Changes in the McGurk effect across phonetic contexts. I. Fusions.

Michelle Hampson, Frank Guenther, and Michael Cohen

November, 1999

Technical Report CAS/CNS-99-031

Boston University Center for Adaptive Systems and
Department of Cognitive and Neural Systems
677 Beacon Street
Boston, MA 02215

Changes in the McGurk effect across phonetic contexts. I. Fusions.

Running title: McGurk fusions across phonetic contexts

Michelle Hampson, Frank H. Guenther, and Michael A. Cohen[1]

Boston University
Center for Adaptive Systems and
Department of Cognitive and Neural Systems
677 Beacon Street
Boston, MA, 02215

Boston University Technical Report CAS/CNS-TR-99-031

Address correspondence to:
Michelle Hampson
Yale University
Department of Diagnostic Radiology
P.O. Box 208042
New Haven, Connecticut 06520-8042
Fax Number: (203) 785-6534
Email: chell@boreas.med.yale.edu

1. Michelle Hampson was supported by the National Institute on Deafness and other Communication Disorders (NIDCD R29 02852). Frank Guenther was supported in part by the Alfred P. Sloan Foundation and the National Institute on Deafness and other Communication Disorders (NIDCD R29 02852). Many thanks to Paulo Gaudiano for loaning us the video equipment necessary to set up these experiments, and to Barbara Shinn-Cunningham and Joseph Perkell for helpful discussions regarding the work.

ABSTRACT

The McGurk effect has generally been studied within a limited range of phonetic contexts. With the goal of characterizing the McGurk effect through a wider range of contexts, a parametric investigation across three different vowel contexts, /i/, /ɑ/, and /u/, and two different syllable types, consonant-vowel (CV) and vowel-consonant (VC), was conducted.

This paper discusses context-dependent changes found specifically in the McGurk fusion phenomenon (Part II addresses changes found in combination percepts). After normalizing for differences in the magnitude of the McGurk effect in different contexts, a large qualitative change in the effect across vowel contexts became apparent. In particular, the frequency of illusory /g/ percepts increased relative to the frequency of illusory /d/ percepts as vowel context was shifted from /i/ to /ɑ/ to /u/. This trend was seen in both syllable sets, and held regardless of whether the visual stimulus used was a /g/ or /d/ articulation.

This qualitative change in the McGurk fusion effect across vowel environments corresponded systematically with changes in the typical second formant frequency patterns of the syllables presented. The findings are therefore consistent with sensory-based theories of speech perception which emphasize the importance of second formant patterns as cues in multimodal speech perception.

1. Introduction

Many studies have shown that vision can assert a strong influence on speech perception. For example, Sumby and Pollack (1954) reported that the intelligibility of speech in noise is higher when viewing the speaker (see also Erber, 1969), and Dodd (1977) found that this benefit persisted despite introduction of a 400 ms auditory delay. Even in the absence of noise, Reisberg, McLean and Goldfield (1987) found that auditory speech perception could be facilitated by watching videos of the speaker.

One of the most striking examples of the important role vision plays in speech perception is the McGurk effect, first reported by McGurk and MacDonald (1976). This effect occurs when viewing the utterance of one consonant while listening to a different consonant. The resulting auditory percept is then affected by the visual input. For example. when watching a video of someone uttering a /ba/ and listening to the syllable /ga/, people will often report hearing /bga/. This type of combination percept tends to occur when subjects are viewing a labial utterance and listening to a velar or alveolar utterance. When the modalities are reversed, a qualitatively different effect occurs, referred to as a fusion. For example, when watching a video of someone speaking /ga/ and listening to a dubbed recording of /ba/, subjects often have an auditory percept of /da/ or /ða/.

The McGurk effect has been studied under many different conditions. Some changes in the effect resulting from a degraded or enhanced acoustic signal are documented in Green and Norrix (1997). These include, for example, an increase in the magnitude of the McGurk effect when the formant transitions are low-pass filtered. Other researchers have manipulated characteristics of the visual signal to determine how they affect perception. Temporal gating of the visual stimulus was found to decrease the McGurk effect in direct proportion to the amount of stimulus removed (Munhall and Tokhura, 1998). This may be due to the loss of dynamic visual information. Static visual information does not appear to be critical to the effect, as a McGurk effect occurs even when the face is replaced by a point-light display that captures the facial dynamics (Rosenblum and Saldana, 1996). There have also been several experiments investigating how temporal incongruencies between the auditory and visual information affect the resulting percept. Although the McGurk effect appears robust to some temporal misalignment (Massaro and Cohen, 1993; Munhall, Gribble, Sacco and Ward, 1996), it is sensitive to mismatches in dynamics across the modalities (Munhall, Gribble, Sacco, and Ward, 1996). These are just a subset of the many studies which have helped to characterize the McGurk effect under different stimulus conditions.

