensory.

Additional facilitation effects were also shown by van Wassenhove, Grant, and Poeppel (2005) for audiovisual speech relative to auditory speech stimuli. This effect was not only reflected in the nonadditive amplitude of evoked responses, but was also shown in the timing of the canonical late auditory evoked responses. Specifically,

[13]Differences in stimuli and experimental methodology could also have contributed to this result. There are reports of multisensory stimuli resulting in response enhancement and as well as suppression depending on electrophysiological technique and experimental task (see Vroomen & Stekelenburg (2010) for a description of these differences).

they showed that the visual information in an audiovisual speech stimulus results in decreased amplitude and reduced latency of the auditory ERP components N1 and P2 relative to the latency to these components for the same stimuli presented only auditorily (see Figure 2.5). Latency effects were found to be "articulator specific" with the greatest facilitation for the bilabial syllable type /pa/, which also had the highest visual-alone identification accuracy. They interpret this speedup in evoked response latencies as evidence for predictive coding in the brain areas responsible for speech processing, and propose a model of audio-visual speech perception where the salience of the visual anticipatory articulation modulates the amount of facilitation in evoked responses to its corresponding auditory signal.

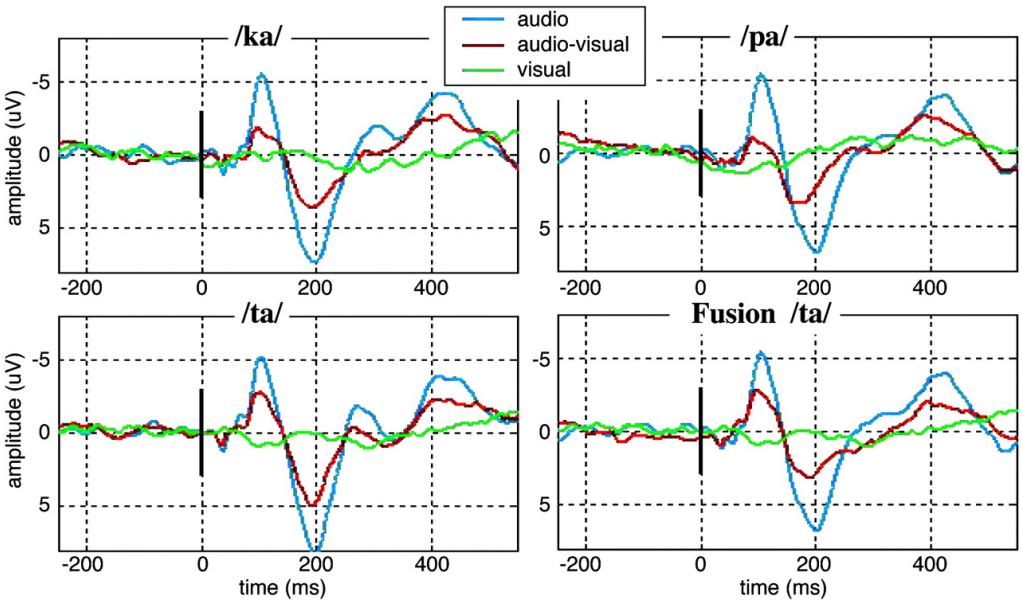


Figure 2.5 Evoked auditory responses to audio, visual, and audio-visual stimuli (from van Wassenhove et al., 2005).

Stekelenburg and Vroomen (2007) found that ecologically valid nonspeech events with visual anticipatory movement (two hands coming together to make a 'clap' sound and a spoon tapping a cup) showed facilitation (reduced amplitude and latency

for N1 and P2 components) relative to the same auditory stimulus presented without the accompanying visual information. Ecologically valid audiovisual events that lacked preceding anticipatory visual movement (a hand abruptly tearing a piece of paper, a handsaw abruptly cutting wood) did not show facilitation in these same electrophysiological responses.

Although Stekelenburg and Vroomen's findings support the observation that electrophysiological audio-visual facilitation effects are not specific to speech, the overall finding that facilitation occurs in the presence of visual anticipatory movements (also confirmed in a follow-up nonspeech experiment (Vroomen & Stekelenburg, 2010)) offers further support for an electrophysiological correlate of the audio-visual advantage. The reduction in latency (and amplitude) of cortical auditory evoked responses suggests that auditory feature analysis can occur before an auditory stimulus actually begins if there is predictive visual information in the signal.

Pilling (2009) replicated Besle et al. (2004) and van Wassenhove et al. (2005) and found amplitude reduction of ERP N1-P2 responses for synchronized audio-visual speech relative to audio-alone speech. However, the amplitude reduction effects were not seen when the audio and visual signals were temporally asynchronous (temporally offset to values beyond the window of audio-visual integration (Dixon & Spitz, 1980; van Wassenhove, et al., 2007), suggesting that this reduction is a marker of

integration rather than simply an attentional byproduct of presenting multimodal signals.

Arnal et al. (2009) replicated the results of van Wassenhove et al. (2005), showing overall latency facilitation for the M100 response (the MEG equivalent of the ERP N1 response) for audio-visual relative to audio-alone stimuli. Furthermore, they found that M100 facilitation was greatest for the syllable types that had the highest identification accuracy when presented visually. Audio-visual M100 facilitation effects were also found for incongruently dubbed syllables, suggesting that the facilitatory processes are at play whenever visual speech information is present, regardless of whether the signals were congruent. Audio-visual stimulus congruence effects were seen in later responses (about 20 ms after the M100 responses), suggesting an initial facilitation effect, followed by an "error detection" response for the incongruent stimuli.

Arnal et al. (2009) also performed a functional connectivity analysis between visual motion and auditory areas, which showed effects that were dependent on the degree of visual predictability (as measured by visual-alone performance) and the congruence of the audio-visual stimuli. They propose a dual-route model where potentially predictive visual anticipatory information provides cortico-cortical facilitation reflected in the M100 response, followed by the error signal generated in STS and fed back to the auditory cortex for stimuli that do not match the visually produced expectations about the auditory event. Because there is a natural lag

between the onset of visual articulatory information and the ultimate auditory event, this system is in place to make predictions about (and later corrections to) predicted auditory features. This again offers support for a multi-stage model of multisensory integration, and shows strong evidence for the model proposed by van Wassenhove et al. (2005).

This set of electrophysiological facilitatory findings is important for several reasons. First, these studies demonstrate a clear neurophysiological difference in processing of auditory compared to audio-visual speech without the drawbacks of ceiling effects in typical undegraded behavioral auditory/audiovisual speech perception behavioral tasks. Second, the early cortical timing and localization of these effects to primary auditory cortical areas (Näätänen & Picton, 1987) gives support for an early integration model of audiovisual speech perception. These effects begin roughly 100ms post-stimulus onset and are localized to early sensory areas rather than "higher level" association areas, which refutes late integration models that claim that each sensory stream is processed individually and then passed on to be combined at a later stage (Schwartz, Robert-Ribes, & Escudier, 1998).

Summary

That visual information can influence auditory processing—behaviorally and neurophysiologically—has now been widely shown. It is clear that in order to have a better understanding of speech perception, visual information should be considered a viable information source. Understanding what particular properties of the visual

speech signal provide facilitation for, or generally influence, auditory speech processing is an open line of research. This thesis aims to investigate some of these properties, including the effects of visual predictability on auditory evoked responses shown by van Wassenhove et al. (2005) and Arnal et al. (2009). I also test the entrainment of neural responses to comodulated audio-visual signals as a candidate mechanism underlying BCMP (Grant & Seitz, 2000).

Chapter 3: Flexibility in the audio-visual speech advantage

Introduction

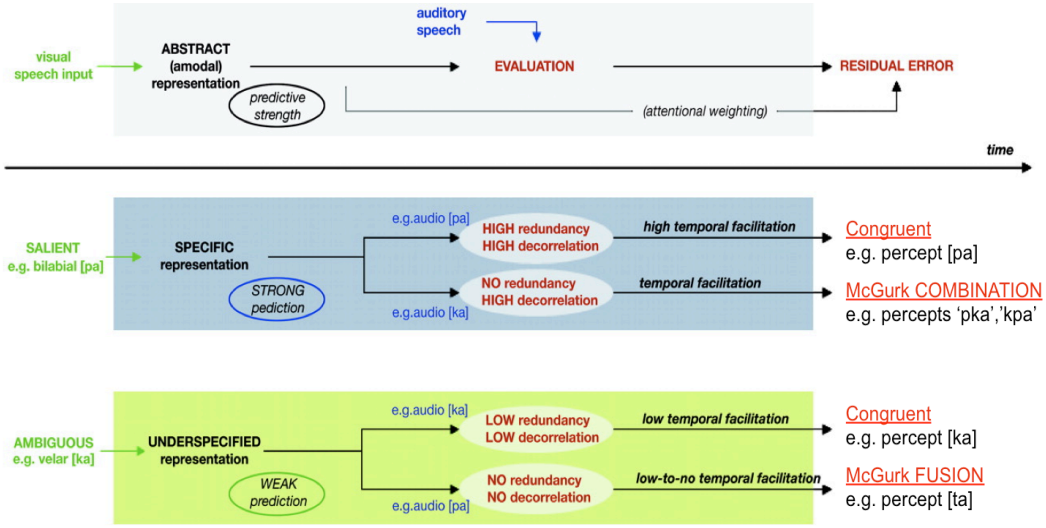
The goal of this chapter is to explore the time course and processing involved in the mapping of multisensory signals onto phonetic and phonological representations by providing behavioral and electrophysiological evidence that the degree of audio-visual facilitation (as measured by reduction in behavioral reaction times and evoked response latencies) in speech perception is modulated by the relative predictive strength of the visual signal rather than an articulator-specific property. I demonstrate that altering the response set available to a perceiver affects the amount of facilitation provided by the visual prearticulatory information. The bilabial consonants—characterized by prominent cues provided by the upper and lower lip making full closure—have previously shown the greatest facilitation because they have been the most distinct among the experimental response set. When increased uncertainty about the bilabial consonants is introduced (by adding a second bilabial to the response set), the behavioral and electrophysiological facilitation effects are diminished, and the non-labial consonant becomes the most facilitated. This demonstrates not only the flexibility of the processes underlying evoked sensory responses, but also informs theories of speech perception that aim to incorporate visual cues into standard auditory feature analysis. Ultimately, I show that behavioral reaction times and

cortical auditory responses are sensitive to general predictive properties rather than specific articulatory features, and can be manipulated by experimental context.

The electrophysiological correlates of the audio-visual advantage (reflected in reduced amplitude and/or latency) were introduced in Chapter 2, and are reviewed briefly here.

Giard and Peronnet (1999) showed that categorization times to audio-visual nonspeech stimuli (auditory pure tones that differed in frequency paired with visual shapes that differed in orientation) were faster than response times to unimodally presented stimuli, which suggested facilitatory processes in object recognition, and also showed increased amplitude to sensory ERP responses to multisensory stimuli. Besle et al. (2004) found *decreased* amplitude of auditory ERP responses to audiovisual speech relative to audio-alone speech (as well as decreased reaction times), which generally supported Giard and Peronnet's (1999) model of AV facilitation effects and extended this facilitation to speech stimuli.

Van Wassenhove, Grant and Poeppel (2005) found amplitude and latency effects on the auditory evoked EEG responses N1 and P2 to audiovisually presented speech syllables relative to audio-alone syllables, with the greatest facilitation seen for the bilabial consonant /pa/, which also had the highest visual alone accuracy in an identification task, leading to the model presented in Figure 3.1.



van Wassenhove V et al. PNAS 2005;102:1181-1186

Figure 3.1 Model for audio-visual speech facilitation proposed by van Wassenhove et al (2005)

This model suggests that the visual information provided by the face of a talker during prearticulatory movements can modulate the responses to auditorily presented speech stimuli. They hypothesized that the syllable /pa/ had the greatest predictive strength because it had the highest visual alone accuracy and therefore provided the greatest facilitation; A non-salient consonant such as /ka/ (which had lower visual alone accuracy) would not provide a strong prediction about the upcoming auditory stimulus, and would not result in a large amount of facilitation. This facilitation has been replicated in both speech and nonspeech domains (Pilling, 2009; Stekelenburg & Vroomen, 2007), showing that visual anticipatory information that is predictive of the timing and/or content of an upcoming auditory stimulus is reflected in the facilitation effects. Arnal et al. (2009) tested these effects in MEG using a different stimulus set,

and also found that the consonant type with the highest visual alone accuracy was the most facilitated[14].

The results discussed so far have shown that audiovisual facilitation is not an all-or-nothing phenomenon; instead, these effects can be modulated by synchrony, by the presence of anticipatory movement, and by the predictive strength of that anticipatory movement. One question that has arisen from these results is whether the "articulator specific" latency facilitation effects (van Wassenhove, et al., 2005) are a function of the visual phonetic structure of particular speech segments, or if this pattern is a consequence of having a highly predictable in the response set of that experiment.

Many studies have shown that the auditory and visual speech channels provide seemingly complementary information: place of articulation is often the most salient linguistic feature in the visual speech signal (Fisher, 1968; Owens & Blazek, 1985; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998; Summerfield, MacLeod, McGrath, & Brooke, 1989). Conversely, place of articulation in the auditory modality is the least resistant to noise degradation and the most confusable (e.g., Miller & Nicely, 1955). The articulator-specific facilitation account would predict that certain speech sounds have the greatest facilitation because they have prominent place of articulation at the front of the mouth/surface of the face (compared to alveolar or velar sounds which are produced further back in the mouth and are less easily

[14] In the Arnal et al. (2009) experiment, the consonant /ʒ/ had the highest visual alone accuracy as well as the greatest facilitation. However, /ʒ/ can also be visually salient, and may be quite labial depending on the talker (and the language—here French).

identified visually). In face-to-face conversation, a talker's lips are highly visible; this anticipatory place of articulation information could be responsible for facilitating auditory feature analysis. Furthermore, the original McGurk and MacDonald (1976) finding that no fusion occurs when a visual bilabial is presented, and the finding that infants are sensitive to labiality suggests that the visible cues provided by the lips are highly accessible (and potentially difficult to override—see Chapter 5).

On the other hand, if the extra facilitation seen for bilabials was truly driven by the predictive strength regarding an upcoming auditory event, as suggested by van Wassenhove et al. (2005), an advantage previously shown for one physically salient consonant could be shifted to a different (less physically salient) consonant in a response set where the anticipatory movements for the bilabial consonants are no longer predictive of a single potential auditory target.

The goal of this study is to explore the nature of the audio-visual latency facilitation effect by evaluating the responses to syllables in two experimental conditions. In one condition, the bilabial initial syllable /ba/ is most visually distinct in the response set of /ba da ga/. This is similar to the response set of van Wassenhove et al. (2005), which showed increased facilitation effects for bilabial /pa/ relative to non-labials /ta ka/. In addition, an additional experimental condition is explored in which more than one bilabial is present (response set /ba pa da/), where /ba/ and /pa/ share visual features, and /da/ becomes the most visually distinct in the response set. In this way, the claim from van Wassenhove et al. (2005) that facilitation effects should vary

based on certainty about the upcoming auditory stimulus that is provided by visual anticipatory articulator movement is directly tested.

If the facilitation effects previously seen for bilabial consonants are truly a consequence of predictive strength (rather than a consequence of optical phonetic salience of the bilabial place of articulation), /ba/ should show greatest facilitation effects when presented in an experimental context that contains only one bilabial and two non-labial response alternatives—in the "bilabial predictive" (BP) response set /ba da ga/—because visual anticipatory movement for the bilabial consonant is unique to the potential auditory token /ba/). Conversely, when more than one bilabial consonant is present in the response set—in an "alveolar predictive" (AP) response set /ba pa da/ —the same visual anticipatory movements could indicate potential auditory token /ba/ *or* /pa/ and facilitatory effects for these consonants should be reduced or eliminated. If, instead, the facilitatory effects seen in van Wassenhove et al. (2005) and Arnal et al. (2009) are a result of inherent distinctness and salience of particular consonants that provide greater predictive strength indicating an upcoming auditory bilabial consonant, audio-visual facilitatory effects for bilabials should be greatest regardless of response set context.

Importantly, these responses are elicited to the same physical stimuli within separate blocks of one experimental session, and in Experiment 2 recorded from the same MEG sensors on the same participants. The crucial difference between experimental contexts is the identity of the token that has the least uncertainty as a result of visual

anticipatory articulation. Modulating the visual distinctness of the auditory stimulus may also have behavioral consequences (reflected in decreased reaction times) in addition to evoked electrophysiological response facilitation. Visual anticipatory articulation should decrease reaction times to visually distinct consonants, despite the fact that the identity of the visually distinct consonant may vary by response set context.

Experiment 1: Behavioral responses to A, V, and AV speech in two response set contexts

Materials and Methods

Stimulus recording[15]: Video and audio materials were recorded concurrently with a Canon DM- XL1 video camera onto digital videotape (mini DV; 29.97 fps). An adult female native speaker of American English was recorded while seated in front of a solid dark blue background. No special effort was made to highlight the mouth of the talker (i.e., no spotlights or any other special lighting or camera angles were used to emphasize the oral cavity). The talker was instructed to start and end from a neutral "resting" mouth position and to avoid blinking during syllable articulation. The material list included 24 randomized CV nonsense syllables (C: /b d g m p t k n/ V: /a i u/) and was repeated five times by the talker. The syllables /ba da ga pa/ were selected from the larger set of CV tokens for the experiments presented in this chapter.

[15] Materials can be obtained at http://files.ling.umd.edu/~arhone/Thesis/Ch3_stimuli/ or by email request: ariane.rhone@gmail.com

Stimulus Selection: To avoid low-level cues that might provide nonlinguistic predictive information for a particular stimulus item, three unique tokens of each syllable type were chosen for inclusion. The selection criteria for choosing the three tokens were as follows: the lips were in a neutral position at least 5 frames before articulation began, lips returned to a resting position after the syllable was produced, and no blinks or other eye movements occurred during the production or for 5 frames before or after articulation. If more than three tokens of each type remained after exclusion criteria were met, the three tokens most similar in duration were chosen (see Appendix I for horizontal and vertical lip aperture by frame for each token).

Video: Digital videotapes were transferred to an Apple MacBook Pro running Windows XP for video segmentation and editing. All video editing was performed in VirtualDub (www.virtualdub.org). Individual files were segmented in the following manner: the in-point of the file was chosen by visually inspecting each token for the first discernable lip movement and then placing a marker 5 frames before that point. The out-point was selected by finding the frame at which the speaker's face returned to a neutral position and then placing a marker 5 frames after that point. The video files were then padded with 5 copies of the in-point frame, and a fade-in filter was applied to the first five frames to minimize visual onset responses. After the out-point frame, five still copies of the out-point were added and faded out, with each stimulus ending in a black frame to lead into the solid black interstimulus portion. The average stimulus duration was 52.7 frames (1758 ms). To reduce the visual complexity of the

stimulus, all files were converted to gray scale and cropped to include the speaker's face, neck, and top of shoulders (see Figure 3.2).

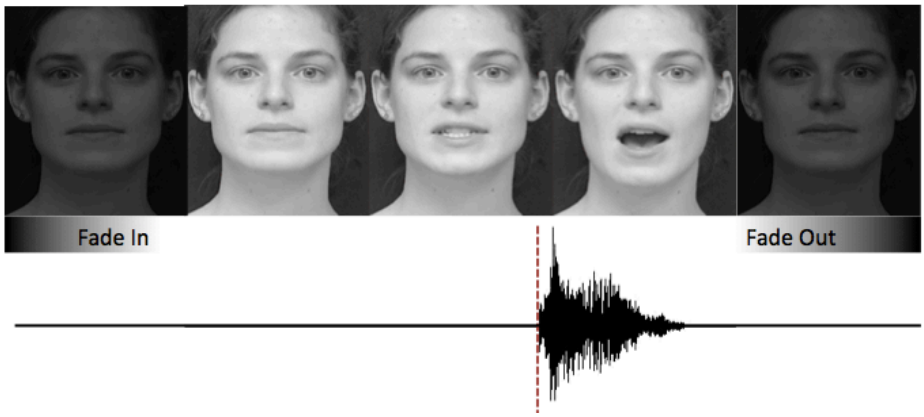


Figure 3.2 Stimulus schematic for the syllable /da/
Selected frames from one visual stimulus and waveform of one auditory stimulus used in this experiment. Dashed red line indicates location of auditory burst used for MEG triggering in Experiment 2. Alignment of audio and visual signals is not to scale (many visual frames not shown)

Audio: To ensure that audio and video durations matched, .WAV files were extracted from the "padded" video described above. Audio files were then resampled at 44.1 kHz in Praat (www.praat.org) and normalized to an average intensity of 70 dB SPL. A 10ms $\cos^2$ ramp was applied to the onset and offset of each audio file to reduce acoustic discontinuities at the edges.

Audio-visual compilation: Three versions of each token were created: visual alone (V) auditory-visual (AV) and auditory alone (A). The AV condition consisted of the original audio and visual signals for that token (processed as described above). For the A condition, a gray rectangle matched in average luminosity to a randomly selected frame from the visual signal was presented with the same visual fade in/fade out parameters and matched in duration to the auditory stimulus. For the V condition,

the audio track was removed from the video file. Audio and video signals were compiled into Audio-Visual Interleave (AVI) files in VirtualDub to avoid timing errors that might occur by compiling at experiment run-time.

Stimulus Presentation and Delivery[16]: Experimental stimuli were presented using a Dell OptiPlex computer with a SoundMAX Integrated HD sound card (Analog Devices, Norwood, MA) via Presentation stimulus presentation software (Neurobehavioral Systems, Inc., Albany, CA). Auditory stimuli were delivered to the subjects binaurally via Eartone ER3A transducers and non-magnetic air-tube delivery (Etymotic, Oak Brook, IL). Videos were presented via InFocus LP850 projector on a screen located approximately 30 cm from participants' nasion. Participants were instructed to fixate on the center of the screen (where all visual stimuli appeared) and to avoid blinking during stimulus presentation.

Task: Participants were asked to identify the syllable that they perceived[17] by pressing one of three buttons (labeled "ba" "da" and "ga" in the BP condition; "ba" "da" and "pa" in the AP condition) as quickly and as accurately as possible. Order of blocks (AP or BP) was counterbalanced across subjects. Each response set condition was divided into five blocks lasting approximately six minutes with each block containing

[16] This experiment was designed and run as a simultaneous behavioral and MEG study; however, low numbers of repetitions per stimulus type resulted in evoked field responses that were difficult to measure. Only behavioral results are considered in this portion of the chapter. A redesigned experiment (Experiment 2) contains both behavioral and neurophysiological data for a larger number of participants.

[17] To avoid emphasis on any one modality, participants were instructed to report what they *perceived* rather than what they *heard* or *saw*.

five repetitions of each token. The interstimulus interval varied pseudorandomly between 750ms-1250ms from the offset of the visual stimulus to the onset of the next visual stimulus. Stimuli were randomized within the blocks, with A, V, and AV stimuli intermixed.

Participants were familiarized with the response buttons and the task during a practice session that lasted approximately 3 minutes. Button configurations were presented to the left and right of the centered visual stimulus for the duration of the practice session. Buttons labeled "ba" and "da" were consistent for both block types (index fingers of the left and right hand, respectively). No feedback was provided during the practice, but participants confirmed that they were comfortable with the task before proceeding to the experimental conditions, at which point the button labels were removed. The testing session lasted approximately 120 minutes.

To motivate participants to give full attention to the identification task, overall percentage correct was reported to them at the end of each test block aggregated over all trials (i.e., no immediate or specific feedback was offered to the participant about performance on particular tokens or types). Participants were given breaks of at least 30 seconds for eye rest between blocks, and one longer break and an additional practice session to familiarize them with new button identities between the BP and AP conditions.

Participants: Four adult native speakers of English participated in this study (4 male; average age: 20.5 years). All had normal hearing and normal or corrected-to-normal vision (20/30 or better acuity verified with a standard Snellen chart), and none reported formal training or experience with lipreading. Presentation of stimuli and biomagnetic recording was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Prior to the start of the experiment, written informed consent was obtained from each participant.

Results

Accuracy analysis

Over all trials, accuracy was 89.2%. Response timeouts (defined as button presses after the trial was complete) comprised <1% of trials and were excluded from further analysis. Participants were highly accurate in the audio (A) and audiovisual (AV) conditions, and showed decreased accuracy in the visual (V) condition (see Figure 3.3).

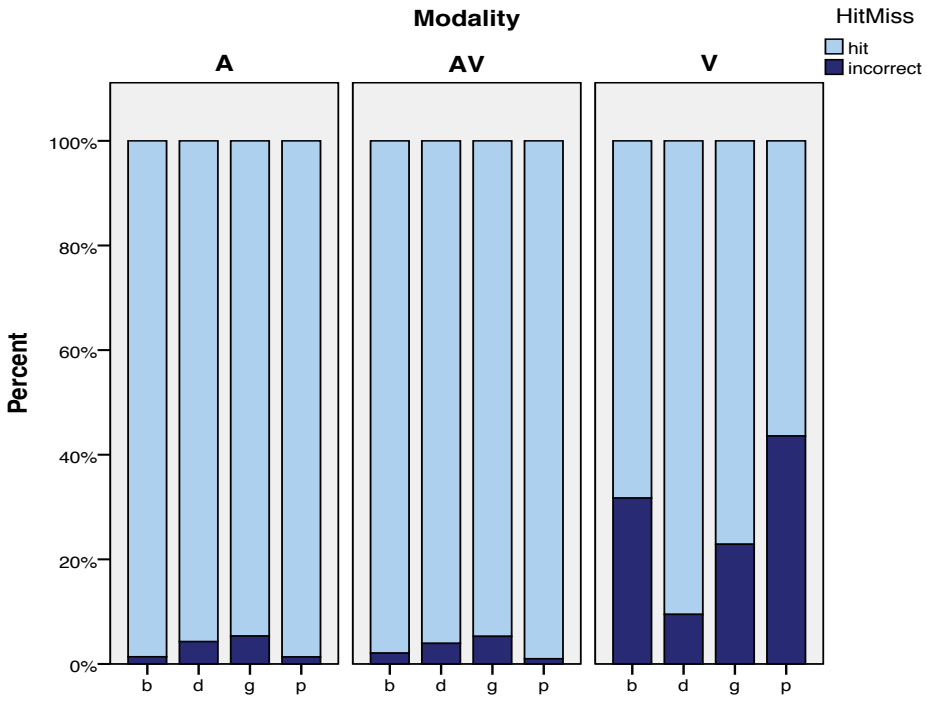


Figure 3.3: Overall accuracy for each modality by syllable type.
Accuracy was reduced in the visual-alone (V) condition relative to the auditory (A) and audio-visual (AV) conditions.