Despite all of this research, there has not been much investigation of the influence of phonetic context on the McGurk effect. Most work has focussed on the /a/ vowel context (or the very similar /ɑ/ and /æ/ contexts) with three major exceptions. First, the perception of acoustic /b/-visual /g/ stimuli in the /i/, /ɑ/, and /u/ contexts was investigated by Green, Kuhl, and Meltzoff (1988)[1]. The frequency of illusory /d/ percepts was found to be high in the /i/ context, moderate in the /ɑ/ context, and very low in the /u/ context. This decrease in /d/ percepts was accompanied by an increase in /b/ percepts as vowel context was changed from /i/ to /ɑ/ to /u/ (Green, 1996). That is, the *magnitude* of the McGurk effect was found to vary across these three phonetic contexts.

Secondly, a series of studies compared the McGurk effect in the two different vowel contexts, /ɑ/ and /i/ (Green, Kuhl, Meltzoff and Stevens, 1991; Green and Gerdeman, 1995; Green and Norrix, 1997). Unlike Green, Kuhl, and Meltzoff (1988), these studies did not find a difference in the magnitude of the effect in these different vowel contexts. However, a qualitative difference in the effect was found: visual /g/-acoustic /b/ stimuli tended to produce more /d/ than /ð/ percepts in the /i/ context, and more /ð/ than /d/ percepts in the /ɑ/ context. It is not clear why these studies yielded different results than the Green, Kuhl, and Meltzoff (1988) study, but it is likely that the particular visual stimuli used play an important role in determining the exact nature of the effect (the studies of Green, Kuhl, Meltzoff and Stevens, 1991, Green and Gerdeman, 1995, and Green and Norrix, 1997, used the same visual stimuli).

One finding that is consistent across these experiments is that illusory /d/ percepts were less frequent in the /ɑ/ context than the /i/ context. As suggested by Green (1996), this decrease in /d/ percepts may be due to differences in the second formant patterns of the consonants in the two vowel contexts. The second formant patterns for /d/ and /b/ are both rising in the /i/ vowel context. In the /ɑ/ context, however, the second formant transition for /d/ is flat or falling, making it less similar to the rising transition of /b/. Green (1996) also notes that the second formant transition for /ð/ is flat or slightly rising in the /ɑ/ context, and in that sense, /ðɑ/ is more similar acoustically to /bɑ/ than /dɑ/ is. Changes in acoustics across contexts thus provide one possible explanation for the findings of Green and colleagues that /d/ percepts were more common in the /i/ context than the /ɑ/ context.

Finally, Jordan and Bevan (1997) examined the McGurk effect in the /i/ and /a/ vowel contexts[2]. In this study, there were as many /d/ percepts reported in the /a/ context as the /i/ context. This is in contrast to the findings of Green and colleagues discussed above. However, such disparity could be due to differences in acoustics between the /a/ context,

1. Although the text of Green, Kuhl, and Meltzoff (1988) refers to the /a/ vowel context, there is some confusion regarding the precise vowel used in this study. All of the studies of Green and colleagues discussed in this introduction (Green, Kuhl, and Meltzoff, 1988; Green, Kuhl, Meltzoff and Stevens, 1991, Green and Gerdeman, 1995, and Green and Norrix, 1997) refer to the /a/ vowel context. However, in the study of Green and Norrix (1997), it appears that this notation is adopted in the text for typographical reasons, as the tables are labelled with the vowel /ɑ/ and the formant patterns of the stimuli (see Table 2 of Green and Norrix, 1997) are typical of the vowel /ɑ/. As several studies of Green and colleagues shared stimuli (see Green and Gerdeman, 1995, and Green and Norrix, 1997), it is possible that /a/ tokens used in Green and Norrix (1997) experiments were also used in the studies of Green, Kuhl, Meltzoff, and Stevens (1991) and Green and Gerdeman (1995). In fact, regardless of whether the acoustic stimuli were shared across these different studies, it is likely that all references by Green and colleagues to the phoneme /a/ are intended to denote the very similar (but typographically more troublesome) phoneme /ɑ/ because the vowel /a/ does not exist in North American English, and the stimuli used by Green and colleagues were all naturally produced utterances. Therefore throughout this paper, all discussion of the papers Green, Kuhl, and Meltzoff (1988), Green, Kuhl, Meltzoff and Stevens (1991), Green and Gerdeman (1995), and Green and Norrix (1997) will assume that references to the phoneme /a/ are actually intended to denote the similar North American English phoneme /ɑ/.
2. Walker, Bruce and O'Malley (1995) also studied the McGurk effect in these two vowel contexts, but their results were not presented separately for the two contexts, so a comparison between the two vowels is not possible.

and the /ɑ/ context. Although these vowels are very similar, /a/ does generally have a higher second formant frequency than /ɑ/, and is more similar to /i/ in that respect. Whether such differences in acoustic patterns would be large enough to account for the differences found in the frequency of /d/ fusion percepts is not clear. The findings in these studies raise an interesting question: do the qualitative characteristics of the McGurk effect across different phonetic contexts depend in predictable ways upon the acoustics associated with those contexts?

One reason it is difficult to answer this question conclusively is that the experiments which have tested the McGurk effect in different phonetic contexts have used only a limited range of stimuli. For example, all of the studies by Green and colleagues discussed above were limited to acoustic /b/-visual /g/ and acoustic /g/ - visual /b/ stimuli. The examination of a wider range of stimuli is more likely to reveal systematic patterns of change across contexts. For this reason, a parametric study of the McGurk effect involving a complete cross of acoustic /b/, /d/, and /g/ stimuli with visual /b/, /d/, and /g/ stimuli in six different phonetic contexts is undertaken here.

The characterization of systematic changes in the McGurk effect across contexts will have implications for theories of speech perception. Currently, there is controversy in the field of speech perception regarding the reference frame in which speech is perceived. One theory is that phonemes (or phonetic units) are identified based on their sensory characteristics (e.g. Massaro, 1987; Diehl and Kluender, 1989a). In this view, the McGurk effect is the result of a general process of multimodal pattern recognition being applied to novel sensory input patterns. As there is a great deal of variability in the acoustics of phonemes across phonetic contexts (see Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967, for overview), this might be expected to result in a qualitatively different McGurk effect in different contexts. Context-dependent changes in the McGurk effect would thus be consistent with theories of sensory-based speech perception *provided* that such changes are systematically related to changes in sensory cues that are known to be important in phoneme perception.

An alternate view of speech perception is that phonemes are identified in terms of the motor actions underlying the spoken utterance (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967; Fowler, 1986). From this theoretical perspective, the McGurk effect arises because different sensory systems are providing conflicting information regarding the articulatory gestures of the speaker (Liberman and Mattingly, 1985; Fowler, 1986; Fowler and Dekle, 1991). This leads to the prediction that changes in the McGurk effect with context, if there are any such changes, should be systematically related to changes in the articulatory gestures across contexts[3]. Examination of the McGurk effect in a set of different contexts may thus provide insight into whether speech perception is based on sensory or motor dimensions.

The primary purpose of this study, therefore, is to characterize the McGurk effect across a range of phonetic contexts. Six phonetic contexts were investigated, representing the three vowel contexts, /i/, /ɑ/, and /u/ and two different syllable types, consonant-vowel and vowel-consonant. The three vowel contexts were chosen to allow comparison with the findings of previous studies by Green and colleagues (Green, Kuhl, and Meltzoff, 1988;

Green, Kuhl, Meltzoff and Stevens, 1991; Green and Gerdeman, 1995; Green and Norrix, 1997) and because they represent the range of English vowels well. The comparison of two different syllable types has not been made in previous studies. The majority of experiments investigating the McGurk effect have used only consonant-vowel (CV) syllables (see, for example, MacDonald and McGurk, 1978; Green and Miller, 1985; Massaro and Cohen, 1983; Rosenblum, Schmuckler and Johnson, 1997). There have been a few studies involving real-word contexts (Easton and Basala, 1982; Dekle, Fowler, and Funnel, 1992; Fuster-Duran, 1996) and some investigating the effect in VCV syllables (Munhall, Gribble, Sacco, and Ward, 1996; Munhall and Tohkura, 1998; Smeele, Hahnlen, Stevens, Kuhl, and Meltzoff, 1995; Siva, Stevens, Kuhl, and Meltzoff, 1995), and it is clear from these studies that the McGurk effect is not limited to CV contexts. However, there is a dearth of information regarding the nature of the McGurk effect in VC contexts, and for this reason, VC as well as CV syllables were used here.

A secondary purpose of this study is to determine the role that linguistic or visual biases (which are present under unimodal testing conditions) play in audio-visual speech perception. Massaro (1998a) emphasized the importance of testing unimodal conditions when performing experiments on audio-visual speech perception. He tested subjects on silent videos of /da/ and /ga/ utterances, and found that subjects were twice as likely to respond /da/ than they were to respond /ga/ to both stimuli. He suggested that we may have a linguistic bias for /da/ over /ga/, because /da/ appears more often in spoken language. This raises the possibility that McGurk "fusions" are not really fusions, but are simply the result of linguistic biases influencing perception. For example, an acoustic /ba/-visual /ga/ stimulus may produce a /da/ percept rather than a /ga/ percept because linguistic biases (or other biases which are present in unimodal visual perception) cause the /ga/ face to be perceived as /da/ (see also Munhall et al, 1996). To address this issue, perceptual tests were run on the unimodal stimuli (Experiment 1), which were used in creating the bimodal, McGurk stimuli (Experiment 2).