When broken down by response set (AP block vs. BP block), a pattern of results emerges for the target syllables /ba/ and /da/ for the visual-alone condition (Figure 3.4). Although visual-alone accuracy for /ba/ in the BP block (unique in response set) was high (96%), in the AP block (which contained two bilabial consonants), the /ba/ stimulus was correctly identified on only 41% of trials. Conversely, the syllable /da/ shows the opposite pattern (although less dramatically), with increased accuracy in the AP (97%) relative to BP (82%) block.

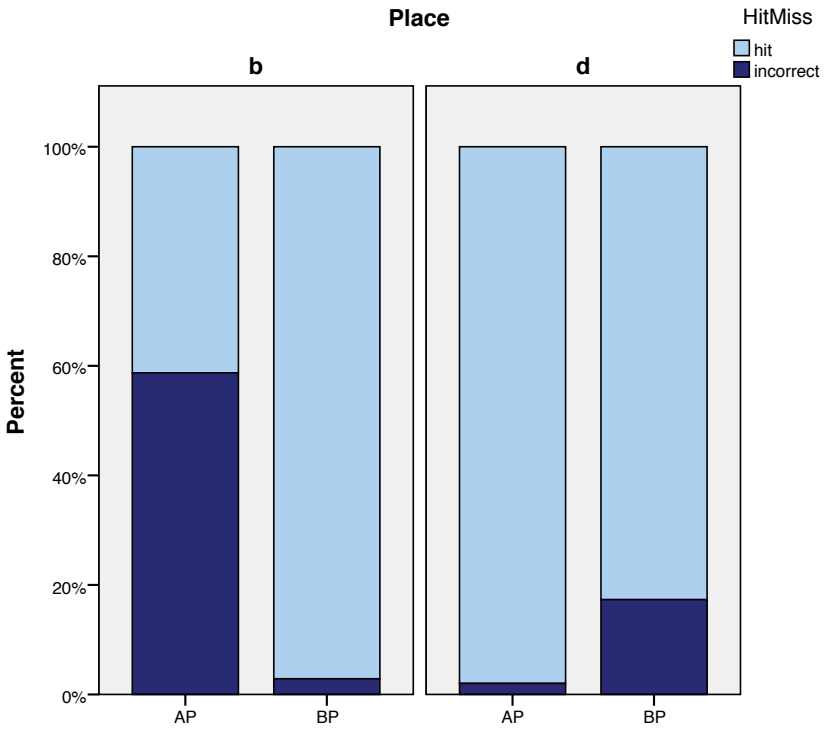


Figure 3.4: Across-subjects accuracy for visual alone stimuli, by response set.
The bilabial stimulus /ba/ has reduced accuracy for the alveolar predictive (AP) condition, while the alveolar /da/ has reduced accuracy in the bilabial predictive (BP) condition.

The pattern of responses to the visual alone stimuli (Table 3.1) shows that the bilabial syllable types in the AP condition (/ba/ and /pa/) were indeed perceptually confusable, while other consonant confusions were less common.

response

stimulus	AP: A	/ba/	/da/	/pa/
	/ba/	296	3	0
	/da/	2	297	1
	/pa/	3	1	295

response

stimulus	BP: A	/ba/	/da/	/ga/
	/ba/	276	5	0
	/da/	3	263	19
	/ga/	2	13	266

response

stimulus	AP: AV	/ba/	/da/	/pa/
	/ba/	290	1	5
	/da/	3	295	2
	/pa/	2	1	297

response

stimulus	BP: AV	/ba/	/da/	/ga/
	/ba/	277	4	2
	/da/	3	266	15
	/ga/	0	15	269

response

stimulus	AP: V	/ba/	/da/	/pa/
	/ba/	123	3	172
	/da/	3	291	3
	/pa/	128	1	167

response

stimulus	BP: V	/ba/	/da/	/ga/
	/ba/	273	7	1
	/da/	3	234	46
	/ga/	1	64	219

Table 3.1: Observed responses by block and modality for A, AV, and V stimuli
In the AP block, the bilabial consonants /ba/ and /pa/ were commonly confused when presented in the visual-alone modality. In the BP block, the non-labial syllable types /da/ and /ga/ had increased confusions in the visual modality relative to the A and AV modalities, but not to the extent seen for bilabials.

Reaction time analysis

Reaction time analysis was limited to correct responses. Statistical comparisons were performed over logarithmic reaction time (logRT). Across participants, one-way repeated measures showed a main effect of modality for logRT [(2,4689), F=157.135, $p < 0.001$]. Reaction times to auditory alone stimuli were significantly longer than to stimuli presented auditorily or audio-visually (see Figure 3.5 and 3.6). Scheffé post hoc tests revealed significant differences ($p < 0.05$) for A versus AV and A versus V contrasts. Because the distributions violated assumptions of equal variance, Mann-Whitney U and 2-sample KS tests confirmed effects of the Scheffé comparisons at the $p < 0.05$ level.

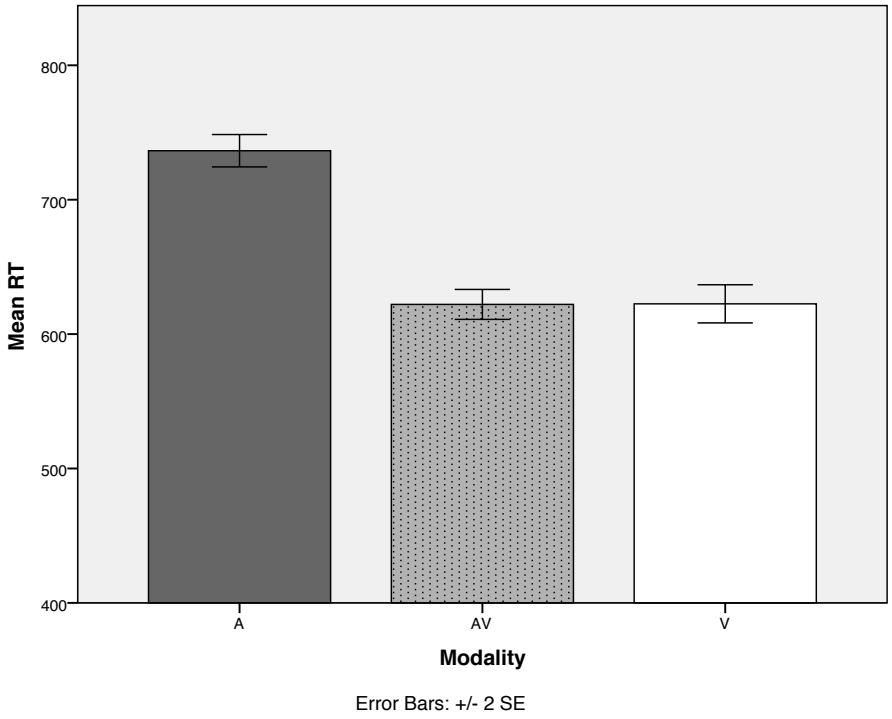


Figure 3.5: Reaction time (in ms) by modality across participants.
Mean reaction time to auditory alone (A) stimuli was significantly longer than RT to audio-visual (AV) or visual alone (V) stimuli.

Reaction time for each syllable type by modality

Collapsed across block, average RT (for correct responses only) to the syllable type /ba/ was shortest to Visual alone stimuli (524.6 ms), slightly higher for AV (599.8ms), and highest for auditory alone stimuli (740.5 ms). The same pattern held for /pa/, with means of 638.7 ms, 644.7 ms, and 738.6 ms for V, AV, and A conditions, respectively. The syllable /da/ showed shortest RT for AV condition (656.1 ms), followed by V (699.4 ms) and A alone (759.6 ms). Syllable /ga/ showed the same pattern as /da/, with mean RTs of 601.9 ms, 573.2 ms, and 676.3 ms for V, AV, and A conditions, respectively (see Figure 3.6). The visual anticipatory articulation that is present in visual and audio-visual conditions offers information

about the identity of a stimulus before the auditory signal begins. This facilitation is reflected in the decreased RT to V and AV stimuli relative to A alone.

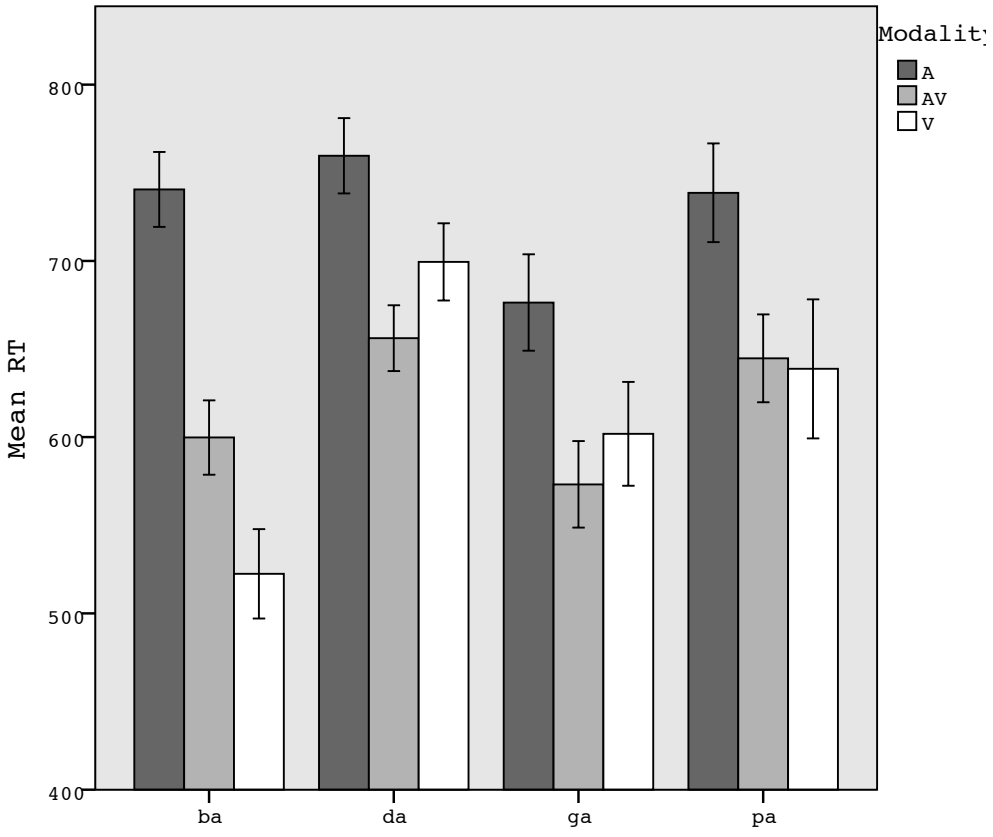


Figure 3.6: Reaction time (in ms) by syllable type and modality.
Error bars indicate +/-2 SEM. All syllable types show significant increase in RT to A relative to V and AV stimuli.

Response set effects

Bilabial consonant /ba/: One-way ANOVA showed significant effects of condition (AP vs. BP) for AV stimuli [(1,565); F=41.423; $p < 0.001$] and V stimuli [(1,392); F=46.767; $p < 0.001$]. Mean RT for AV-AP /ba/ was 653.9, AV-BP /ba/ was 543.1 ms. Mean RT for V-AP /ba/ was 642 ms, V-BP /ba/ was 471.3. No effect of condition

was shown for the responses to A-alone stimuli [(1,570); F=1.432] (A-AP /ba/: 753.3 ms, A-BP /ba/ 726.3 ms). For all significant differences, reaction time for the syllable /ba/ in the BP condition was shorter than in the AP condition (see Figure 3.7).

Alveolar consonant /da/: One-way ANOVA showed significant differences for AP vs. BP for all three modalities (A: [(1,558); F=7.607; $p < 0.05$]; AV: [(1,559); F = 11.817; $p < 0.001$]; V: [(1,523); F=40.426; $p < 0.001$]), with shorter latencies for AP than BP condition (A-AP /da/ 735.5 ms, A-BP /da/ 787 ms; AV-AP /da/ 627.9 ms, AV-BP /da/ 687.5 ms; V-AP /da/ 646.8 ms, V-BP /da/ 764.8 ms).

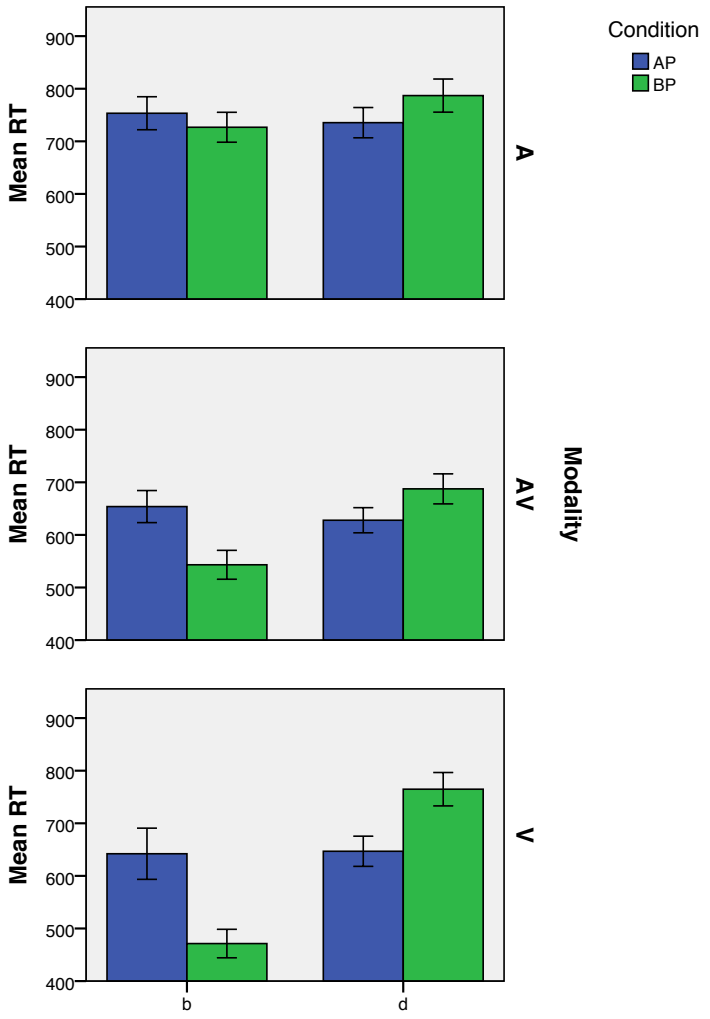


Figure 3.7: Mean reaction time (in ms) for syllables of interest (/ba/ and /da/) by block and modality.
Error bars indicate +/-2 SEM. Significant block effects were shown for AV and V modalities for both syllable types; significant block effect was also shown for /da/ in the A modality.

Interim Discussion

Although the number of participants for this experiment was small, predicted effects were shown for reaction times to speech syllables in two response contexts, for auditory-alone, visual-alone, and audio-visual stimuli. Significant effects of modality and block were found, and confirmed that the bilabials in the AP block were less

visually salient than the bilabial in the BP block (as shown by reaction time and accuracy comparisons).

However, the neurophysiological facilitation effects of response set could not be examined with this experimental design. Practical time limitations for durations of MEG recording that are acceptable for participants limited the number of trials that could be presented per condition to only 75 per type per modality, and the visually demanding audio-visual stimuli introduced an additional confound: a large number of eye blinks during trials required a large number of epoch rejections, which decreased the signal-to-noise ratio of the recorded magnetic field response even further.

These issues prevented the analysis of evoked responses to these materials, but provided visual-alone accuracy and showed the predicted block and modality effects. In order to test the predictability/salience hypothesis and its effects on the auditory evoked responses, the Visual modality condition was removed for Experiment 2 and an increased number of trials per condition were used to provide more robust onset responses in addition to reaction time information.

Experiment 2: Behavioral & electrophysiological responses to A and AV speech in two response set contexts

Materials and Methods

Stimulus materials were the same as those in Experiment 1. However, this experiment did not include any visual-alone (V) tokens (only A and AV conditions were presented). The within-subjects design of the experiments presented in this chapter required that all conditions be presented to all participants, and with three modalities, two blocks, and three syllable types per block that had to be averaged independently, the number of usable trials per type that could be averaged for MEG analysis after artifact rejection was too low to obtain reliable evoked fields. Removing the visual-alone condition allowed more repetitions per type for the audio-alone and audio-visual modalities without extending the overall duration of the experiment, which resulted in more robust auditory evoked fields.

Participants

Twelve native speakers of American English were recruited from the University of Maryland, College Park community and received course credit or monetary compensation ($10 per hour) for their participation. Presentation of stimuli and biomagnetic recording was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Prior to the start of the experiment, written informed consent was obtained from each participant. All participants were right-handed (Oldfield, 1971) and had no self-reported history of speech or hearing deficits and reported normal vision at test time.

No participant reported formal lipreading experience or training. Two participants were excluded from analysis (one participant failed to complete the experiment, one participant had excessive movement during MEG recording). Data from the 10 remaining participants (8 female; average age 21.5 years) are included in subsequent results and discussion.

Stimulus Presentation and Delivery: Experimental stimuli were presented using a Dell OptiPlex computer with a SoundMAX Integrated HD sound card (Analog Devices, Norwood, MA) via Presentation stimulus presentation software (Neurobehavioral Systems, Inc., Albany, CA). Auditory stimuli were delivered to the subjects binaurally via Eartone ER3A transducers and non-magnetic air-tube delivery (Etymotic, Oak Brook, IL). Videos were presented via InFocus LP850 projector on a screen located approximately 30 cm from the participants' nasion. Participants were instructed to fixate on the center of the screen (where all visual stimuli appeared) and to avoid blinking during stimulus presentation.

Task: As in Experiment 1, participants were asked to identify the syllable that they perceived by pressing one of three buttons as quickly and as accurately as possible. The order of conditions was counterbalanced across participants. Each condition was divided into five blocks lasting approximately six minutes with each block containing 7 repetitions of each token (21 of each of each syllable type per modality). Interstimulus interval (ISI) varied pseudorandomly between 750ms-1250ms from the

offset of visual stimulus to the onset of the next visual stimulus. Stimuli were randomized within the blocks, with A-alone and AV stimuli intermixed.

Participants were familiarized with the response buttons and the task during a practice session that lasted approximately 3 minutes with button identities provided on the screen. No feedback was provided during the practice, but participants confirmed that they were comfortable with the task before proceeding to the experimental conditions, where button identities were removed. The entire testing session lasted approximately 120 minutes.

To motivate participants to give full attention to the identification task, overall percentage correct was reported to them at the end of each test block aggregated over all trials in the block (i.e., no immediate or specific feedback was offered to the participant about performance on particular tokens or types). Participants were given breaks of at least 30 seconds for eye rest between blocks, and one longer break with an additional practice session to familiarize them with new button identities between the BP and AP conditions.

MEG recording: Data were acquired using a 160-channel whole-head biomagnetometer with axial gradiometer sensors (KIT System, Kanazawa, Japan). Recording bandwidth was 1-200 Hz, with a 60 Hz Notch filter, at 1000 Hz sampling rate. Data were noise reduced using time-shifted PCA (de Cheveigné & Simon, 2007) averaged offline (triggered to auditory onset – see Figure 3.1 for schematic) with

epochs from 100ms pre-stimulus onset to 400ms post stimulus onset, artifact rejected at ± 2.5 pT, low pass filtered at 20 Hz and baseline corrected over the 100 ms pre-stimulus interval.

Analysis

Sensor selection: An auditory pure tone localizer was administered for each participant before participation in the experiment. The 10 strongest channels per hemisphere (5 from the magnetic field source and 5 from the magnetic field sink for each hemisphere) were selected from M100 elicited by 1kHz pure tone pretest for each subject. M100s were found consistently in the left hemisphere (LH) for all participants.

Component selection: Because previous results have shown that the visual signal modulates responses generated in auditory cortex (Besle, et al., 2008; Sams, et al., 1991; van Wassenhove, et al., 2005), the canonical cortical auditory evoked fields (M50, M100, M150) were measured for each subject. However, because of variability across the participants' neuromagnetic response profiles (i.e., many participants did not show all three auditory responses to the pretest and/or experimental stimuli) the M100 was selected as the most reliable evoked response across participants and is included in subsequent results and discussion.

Latency analysis: The root-mean-square (RMS) of the magnetic field deflections from the 10 selected channels was obtained using MEG160 (KIT, Kanazawa). The latency corresponding to the peak amplitude of the RMS wave was then obtained for each condition. Visual inspection of the contour plot displaying field strength at each sensor at the time point of peak RMS amplitude confirmed typical auditory M100 response distribution in all participants.

Results

Identification Accuracy: Participants correctly identified stimuli on 95.6 percent of all trials[18]. Table 3.2 provides details by syllable type, modality, and response set (AP or BP blocks), and observed syllable confusions. Response timeouts comprised <1% of all trials and were excluded from analysis.

[18] The overall percent correct in Experiment 2 is higher than Experiment 1 because of the lack of visual-alone stimuli.

response

stimulus	AP: A	/ba/	/da/	/pa/
	/ba/	985	19	9
	/da/	14	1004	3
	/pa/	4	9	1000

response

stimulus	BP: A	/ba/	/da/	/ga/
	/ba/	1017	14	2
	/da/	9	975	58
	/ga/	3	35	999

response

stimulus	AP: AV	/ba/	/da/	/pa/
	/ba/	989	13	8
	/da/	4	1018	4
	/pa/	11	3	1006

response

stimulus	BP: AV	/ba/	/da/	/ga/
	/ba/	1026	9	0
	/da/	9	987	45
	/ga/	3	18	1014

Table 3.2: Observed response patterns by block and modality for A and AV stimuli for Experiment 2

Reaction Time Analysis: Collapsing all blocks and syllable types, reaction time audiovisually presented syllables (613.6 ms) was significantly shorter[19] than audio-alone syllables (674.7 ms) as assessed with one-way ANOVA [(1,12002); F=284.029; $p < 0.001$]. A significant effect of modality was found in for comparisons within each syllable type (i.e., ba-AV was significantly faster than ba-A, da-AV faster than da-A, etc.) at the $p < 0.05$ level (see Figure 3.8). As in Experiment 1, The visual anticipatory information provided in the frames prior to auditory stimulus onset provided enough information for participants to correctly classify the stimuli faster than when visual anticipatory information was not available for the same tokens.

[19] As in Experiment 1, statistical comparisons were performed on logRT

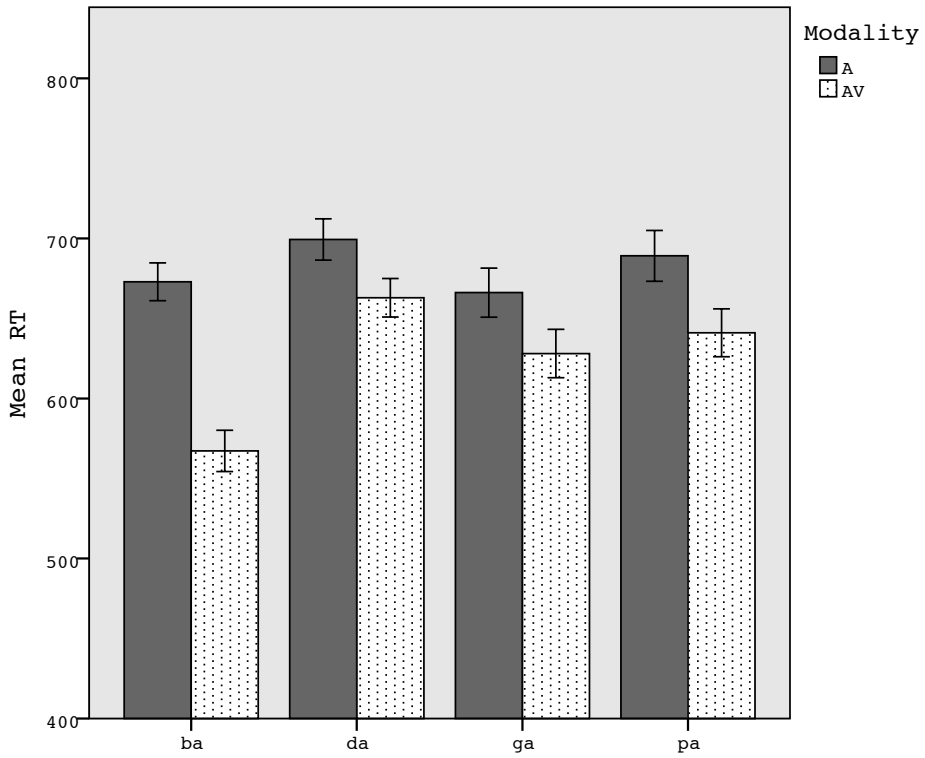


Figure 3.8: Mean RT by modality for all syllable types across both blocks. Error bars : +/-2 SEM

Comparisons of the logRT to the same stimulus across conditions showed significant effects of response set for audiovisual /ba/ [(1,1998); F=169.311; $p < 0.001$)] and audiovisual /da/ [(1,2003); F=4.867, $p < 0.05$]. As predicted (from results from Experiment 1), RT to audiovisual /ba/ in the BP condition was significantly faster (mean: 501 ms) than audiovisual /ba/ in the AP condition (mean: 622 ms). Syllable type /da/ showed the opposite pattern, with shorter RT in the AP (646 ms) than the BP condition (666 ms). For auditory-alone conditions, /ba/ was significantly faster in the BP than AP condition [(1,2000); F=6.723; $p < 0.05$], and /da/ showed a trend [(1,1977); F=3.156; p=0.076] toward shorter RT for the AP condition than the BP condition. Figure 3.9 shows the reaction time effect by block for each modality.

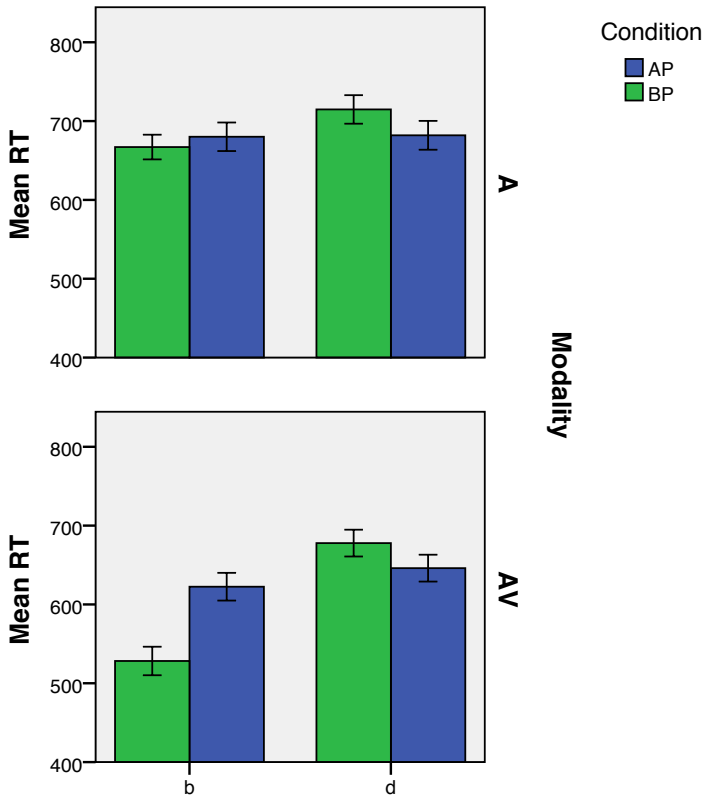


Figure 3.9: Mean RT as a function of response set for the two syllable types occurring in both conditions. Error bars : +/-2 SEM

Although these stimuli were physically identical in each block, the change in degree of relative visual predictability provided by the different response sets modulated participants' reaction time in identifying the syllable type, as was also shown in Experiment 1. In particular, the /ba/ syllable shows a marked increase when an additional bilabial is present in the response set (the AP block), despite the fact that the optical phonetic structure of the syllable is held constant. The /ba/ syllable in the AP block is no longer uniquely predicted by anticipatory bilabial motion when the syllable /pa/ is also present in the response set, and this uncertainty is reflected in the increased reaction time. The amount of facilitation for a particular syllable type

depends on the relative certainty about the upcoming auditory stimulus that is provided by visual anticipatory articulation.

MEG Results:

Because the latency of the auditory M100 is know to be sensitive to acoustic properties of the stimulus (Obleser, Lahiri, & Eulitz, 2003; Roberts & Poeppel, 1996; Sharma & Dorman, 2000), latency comparisons are limited to the same physical stimulus across the different response set blocks. Figure 3.10 shows an example auditory evoked response to the syllable /ba/ for each stimulus modality and by response set block (see *Methods* for details). Latency facilitation effects were evaluated by subtracting the M100 latency to AV stimuli from the M100 latency to A-Alone stimuli for each subject.

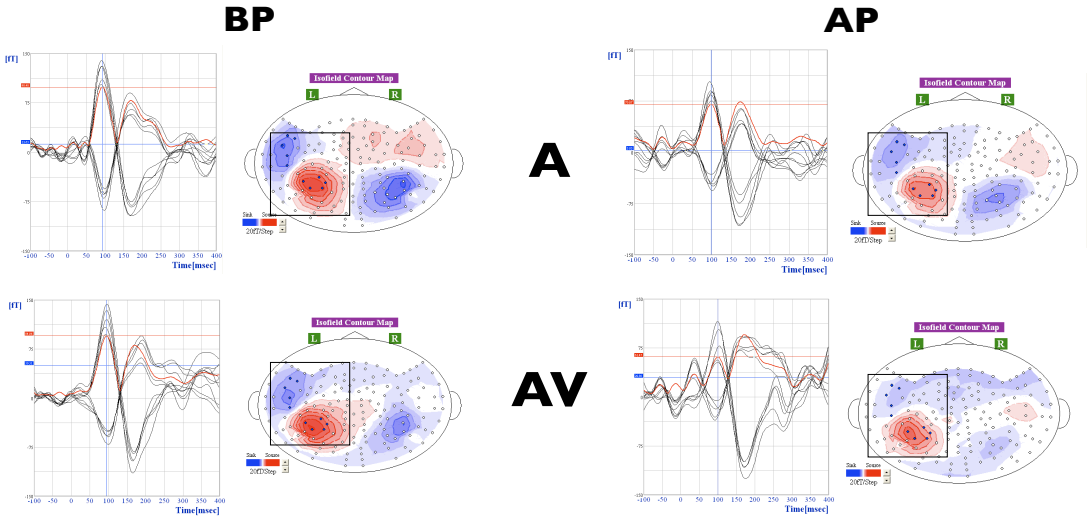


Figure 3.10 Example M100 waveforms and contour plots for the syllable /ba/
Auditory-alone (A) and audio-visual (AV) responses to the syllable /ba/ from a representative participant in each predictability condition. 10 LH channels (5 sink, 5 source) were selected for each participant based on 1kHz scout test.

In the BP response set, significant differences were seen for all three syllable types. Syllable types /ba/ and /da/ patterned as predicted, with significant facilitation effects for AV relative to A alone stimulus presentation (Wilcoxon Signed-Ranks test /ba/: $V = 45$, $p < 0.01$; /da/: $V = 47.5$, $p < 0.05$). Syllable type /ga/ showed *increased* latency for AV relative to A stimuli ($V = 4$, $p < 0.05$) see *Discussion* for further consideration of this difference). In the AP response set, a significant facilitation effect was found only for the alveolar syllable type /da/ ($V = 53$, $p < 0.01$). The bilabial consonants /ba/ and /pa/ did not show significant latency facilitation effects in this response set condition (see Figure 3.11 for facilitation by syllable type and response set). Direct comparisons of latency facilitation by block (AP vs. BP) showed no significant differences. No significant differences in amplitude for any comparison were observed.

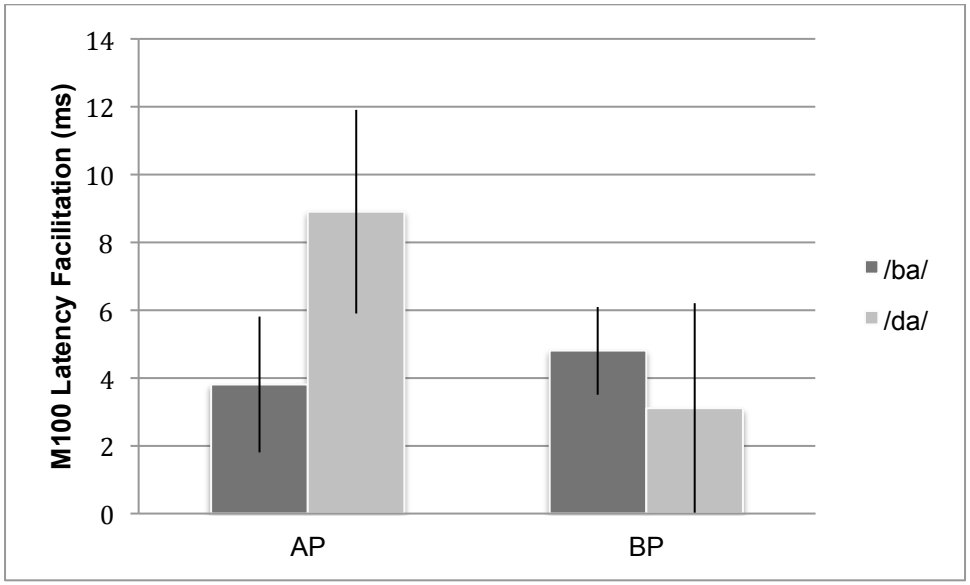


Figure 3.11 M100 facilitation (A – AV) by syllable type and block. Error bars : +/-2 SEM

Discussion:

In addition to demonstrating general audio-visual response facilitation (demonstrated by reduced reaction time to AV relative to A-alone stimuli), these results also demonstrate that the response set/experimental context affect auditory M100 latency facilitation (Figure 3.11). In the Bilabial Predictive condition, the bilabial /ba/ shows facilitation effects; however, this facilitation is eliminated in a response set containing another bilabial (AP condition). Reaction times to the bilabial consonant also increase when presented in the AP response set, while reaction times to the non-bilabial /da/ decreases (Figure 3.8).

General Discussion

This chapter showed that the advantage that has previously been discussed loosely in terms of predictive strength (van Wassenhove et al., 2005; Campbell, 2008), but which has thus far only been determined by correlation between visual-alone intelligibility, can be thought of as true facilitation based on the relative certainty provided by the anticipatory visual input. By dissociating the optical phonetic structure from relative uncertainty (e.g., by making the visually salient bilabial consonants the least informative), the facilitation provided by visual speech information can be more directly attributed to the strength of the prediction about an upcoming auditory token. This supports the model of van Wassenhove et al. (2005), where the salience of the visual stimulus modulates the strength of the prediction about an upcoming auditory signal, which results in facilitation at the neurophysiological as well as the behavioral level (see Figure 3.1). Having low

uncertainty about the identity of an upcoming sound based on the relative uniqueness of the visual articulation preceding the onset of the auditory stimulus results in greater facilitation. Having high uncertainty (e.g. when visual articulation does not uniquely correspond to one syllable type) does not allow a strong prediction to be made about an upcoming sound, and results in decreased facilitation.

By presenting the same physical stimuli across modalities (auditory-alone and audio-visually, as well as visual-alone for Experiment 1) and across two response set blocks, we test whether the previously shown articulator-specific facilitation effects for inherently salient consonants—such as those with some kind of labial feature—could be shifted to a syllable produced at a different place of articulation. In one response set, where /ba da ga/ are potential auditory stimuli that the participant may encounter, the visual anticipatory articulation of lips coming together to produce a bilabial consonant uniquely predicts the syllable /ba/. In the alternate response set, where /ba pa da/ are potential auditory stimuli, the anticipatory bilabial articulation no longer uniquely predicts one syllable type. Instead, bilabial anticipatory articulation increases the uncertainty about the upcoming auditory stimulus (because the anticipatory movements for /ba/ and /pa/ are visually indistinguishable to most participants (see Experiment 1; Owens & Blazek, 1985). In this alternate response scenario, the non-labial consonant (here the alveolar /da/, which is produced further back in the mouth and is generally classified in a larger set of visually perceptually confusable consonants than bilabials are) becomes highly predictable given the visual

anticipatory movements (where lip separation indicates that the upcoming auditory stimulus is *not* a bilabial consonant).

The results from this study suggest that the extra facilitation for particular consonant types reflects an advantage provided by relative certainty within a response set rather than a physical articulator specific advantage for the visually salient consonants. Behavioral data show that there is a significant difference in reaction time dependent on the current response set. When a the bilabial syllable /ba/ is uniquely predicted by anticipatory motion (the BP condition), bilabials show faster RT relative to responses to the same audio-visual stimulus when it is not uniquely predicted by anticipatory bilabial articulation. Conversely, when the response set contains two bilabials /pa/ and /ba/ (AP condition), the syllable /da/ is uniquely predicted by non-labial anticipatory articulation and shows significantly reduced RT relative to the BP condition. This is in addition to the overall RT reduction for audio-visually presented syllables relative to syllables presented only auditorily.

The timing of the auditory evoked magnetic field (M100) is also manipulated by response set. In the BP condition, both /ba/ and /da/ show facilitation for audiovisual relative to auditory alone stimuli. Although /ba/ was predicted to be the most visually salient, it is possible that for this particular talker, the syllable /da/ was equally salient and uniquely determined from anticipatory visual cues (see the visual alone confusion matrices in Table 3.1). Variability in consonant confusability for different talkers has been demonstrated in previous studies (Gagné, Masterson, Munhall, Bilida, &

Querengesser, 1994), and it is possible that for this talker, cues to the alveolar syllable /da/ were present in the pre-auditory anticipatory movement.

The observation that reaction times vary by experimental response set manipulations offers further evidence that listeners/seers are sensitive to the predictability of the visual consonant, even within the same experiment, and M100 latency facilitation effects in both blocks supports the suggestion that listeners (and their brains) make use of any and all visual cues available to them when making identification judgments about a particular token.

The syllable /ga/ fails to show audiovisual facilitation effects, perhaps because the visual stimuli were more variable in timing (see Appendix I) or potentially because /da/ may have been a default non-labial prediction (see Massaro 1998). If /da/ were the default prediction for all non-labial syllables, perhaps the anticipatory movement of the non-labial articulators may have hindered, rather than helped, the preliminary feature analysis that may be indexed by the M100 response. Further investigations into this discrepancy should address this issue to see if the lack of facilitation was simply due to the wider variation in the dynamic structure of the stimuli or if there is a principled cause for this effect.

Conclusion

The present experiment was designed to test whether "articulator specific" audiovisual facilitation effects that had been previously observed are specific to

particular phonetic features (such as place of articulation) which should be incorporated into linguistically motivated models of speech perception, or are alternatively a result of visual predictability in a given experimental response set paradigm. The findings of this study support the latter. When a non-labial (despite not being traditionally thought of as visually salient) syllable is highly predictable in a given response set, the auditory processing load is lightened. This facilitation in processing is reflected in decreased reaction times and M100 latency differences to audiovisual stimuli relative to auditory alone stimuli. The results reported here suggest that visual predictability does modulate the M100 auditory response, possibly because of reduced processing demands for incoming auditory stimuli that have been previously specified by non-auditory predictive information present in the visual speech signal.

Chapter 4: Neural entrainment to speech like audiovisual signals

Introduction

By this point, a large number of studies—including those presented in this thesis—have shown that visual information can affect speech perception. Behaviorally, the information provided by the face of a talker has been shown to improve detection, identification, and reaction times, and discordant visual information can also influence the perception of an auditory stimulus. Neurophysiologically, the addition of visual anticipatory motion preceding an auditory target has been shown to affect latency and amplitude of evoked responses recorded from auditory cortex and surrounding areas.

Although audio-visual facilitation effects are now well established, the mechanism underlying this audio-visual advantage is a topic of current investigation. *How* these effects are implemented in the brain is of considerable interest for establishing a more detailed account of multisensory interactions and integration, with the ultimate goal of making neurobiologically grounded theories of speech perception. This chapter presents a set of experiments that aim to test how one audiovisual facilitation effect—bimodal coherence masking protection—is potentially implemented in human brain.

One of the proposed explanations for audio-visual advantages in speech detection is the phenomenon known as Bimodal Coherence Masking Protection, or BCMP (Grant

& Seitz, 2000). BCMP has roots in classic auditory experiments that have shown that cross-frequency correlation in modulations results in significant benefits for detecting a target signal in noise.

Hall et al. (1984) showed that target tones are more easily detected when the envelope of a masker noise was correlated across several auditory filter banks (Hall, Haggard, & Fernandes, 1984; Hall & Grose, 1988). Although a wider band masker creates greater mechanical energy, the presence of AM noise correlated across *several* critical bands apparently helps to establish the masker as an auditory object, and groups the masker better. This, in turn, allows the signal to stand out relative to the masker. A decrease in threshold, then, appears to result from having auditory cues present across multiple frequencies, as long as they are correlated, and is known as Comodulation Masking Release (CMR).

In a similar spirit, Gordon (1997a, 1997b) found that increasing coherence within speech sounds also resulted in improved thresholds in a masked detection task. Rather than examining coherence across maskers, Gordon (1997b) investigated *protection* from masking for speech sounds, by presenting synthetic vowels in low-passed noise. The masker was limited to the first formant (F1) region, which contained the distinguishing acoustic information about the identity of the vowels that were tested (/ɪ/ and /ɛ/). Relative to stimuli that contained only the masked F1 region, identification thresholds improved when stimuli contained F2 and F3, despite the fact that these formants were held constant across vowel types. As with CMR, this

protection from masking was attributed to increased information available for grouping an auditory object. When onsets and offsets were misaligned temporally, the coherence masking protection effect was eliminated.

The fundamental finding of these studies is this: when signals or maskers are concurrently modulated, the listener has more information to group signals or noise crossing several channels, and this makes them easier to detect and identify. These studies were relevant for theories of perceptual grouping and auditory object identification (Darwin, 1984; Bregman, A.S., 1990) and have spawned numerous follow-ups to examine the limits of the phenomenon.

In the audio-visual domain, Grant and Seitz (2000) established BCMP by showing that the presence of matching visual input improved auditory detection for normal hearing listeners. They presented three spoken sentences in three conditions: auditory only, audio-visual matched, and audio-visual mismatched (with audio from one sentence dubbed to video from a different sentence), and found decreased improved detection thresholds only when audio and visual signals matched; mismatched audio-visual stimuli and audio-alone stimuli did not differ.

Although the crucial result of the set of studies presented in Grant & Seitz (2000) is that having visual speech information presented concurrently with auditory information improves detection of auditory stimuli, in a second experiment they also found an improvement when participants were orthographically presented with the

upcoming auditory stimuli orthographically. The threshold improvement was less than what was shown for natural (matched) audio-visual speech stimuli (0.5 dB and 1.6 dB, respectively), but the addition of information that cued participants to the content of the auditory stream was beneficial in the detection task. Put simply, knowing what you are about to hear helps you hear it better (at least for the sentences they used). Although orthographic-auditory relationships are learned associations and not a byproduct of the articulation of the utterance, this offers further support for a model of speech perception where decreased uncertainty results in improved performance.

Their third experiment linked performance on the AV detection task with the degree of correlation between the auditory envelope and the area of the lip opening, with the greatest correlations found between lip aperture and higher frequency speech envelope, in the bands that they consider to be in the F2-F3 range (800-2200 Hz for F2, 2200-6500 Hz for F3). However, correlations were also found for the overall (i.e., wideband) envelope as well as the lower frequency F1 band. Although it is tempting to credit F2 and F3 for the improvement in speech detection in noise in this paradigm, the connection between the auditory bands and the visual lip area is perhaps better interpreted as a general, overall benefit provided by visual information during auditory speech perception; although the correlation coefficient was higher for the F2 and F3 regions, there is no evidence to suggest that this is the *dominant* information used by the perceiver.

Furthermore, the lag constraint that they built into their analysis may have influenced this correlation. They noted that their correlations improved with a 1-3 frame (33 -100 ms) lag built in, but the physiology of the auditory and visual systems introduce their own lag, which is not accounted for here, and the optimal audio-video correlation (in the technological sense) may not be the same as the 'brain's eye view' correlations. Auditory and visual information are transmitted to the human cortex at different rates, and even though an offset of a few tens of milliseconds does not seem to disrupt audio-visual integration, the addition of this constraint may have introduced unanticipated byproducts for the correlational analysis. This is not to say that this correlation does not occur or does not play a role in the audiovisual speech advantage; however, care must be taken in interpreting correlations across "optimal" stimulus timing parameters.

Despite these shortcomings, Grant & Seitz (2000) do show that the addition of visual information is beneficial in speech detection, and that there is some relationship between the amount of BCMP and the correlation between auditory and visual information. In addition, knowing what to listen for appears to be an important cue underlying the audiovisual advantage, because providing the participant with an orthographic representation of an upcoming stimulus also resulted in improved detection thresholds.

Bernstein, Auer, and Takayanagi (2004) followed up on Grant & Seitz (2000) with a number of critical comparisons to test the specificity and the source of the detection

advantage. They compared audio-alone thresholds with the following multisensory stimuli: auditory speech + natural visual speech; auditory speech + dynamic Lissajous figure; auditory speech + dynamic rectangle; and auditory speech + static rectangle. They found a significant improvement for all multimodal stimuli, but found the most improvement for the natural speech tokens. The dynamic information provided by the Lissajous figure and dynamic rectangle (which changed in size and were correlated with the envelope of the auditory signal) did not result in significant improvement in detection thresholds compared with the static rectangle, which would have been expected if the advantage were driven by purely correlational aperture-acoustic information.

In a second experiment, Bernstein et al. (2004) found that when the anticipatory motion was unreliable (because some stimuli were created to have 20 static frames of the speaker's face rather than the natural visual lead-in), the detection advantage for all natural speech types was similar to the advantage for non-speech stimuli that were presented in their first experiment. This is an additional example of the flexibility of the audio-visual advantage, where the amount of relative uncertainty affects performance on a particular task (see Chapter 3).

Based on their results, Bernstein et al. (2004) conclude that the detection improvement effects for audio-visual speech perception are not a result of highly correlated audio-and visual signals. However, the pairing of an arbitrary visual signal with an ecologically natural auditory stimulus like speech, which listeners have had

extensive cross-modal experience, could have affected this result. Furthermore, the variability across participants was large. Most participants did show consistent best thresholds on AV speech stimuli (7/10), and worst thresholds on auditory alone stimuli (9/10); however, relative thresholds for the nonspeech audiovisual stimuli were variable for each participant, with only two of 10 participants displaying the same relative pattern as the group means reported.

Regardless of interpretation of the nonspeech stimuli, Bernstein et al. (2004) do show that detection thresholds to audio-visual speech is improved relative to audio-alone speech, in support of Grant & Seitz (2000) and Grant (Grant, 2001). With these findings in mind, an open question is how this reduction in detection threshold is instantiated in the brain. This detection improvement for matched audiovisual speech may play a key role in the audiovisual speech advantage, and therefore should be investigated further to understand how such an advantage is implemented neurophysiologically. If envelope correlation is evaluated across incoming sensory modalities and provides a boost in detection tasks, it may be possible to find neural correlates of this boost.

One electrophysiological approach that has been used extensively for testing sensory responses is the steady-state response (SSR). The SSR is a peak in neural activity occurring at a frequency that corresponds to the repetition or modulation rate of an input signal, and reflects entrainment to the temporal properties of the stimulus (Mäkelä, 2007). This response has been documented for both visual and auditory

signals and has been used extensively for clinical and diagnostic purposes (Sohmer, Pratt, & Kinarti, 1977), most commonly in EEG but also in MEG (Müller et al. 1997; Ross et al. 2000). Auditory SSRs are generally elicited by amplitude or frequency modulated signals, or both (T. W. Picton, John, Dimitrijevic, & Purcell, 2003; Luo, Wang, Poeppel, & J.Z. Simon, 2006) at modulation frequencies below about 80 Hz, while visual SSRs (also called Steady State Visual Evoked Potentials (SSVEPs)) are typically elicited by transient high-contrast stimuli such as checkerboard reversals or luminance flicker, and are typically elicited above about 4 Hz (Di Russo et al., 2007).

Although the SSR has typically been used clinically to test sensory function in both the auditory and visual modalities, it has also been adopted as an experimental tool. The amplitude of the SSVEP response has been shown to increase for attended compared to unattended stimuli (e.g. Müller & Hillyard, 2000), and this finding has been exploited to test attentional limits in multisensory perception. For example, Talsma et al. (2006) found that the amplitude of SSVEPs elicited by rapidly presented serial presentation of letter streams was reduced when attention was directed to a concurrent stream of visual objects versus a concurrent stream of auditory objects, and that the amplitude for visual and audio-visual attention conditions did not differ, suggesting that attentional capacity is increased for cross-modal stimuli relative to unimodal stimuli.

However, these experiments have not directly assessed whether steady state responses to multimodal stimuli are qualitatively or quantitatively different than steady state

responses to each modality alone. The purpose of the experiments presented here[20] is to evaluate SSRs elicited to multisensory stimuli that share properties with speech as a potential mechanism for bimodal masking coherence protection. If increased neural synchrony is present when signals are coherently modulated across modalities, which would be reflected in increased power of the SSR at the modulation frequency, this could reflect a correlate of the detection threshold improvement seen in BCMP tasks.

This chapter explores responses to multimodal stimuli consisting of modulated auditory and visual components within the frequency range of the speech envelope. By building on results investigating SSRs to auditory and visual stimuli presented alone, we assess the SSR to bimodal audio-visual signals with the hypothesis that increased SSR power will be found for audio-visually modulated signals relative to the power for either modality modulated alone.

We expect the cortical responses in auditory and visual cortex to be sensitive to this manipulation because the auditory speech signal contains both relatively rapid frequency fluctuations in the spectral domain, along with slower (2-16 Hz) amplitude modulations corresponding to the syllabic envelope (Steeneken & Houtgast, 1980). The temporal envelope of the auditory signal—corresponding to amplitude fluctuations at roughly syllabic rate—is related to the dynamics of the visual articulators such as the lips and mandible, and the presence of this cross-modal relationship could increase the power of the neural response. Furthermore, intrinsic

[20] Portions of this chapter (plus additional analyses) published as: (Jenkins, Rhone, Idsardi, J.Z. Simon, & Poeppel, 2011).

cortical oscillations are particularly sensitive to speech frequencies in the range of 4–16 Hz (Luo & Poeppel, 2007; Howard & Poeppel, 2010), and auditory cortical responses have shown correlations with the auditory speech envelope modulations between 2-20 Hz (Aiken & Picton, 2008).

In addition, we test three envelope phase relationships to determine whether the multi-sensory SSR is sensitive to synchrony across modalities. If correspondence between auditory and visual envelopes is necessary for the benefits seen in audiovisual speech detection tasks, and the SSR is a potential index of these effects, then stimuli with offset envelopes should show decreased power relative to audio-visual stimuli that are synchronously modulated.

Experiment 3: establishing bimodal SSR

Materials and Methods

Participants*:* Thirteen right-handed (Oldfield 1971) adult subjects (seven female; mean age 27.08 years) with normal hearing and normal or corrected-to-normal vision underwent MEG scanning. One participant was excluded from all analyses due to excessive motion artifacts during MEG recording. Participants were either compensated for their participation ($10/hour) or earned course credit in an introductory linguistics course. Presentation of stimuli and biomagnetic recording was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Prior to the start of the experiment, written informed consent was obtained from each participant.

Stimuli[21]:

To control for low-level sensory activity, all signals of interest consisted of both auditory and visual components. Envelope phase relationships across modalities were manipulated for comparisons.

The experimental stimuli consisted of five types of audio-visual signals presented at two modulation frequencies, for a total of ten signals. The unimodally modulated stimuli included amplitude-modulated sine waves presented concurrently with a static white square on black background ("audio alone"), and a radius-modulated white disc on black background concurrently presented with approximately Gaussian white acoustic noise ("visual alone"). Comodulated Stimuli included a radius-modulated disc and an amplitude modulated (AM) sine wave at one of three phase relationships (in phase, $\pi/2$ radians envelope shift, π radians envelope shift—see Figure 4.1 for schematic).

[21] Materials can be obtained at http://files.ling.umd.edu/~arhone/Thesis/Ch4_stimuli/ or by email request: ariane.rhone@gmail.com

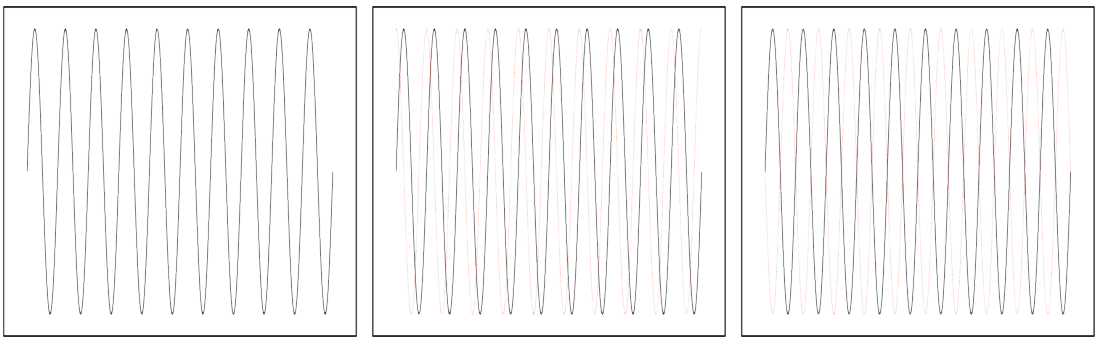


Figure 4.1 Schematic of phase relationships for comodulated conditions (Fm = 2.5 Hz).
Left panel: modulation envelopes in synch; middle panel: pi/2 envelope offset; right panel pi offset. Period of amplitude modulation in this condition is 400ms, duration is 4 s (10 cycles per trial).

The amplitude-modulated sine waves and radius-modulated discs were modulated at 2.5 Hz or 3.7 Hz with 24 percent modulation depth. All stimuli were 4s duration. These values were chosen after extensive piloting revealed that higher visual modulation frequencies were uncomfortable for participants to view for extended periods of time. Two frequencies were chosen to establish SSR effects at distinct, non-harmonically related modulation frequencies. For the comodulated conditions, the auditory and visual signal components had the same onset and offset, with the auditory component reaching the maximum value of the modulation envelope first for out-of-phase conditions.

Auditory signal components were generated with MATLAB (v2007b) and consisted of a sine wave envelope (either 2.5 Hz or 3.7 Hz modulation frequency) applied to an 800 Hz sine wave carrier signal with 6 ms $\cos^2$ onset and offset ramps presented at approximately 65 dB SPL. The signals were sampled at 44.1 kHz with 16-bit resolution. Signals were generated using the sine function to eliminate undesired phase effects on onset responses. Visual signal components were generated using Gnu

Image Manipulation Program (www.gimp.org). The radius-modulated white discs were centered on a 640 x 480 pixel black background, and ranged from 2.5° visual angle minimum diameter to 4°visual angle maximum diameter (see Figure 4.2 for stimulus schematic).

Figure 4.2 Schematic of audio-visual pairing for radius modulated circles and amplitude modulated pure tone.

The individual frames were compiled into Audio-Video Interleave (AVI) format using VirtualDub (www.virtualdub.org) for presentation. Stimulus timing/frequency was verified with an oscilloscope. The visual components were projected on a screen approximately 30 cm from the participant's nasion. Participants were supine in the MEG scanner for the duration of the experiment.

Experimental stimuli were presented in nine blocks, with three repetitions per signal per block (27 total per condition). Presentation of conditions was randomized within blocks. The experimental materials were passively attended to; however, a distracter task was incorporated to encourage participant vigilance. A target audio-visual crosshair combined with approximately Gaussian white noise (500 or 1500 ms duration) was pseudorandomly presented with the experimental signals (~17% of

total trials). Participants were instructed to press a button when they detected the target signal; these trials were excluded from analysis.

Delivery: All experimental stimuli were presented using a Dell OptiPlex computer with a SoundMAX Integrated HD sound card (Analog Devices, Norwood, MA) via Presentation stimulus presentation software (Neurobehavioral Systems, Inc., Albany, CA). Stimuli were delivered to the subjects binaurally via Eartone 183 ER3A transducers and non-magnetic air-tube delivery (Etymotic, Oak Brook, IL). The inter-stimulus interval varied pseudo-randomly between 2500 and 3500 ms.

Recording: Data were acquired using a 160-channel whole-head biomagnetometer with axial gradiometer sensors (KIT System, Kanazawa, Japan). Recording bandwidth was DC-200 Hz, with a 60 Hz Notch filter, at 1000 Hz sampling rate. Data were noise reduced using time-shifted PCA (de Cheveigné & Jonathan Z Simon, 2007) trials averaged offline (artifact rejection ± 2.5 pT), bandpass filtered between .03 - 25 Hz (161 point Hamming window) and baseline corrected over the 700 ms pre-stimulus interval.

Data Analysis

The analysis was performed in sensor space, not source space, to stay as close as possible to the recorded data without making source configuration assumptions. All analyses—pre-experiment localization parameters, waveform assessment, and the calculation of the magnitude and phase of the SSR as well as significance values –

were performed in MATLAB. Statistical analysis of SSR parameters was evaluated using the statistical and probability distribution functions in MATLAB's Statistics Toolbox.

Sensor selection from pre-test: Determination of maximally responsive auditory and visual channels was performed in separate unimodal pre-tests. The auditory pre-test consisted of amplitude-modulated sinusoidal signals with 800 Hz sinusoidal carrier signal, modulation frequency 7 Hz, 100 percent modulation depth, 11.3 s duration. The visual pre-test consisted of a checkerboard flicker pattern (Fm = 4 Hz, 240 s duration).

The sensor space was divided into quadrants (see Figure 4.3) to characterize the auditory response and sextants to characterize the visual response based on the expected peak and trough field topography recorded from axial gradiometers for each modality. Sensor channel designations were: anterior temporal (front of head), posterior temporal (rear quadrants/ middle of head) and occipital (back of head overlying occipital lobe). Five channels from source and sink from each sensor division (i.e. ten channels for auditory response and five channels for visual response per hemisphere; 15 channels per hemisphere total) with the maximum measured magnetic field deflection were used for subsequent analyses.

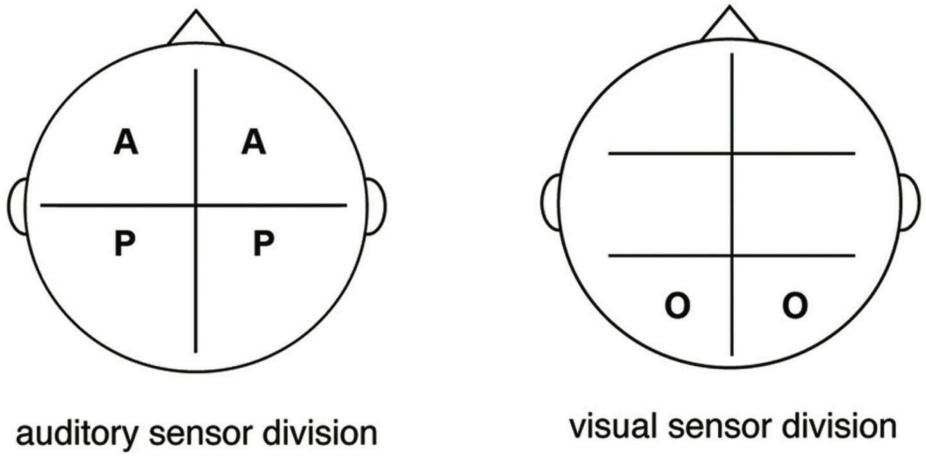


Figure 4.3 Sensor divisions for Experiments 3 and 4.
Auditory sensors (left) are divided into anterior temporal (A) and posterior temporal (P) quadrants. Visual sensors (right) were divided into sextants, with the most posterior sextants (O) used for analysis.

The auditory pre-test response was characterized using two methods. The first analysis examined the power spectral density (PSD) of the response and selected the sensors with the strongest response (Fourier Transform window: 3 to 5 s), at the modulation frequency. The second analysis examined the maximum field deflection of the M100 response (search window: 80 to 130 ms post stimulus onset) and selected the channels with the maximum response amplitude (both source and sink). The pretest visual response was characterized only using the PSD, at twice the modulation frequency (the reversal rate), because the low number of trials did not provide a sufficient signal-to-noise ratio for standard averaged onset analysis.

Because the data were analyzed in sensor space rather than source space, special care was taken to avoid having overlap in posterior temporal and occipital sensors. This ensured that no sensor was contributing to more than one analysis grouping. When

posterior temporal and occipital sensors did overlap as one of the strongest sensors in both unimodal pretests, the overlapping sensor was removed from the posterior temporal division and was replaced with the next strongest non-overlapping posterior temporal sensors.

Onset response evaluation and PCA: The signal evaluation window (for averaged and filtered sensor data) ranged from 700 ms pre-trigger to 3999 ms post-trigger. Onset peak root-mean-square (RMS), RMS latency, magnetic field deflection and magnetic field deflection latency responses corresponding to the M100 (auditory; search window: 80 to 130 ms after stimulus onset) and M170 (visual; 145 to 195 ms after stimulus onset) for each hemisphere for each condition were collected and averaged across subjects for each stimulus and were plotted topographically to examine the response. The number of trials averaged ranged from 12-27.

SSR analysis: The magnitude and phase spectra of the SSR were determined using the Fast Fourier Transform (FFT) of the baseline corrected and filtered channel data. The FFT was calculated from stimulus onset (0 ms) to the end of the signal evaluation window (3999 ms). Prior to FFT calculation, the data was multiplied by a Kaiser window (length 4000 samples, beta = 13) to minimize onset and offset responses to the audio-visual signals and to minimize spurious frequency contributions.

The magnitude of the response was calculated using the RMS of the FFT across channels. The phase response was determined by calculating the mean direction as

described by Fisher (1996) based on the phase angle of the Fourier-transformed data. The across subject response power was determined by calculating the mean of the individual subject power vectors. To determine the across subject phase response, the mean direction of the individual mean directions was calculated.

SSR cross-modal control analysis: To determine the validity of the sensor selection from the pre-experiment localization, unimodal modulation data were analyzed using the sensors from the other modality (e.g., unimodal auditory response was evaluated using the occipital sensors, and unimodal visual response was evaluated using the anterior and posterior temporal sensors). This analysis confirmed that the responses recorded from the unimodal modulation conditions truly reflected that particular modality.

Across-subject response averaging: Across-subject responses were computed by collecting individual subject field deflections (source and sink field deflections and RMS) and calculating the mean response amplitudes and the RMS of the subject RMS values. The aggregate waveforms peaks and latencies were characterized in the same search windows as described above. Individual subject vectors for response power (squared magnitude) and phase were also collected for statistical analyses.

Statistical analyses:

The significance of the SSR amplitude at a specific frequency was analyzed by performing an *F* test on the squared RMS (power) of the Fourier transformed data

(Dobie and Wilson 1996). The signal evaluation window resulted a frequency resolution of 0.25 Hz and gave the exact response at Fm = 2.5 Hz, but not at 3.7 Hz. To evaluate the response at Fm = 3.7 Hz, the bin closest in frequency (3.75 Hz) was used. The significance of the phase of the response was assessed using Rayleigh's phase coherence test on the mean direction (Fisher, 1996). Individual subject responses at each modulation frequency for each condition were assessed using an *F* test to determine if the response was significant and whether or not a particular subject should be excluded due to lack of a response or exhibiting a response other than at the modulation frequencies and harmonics of interest. For the across-subject data, *F* tests were performed on the power of the SSR at the modulation frequency, two subharmonics, and the second and third harmonics; these harmonics may relate to functionally significant bands (see e.g., Jones and Powell (1970); Senkowski et al. (2008) for review of frequency band descriptions;).

The power at individual harmonic components of the modulation frequency at the subharmonics and harmonics across conditions was compared using Wilcoxon signed-rank tests (Matlab v7). Two sets of signed-rank tests were performed: the first compared the mean unimodal modulation magnitudes against the mean comodal modulation magnitudes for a given sensor area (e.g. LH anterior temporal unimodal auditory vs. LH anterior temporal comodal, $\Phi = \pi$[22]) and the second compared the comodulated conditions (e.g. RH occipital, Φ = zero vs. RH occipital, $\Phi = \pi/2$. A

[22] The symbol Φ (phi) denotes the phase shift of the envelope across modalities.

mixed effect ANOVA implemented in R (Baayen, 2008) assessed any possible differences in modulation frequency and hemisphere.

Results

Across-Subject Power Analysis

Figure 4.4 displays the across subject response power for Fm = 3.7 Hz, plotted with a linear scale for frequency and a logarithmic scale for response power, shown here for right hemisphere sensors only. Across conditions, anterior channels do not show substantial SSR power; posterior temporal and occipital channels reveal clear peaks in activity at the modulation frequency and its harmonics.

Response power was concentrated at the modulation frequency and the second harmonic, with some activity entered also around 10 Hz. Response power significance for all bimodal conditions (as determined by Wilcoxon signed-rank tests) compared to the unimodal modulation conditions show that the responses are significantly greater in bimodal than unimodal responses.

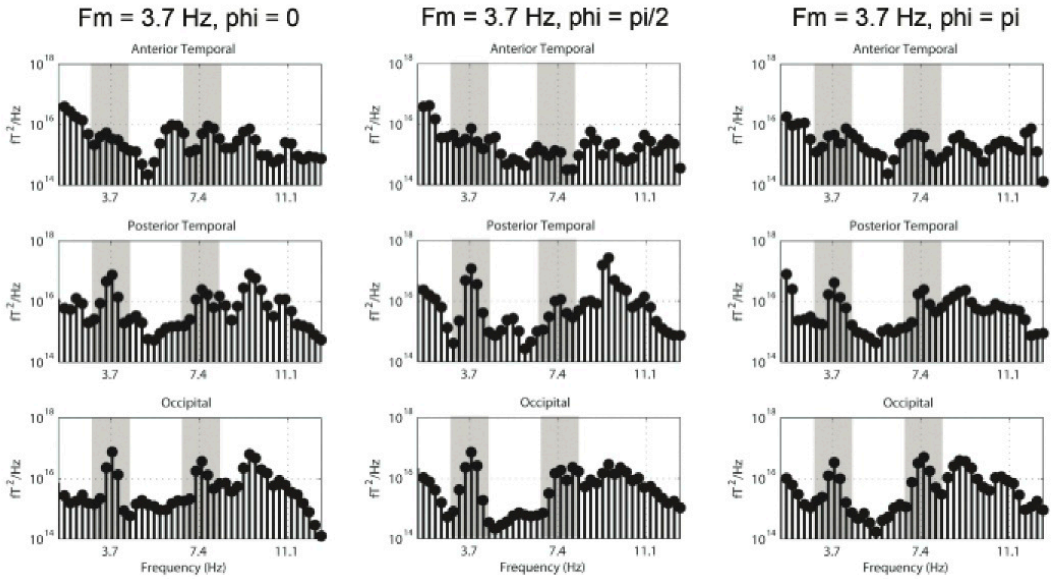


Figure 4.4 Across subject response power Fm = 3.7 Hz for RH sensors only
Peaks in activity within the shaded bands indicate power at frequencies of interest (modulation frequency and second harmonic)

Several results merit highlighting: first, the majority of the activity is reflected in the sensors overlying the posterior temporal lobes and occipital lobes; second, for the AV comodulated condition in which the signal envelopes are at the same initial phase, the response power is greatest at the modulation frequency, localized to the sensors overlying the posterior temporal lobes; third, as the difference in the relative phase increases, the response power decreases, although the response is still greater than that of unimodal modulation condition (see Figure 4.5).

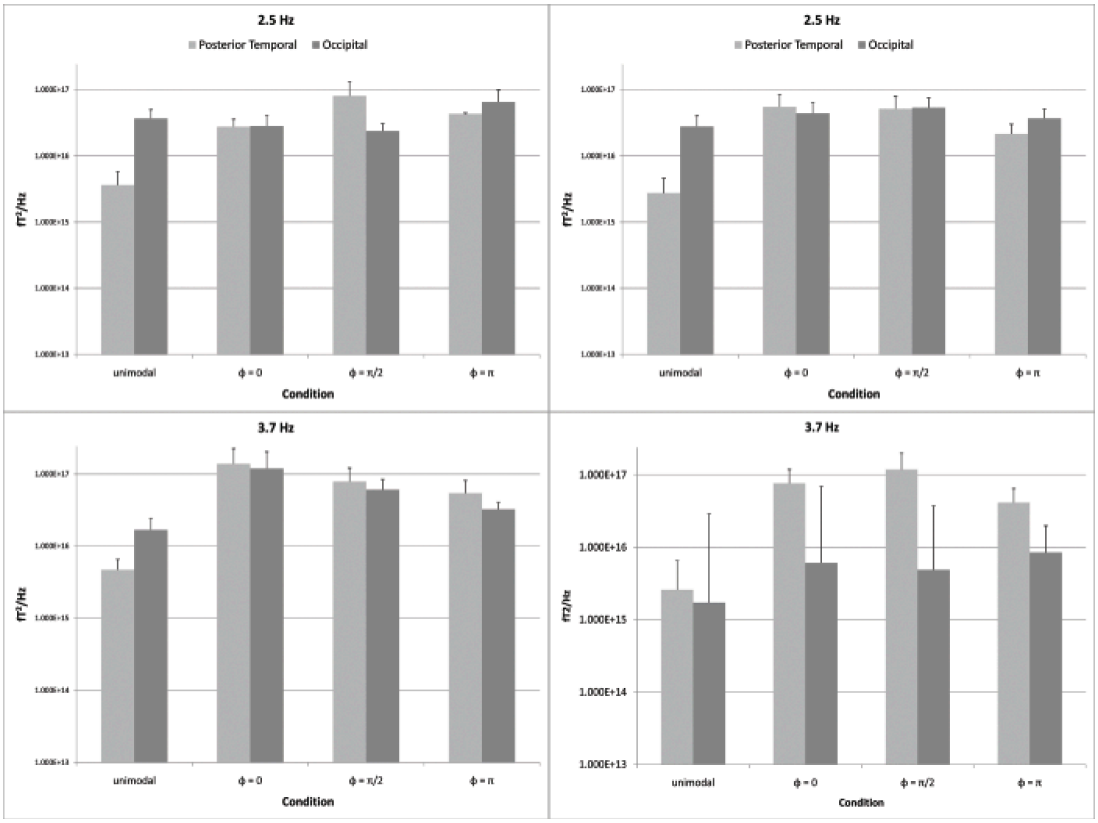


Figure 4.5 Across-subject response power for all conditions at Fm, by hemisphere
(top panels: 2.5 Hz Fm, bottom panels: 3.7 Hz Fm; left panels: LH, right panels: RH).

Statistical summary

The significance of the SSR was calculated at the modulation frequency, as well as two subharmonics, and the second and third harmonics. Significance was determined by means of an *F* test on the power of the SSR at each frequency as described by Dobie and Wilson ((1996) - see *Methods*) and takes into account both amplitude and phase of the response (Valdes et al., 1997; T. W. Picton, John, et al., 2003). All subjects elicited a statistically significant response for the SSR at each envelope modulation frequency. Within-subject response significance was restricted to evaluation at the modulation frequency (see *Methods*) with degrees of freedom (*df*) (2,12) and $\alpha = 0.05$. The across subject significance for subharmonics was assessed

using *df* = (2,4) and significance for the modulation frequency and second and third harmonics were assessed using *df* (2,12).

Statistically significant responses were observed at the modulation frequency, as well as second and third harmonics. SSR power at subharmonics was not statistically significant. The difference between the observed statistical significance for subharmonics and the second and third harmonics may be attributable to the decreased degrees of freedom for this comparison.

Results of Rayleigh's test on the mean direction of the SSR vectors (at the frequencies observed to be significant by the *F* test) found the phase angle directions to be statistically significant at $\alpha = 0.05$.

SSR power comparisons

For both modulation frequencies, several statistically significant responses are held in common. First, responses to both Fm =2.5 and Fm 3.7 exhibit statistically significant responses power at the modulation frequency for all comodulated conditions, and this interaction is largely limited to the sensors overlying the posterior temporal lobe for both hemispheres. Second, there were significant interactions at the second harmonic for $\Phi = 0$ and $\Phi = \pi$; both modulation frequencies held this interaction in common in the LH sensors overlying the posterior temporal lobe. One last interaction was common to both modulation frequencies for the third harmonic for $\Phi = 0$ in the LH sensors overlying the occipital lobe.

Several other statistically significant interactions were found to be unique to each modulation frequency; these perhaps inconsistent interactions may be a result of true variance in the modulation frequencies tested, or could be an artifact of analysis techniques. No statistical difference was observed for SSR power between the three bimodal conditions. Linear mixed effects models with modulation frequency and hemisphere as factors found no significant statistical interactions. Wilcoxon signed-rank tests were performed on the incidental power centered around 10 Hz to determine if it was significant. Results of the tests across conditions yielded no significant results[23].

Overall, we find that redundant information present in comodulated audio-visual stimuli resulted in increased response power relative to unimodally-modulated conditions, regardless of phase incongruities. However, there is a more pronounced effect for presumably primary auditory cortical neuronal population.

Interim Discussion

Experiment 3 established that a multisensory steady state response (SSR) could be elicited at unrelated modulation frequencies using non-traditional stimulus types (looming-receding circles and low frequency amplitude modulated sine waves). However, no effect of phase envelope was shown in the comparison of the three

[23] Power in this frequency band (near 10 Hz) could be attributed to endogenous alpha activity, related to the attentional states of the participants.

comodulated conditions. As with any non-effect, it was unclear whether there was truly no difference between envelope phase relationship conditions or if our analysis was not powerful enough to reveal it statistically. In particular, the low number of epochs averaged for each condition do not provide an ideal signal-to-noise ratio for analysis of averaged data, even in the frequency domain (see Ross et al., (2000).

Furthermore, the goal of this study is to examine potential mechanisms for the enhancement effects seen in speech perception--and although the two modulation frequencies used in Experiment 3 fell within the envelope ranges of speech, the pure-tone carrier and visual circle are, admittedly, poor approximations of actual audio-visual speech stimuli.

With this in mind, Experiment 4 was designed to elicit steady state responses to a still highly controlled, but more speech-like signal. The auditory signal was changed from a pure tone carrier to a broader band (filtered pink noise) carrier, and the visual signal was changed from a radius-modulated circle to an ellipse shape that was modulated on its minor axis (to simulate opening and closing of the mouth). The stimulus duration was slightly shortened, and only one modulation frequency was used so that the number of trials that were presented for each condition could be increased without increasing total recording time.

Experiment 4: SSR to more 'speechlike' stimuli

Materials and Methods

Participants: Fourteen participants (thirteen right-handed; one ambidextrous, as tested by the Edinburgh Handedness Inventory (Oldfield, 1971); six female) with normal hearing and normal or corrected-to-normal vision underwent MEG scanning. Data from two participants were excluded due to an insufficient signal-to-noise ratio for all conditions. Age range was 18–27 (mean 20.1 years). Participants were compensated for their participation. Presentation of stimuli and biomagnetic recording was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Prior to the start of the experiment, written informed consent was obtained from each participant.

Stimuli

As with Experiment 3, all signals were presented bimodally to control for low-level sensory activity. Unimodally modulated conditions were an amplitude-modulated three-octave pink noise presented concurrently with a static white rectangle on a black background ("audio alone") and a radius-modulated white ellipse on a black background concurrently presented with approximately Gaussian white acoustic noise ("visual alone"). The same three envelope phase relationships were examined (0, $\pi/2$ radians, π radians) for comodally modulated conditions. The amplitude-modulated

three-octave pink noise and radius-modulated ellipses were modulated at 3.125 Hz[24] with a modulation depth of 25% of peak amplitude and radius for audio and visual signals, respectively. The SSR- inducing signals were 3.520 s in duration, for a total of 11 cycles of "opening" and "closing" per trial. For the comodulated conditions, the auditory and visual signal components had the same onset and offset, with the auditory component reaching the maximum value of the modulation envelope first for out-of-phase conditions (see Figure 4.6).

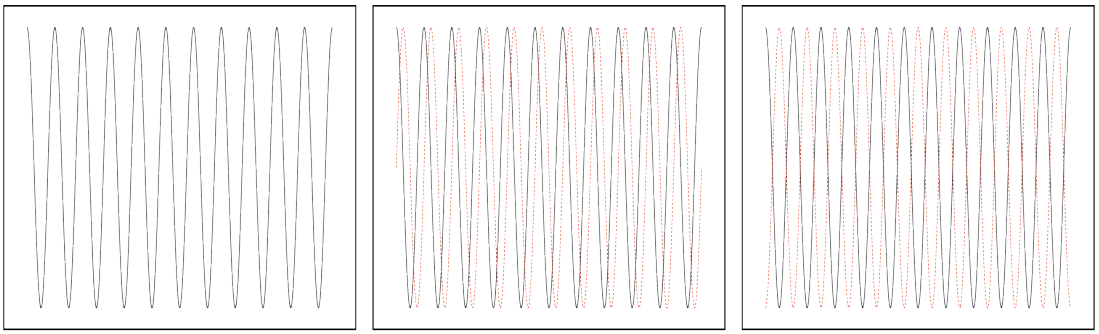


Figure 4.6 Envelope phase relationships (Fm = 3.125 Hz)
Left panel: synchronous envelopes; middle panel: π /2 offset; right panel: π offset.

Auditory signal components were generated with MATLAB (R2009a, The Mathworks, Natick, MA) and consisted of a cosine wave envelope (3.125 Hz modulation frequency) applied to a three-octave pink noise carrier signal with 6 ms $\cos^2$ onset and offset ramps presented at approximately 65 dB SPL. The cosine function was chosen to maximize onset responses. The three-octave pink noise contained a lowest frequency of 125 Hz and was generated using the NSL Toolbox

[24] A modulation frequency of 3.125 Hz was chosen because it falls within the range of speech envelope rates and is contained within a single bin for Fourier analysis, eliminating the need for windowing and filtering the SSR data that was done for Experiment 3.

for MATLAB (Chi and Shamma, http://www.isr.umd.edu/Labs/NSL/Software.htm). These parameters cover the fundamental frequency range of the human voice as well as the frequency region where most of the energy arising from the first formant tends to be concentrated. The signals were sampled at 44.1 kHz with 16-bit resolution.

Visual signal components were generated using Gnu Image Manipulation Program (www.gimp.org). The radius-modulated white ellipses were centered on a 640 x 480 pixel black background, and ranged from 0.84° to 1.68° visual angle for the minor radius and 3.71° visual angle for the major radius (see Figure 4.7 for Experiment 4 stimulus schematic). The individual frames were compiled into Audio–Video Interleave (AVI) format using Virtual Dub (www.virtualdub.org). Stimulus timing and modulation frequency was verified with an oscilloscope. The visual components were projected on a screen approximately 30 cm from the participant's nasion. Participants were supine in the MEG scanner for the duration of the experiment.

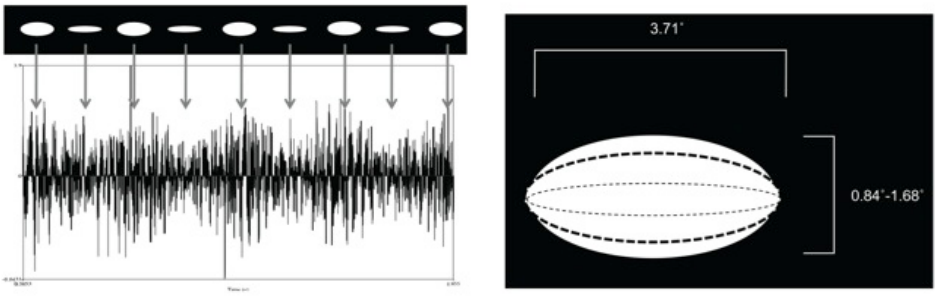


Figure 4.7 Stimulus schematic for Experiment 4

To maintain vigilance to both modalities, brief targets (500 ms) were pseudorandomly interleaved throughout the experimental trials. To encourage attention to both modalities, audio-only (white noise burst), visual-only (crosshair), and audio-visual targets (noise + crosshair) were used.

Experimental stimuli were presented in six blocks, with 15 repetitions per signal per block, for a total of 90 trials per condition. Presentation of conditions was randomized within blocks. The SSR-inducing materials were passively attended to; target signals (38% of trials) required a button press.

Delivery

Stimuli were presented using a Dell OptiPlex computer with a M-Audio Audiophile 2496 soundcard (Avid Technology, Inc., Irwindale, CA) via Presentation stimulus presentation software (Neurobehavioral Systems, Inc., Albany, CA). Stimuli were delivered to the participants binaurally via Eartone ER3A transducers and non-magnetic air-tube delivery (Etymotic, Oak Brook, IL). The interstimulus interval varied pseudo-randomly between 980 and 2000ms.

Recording and Filtering

Data were acquired using a 160-channel whole-head biomagnetometer with axial gradiometer sensors (KIT System, Kanazawa, Japan). Recording bandwidth was DC-200 Hz, with a 60 Hz Notch filter, at 1000 Hz sampling rate. The data were noise

reduced using time-shifted PCA (de Cheveigné & Jonathan Z Simon, 2007) and trials were averaged offline (artifact rejection ± 2.5 pT) and baseline corrected.

Participant Head Location

Head position measurements using sensors at standard anatomical fiducial points were taken prior to and after experimental completion to determine proper head placement within the dewar, that the sensors were recording from the entire head (occipital, posterior temporal/parietal, anterior temporal/frontal areas), and to ensure that participants did not have significant head movement during the recording session.

Analysis

Determination of maximally responsive auditory and visual sensors was performed in separate pre-tests for each modality (see Experiment 3 for materials and methods for pre-test sensor selection).

Onset and Dipole Analyses

The higher number of epochs that were averaged for a particular condition made onset analysis more viable for this experiment (relative to Experiment 3). However, large variation across participants precluded an extensive group analysis of onset effects. Dipole analysis was also unsuccessful.

SSR Analysis

The magnitude and phase spectra of the SSR were determined using the Fast Fourier Transform (FFT) of the baseline corrected channel data. The FFT was calculated from 320 ms post-stimulus onset to the end of the signal evaluation window (3519 ms) for a total of 3200 samples; this yielded frequency bins commensurate with the modulation frequency and its harmonics. The magnitude of the response was calculated using the RMS of the FFT across channels. The phase response was determined by calculating the mean direction as described by Fisher (1996) based on the phase angle of the Fourier transformed data. The across participant response power was determined by calculating the mean of the individual participant power vectors. To determine the across participant phase response, the mean direction of the individual mean directions was calculated.

Statistical Analyses

The significance of the SSR amplitude at a specific frequency was analyzed by performing an F test on the squared RMS (power) of the Fourier transformed data using the MATLAB Statistics Toolbox. For the across-participant data, F tests were performed on the power of the SSR at the modulation frequency and the second harmonic. Responses at the third harmonic were not statistically different from background noise. The response power in linear values and decibels (dB) was assessed using ANOVAs as well as General Linear Models (GLMs) using the ‘‘languageR’’ statistical package (Baayen, 2008). Factors for both sets of statistical tests were Hemisphere, Harmonic, Condition, and Sensor Area, with Participant as a random effect. To determine the separation of densities, distributions of the responses

for each hemisphere, harmonic, condition and area were compared using Kolmogorov–Smirnov tests.

Additionally, we compared response additivity using the AV versus (A + V) model, using the complex representation from the Fourier transform of the data[25] on the frequency bins containing the first and second harmonic. Statistical differences were assessed using Wilcoxon signed-rank tests in order to decrease the assumptions concerning the distribution of the data recorded between pairs of conditions.

Results

SSR responses were reliably generated. The response pattern showed no difference between hemispheres in the power of the response (as measured using ANOVAs and GLMs as well as data visualization), and also showed that the posterior temporal and occipital channels had the greatest response, as in Experiment 3. SSR power was analyzed using decibel (dB), rather than linear, power values due to the effectively normally distributed nature of dB power measurements (Dobie & Wilson, 1996). Data visualization of power densities was performed using the ‘‘ggplot2’’ package for R (Wickham, 2009). The dB values provide to a more robust and comprehensible statistical analysis.

[25] If the additivity were assessed using RMS, this would assume that a single source is generating the response for each condition, which is not appropriate for the multi-modal nature of the stimulus

Across-Participant Power Analysis

As in Experiment 3, most of the response power was generated in the sensors overlying the posterior temporal and occipital areas. Response power was concentrated at the modulation frequency and the second harmonic, and the power values at those frequencies were used for the subsequent statistical analyses. Statistical significance was assessed using F tests with 2 and 12 degrees of freedom (df = 2, 12, $\alpha = 0.05$) and was confirmed by comparing the average power of the background noise (surrounding frequency bins) with the bin containing the modulation frequency. On average, the frequency bins containing the frequencies of interest were an order of magnitude (~10 dB) greater than the background, except for responses measured at anterior temporal sensors, as in Experiment 3 (see Figure 4.8).

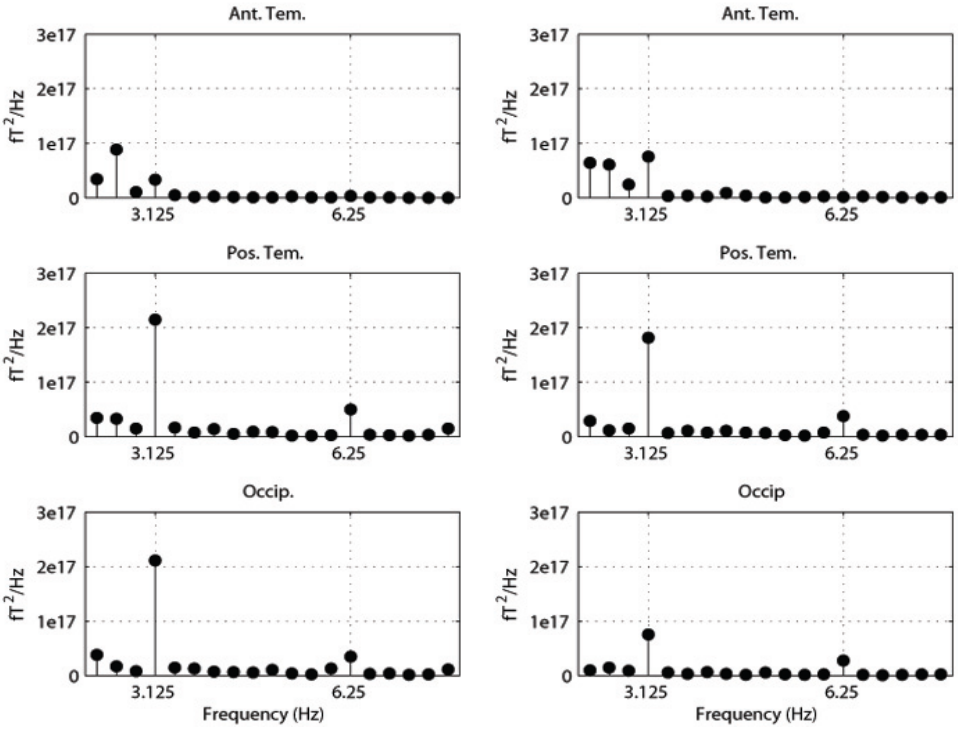


Figure 4.8 Grand averaged response power for all participants, ϕ=0 condition

For the unimodal modulation conditions, statistically significant F ratios were found at the modulation frequency for the occipital sensors in both hemispheres (LH: $F = 37.441$, $p < 0.01$; RH: $F = 10.539$, $p < 0.01$), but not for the anterior and posterior temporal sensors; the second harmonic F ratio was significant only in the RH occipital sensors ($F = 7.853$, $p < 0.01$).

For the $\Phi = 0$ comodulated condition at the modulation frequency, significant F ratios were found for the posterior temporal and occipital sensors in the LH ($F = 7.822$, $p < 0.01$ and $F = 60.107$, $p < 0.01$, respectively); the RH occipital sensors F ratio was marginally significant ($F = 4.113$, $p < 0.05$); this same pattern held for the second harmonic ($F = 4.839$, $p < 0.05$; $F = 4.733$,$p < 0.05$; $F = 4.061$, $p < 0.05$, respectively).

For the $\Phi = \pi/2$ condition, significant F ratios were found for the occipital sensors in both hemispheres at the modulation frequency (LH: $F = 74.436$, $p < 0.01$; RH: $F = 10.04$, $p < 0.01$) and the LH occipital sensors for the second harmonic ($F = 37.351$, $p < 0.01$). For the $\Phi = \pi$ condition, significant F ratios were found for the posterior temporal (LH: $F = 16.833$, $p < 0.01$; RH: $F = 7.358$, $p < 0.01$) and occipital sensors (LH: $F = 23.954$, $p < 0.01$; RH:$F = 12.864$, $p < 0.01$) at the modulation frequency; at the second harmonic significant F ratios were found for the occipital sensors (LH: $F = 12.663$, $p < 0.01$; RH:$F = 8.127$, $p < 0.01$) and the RH posterior temporal sensors ($F = 3.901$, $p < 0.05$).

Statistical Summary

Separate ANOVAs were calculated with the following interactions: (i) Hemisphere (two levels) x Harmonic (two levels) x Condition (four levels) x Sensor Area (three levels), (ii) Harmonic x Condition x Sensor Area and (iii) Condition x Sensor Area. For the first ANOVA, significant interactions were found for Harmonic ($F(1,13) = 148.053$, $p < 0.001$), Sensor Area ($F(2,13) = 134.441$,$p < 0.001$), and Condition x Sensor Area ($F(6,13) = 4.208$, $p < 0.001$); the interaction Hemisphere x Sensor Area was marginally significant ($F(2,13) = 3.013$, $p = 0.049$). For the second ANOVA, significant interactions were found for Harmonic ($F(1,13) = 150.546$, $p < 0.001$), Sensor Area ($F(2,13) = 136.705$, $p < 0.001$) and Condition x Sensor Area ($F(6,13) = 4.279$, $p < 0.001$). For the third ANOVA, significant interactions were found for Sensor Area ($F(2,13) = 111.093$, $p < 0.001$) and Condition x Sensor Area ($F(6,13) = 3.477$, $p < 0.05$).

Two-sample Kolmogorov–Smirnov tests indicated that the power distributions for the harmonics ($D = 0.324$, $p < 0.001$), anterior and posterior temporal sensors ($D = 0.455$, $p < 0.001$), anterior temporal and occipital sensors ($D = 0.4821$, $p < 0.001$) and posterior temporal and occipital sensors ($D = 0.134$, $p < 0.05$) differed significantly.

Post hoc analyses on the posterior temporal channels found significant interactions of Harmonic ($F(1,13) = 49.199$, $p < 0.001$; $F(1,13) = 50.157$, $p < 0.001$) and Condition ($F(3,13) = 10.103$, $p < 0.001$; $F(3,13) = 10.300$, $p < 0.001$) for the triple- and double-

factor ANOVAs and Condition ($F(3,13) = 8.348$, $p < 0.001$) for the single-factor ANOVA.

SSR Power Comparisons

Figure 4.9 illustrates the differences in overall power between harmonics for each condition for the entire dataset for all sensor divisions (collapsed across hemispheres). Plots of the mean dB power show there is no statistical difference in power between conditions, but there is a difference in the power between harmonics, with the modulation frequency exhibiting greater power for each condition than the second harmonic (a typical SSR response property).

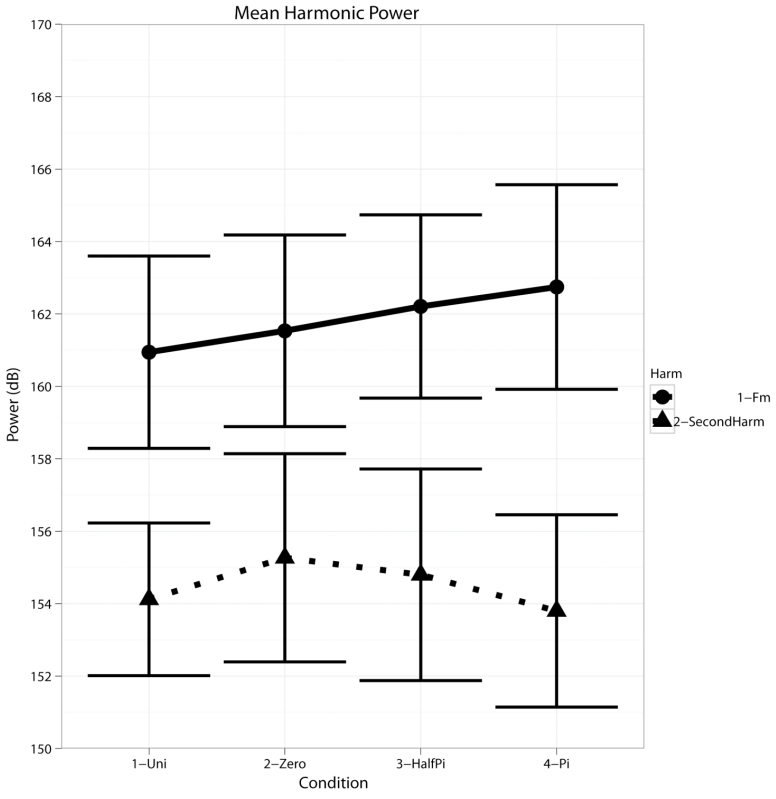


Figure 4.9 Mean harmonic power by condition, collapsed across all sensor areas

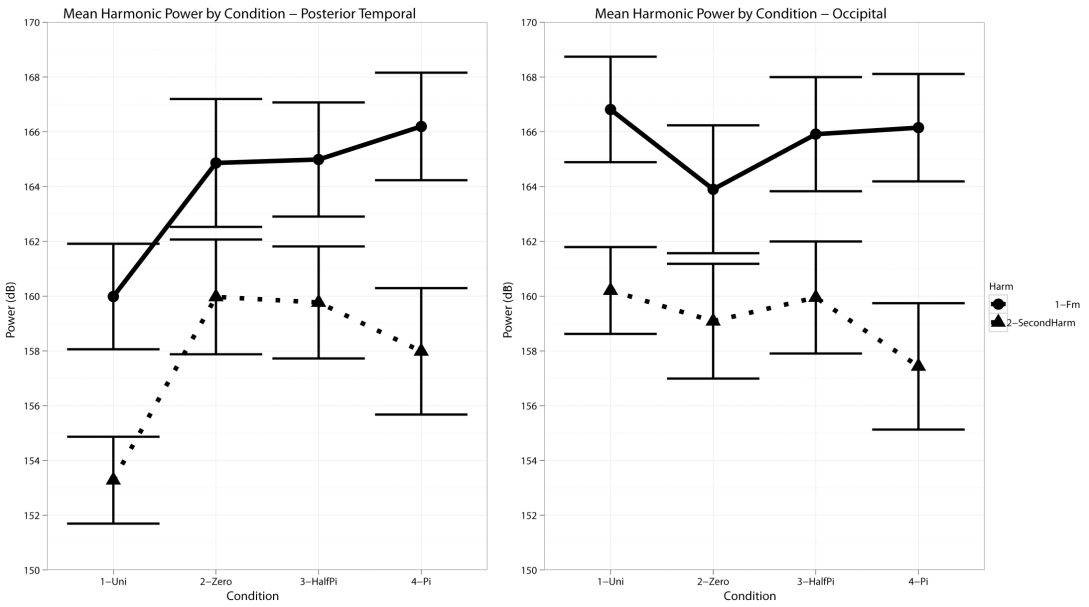


Figure 4.10 Mean harmonic power for Posterior Temporal and Occipital Sensors by condition. Posterior temporal sensors show a significant effect of comodulation at both the modulation frequency and the second harmonic.

Figure 4.10 shows changes in response power for the posterior temporal (left panel) and occipital (right panel) sensors across comodulation conditions. Several trends can be observed. First, there is greater power at the modulation frequency than at the second harmonic. Second, the comodulated conditions exhibit greater power than the unimodally modulated conditions. Third, and most importantly, the difference in power between unimodal and comodal conditions seems to be directly attributable to posterior temporal sensors. No difference in power for either harmonic across conditions is observed in the occipital sensors.

Figure 4.11 and Figure 4.12 illustrate the grand average topography at the modulation frequency and the second harmonic, respectively, in the form of phasor plots, which show the sink-source distribution and the phase of the response (J.Z. Simon & Wang, 2005). The sink-source distribution (and phase distribution) at the modulation

frequency (Fig. 4.10) for all conditions resembles that of a visual response recorded from axial gradiometer sensors. This supports the results from the power analyses, which showed that the occipital sensors generated larger responses than the anterior and posterior temporal sensors.

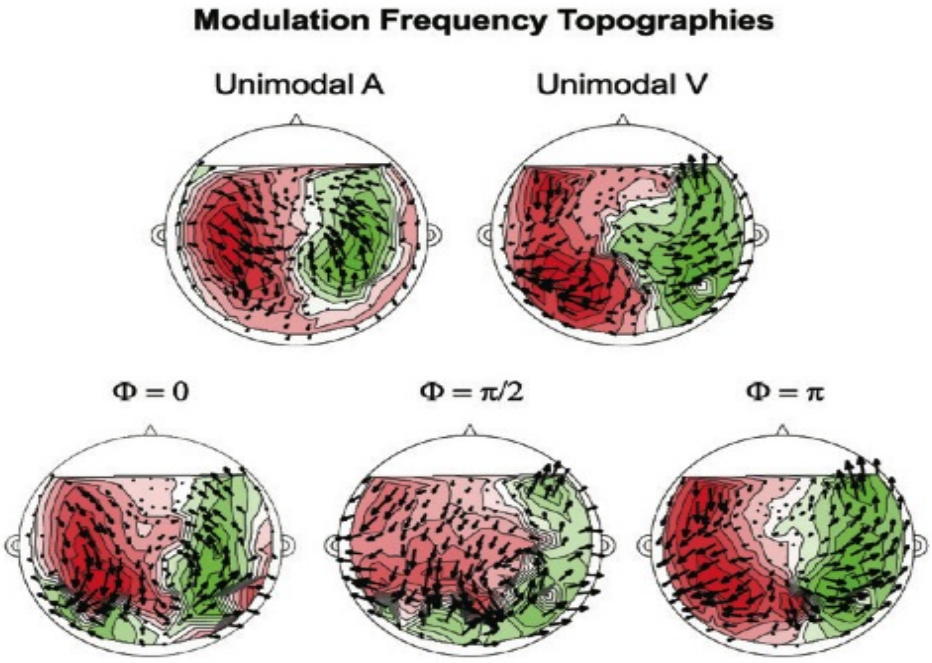


Figure 4.11 Response topographies at modulation frequency

For the response at the second harmonic (Fig. 4.12), the topographies are less straightforward. For the unimodal auditory condition, the sink-source distribution reflects responses typically recorded from auditory cortex. For the unimodal visual condition, the sink-source distribution appears mixed. The sink-source distribution for the comodal conditions suggests (i) the degree of synchronicity and integration between the signal components and (ii) the contribution of the posterior temporal sensors (with contributions from auditory cortex and/or parietal lobes).

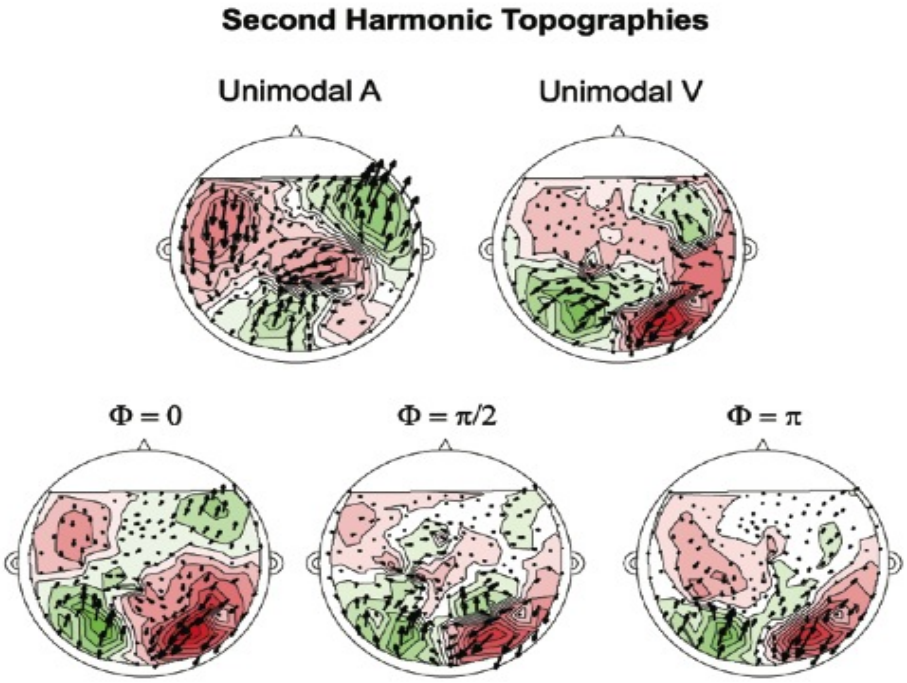


Figure 4.12 Response topographies at second harmonic

For the Φ = 0 condition, a typical auditory sink-source distribution is observed. For the Φ = π/2 and Φ = π conditions, especially for sensors overlying the posterior of quadrant, the sink- source distribution reflects the posterior auditory field topography, while for the remaining sensors the magnetic field distribution is not easily interpretable. Taken with the results of the statistical analyses, it is hypothesized that the changes in the response topographies and response power are due to activity the second harmonic frequency and reflect activity generated in the posterior temporal lobes (and/or auditory cortex) and possibly parietal lobes (Howard & Poeppel, 2010).

General Discussion

This set of experiments was designed to test neural entrainment to speech-like multisensory stimuli. In Experiment 3, we showed that the steady state response (reflecting neural entrainment) could be elicited to gradually changing radius modulated visual stimuli combined with amplitude modulated pure tones at two modulation frequencies that were not harmonically related. However, no effect of envelope phase offset was shown in Experiment 3, and because of the low number of trials it was unclear whether the non-difference was an accurate reflection of the neural response or if we were unable to capture a potentially small effect using the techniques that we used. However, peaks in the Fourier transformed averaged data at the modulation frequency were found, and this motivated a follow up experiment with increased trials and more speech-like stimuli.

Experiment 4 also showed significant SSR at the modulation frequency and at the second harmonic. As in Experiment 3, the overall power of the SSR to comodulated stimuli was greater than the power to unimodally modulated stimuli; but, once again, no difference was seen for the three envelope phase relationships that were tested. In conditions where only one modality was modulated, the response power was significantly lower than conditions where coherent modulations were present across modalities.

Although we did not use natural speech as a stimulus, we did use novel audio-visual pairings rather than combining natural auditory speech with an artificial visual signal.

In this sense, the findings reported here do not necessarily conflict with those found by Bernstein et al. (2004), because participants were presumably not familiar with signals presented in either modality (contrast this with speech, which is extremely familiar).

However, the lack of differences for envelope phase shifted conditions poses problems for the BCMP account, because the peaks in amplitude of one modality did not necessarily occur concurrently with the other modality. Although our stimuli were shifted well beyond the proposed window of integration for multisensory stimuli presented by van Wassenhove et al. (2007), we did not find significant differences between the completely in-phase stimuli relative to the shifted envelope stimuli.

It is possible that this effect was not shown because the stimuli shared the same onset and offset. In the traditional coherence masking paradigm, stimulus onsets were an important cue for auditory grouping (Gordon 1997b). Perhaps the presence of any cross-modality modulation in addition to synchronous onsets and offsets provided enough information to increase neural coherence. However, in the unimodal conditions, we did not find increased power despite the fact that onset and offset were aligned across modalities, but these stimuli lacked concurrent modulation.

Furthermore, there *is* a statistical relationship/correlation between the audio-visual stimuli, even at offset envelope phase relationships. Whether the envelopes were offset by 0 ms, 90ms or 180ms, the relative shift between envelopes was consistent

for the entire duration of a particular trial, and with the pseudo-speech stimuli that were used for these experiments it is possible that arbitrary associations were formed for the observer. This consistent relationship between envelopes may have been sufficient for providing an overall increase in power of the response relative to unimodally modulated signals of either modality, and this may be enough to boost the perception of multisensory stimuli for the observer. Varying the onset asynchronies between auditory and visual signals may be one way to establish the importance of aligned onsets in this type of paradigm. If similar results were found when the stimuli were misaligned temporally at the onset and offset, this would argue against the steady-state response as a potential indicator of mechanisms underlying audio-visual integration. On the other hand, if effects of temporally shifted onsets resulted in a decrease in the power of the SSR for out of phase conditions, the cross-modality envelope correlations that are purportedly driving the BCMP effects may be crucial at the level of neural entrainment as well.

Conclusion

This chapter showed that neural entrainment to audio-visual stimuli that share spectral and temporal properties of speech is increased when the modalities are concurrently modulated, regardless of the phase relationship between AM envelopes. The increase in SSR power at the modulation frequency and second harmonic is likely a result of increased phase-locked neural responses to the coherently modulated signals, relative to stimuli that were modulated in only one modality. It is possible that the increase in power shown here is a potential mechanism underlying the

behavioral finding that auditory detection thresholds are decreased when auditory speech is accompanied by a concurrently modulated visual speech envelopes. The increase in neural synchrony that the increased power of the SSR reflects may serve as a marker of concurrent modulation across modalities, which may be reflected in the improved detection thresholds for audiovisual relative to audio-alone stimuli. Further exploration of this paradigm is warranted to determine whether asynchronous onsets and offsets across modalities would diminish these effects.

Chapter 5: An investigation of lexical, phonetic, and word position influences on the McGurk effect

Introduction

Since the McGurk effect was first reported (McGurk and MacDonald, 1976), many variations on the basic finding have been explored, with more than 3000 PubMed citations as of the writing of this thesis. As such, it has become the default example for introductory linguistics and psychology courses to show that perception does not necessarily correspond physical properties of the stimulus, and that speech is 'more than meets the ear.' However, it is an almost impossible type to encounter in nature; with the exception of badly dubbed voiceovers in foreign language films, the likelihood of encountering incongruent audio-visual mismatches in actual perceptual environments is very slim. Although the phenomenon has been used widely as an example, the number of studies that have tested the limits of this effect are much more restricted.

Despite its ubiquity, almost all the studies that have actually utilized the McGurk-MacDonald paradigm have used nonsense syllables and/or limited phonetic contexts. An examination of the illusion in larger response sets, larger participant pool, and with stimuli that are real words can be informative for understanding this effect,

which has long been used as evidence for or against particular hypotheses about audiovisual integration.

If McGurk effects do extend beyond simple consonant + vowel nonsense syllables to actual lexical items, this would not only suggest that the illusion is not a simple 'party trick', but could also be utilized to test theories of lexical access and questions of lexical representation. For example, if words are stored as abstract, amodal representations in the lexicon, the presentation of a McGurk fusion stimulus [pick]+<kick>={tick} would be expected to facilitate a congruent target of /tick/ in a medium-lag repetition priming paradigm, despite the fact that {tick} was not physically present in the priming stimulus. Alternatively, if the congruent target /pick/ was primed in this scenario (assuming the reported percept was {tick}), it would suggest that the acoustic event may be stored in the lexicon instead.

However, before McGurk percepts can be used as a critical manipulation to ask this type of questions, it must be established that real word stimuli can elicit these illusions. The question of whether or not it is possible to obtain a McGurk effect with actual lexical items has received limited attention in the literature; the studies that have addressed this question have used inconsistent methods of stimulus creation, and have subsequently shown inconsistent results.

Easton and Basala (1982) tested the effects of discrepant audio-visual information on the perception of real words and found few visually influenced responses (99% of

responses corresponded to the auditory stimulus for participants who watched the screen and were instructed to listen to what was said). They concluded that visual speech has little or no influence on the perception of real words, in contrast with the McGurk and MacDonald (1976) and MacDonald and McGurk (1978) findings for nonsense syllables.

Although they do not provide their entire 30 pair stimulus list, the example stimuli that they do provide do not conform to McGurk-type parameters. For example, one of their pairs was auditory "mouth" + visual "teeth" ([maʊθ]+<tiθ>) – which differ not only in their initial consonant place of articulation, but also several features of the vowel which were likely visually available to the participants. Depending on the speaker, the /i/ vowel in "teeth" can have considerable lip spread, exposing the front teeth to the viewer; the vowel /aʊ/ in "mouth" starts with a wider jaw aperture and can have considerable rounding that obscures the teeth. It is possible that such a large feature mismatch between the visual signal (spread lips) and the auditory signal (rounded vowel) resulted in lower fusion effects simply because the auditory and visual events were too different to merge. In addition, it is unclear what the authors predicted the response to be, given that fusion responses for stimuli that have different vowels have been largely unsuccessful (Green & Gerdeman, 1995). In other cases, it seems that the visual and auditory signals would not have been distinct enough, such the pair [whirl] + <word> (the final consonant does not differ in visual place of articulation), and again it is unclear what the expected visually influenced response would have been.

The conclusion that the McGurk effect does not occur for real words was criticized by Dekle et al. (1992). Dekle et al. (1992) also used real words as auditory and visual input, and in contrast to Easton and Basala (1982), did find a high proportion of McGurk fusion responses, which suggested visual influences can be observed in real-word contexts. With fusion (or visually influenced) responses as high as 79%, they conclude that the lack of McGurk effects in the Easton and Basala (1982) could be attributed to the stimulus list that was used. However, the Dekle et al. (1992) stimuli were also nontraditional. They used a low number of stimuli (9 pairs, see Table 5.1) and the words that they did use were non-minimal pairs that were primarily bilabial audio + dental video (labio-dental or alveolar, e.g. [bent] + <vest> = {best}).

Audio	**Video**	**Expected McGurk**
bat	vet	vat
bet	vat	vet
bent	vest	vent
boat	vow	vote
might	die	night
mail	deal	nail
mat	dead	gnat
moo	goo	new
met	gal	net

Table 5.1 List of audio-visual stimuli used by Dekle et al. (1992) and their expected percepts

Although their results do suggest that visual information can influence the perception of multimodally presented lexical items, only one of nine stimulus pairs used by Dekle et al. (1992) was a traditional bilabial+velar minimal pair McGurk dub ([moo] + <goo>). One of the reasons that the McGurk effect is so remarkable is that the

perceived stimulus can correspond to *neither* of the physical inputs. When the responses that are considered McGurk are actually present in the stimulus, it is a less compelling example of the reported phenomenon because there is not abstraction away from the physical input.

Sams et al. (1998) examined traditional [bilabial] + <velar>McGurk stimuli in Finnish nonsense syllables, words, nonwords, and words/nonwords in sentential contexts. They reasoned that if audio and visual information is fused at a relatively late stage (after phonetic processing), the proportion of real-word responses should be greater than non-word responses, especially in constrained sentence contexts. Sams et al. (1998) report very robust McGurk responses (>90% of responses were visually influenced), but found no effect of lexical status (whether the expected fusion consonant formed a lexical item or a nonword), or sentential context. Because they did not find a higher proportion of word than nonword responses, they conclude that the audiovisual integration occurs at the phonetic level, before lexical access. This finding seemingly conflicts with the results of Brancazio (2004), which did show effects of lexical status on the response (discussed in Chapter 2), but because the stimuli were composed differently (Brancazio (2004) used [bilabial]+<alveolar> stimuli), and because the experimental task was dissimilar (Sams et al. (1998) use a larger stimulus set and a multi-part nonsense syllable, word, and sentence task), it is difficult to directly compare the results across these studies. Furthermore, Sams et al. (1998) embedded their auditory stimuli in noise, while Brancazio (2004) presented unaltered auditory stimuli, which may have affected the response.

Barutchu et al. (2008) manipulated the lexical status of the input stimuli that formed the McGurk stimuli, rather than the resulting percept (i.e., whether or not the bilabial auditory stimulus and velar video stimulus were real words) and the position of the audio-visual discrepancy. They found a lexical effect only for discrepancies in the word-final consonant, and interpreted this as evidence for higher-level word knowledge affecting phonetic processing. However, their stimulus list was relatively small (5 items per condition, see Table 5.2), and, crucially, what they consider to be the source of the "lexical" effects is the source stimuli, not the resulting percept.

		Audio	Video	(potential percept)
Real Words	**Onset**	bay	gay	day
		bill	gill	dill
		bet	get	debt
		bod	god	*dod*
		bun	gun	done
	Offset	lab	lag	lad
		rib	rig	rid
		dob	dog	*dod*
		rub	rug	*rud*
		hub	hug	*hud*
Nonwords	**Onset**	bip	gip	dip
		bez	gez	*dez*
		bov	gov	dove
		bup	gup	*dup*
	Offset	yab	yag	*yad*
		seb	seb	said
		vib	vig	vid
		pob	pog	pod
		zub	zug	*zud*

Table 5.2 Stimulus pairs used by Barutchu et al. (2008)

It is an interesting observation that nonword stimuli can create McGurk-type percepts, but the conclusion that this is a reflection of lexical-phonetic interactions is tenuous. Consider their real word-offset condition, which was the source of the "lexical" effects. A real-word percept would be expected (for traditional McGurk illusions) in only the first two of these five stimulus pairs. In contrast, four of five real word word-onset stimulus items would be expected to form real-word McGurk percepts. Furthermore, it is possible that lexical properties of the source stimuli and the illusory percepts could bias responses, but it appears that they do not consider this factor. For example, the stimulus [bod]+<god>, which has a high frequency visual stimulus and a low frequency auditory stimulus, was, perhaps unsurprisingly, overwhelmingly perceived as {god} or {odd}. Although Barutchu et al. (2008) do show that both words *and* nonwords with discrepancies in initial *or* final position can elicit McGurk effects, other conclusions should be made with caution.

In this chapter, I aim to test the McGurk effect using minimal pairs that are both lexical items, similar to the solitary Finnish McGurk stimuli used by Sams et al. (1998). The expected McGurk percept could result in an actual lexical item, or could result in a nonsense word. The consonants occurred in a variety of phonological contexts, and could occur word initially or word finally. In this way, an estimation of the traditional McGurk effect—where the expected percept does not correspond to the physical content of either of the input modalities—can be extended to real word sources and real and non-word illusions.

Experiment 5

Materials and Methods[26]

The stimulus list for this experiment was generated by searching the COBUILD corpus for English word pairs that differed in place of articulation. The bilabial consonants [b p] were candidates for auditory stimuli, and velar consonants <g k> were candidates for visual stimuli, in either word initial or word final position. In line with traditional McGurk paradigm, dubbing of each of these stimulus pairs was expected to produce a third, non-physically present percept containing the fusion consonants {d t}. In some cases, the McGurk percept formed an actual lexical item, for example [pick] + <kick> = {tick} (see Appendix II for complete stimulus list); however, a proportion of the stimulus list contained dubbings that would result in McGurk percept nonwords, such as [best] + <guest> = *{dest}. The voicing of the all pairs was matched across modalities (i.e., only [b] + <g> and [p] + <k> pairs were included). Paired t-tests on log normalized orthographic frequencies from COBUILD showed no significant difference of word frequency for the physical stimulus pairs (i.e., the bilabial and velar source words). Orthographic frequency for McGurk percepts was not included in this comparison because of the large number of nonword percepts in the stimulus list; however, evaluation of frequencies of the real-word McGurk (alveolar) potential percepts were obtained and used as exclusion criteria for stimulus list creation (e.g., the pair [bag]+<gag> resulted in the low frequency word {dag} and was excluded from the stimulus list).

[26] Example stimuli can be obtained at http://files.ling.umd.edu/~arhone/Thesis/Ch5_stimuli/ or by email request: ariane.rhone@gmail.com

Video and audio were concurrently recorded using a Canon DM- XL1 video camera onto digital videotape (mini DV; frame rate 29.97 frames/second). An adult female native speaker of American English was recorded while seated in front of a solid dark black background. The stimulus list[27] was randomized and presented on a screen behind the camera. The talker was instructed to start and end articulations from a neutral mouth position, and to minimize blinks and head and eye movements during recording. The list was repeated three times.

In addition to audio recorded from the camera microphone, high quality audio was recorded using an external microphone positioned approximately 15 inches from the talker's mouth (but out of the field of view of the camera) to a memory card. Digital video and audio from the DV tape were imported to a Dell Inspiron running Windows XP and segmented using VirtualDub for further processing. External microphone audio files were also imported and segmented in Praat.

The video tokens were converted to gray scale and a fade in/out filter was applied (5 frames each) in VirtualDub. At least 5 frames of a neutral face before articulatory onset was included; if 5 frames were not available, the video file was padded using still frames from the prearticulatory period. The audio track for each video token was extracted to use as a reference for dubbing (see below).

[27] The stimulus list presented here was a subset of materials recorded in this session. Approximately 250 word and nonword fillers for a related experiment were intermixed with the McGurk eliciting stimuli reported here.

Prior to dubbing, the auditory tokens of all stimuli were tested in a pilot experiment (n = 5, all adult native speakers of English who received course credit) to determine whether any auditory stimuli were ambiguous. Overall performance on the audio-alone stimuli was greater than 99%, and no consistent patterns of errors were shown for any stimulus item[28] so no auditory items were excluded on this basis.

Auditory stimuli that contained noise and video stimuli that contained excessive head movements or non-neutral starting/ending positions were excluded from dubbing. A number of auditory items were excluded for idiosyncratic differences between the word pairs, or for generally not conforming to predicted pronunciations (e.g., the word "pool" was pronounced with two syllables on two of the three repetitions, but the velar "cool" was consistently monosyllabic, leading to discrepant audio-visual offsets, despite the fact that the audio for "pool" was perceived correctly in the pilot experiment).

Dubbing: The timing of the stop burst was determined, via visual inspection of the waveform, from the audio extracted from each video token, and segmented audio from the external microphone was edited to match this timing (pre-stimulus samples were added or removed, as necessary). Edited audio files were normalized in Praat to average RMS of 70 dB SPL, and 10ms $\cos^2$ onset and offset ramps were applied to each file.

[28] No confusions in the crucial consonants were reported, with the exception of a small number of typographic errors (e.g. "[in" for "pin")

In VirtualDub, the normed, ramped, and edited microphone audio was dubbed onto a target video file. Both congruent and incongruent dubs were created so that the resulting dubs were congruent bilabial (e.g., [pick] + <pick>), congruent velar (e.g. [kick]+<kick>), McGurk-type (e.g., [pick]+<kick>), or congruent alveolar ([tick]+<tick>). All video and audio files that met inclusion criteria described above were combined from all repetitions (e.g. each repetition of [pick] was dubbed to each repetition of <kick>, each video of <tick> was dubbed to each audio of [tick], etc.). A total of 1100 dubbed videos were created in AVI format. All McGurk dubs (374) and a subset of the congruent dental (22), bilabial (7), and velar (9) dubs were included in the stimulus list for this experiment[29].

Resulting Stimuli: Because a full crossing of repetitions was performed in the dubbing process, some types were represented by more than one token (for McGurk stimuli only). This resulted in an unequal number of tokens per type, but allowed for a better understanding of the factors influencing illusion.

Of the McGurk-type dubs that were presented to participants, 71.9% occurred in word initial position (e.g. [pick]-<kick>), and 21.8 occurred in final position (e.g. [lip]-<lick>); 66.8% of the items were expected to result in real word percepts ([pick]+<kick>={tick}) and 33.2% were expected to result in nonword percepts

[29] The original purpose of these stimuli was to test lexical access in a medium-lag repetition priming design, as described in the introduction. The stimuli reported here were tested to determine which stimulus pairings resulted in the strongest McGurk effect in an effort to select the most compelling stimuli for the priming study.

([best]+<guest>={dest}). The target consonants occurred in a variety of phonological environments. Of note, 15.2% of items had the crucial consonant adjacent to an /r/ and formed a consonant cluster (e.g., [brain]+<grain> = {drain}).

All stimuli were presented on a Dell OptiPlex 320 running Windows XP with a SoundMax Integrated Digital HD Audio sound card and an ATI Radeon Xpress 1100 video card. Video was presented on a standard computer CRT monitor at a distance of approximately 18 inches; auditory stimuli was delivered via Sennheiser HD 580 Precision over-ear headphones at a comfortable listening level determined by the participant.

Task: Stimuli were randomized and responses obtained using Alvin experimental control software (Hillenbrand & Gayvert, 2005). Participants were naïve to the presence of mismatched audio-visual pairings, and were debriefed after participation regarding the nature of the stimuli. Participants were told that they would be shown short video clips of a person saying real and nonsense words, and were instructed to report what was said by typing their response into a text field. An open set response, rather than forced choice, was utilized to minimize emphasis on the dubbings and to gather information about responses that might not conform to the bilabial-alveolar-velar responses that were expected. Participants were instructed to guess if unsure, and were told to make up spellings for nonsense words (they were given an auditory example during the instructions: “If the person said /gut/, you might spell it as “goot” or “gute”).

Participants were also told that on a small number of trials there would be a small dot on or near the mouth of the talker, and were instructed to identify these catch items by typing the color of the dot after their response. Fifty-two catch trial items were created by superimposing a black or white dot with a 12-pixel radius around the speaker's lip and mouth area using GIMP. Catch trial items were selected from all types of stimuli (congruent bilabial, congruent dental, congruent velar, and McGurk-type were all included in distractor items). The dot could appear at any point during the video stimulus and lasted between 4-6 frames (133-200 ms). Placement of the dot varied for each stimulus item. To encourage attention to the entire video, they were told that the dot could occur at any time while a token was on the screen.

Participants could repeat a stimulus item up to one time, but were encouraged not to use that option unless they missed a token because of computer error or external distraction (<0.01% of all trials were repeated). The testing session lasted approximately 45 minutes.

Participants: Thirty-one adult participants were recruited from the University of Maryland College Park community and received course credit in an introductory linguistics course for their participation. Presentation of stimuli and response collection was performed with the approval of the institutional committee on human research of the University of Maryland, College Park. Prior to the start of the experiment, written informed consent was obtained from each participant. All

participants had normal or corrected-to-normal vision and hearing (self reported). Six participants were native speakers of a language other than American English.

Exclusion criteria: Because attention to the visual stream is crucial for the McGurk illusion, participants were excluded if they failed to detect more than 1/3 of the catch trials described above. Nineteen of 25 native English speakers and five of seven nonnative speakers met inclusion criteria and are included in the following analysis.

Results

Congruent audio-visual dubs:

For congruent audio-visual stimuli, participants were highly accurate (> 99%) at identifying the stimulus item. For nonword stimuli, spellings showed some variation in orthographic representation of the vowel (e.g., for the congruent alveolar A-V stimulus /draɪd/ the responses “drade”, “draid”, and “drayed” were all reported responses), but target consonant orthography was consistent. Table 5.3 shows all reported errors on the congruent audiovisual stimuli.

	stimulus	incorrect responses
congruent bilabial	flab	flagb, slap
congruent alveolar	quit tatch trit	quite patch trip, trir
congruent velar	shock	sock

Table 5.3 All incorrect responses to congruently dubbed bilabial, alveolar, and velar tokens

Incongruent audio-visual (McGurk) dubs:

Alveolar responses, which were expected to make up the greatest proportion given the McGurk dubbings that were used, comprised only 3.93% of responses overall. The bilabial response category (corresponding to the identity of the auditory stimulus) was the most common response for all stimuli (90.62% overall), with velar responses (corresponding to the identity of the visual stimulus) occurring on 4.17% of trials. Non-prototypical fusions comprised 1.08% of responses, and included labiodentals, interdentals, and [h]. Responses that could not be coded for a particular response category (e.g. [par] + <car> = {;ar}) comprised < 1% of responses.

Although some variability was expected with the large number of tokens and types that were presented, the extremely high proportion of responses corresponding to the auditory stimulus was surprising.

Seventeen of the 19 native English speakers showed at least one expected McGurk fusion, but a large proportion of the non-bilabial (non-auditory) responses were

obtained from a small number of participants (see Table 5.4 and Figure 5.1). For example, participant E1 contributed 68 of the 243 total McGurk responses (27.98%), and participant E16 contributed 149 of the 346 velar responses (43.06%).

SubID	bilabial	alveolar	velar	other	unknown	total count
E1	299	68	5	1	1	374
E2	355	2	17	0	0	374
E3	374	0	0	0	0	374
E4	353	4	16	0	1	374
E5	368	1	2	1	2	374
E6	365	4	5	0	0	374
E7	371	2	0	1	0	374
E8	363	4	7	0	0	374
E9	254	40	60	16	4	374
E10	356	8	9	1	0	374
E11	345	9	7	12	1	374
E12	369	0	4	0	1	374
E13	363	4	3	4	0	374
E14	331	16	21	5	1	374
E15	356	16	1	0	1	374
E16	197	10	149	18	0	374
E17	358	4	7	4	1	374
E18	337	28	4	5	0	374
E19	313	23	29	6	3	374
Total%	**90.44%**	**3.42%**	**4.87%**	**1.04%**	**0.23%**	

Table 5.4 Counts for each response category by participant (English speakers)
The "bilabial" category corresponds to the identity of the auditory stimulus, "alveolar" corresponds to McGurk-type percepts, and "velar" corresponds to the identity of the visual stimulus. The "other" category contains alternate fusion responses including {f v ð θ h}, and "unknown" contains responses that could not be categorized (e.g., non-alphabetic response entries such as ";").

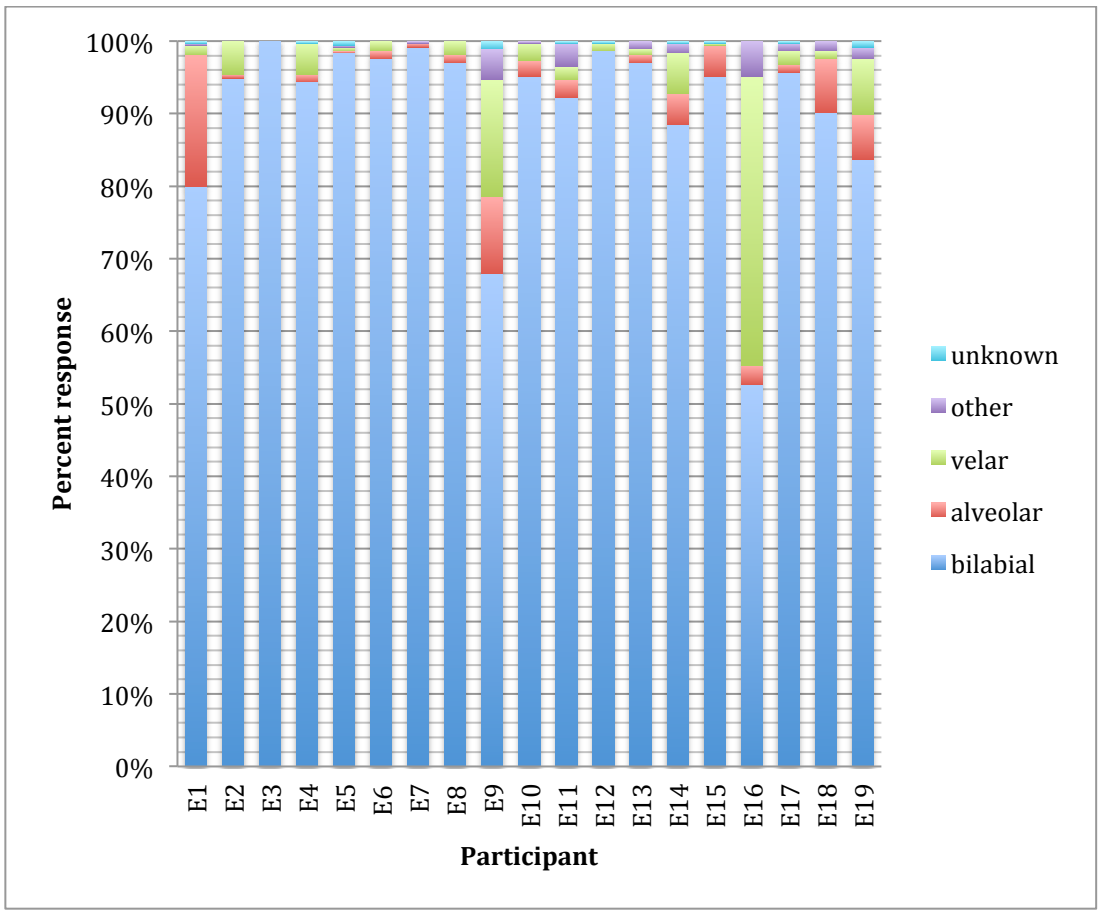


Figure 5.1 Proportion of responses categories reported for each participant (native English speakers)

Non-native English speakers showed slightly higher alveolar (McGurk), and lower velar (visual input) responses (see Table 5.5, Figure 5.2), but the group percentages should be interpreted with care, because the five non-native speakers also showed highly variable response patterns just as the native English speakers (e.g., NNS5 contributed 60% of total McGurk responses). Furthermore participants were native speakers of four different languages (2 Spanish, 1 French, 1 Romanian, and 1

Japanese) and it is possible that individual perceptual biases, rather than native language of the speaker, are responsible for this difference.

SubID	bilabial	alveolar	velar	other	unknown	total count
NNS1	359	7	6	2	0	374
NNS2	360	5	7	2	0	374
NNS3	345	17	2	9	1	374
NNS4	356	15	1	1	1	374
NNS5	287	66	12	9	0	374
Total%	**91.28%**	**5.88%**	**1.50%**	**1.23%**	**0.11%**	

Table 5.5 Counts for each response category by participant (non-native English speakers) Category "bilabial" = auditory stimulus, "alveolar" = McGurk percept, "velar" = visual stimulus, "other" = alternate fusions, "unknown" = not categorizable (see Table 5.4).

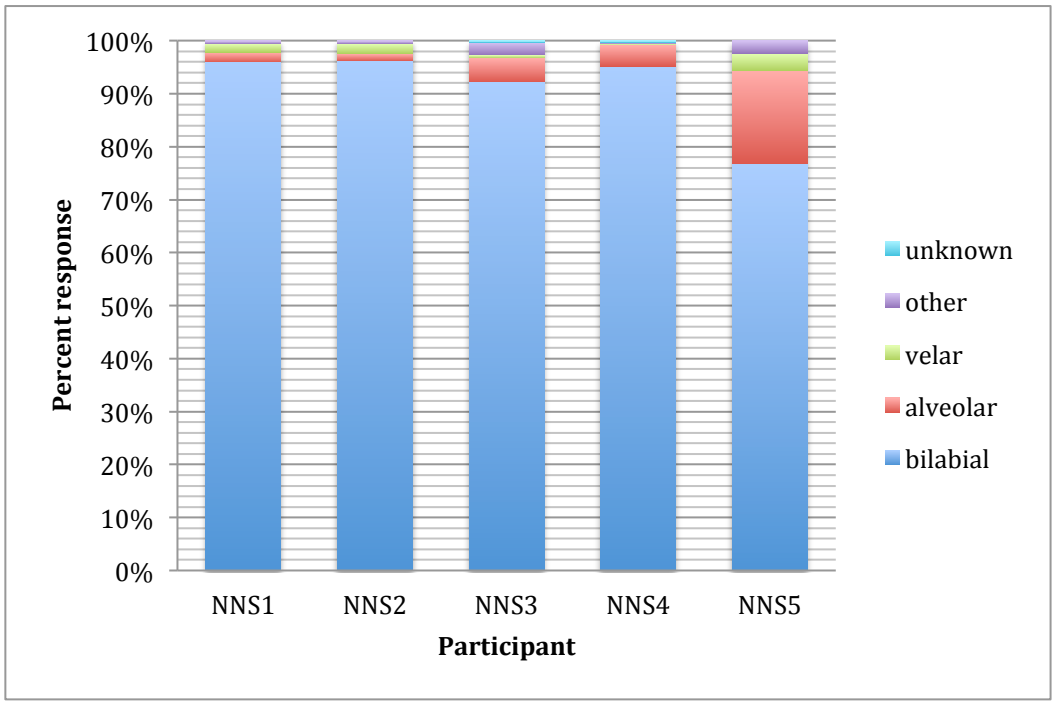


Figure 5.2 Proportion of responses categories reported for each participant (non-native speakers)

The observed patterns for known stimulus parameters are described below. Although I present the proportions separately for each parameter for descriptive purposes, it is important to consider that each of these properties was not manipulated independently. Proportions are reported here because the stimulus list did not contain equal numbers of items per type.

Word position:

English Speakers	bilabial	alveolar	velar	other	unknown
Initial	92.27%	3.80%	2.50%	1.35%	0.08%
Final	85.76%	2.46%	10.93%	0.25%	0.60%

Non-Native Speakers	bilabial	alveolar	velar	other	unknown
Initial	90.26%	7.06%	1.34%	1.26%	0.07%
Final	93.90%	2.86%	1.90%	1.14%	0.19%

Table 5.6 Percentage of response categories perceived, by position of critical consonant

Native English speakers show increased velar category responses in word-final, relative to word-initial, position. Alveolar percepts showed similar proportions in both positions. Non-native speakers did not show large differences in velar proportion as a function of word position, but fusion (alveolar) responses were considerably higher in initial position than in final position.

Voicing

English Speakers	bilabial	alveolar	velar	other	unknown
Voiced	89.69%	4.87%	2.89%	2.46%	0.09%
Voiceless	90.81%	2.72%	5.83%	0.36%	0.29%

Non-Native Speakers	bilabial	alveolar	velar	other	unknown
Voiced	88.20%	8.03%	0.33%	3.44%	0.00%
Voiceless	92.78%	4.84%	2.06%	0.16%	0.16%

Table 5.7 Percentage of response categories perceived, by critical consonant voicing

Overall, both native and nonnative English speakers showed high bilabial proportions, with slightly higher non-auditory responses for voiced stimuli relative to voiced stimuli. Alveolar (McGurk) percepts were greater in the voiced condition, while velar (video) responses were greater for voiceless consonants for both language groups.

Lexical status

English Speakers	bilabial	alveolar	velar	other	unknown
McGurk Word	89.20%	4.65%	4.61%	1.31%	0.23%
McGurk Nonword	92.95%	0.93%	5.39%	0.51%	0.21%

Non-Native Speakers	bilabial	alveolar	velar	other	unknown
McGurk Word	88.88%	8.48%	0.96%	1.52%	0.16%
McGurk Nonword	96.13%	0.65%	2.58%	0.65%	0.00%

Table 5.8 Percentage of response categories perceived, by lexical status of the expected McGurk (alveolar) percept

The occurrence of alveolar responses for audio-visual pairings that were expected to elicit nonwords was reduced relative to pairings that were expected to elicit real-word illusions, for both native speakers of English and non-native speakers (see Table 5.5). Examination of all responses showed that the responses to nonword filler items (true alveolars such as [tave]+<tave>={tave}) were consistent with participants following the instructions to make up spellings for nonword items. In addition, a small number

of nonwords were also perceived to real-word audio, video, and McGurk responses. For example, for the incongruent stimulus [boast]+<ghost>, the response {thost} (nonword) was recorded three times, when {dosed}—a real word—was the expected alveolar fusion response.

Cluster status:

English Speakers	bilabial	alveolar	velar	other	unknown
Cluster	97.69%	0.00%	1.94%	0.18%	0.18%
No Cluster	89.14%	4.03%	5.40%	1.20%	0.23%

Non-Native Speakers	bilabial	alveolar	velar	other	unknown
Cluster	99.30%	0.00%	0.35%	0.35%	0.00%
No Cluster	89.84%	6.94%	1.70%	1.39%	0.13%

Table 5.9 Percentage of response categories perceived, by cluster status of critical consonant

Few non-auditory responses were observed for dubbed consonants in an /r/ cluster environment. Of note, zero alveolar (McGurk) percepts were reported for source stimuli that contained clusters. The small number of "other" responses that were reported in the Cluster condition were all the labiodental {f}.

Although no alveolar responses were found for this stimulus set, the question of whether or not conflicting audio-visual consonant clusters are simply not combinable cannot be determined. For this particular talker, the visual < kr > and < gr > had considerable rounding, although the lips did not make full contact (see Figure 5.3 for

example frames from a <kr> cluster). It is possible that visual stimuli from a different talker could have resulted in fusion responses.

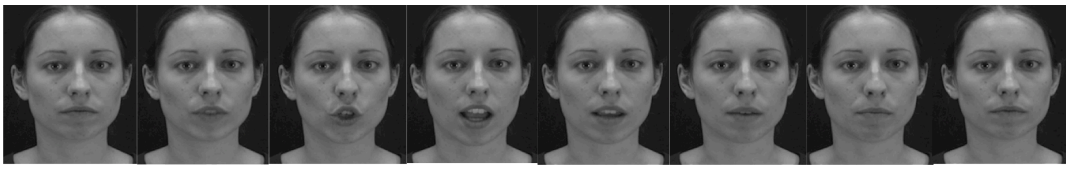


Figure 5.3 Selected frames from the visual stimulus <crime>
Note the lip rounding in the third frame presented. This video stimulus, dubbed with the bilabial auditory stimulus [prime], resulted in 47 bilabial responses out of 48 presentations across participants (1 velar response).

Response patterns by item

Items that elicited McGurk percepts on at least 10% of trials across all participants are listed in Table 5.10 (see Appendix II for full item list). The stimulus [big]+<gig> elicited the highest percentage of McGurk responses ({dig}: 21.35%). As previously discussed, no fusion responses were shown for any item with the critical consonant adjacent to an /r/ (in a consonant cluster).

Audio	Video	Expected Percept	# of tokens	Percent McGurk	Percent Velar	Percent Other
big	gig	dig	x8	21.35%	0.52%	0.52%
beer	gear	dear	x7	20.83%	0%	1.79%
pear	care	tear	x5	18.33%	0%	3.33%
bash	gash	dash	x7	15.48%	0%	4.76%
rib	rig	rid	x3	13.89%	2.78%	1.39%
pick	kick	tick	x10	13.33%	3.75%	0%
bait	gate	date	x8	13.02%	0%	5.73%
pan	can	tan	x4	11.46%	4.17%	0%

Table 5.10 Stimulus types showing highest percentage of alveolar (McGurk) responses.

Audio	Video	Expected Percept	# of tokens	Percent McGurk	Percent Velar	Percent Other
pearl	curl	*turl*	x6	0%	27.78%	2.08%
cheap	cheek	cheat	x4	3.13%	21.88%	0%
lab	lag	lad	x4	2.08%	20.83%	4.17%
robe	rogue	rode	x3	0%	18.06%	0%
lap	lack	*lat*	x7	0%	16.67%	0%
shop	shock	shot	x7	4.17%	15.48%	0%
poach	coach	*toach*	x4	1.04%	11.46%	0%
ape	ache	ate	x3	1.39%	11.11%	0%
nip	nick	knit	x3	0%	11.11%	0%
reap	reek	*reet*	x3	0%	11.11%	0%
snap	snack	*snat*	x2	0%	10.42%	0%

Table 5.11 Stimulus types showing highest percentage of velar (video) responses

Stimulus pairs that resulted in >10% velar responses (corresponding to the video input signal) are reported in Table 5.11. Most of the top velar percepts are word final, but word initial [pearl]+<curl> had the highest proportion of velar responses overall, and zero McGurk fusions ({hurl} was the most common "other" response for this item).

Alternate fusion responses were reported for 25 stimulus pair types, and are listed in Table 5.12. The pair [bye]+<dye> resulted in 12.5% of non-standard fusion responses, with {thy} as the most common reported percept. Other common fusions were [bun]+<gun>={*thun*}, [bet]+<get> = {vet}, [bash]+<gash>={*thash*}, [pear]+<care>={hair}, and [sip]+<sick> = {*sith*}.

Audio	Video	Expected Percept	# of tokens	Percent McGurk	Percent Velar	Percent Other
bye	guy	dye	x6	2.78%	0%	12.50%
bait	gate	date	x8	13.02%	0%	5.73%
bun	gun	done	x4	1.04%	0%	5.21%
bash	gash	dash	x7	15.48%	0%	4.76%
bet	get	debt	x8	0.52%	0%	4.69%
sip	sick	sit	x3	5.56%	5.56%	4.17%
boast	ghost	dosed	x3	4.17%	2.78%	4.17%
lab	lag	lad	x4	2.08%	20.83%	4.17%
barter	garter	*darter*	x3	0%	2.78%	4.17%
pear	care	tear	x5	18.33%	0%	3.33%
flab	flag	*flad*	x3	2.78%	8.33%	2.78%
brace	grace	*drace*	x5	0%	0%	2.50%
pearl	curl	*turl*	x6	0%	27.78%	2.08%
puddle	cuddle	*tuddle*	x4	0%	3.13%	2.08%
boat	goat	dote	x4	0%	2.08%	2.08%
pill	kill	till	x4	0%	1.04%	2.08%
pall	call	tall	x2	0%	0%	2.08%
beer	gear	dear	x7	20.83%	0%	1.79%
rib	rig	rid	x3	13.89%	2.78%	1.39%
pub	cub	tub	x3	4.17%	1.39%	1.39%
bust	gust	dust	x3	4.17%	0%	1.39%
tab	tag	tad	x4	6.25%	5.21%	1.04%
post	coast	toast	x9	4.63%	4.17%	0.93%
pamper	camper	tamper	x6	4.86%	0%	0.69%
big	gig	dig	x8	21.35%	0.52%	0.52%

Table 5.12 All stimuli eliciting "other" responses

Statistical analysis of response patterns for known stimulus parameters were performed (using GLM function in SPSS), with fixed factors Voicing, Word Position, Lexical Status, and Cluster Status, with Participant as a random effect. For native speakers of English, significant effects of Cluster Status ($F = 23.041$, $p < 0.001$), and a marginal effect of Lexical Status ($F = 4.135$, $p = 0.057$) were found. Significant interactions was found for Lexical Status * Cluster Status ($F = 5.677$, $p=0.028$) and a marginal interaction for Voicing * Lexical Status ($F=3.851$, $p = 0.065$). No other significant effects or interactions were shown.

Discussion

This experiment examined the McGurk effect (McGurk & MacDonald, 1976; MacDonald and McGurk, 1978) using real words differing only in place of articulation of the critical consonant in word initial or word final position (bilabial in the auditory modality, and velar in the visual modality). Overwhelmingly, the participants in this study reported that they perceived the auditory stimulus, rather than a fused McGurk percept or the physical velar stimulus provided by the video.

Although we did not find a high proportion of typical McGurk-type responses in this experiment, we did find several interesting patterns of responses related to audio-visual combination of real English words. First, the response profile of individual participants was highly variable, with some participants showing exclusively auditory responses and others more likely to report visual or fusion responses. There was considerable variability across participants in the sample reported here. Some individuals were highly inclined to perceive the auditory stimulus (despite performing well on the visual catch trials), while others were more likely to report fusion percepts, and others commonly reported the visual stimulus. Although responses to McGurk-type stimuli as a function of individual variability have been examined (J.-L. Schwartz, 2010), the response bias could not be measured with the particular experimental design that was reported here. A large variation in number of repetitions per type and an imbalanced stimulus list (with respect to position, voicing, cluster status, and lexical status) may have limited the statistical analyses that can be performed on the response patterns, but the overwhelming proportion of responses

that corresponded to the auditory stimulus does suggest that the McGurk effect may not be as robust of a phenomenon as has been previously described (at least for these stimuli). However, without having separate evaluations of these participants' performance on the typical CV McGurk battery, it is difficult to draw conclusions about how they would have performed on a more canonical task.

In addition, the presence of cluster resulted in fusion responses to dubbed [bilabial]+<velar> stimuli. It is likely that the significant lip rounding in the visual articulation for this talker's /r/ clusters was compatible with the bilabial auditory signal [b] or [p] that they were dubbed to. However, true alveolar distractor items (e.g., [draze]+<draze>={draze}, [trit]+<trit>={trit}) were never categorized as bilabial by any participants (and also contained considerable /r/ rounding), which suggests that this is constrained to instances of incongruent audio-visual stimuli. Even participants who showed fusion responses on a relatively high proportion of trials did not fuse these items. Alternatively, participants may not have shown fusion percepts for stimuli containing word-initial clusters on the basis of phonological expectations. In many dialects of English, /tr/ and /dr/ clusters in syllable-initial position become affricated (e.g., the initial alveolar stop /t/ in the word "tree" is often pronounced as a post-alveolar affricate [tʃ]). The bilabial auditory component of the dubbed items does not contain the acoustic correlates of affrication, which may have violated participants' expectations about what the /tr/ and /dr/ clusters should sound like. The observation that some cluster stimuli did result in velar percepts offers some support for this explanation, because a velar clusters are not typically affricated.

The complete lack of fusion for dubs containing /r/ clusters could also be a talker-specific result, and could be investigated further by testing /r/ clusters spoken by a different talker, in more controlled stimulus types (e.g., all nonsense syllables such as [pra]+<kra>={tra}).

The effect of lexical status on the response categories for the McGurk-type stimuli presented here offers additional support for the findings of Brancazio (2004), in that dubs that formed actual lexical items when combined were more frequently fused than dubs that formed nonwords. However, unlike the materials used by Brancazio (2004), all the physical stimuli (both acoustic and optical signals) were actual lexical items, and the lexical status of the potential fusion percept was manipulated.

Sams et al. (1998) did not find lexical effects for the Finnish stimuli that they presented, which were similar in structure to the materials presented here but used a smaller number of items for comparison. There is some difficulty in interpretation of this result and the results of Sams et al. (1998), because the proportion of words vs. nonwords that will actually be perceived within the experiment is difficult to determine. Incongruously dubbed stimuli may or may not be perceived as expected for each stimulus token and for each individual, which can result in large variability across the participant sample and within the experiment itself. In this study, our stimulus list contained 33% of McGurk items that were expected to result in nonword percepts (if they were perceived as alveolar). However, the reported percepts were

predominantly bilabial (for all stimuli and all known stimulus parameters), and alveolar responses to this category comprised less than 1%.

Unlike Barutchu et al., (2008) we did not find effects of word position on the response category for word compared with nonword items (but see introduction above regarding their definition of a "lexical" effect). However, only one of the ten most frequently fused stimulus pairs occurred in word-final position. It is important to note that the composition of the stimulus list was considerably different in this study relative to other McGurk-type experiments that have tested lexical effects. Barutchu et al. (2008) manipulated the lexical category of the stimulus items, rather than the expected response. Furthermore, they used a limited stimulus set that may have differed on other potentially relevant parameters (e.g., word frequency effects), and did not offer detailed breakdown of the responses that were observed.

As discussed in the introduction to this chapter, it is important to consider whether the McGurk effect can be extended to real word stimuli before these stimuli can be used in interesting ways to address larger issues in psycholinguistics, such as episodic vs. abstract storage of words in the lexicon. Although the stimuli presented in this experiment were designed to address questions of lexical representation, the reported percepts from this set of items and this talker did not result in robust illusory responses, which precludes further direct use of this stimulus set for higher-level studies. However, the question of why these stimuli did not result in consistent McGurk percepts (at least for the participant sample tested here) is still open.

Conclusion

Although the McGurk effect has been widely cited and often used as a tool for exploring audio-visual interactions and integration in speech perception, the structure of the stimulus types that have been used has been limited. This experiment used real English words as the physical input in each modality, and explored the patterns for dubs that differed in voicing, word position, lexical status of the expected percept, and syllable structure (specifically, the presence or absence of a consonant cluster).

Over 90% of the responses reported corresponded to the auditory stimulus identity. We found a low proportion of fusion responses, and a low proportion of responses corresponding to the visual input. Despite the failure to elicit robust McGurk-type effects, we did find several differences that could provide information about which stimulus types are more likely to be fused. Audio-visual dubbings that resulted in a real-word percept were more likely to result in fusion response than dubbings that resulted in nonwords. Also, the environment that the critical consonant is in was shown to affect the proportion of fusion responses. Specifically, we found that when the critical consonant occurred in an /r/ cluster, no fusion responses were reported. A diverse set of audiovisual stimuli—approximately matched in word frequency and all minimal pair lexical items—were used in an effort to understand factors that may influence the audio-visual integration of real words, but we ultimately failed to elicit robust McGurk effects. However, the difficulty in obtaining consistent fusion responses for this stimulus set does not necessarily mean the McGurk effect is not an

interesting perceptual illusion; instead, further investigation could help clarify why the effect was so difficult to produce with these stimuli. Expanding on this study using different talkers and participants, as well as establishing baseline McGurk effects for each participant by testing responses to canonical CV nonsense syllables, is crucial for establishing whether or not real-word stimuli can elicit robust illusory percepts.

Chapter 6: General Discussion

There has recently been an increase in interest in exploring the influence of visual speech information on auditory speech perception. Along with the observation that visual information affects auditory perception, the question of where, when, and how multisensory interactions occur in the human brain has also recently gained attention. In particular, researchers have begun to explore the behavioral and neurophysiological consequences of multisensory perception (see Ghazanfar and Schroeder (2006) for a review). By now, effects of visual information on speech perception have been shown to occur at various stages of the processing stream. For example, at a pre-categorical level, thresholds for detecting and audio-visual stimulus are improved relative to thresholds for auditory alone stimuli. At the level of phonetic processing, speech syllables are identified faster (and more accurately) with the presence of visual speech information. Neurophysiologically, cortical networks involved in cross-sensory binding have been proposed, and the notion of "unimodal" cortices is falling out of favor. Many studies have focused on determining which brain areas are responsible for multisensory binding, and the discovery that cortical areas once thought to be dedicated to auditory perception are also implicated in multisensory processing has paved the way for further exploration into the mechanisms responsible for this effect. Additional studies examining the time course of integration for ecologically valid multisensory stimuli (such as speech) have

informed models of speech perception by suggesting that visual predictive information can facilitate auditory processing, possibly by way of preliminary feature analysis, at the level of responses generated in auditory cortex.

The studies presented in this thesis offer further support for the influence of visual information on auditory speech perception, from potential neurophysiological mechanisms for tracking envelope relationships across modalities (Chapter 4), to understanding more about the nature of predictive information at the visual-phonetic level (Chapter 3).

Chapter 3 showed that facilitation effects for audio-visual speech relative to audio-alone speech can be attributed to the relative predictive strength about an upcoming auditory event, rather than a general facilitatory effect based on the physical salience of the input. This offers both support for and clarification of the audiovisual speech perception model of van Wassenhove et al. (2005), where predictive strength modulates the degree of facilitation.

I tested this by showing that responses to the same stimulus can differ as a function of the other members of the response set. When bilabial anticipatory motion no longer uniquely predicted the /ba/ syllable type, the M100 latency facilitation effects for the syllable /ba/ were no longer seen. In this situation, the non-labial syllable type /da/ was the only response candidate that was predictable by non-labial anticipatory movements, and so it was facilitated. This finding suggests that the previously

observed “articulator specific” facilitation is not, in fact, articulator specific at all. When the upcoming auditory stimulus is highly predictable (regardless of which articulators are involved—bilabial or not), auditory evoked responses are facilitated. The potential for pre-auditory onset feature analysis based on visual predictive information is one more demonstration of the flexibility of the human brain. Furthermore, reaction times to these stimuli also varied by response set, indicating that the behavioral facilitation for audiovisual identification is also a flexible process. An effect of response set was shown for the two syllable types that were present in both conditions; however, real-world audio-visual speech perception takes place outside of a well-defined response set, but the benefit of seeing a talker relative to hearing alone still exists. It is likely that a combination of contextual information, visual predictive information, and general knowledge contribute to this benefit, and the cues available for providing this effect should continue to be explored.

Other behavioral advantages, such as improvement in audio-visual detection relative to auditory-alone detection, have also been previously shown. Bimodal coherence masking protection (BCMP) is one suggested as a mechanism underlying the audio-visual detection advantage, because stimuli that are correlated across modalities seem to show the greatest detection improvement. The question of how this envelope cross correlation may be implemented in the brain was the focus of Chapter 4. Taking advantage of a neural entrainment paradigm that has been used extensively in evaluations of unimodal sensory processing, we showed that entrainment to multisensory stimuli was possible, and that steady state responses were enhanced at

frequencies of interest for speech perception for the multimodal relative to unimodally modulated stimuli. Contrary to our hypothesis, we did not find significant effects of envelope phase shifts on the power of the steady state response.

However, it is possible that the simultaneous onsets and offsets for the multisensory modulated stimuli reduced the perception of asynchrony, since the correlation between envelopes—even when shifted—stayed constant throughout a given trial. Because these stimuli were abstractions of speech and not likely interpreted as lips and a voice, it is possible that these signals were represented as novel multisensory objects that contained an intrinsic lag. Whether onset asynchrony would disrupt the pattern of responses reported here is an area of future investigation.

And, although a multimodal SSR was elicited—and showed differences relative to unimodal SSRs—the use of nonspeech stimuli limits the extension of these findings to real-world audio-visual speech perception. Future studies should explore whether modulation differences across modalities for natural speech stimuli shows similar effects to those reported here. Although the audio-visual pairings that were used were not natural speech tokens, the stimuli in Experiment 4 did share some critical attributes of the audio-visual speech signal, and were modulated at an approximate speech envelope rate. We hope that this paradigm could be explored further with natural speech stimuli, and hypothesize that correlation across modalities does have measurable neural consequences.

In Chapter 5, my goal was to test a large number of McGurk-type stimulus pairs that were all lexical items and that differed on several parameters (such as voicing, word position, etc.) to test which stimuli would elicit strong McGurk percepts. Disappointingly, an overwhelming number of these stimuli were not visually influenced (at least at the level that we can detect using open-set response tasks. Although most participants did report fusion percepts on some proportion of trials, the overwhelming majority of responses matched the physical auditory stimulus.

When a research community latches on to a particular effect, it is easy to assume. Throughout this thesis, I have advocated for the inclusion of visual information into theories of speech perception; however, the practice of basing theories of perceptual integration primarily on results from McGurk-type experiments (Dodd & Campbell, 1987; Campbell et al., 1998) seems misguided (Massaro, 1998), considering the overwhelming lack of fusion responses found in this experiment. That being said, the variation in audio-visual integration effects for the various stimulus parameters that we tested does offer further support for the flexible nature of audio-visual speech processing. Additional testing of real-word McGurk percepts in additional experimental paradigms could clarify whether the findings reported here are an interesting non-effect, or if the combination of talker attributes, an imbalanced stimulus list and a high-demand open-set response task (possibly in conjunction with a participant sample that was less likely to fuse tokens) combined to diminish fusion percepts overall.

These experiments, taken together, reinforce the idea that it is important to explore multisensory interactions at various levels, from the low-level sensory integration of audio-visual signals to determining the properties of predictive visual cues that are responsible for auditory response facilitation, to addressing stimulus parameters that potentially influence the fusion rates in McGurk dubbings of real word stimuli.

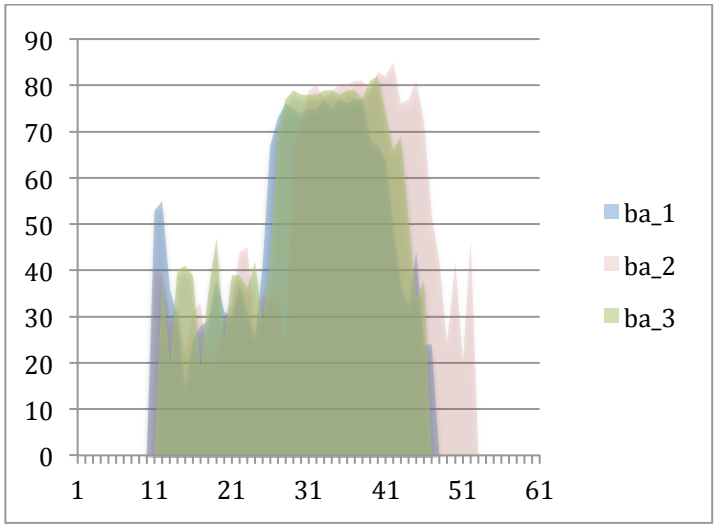
Each of the experiments presented in this thesis could be built on and extended to more accurately assess mechanisms for multisensory integration in natural speech settings. Establishing audio-visual effects in highly controlled experimental designs is a critical first step in understanding where, when, how, and why these interactions may be occurring, but modifying these studies to make them more realistic (e.g., utilizing real speech stimuli to test envelope entrainment with the SSR paradigm described in Chapter 4) is necessary to make strong claims about real-world implications of the results reported here.

Appendices

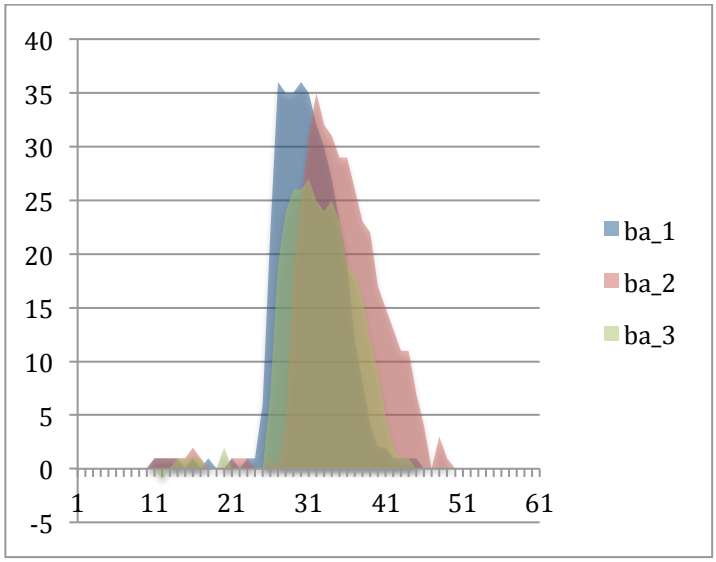
Appendix I: Chapter 3 visual stimulus details

Lip aperture by frame for each stimulus. X-axis: frame number; Y-axis: Aperture (in pixels).

Stimulus /ba/: Horizontal distance between lip corners by frame for each token.

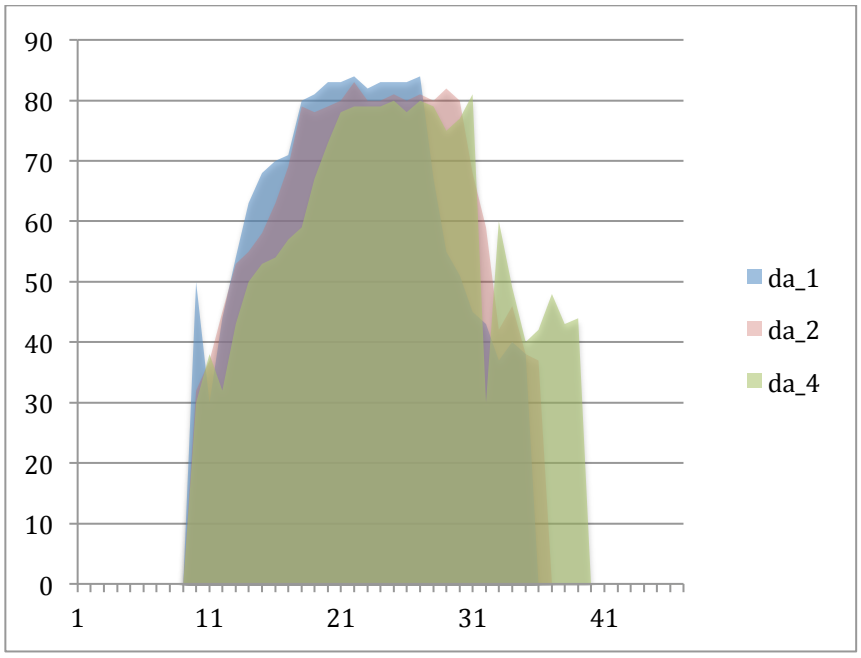


Stimulus /ba/: Vertical distance between lip midpoints by frame for each token

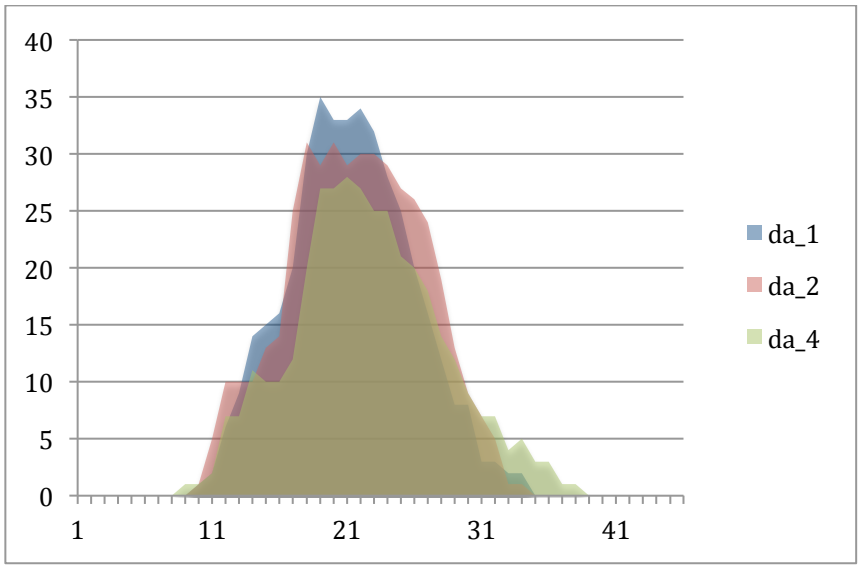


Lip aperture by frame for each stimulus. X-axis: frame number; Y-axis: Aperture (in pixels).

Stimulus /da/: Horizontal distance between lip corners by frame for each token.

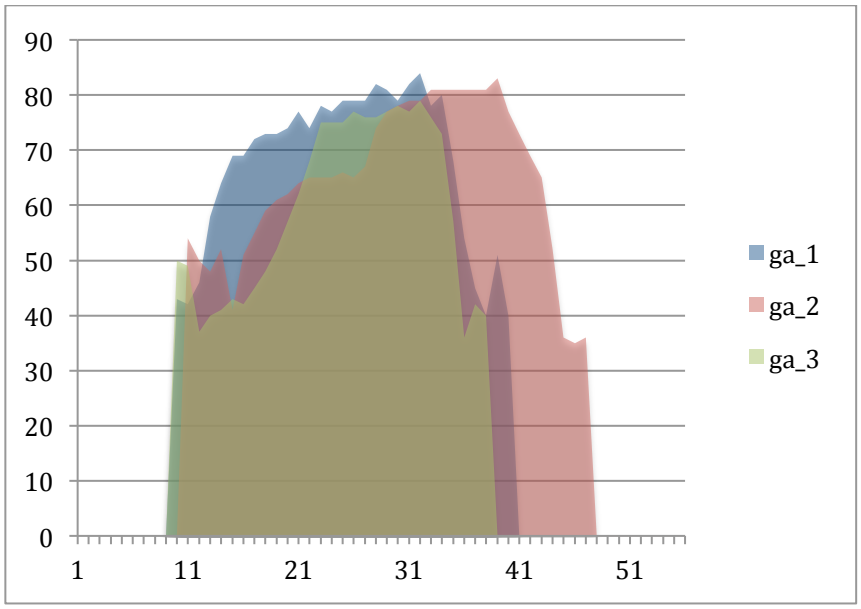


Stimulus /da/: Vertical distance between lip midpoints by frame for each token.

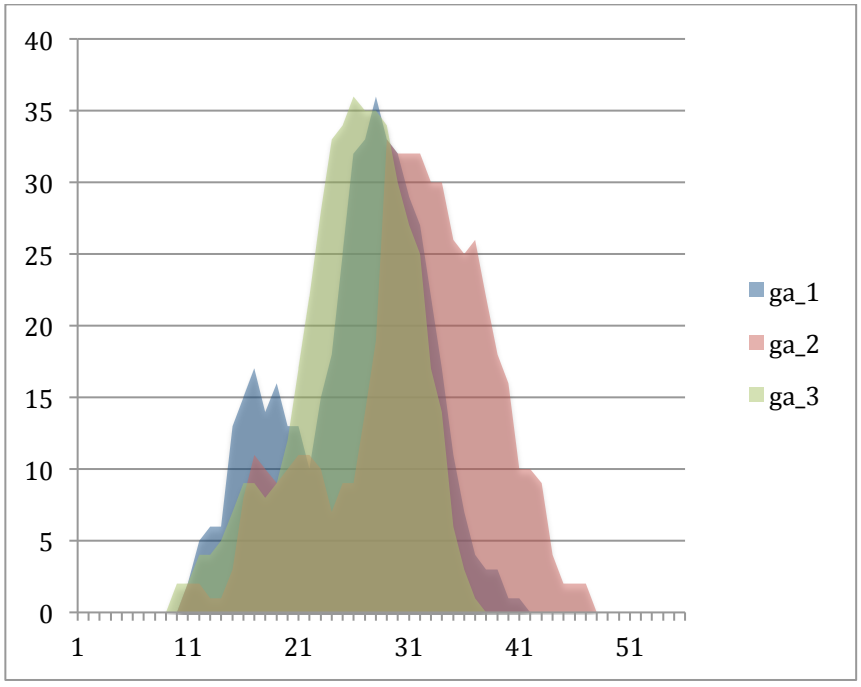


Lip aperture by frame for each stimulus. X-axis: frame number; Y-axis: Aperture (in pixels).

Stimulus /ga/: Horizontal distance between lip corners by frame for each token.

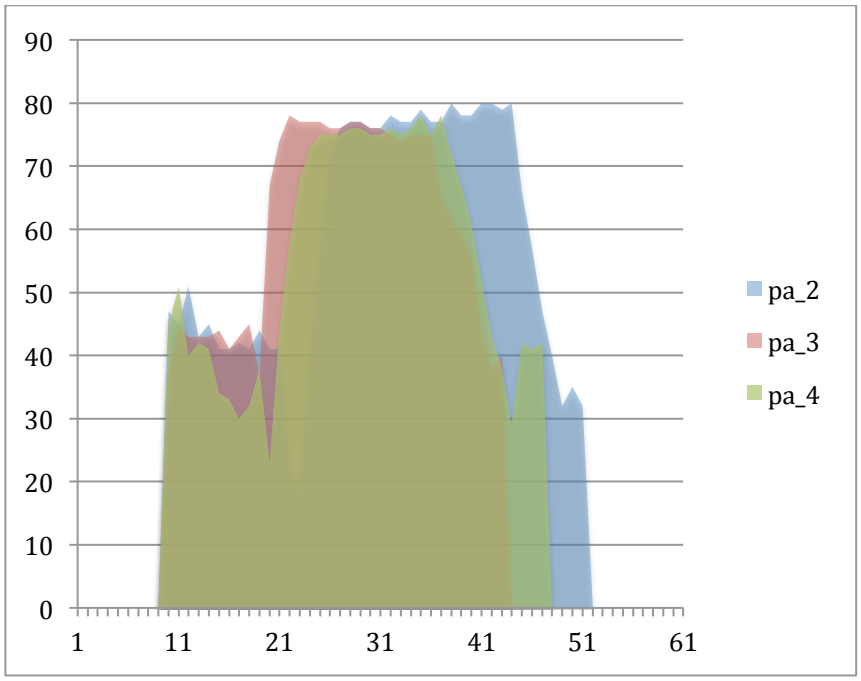


Stimulus /ga/: Vertical distance between lip midpoints by frame for each token.

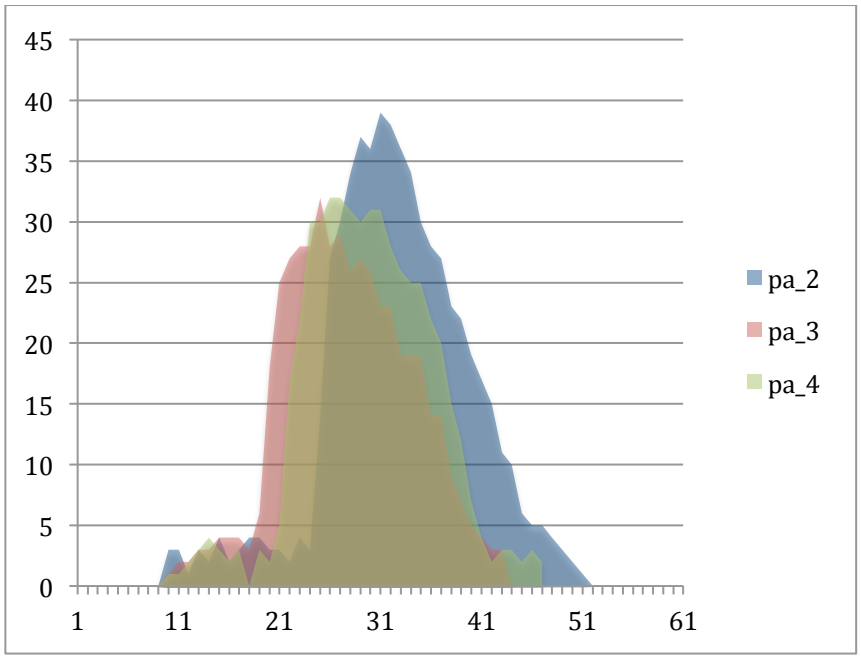


Lip aperture by frame for each stimulus. X-axis: frame number; Y-axis: Aperture (in pixels).

Stimulus /pa/: Horizontal distance between lip corners by frame for each token.



Stimulus /pa/: Vertical distance between lip midpoints by frame for each token.



Appendix II: Chapter 5 stimulus list and response proportions

Italics indicate nonword McGurk percepts

Audio	Video	Expected Percept	# of tokens	Percent McGurk	Percent Velar	Percent Other
ape	ache	ate	3	1.39%	11.11%	0%
bait	gate	date	8	13.02%	0%	5.73%
barter	garter	*darter*	3	0%	2.78%	4.17%
bash	gash	dash	7	15.48%	0%	4.76%
beep	beak	beat	2	8.33%	6.25%	0%
beer	gear	dear	7	20.83%	0%	1.79%
best	guest	*dest*	2	4.17%	0%	0%
bet	get	debt	8	0.52%	0%	4.69%
big	gig	dig	8	21.35%	0.52%	0.52%
boast	ghost	dosed	3	4.17%	2.78%	4.17%
boat	goat	dote	4	0%	2.08%	2.08%
brace	grace	*drace*	5	0%	0%	2.50%
braid	grade	*draid*	4	0%	0%	0%
brain	grain	drain	8	0%	0%	0%
braise	graze	*draze*	2	0%	0%	0%
brass	grass	*drass*	2	0%	4.17%	0%
brave	grave	*drave*	3	0%	0%	0%
brim	grim	*drim*	2	0%	0%	0%
brunt	grunt	*drunt*	3	0%	5.56%	0%
bum	gum	dumb	3	1.39%	0%	0%
bun	gun	done	4	1.04%	0%	5.21%
bust	gust	dust	3	4.17%	0%	1.39%
butter	gutter	*dutter*	4	0%	0%	0%
bye	guy	dye	6	2.78%	0%	12.50%
cheap	cheek	cheat	4	3.13%	21.88%	0%
flab	flag	*flad*	3	2.78%	8.33%	2.78%
flap	flak	flat	2	2.08%	4.17%	0%
hip	hick	hit	8	1.04%	7.81%	0%
hype	hike	height	5	3.33%	5.00%	0%
job	jog	*jod*	2	0%	8.33%	0%
lab	lag	lad	4	2.08%	20.83%	4.17%
lap	lack	*lat*	7	0%	16.67%	0%
lip	lick	lit	2	4.17%	4.17%	0%
lop	lock	lot	5	0%	6.67%	0%
nip	nick	knit	3	0%	11.11%	0%
page	cage	*tage*	5	1.67%	0%	0%
palace	callous	*talace*	3	0%	0%	0%

pall	call	tall	2	0%	0%	2.08%
pamper	camper	tamper	6	4.86%	0%	0.69%
pan	can	tan	4	11.46%	4.17%	0%
par	car	tar	8	1.56%	0.52%	0%
paste	cased	taste	7	5.95%	0%	0%
patch	catch	*tatch*	4	0%	0%	0%
pause	cause	taws	2	2.08%	0%	0%
pave	cave	*tave*	3	1.39%	0%	0%
pear	care	tear	5	18.33%	0%	3.33%
pearl	curl	*turl*	6	0%	27.78%	2.08%
peg	keg	*teg*	2	8.33%	2.08%	0%
petal	kettle	*tettle*	3	0%	0%	0%
pick	kick	tick	10	13.33%	3.75%	0%
pill	kill	till	4	0%	1.04%	2.08%
pin	kin	tin	3	2.78%	1.39%	0%
poach	coach	*toach*	4	1.04%	11.46%	0%
poll	coal	toll	9	0.46%	2.31%	0%
pool	cool	tool	3	0%	4.17%	0%
pop	cop	top	3	1.39%	0%	0%
pork	cork	torque	6	1.39%	3.47%	0%
post	coast	toast	9	4.63%	4.17%	0.93%
poster	coaster	toaster	11	3.03%	2.27%	0%
pour	core	tore	12	5.56%	6.94%	0%
prank	crank	*trank*	3	0%	2.78%	0%
preacher	creature	*treacher*	4	0%	3.13%	0%
prime	crime	*trime*	2	0%	2.08%	0%
prop	crop	*trop*	4	0%	2.08%	0%
proud	crowd	*troud*	5	0%	0%	0%
prude	crude	*trude*	1	0%	0%	0%
pry	cry	try	6	0%	2.08%	0%
pub	cub	tub	3	4.17%	1.39%	1.39%
puddle	cuddle	*tuddle*	4	0%	3.13%	2.08%
puff	cuff	tough	4	9.38%	1.04%	0%
quip	quick	quit	2	0%	6.25%	0%
reap	reek	*reet*	3	0%	11.11%	0%
rib	rig	rid	3	13.89%	2.78%	1.39%
robe	rogue	rode	3	0%	18.06%	0%
shape	shake	*shate*	3	2.78%	4.17%	0%
sharp	shark	*shart*	3	0%	6.94%	0%
shop	shock	shot	7	4.17%	15.48%	0%
shrub	shrug	*shrud*	2	0%	6.25%	0%
sip	sick	sit	3	5.56%	ß5.56%	4.17%
sleep	sleek	sleet	2	6.25%	4.17%	0%

slip	slick	slit	4	3.13%	5.21%	0%
snap	snack	*snat*	2	0%	10.42%	0%
soap	soak	*sote*	4	4.17%	1.04%	0%
stab	stag	*stad*	3	0%	5.56%	0%
tab	tag	tad	4	6.25%	5.21%	1.04%
trip	trick	*trit*	4	4.17%	4.17%	0%
weep	week	wheat	2	0%	8.33%	0%

Bibliography

Aiken, S. J., & Picton, Terence W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, *29*(2), 139-157.

Amano, J., & Sekiyama, K. (1998). The McGurk effect is influenced by the stimulus set size. *AVSP-1998* (pp. 43-48). Presented at the Auditory-Visual Speech Processing, Sydney, Australia.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A.-L. (2009). Dual Neural Routing of Visual Facilitation in Speech Processing. *Journal of Neuroscience*, *29*(43), 13445-13453. doi:10.1523/JNEUROSCI.3194-09.2009

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *Br J Psychol*, *92*, 339-355.

Baayen, R. H. (2008). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics", *R package version 0.953*.

Baier, R., Idsardi, W. J., & Lidz, J. (2007). Two-month olds are sensitive to lip rounding in dynamic and static speech events. *International Conference on Auditory-Visual Speech Processing*. Hilvarenbeek, The Netherlands.

Barutchu, A., Crewther, S., Kiely, P., Murphy, M., & Crewther, D. (2008). When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology*, *20*(1), 1-11. doi:10.1080/09541440601125623

Bernstein, L. E., Auer Jr, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Commun*, *44*, 5-18.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (1998). What makes a good speechreader? First you have to find one. In R. Campbell, B. Dodd,, & D. K. Burnham (Eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual speech* (pp. 211-228). East Sussex, UK: Psychology Press Ltd.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur J Neurosci*, *20*(8), 2225-34.

Brancazio, L. (2004). Lexical Influences in Audiovisual Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 445-463. doi:10.1037/0096-1523.30.3.445

Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press.

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

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., et al. (1997). Activation of Auditory Cortex During Silent Lipreading. *Science*, *276*(5312), 593-596.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos Trans R Soc Lond B Biol Sci*, *363*, 1001-1010.

Campbell, R., Dodd, B., & Burnham, D. (1998). *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech*. Hove, East Sussex, UK: Psychology Press Ltd.

de Cheveigné, A., & Simon, Jonathan Z. (2007). Denoising based on time-shift PCA. *J Neurosci Methods*, *165*(2), 297-305.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *The Journal of the Acoustical Society of America*, *76*(6), 1636. doi:10.1121/1.391610

Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics*, *51*(4), 355-362. doi:10.3758/BF03211629

Dobie, R. A., & Wilson, M. J. (1996). A comparison of t test, F test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion product otoacoustic emissions, or other sinusoids. *J Acoust Soc Am*, *100*(4), 2236-2246.

Dodd, B., & Campbell, R. (Eds.). (1987). *Hearing by eye : the psychology of lip-reading*. London;Hillsdale N.J.: Lawrence Erlbaum Associates.

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Percept Psychophys*, *32*(6), 562-570.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423-425.

Erber, N. P. (1975). Auditory-Visual Perception of Speech. *J Speech Hear Disord*, *40*, 481-492.

Fisher, N. I. (1996). *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*, *52*(6), 525-532. doi:10.1016/j.specom.2010.02.005

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

Ganong, W. F., 3rd. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110-125.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn Sci*, *10*(6), 278-285.

Giard, M. H., & Peronnet, F. (1999). Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study. *J Cogn Neurosci*, *11*(5), 473-490.

Gordon, P. C. (1997a). Coherence masking protection in brief noise complexes: Effects of temporal patterns. *J Acoust Soc Am*, *102*(4), 2276-2282.

Gordon, P. C. (1997b). Coherence masking protection in speech sounds: The role of formant synchrony. *Percept Psychophys*, *59*(2), 232-242.

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *J Acoust Soc Am*, *109*(5 Pt 1), 2272-5.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *J Acoust Soc Am*, *104*(4), 2438-50.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am*, *108*(3 Pt 1), 1197-208.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *J Acoust Soc Am*, *103*(5 Pt 1), 2677-90.

Green, K. P., & Gerdeman, A. (1995). Cross-Modal Discrepancies in Coarticulation and the Integration of Speech Information: The McGurk Effect With Mismatched Vowels. *J Exp Psychol Hum Percept Perform*, *21*(6), 1409-1426.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Percept Psychophys*, *50*(6), 524-536.

Hall, J. W., & Grose, J. H. (1988). Comodulation masking release: Evidence for multiple cues. *J Acoust Soc Am*, *84*(5), 1669-1675.

Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *J Acoust Soc Am*, *76*(1), 50-56.

Hillenbrand, J. M., & Gayvert, R. T. (2005). Open Source Software for Experiment Design and Control. *J Speech Lang Hear Res*, *48*(1), 45-60. doi:10.1044/1092-4388(2005/005)

Howard, M. F., & Poeppel, D. (2010). Discrimination of Speech Stimuli Based on Neuronal Response Phase Patterns Depends On Acoustics But Not Comprehension. *J Neurophysiol*.

Jenkins, J. 3rd, Rhone, A. E., Idsardi, W. J., Simon, J.Z., & Poeppel, D. (2011). The elicitation of audiovisual steady-state responses: multi-sensory signal congruity and phase effects. *Brain Topography*, *24*(2), 134-148. doi:10.1007/s10548-011-0174-1

Jones, E. G., & Powell, T. P. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain*, *93*(4), 793-820.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138-1141.

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences*, *105*(32), 11442 -11445. doi:10.1073/pnas.0804275105

Lachs, L., Pisoni, D. B., & Kirk, K. I. (2001). Use of Audiovisual Information in Speech Perception by Prelingually Deaf Children with Cochlear Implants: A First Report. *Ear Hear*, *22*(3), 236-251.

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*, 1001-1010.

Luo, H., Wang, Y., Poeppel, D., & Simon, J.Z. (2006). Concurrent Encoding of Frequency and Amplitude Modulation in Human Auditory Cortex: MEG Evidence. *J Neurophysiol*, *96*, 2712-2723.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLoS ONE*, *4*(3), e4638.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Percept Psychophys*, *24*(3), 253-257.

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: how much spatial degradation can be tolerated. *Perception*, *29*, 1155-1168.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluate, and recommendations for use. *Br J Audiol*, *24*(1), 29-43.

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P., Williams, S. C., et al. (2000). Silent speechreading in the absence of scanner noise: an event-related fMRI study. *Neuroreport*, *11*(8), 1729-1733.

MacSweeney, M., Campbell, R., Calvert, G. A., McGuire, P. K., David, A. S., Suckling, J., Andrew, C., et al. (2001). Dispersed activation in the left temporal cortex for speech-reading in congenitally deaf people. *Proceedings. Biological sciences / The Royal Society*, *268*(1466), 451-457. doi:10.1098/rspb.2000.0393

Mäkelä, J. P. (2007). Magnetoencephalography: Auditory evoked fields. *Auditory Evoked Potentials: basic principles and clinical application* (pp. 525-545).

Massaro, D. W. (1998). Illusions and Issues in Bimodal Speech Perception. *Proceedings of Auditory Visual Speech Perception 1998* (pp. 21-26). Presented at the AVSP, Sydney, Australia.

Massaro, D. W. (n.d.). Speech Perception by Ear and Eye. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53-83). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(5), 753-771. doi:10.1037/0096-1523.9.5.753

Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. *J Phonetics*, *21*(4), 445-478.

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

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121*, 1013-1052.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *J Acoust Soc Am*, *82*(6), 2145-2147.

Müller, M. M., & Hillyard, S. (2000). Concurrent recording of steady-state and transient event-related potentials as indices of visual-spatial selective attention. *Clinical Neurophysiology*, *111*(9), 1544-1552. doi:16/S1388-2457(00)00371-0

Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology*, *24*(4), 375-425.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychol Res*, *71*, 4-12.

Neale, M. (1999). *Neale analysis of reading ability* (3rd ed.). Melbourne Vic.: Australian Council for Educational Research Ltd.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3T. *NeuroReport*, *16*(2), 125-128. doi:10.1097/00001756-200502080-00010

Picton, T W, Woods, D. L., Baribeau-Braun, J., & Healey, T. M. (1976). Evoked potential audiometry. *The Journal of Otolaryngology*, *6*(2), 90-119.

Picton, T. W., John, M. S., Dimitrijevic, A., & Purcell, D. (2003). Human auditory steady-state responses. *Int J Audiol*, *42*(4), 177-219.

Pilling, M. (2009). Auditory Event-Related Potentials (ERPs) in Audiovisual Speech Perception. *J Speech Lang Hear Res*, *52*, 1073-1081.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Richie, C., & Kewley-Port, D. (2008). The Effects of Auditory-Visual Vowel Identification Training on Speech Recognition Under Difficult Listening Conditions. *J Speech Lang Hear Res*, *51*, 1607-1619.

Rosenblum, L. D., Johnson, J., & Saldaña, H. M. (1996). Visual kinematic information for embellishing speech in noise. *J Speech Hear Res*, *39*(6), 1159-1170.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Percept Psychophys*, *59*(3), 347-357.

Ross, B., Borgmann, C., Draganova, R., Roberts, L. E., & Pantev, C. (2000). A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude modulated tones. *J Acoust Soc Am*, *108*(2), 679-691.

Di Russo, F., Pitzalis, S., Aprile, T., Spitoni, G., Patria, F., Stella, A., Spinelli, D., et al. (2007). Spatiotemporal analysis of the cortical sources of the steady-state visual evoked potential. *Human Brain Mapping*, *28*(4), 323-334. doi:10.1002/hbm.20276

Sams, M., Aulanko, R., Hämäläinen, 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. *Neurosci Lett*, *127*(1), 141-145.

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, *26*(1-2), 75-87. doi:16/S0167-6393(98)00051-X

Sams, M., Paavilainen, P., Alho, K., & Näätänen, R. (1985). Auditory frequency discrimination and event-related potentials. *Electroencephalography and Clinical Neurophysiology*, *62*(6), 437-448.

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proc Natl Acad Sci U S A*, *102*(51), 18748-50.

Schwartz, J.-L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *The Journal of the Acoustical Society of America*, *127*(3), 1584. doi:10.1121/1.3293001

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, *59*(1), 73-80. doi:10.3758/BF03206849

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797. doi:10.1121/1.401660

Senkowski, D., Schneider, T. R., Foxe, John J, & Engel, A. K. (2008). Crossmodal binding through neural coherence: implications for multisensory processing. *Trends Neurosci*, *31*(8), 401-409.

Simon, J.Z., & Wang, Y. (2005). Fully complex magnetoencephalography. *J Neurosci Methods*, *149*, 64-73.

Sohmer, H., Pratt, H., & Kinarti, R. (1977). Sources of frequency following response (FFR) in man. *Electroencephalogr Clin Neurophsyiol*, *42*(5), 656-664.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-Visual Speech Perception and Auditory Visual-Enhancement in Normal-Hearing Younger and Older Adults. *Ear Hear*, *26*(3), 263-275.

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 580-587. doi:10.1037/a0013483

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *J Acoust Soc Am*, *67*(1), 318-326.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J Cogn Neurosci*, *19*(12), 1964-73.

Strelnikov, K., Rouger, J., Barone, P., & Deguine, O. (2009). Role of speechreading in audiovisual interactions during the recovery of speech comprehension in deaf adults with cochlear implants. *Scandinavian Journal of Psychology*, *50*(5), 437-444. doi:10.1111/j.1467-9450.2009.00741.x

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *J Acoust Soc Am*, *26*(2), 212-215.

Summerfield, Quentin. (1992). Lipreading and Audio-Visual Speech Perception. *Philos Trans R Soc Lond B Biol Sci*, *335*(1273), 71-78.

Talsma, D., Doty, T. J., Strowd, R., & Woldorff, M. G. (2006). Attentional capacity for processing concurrent stimuli is larger across sensory modalities than within a modality. *Psychophysiology*, *43*, 541-549.

Valdes, J. L., Perez-Abalo, M. C., Martin, V., Savio, G., Sierra, C., Rodriguez, E., & Lins, O. (1997). Comparison of Statistical Indicators for the Automatic Detection of 80 Hz Auditory Steady State Responses. *Ear and Hearing*, *18*(5), 420-429.

Walden, B. E., Erdman, S. A., Montgomery, A. A., Schwartz, D. M., & Prosek, R. A. (1981). Some Effects of Training on Speech Recognition by Hearing-Impaired Adults. *J Speech Hear Res*, *24*, 207-216.

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *J Acoust Soc Am*, *124*(3), 1716-1726.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci U S A*, *102*(4), 1181-6.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598-607.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language*, *50*(2), 212-230. doi:16/j.jml.2003.10.001