2. Experiment 1: Perception of Unimodal Stimuli

This experiment tested the auditory and visual stimuli used for Experiment 2. This was important for ensuring that the unimodal stimuli were perceptually clear, and for identifying any intrinsic biases in visual perception.

3. The revised motor theory asserts that "(articulatory) gestures do have characteristic invariant properties" (Liberman and Mattingly, 1985), and direct realism hypothesizes that "the organization of the vocal tract to produce a phonetic segment is invariant over variation in segmental and suprasegmental contexts" (Fowler, 1986). If such gestural invariance is assumed to exist at a phonemic level, then a lack of context-dependent variability in the McGurk effect could be consistent with the two major motor-based theories of speech perception. Similarly, a lack of context-dependent variability in the McGurk effect could also be consistent with a sensory-based theory of perception which assumes that invariant sensory features are used in phoneme perception. Therefore, if the McGurk effect does not change across contexts, this will not allow us to discriminate between sensory and motor representations.

2.1 Stimuli

A female speaker (the experimenter) was taped uttering the nine syllables /gi/, /di/, /bi/, /gɑ/, /dɑ/, /bɑ/, /gu/, /du/, and /bu/ several times each. Video clips of these utterances were captured for playing on the computer at 15 frames per second. The audio track was digitized with a 44 kHz sampling rate. One perceptually robust acoustic recording of each syllable was selected and saved as an audio file. One clean video recording of each syllable was also chosen, and saved without its audio track. In the same way, acoustic recordings and silent video clips of the corresponding nine VC syllables (/ig/, /id/, /ib/, /ɑg/, /ɑd/, /ɑb/, /ug/, /ud/, and /ub/) were obtained.

2.2 Subjects

Ten adult subjects were recruited by flyers placed around the Boston University campus. They all had English as their first language, and normal or corrected-to-normal vision. None of the subjects reported any history of a speech or hearing disorder.

2.3 Procedure

Subjects were tested alone in a dimly-lit room. They were seated approximately 2 feet in front of a computer monitor with speakers on either side, and a keyboard in front of them. The experiment was response-paced, with a new video being played immediately following the previous response. Directions were given verbally prior to initiation of the session, after which the experimenter turned off the lights and left the room, and the subject began the test.

The stimuli were separated into four separate blocks: CV audio only, CV silent video, VC audio only, VC silent video. Each subject was given practice trials of all four types of stimuli (from the four different blocks) before the testing began. Four counterbalanced sequences of the four blocks were created (using a Latin square given in Keppel, 1991) and subjects were randomly assigned to one of these sequences. Within each block, the nine stimuli were each played ten times, in a random order. Directions were presented on the screen prior to each block. These instructions indicated what type of stimuli would be played during that block. Subjects were instructed to respond according to the consonant they heard during audio-only blocks, and according to the consonant they believed the speaker was uttering during the video only blocks. A prompt appeared after each video clip was played, and subjects entered their responses by typing in the letter (or letters) corresponding to their consonant percept, and pressing return. They were told that multiple letters could be entered, such as "ch" as in "chew", or "pk" if they heard (or saw) a "p" followed by a "k". If they did not know what consonant they perceived, subjects were instructed to enter a "?".

2.4 Results

Auditory only tests.

Table 1: Results from auditory only test

auditory stimulus			response percentages
syllable type	vowel	consonant	
consonant-vowel	i	b	b 93, p 6, d 1
		d	d 98, t 2
		g	g 96, k 3, gk 1
	ɑ	b	b 98, p 2
		d	d 100
		g	g 95, k 5
	u	b	b 91, p 4, bl 3, g 1, v 1
		d	d 100
		g	g 100
vowel-consonant	i	b	b 100
		d	d 100
		g	g 100
	ɑ	b	b 100
		d	d 90, n 10
		g	g 100
	u	b	b 100
		d	d 100
		g	g 100

All auditory stimuli were correctly identified at least 90% of the time. The errors which did occur were generally a confusion of manner and did not involve place of articulation, which is the primary dimension of interest in this study. A summary of the responses to the auditory tests is provided in Table 1. Most of the incorrect consonant identifications were due to two subjects. The ten percent error in identification of the syllable /ɑd/ is the result of one subject who consistently perceived this syllable as /ɑn/, although all other subjects perceived it consistently as /ɑd/. Another subject experienced confusions between voiced and voiceless CV syllables which resulted in over 75% of the errors in the CV syllable set. The perceptual data from each individual in the auditory-only condition are available in Appendix B of Hampson (1999).

Visual only tests

The totals across all subjects for consonant-vowel syllables in the /ɑ/ vowel context are shown below in Table 2 (individual data are available in Appendix B of Hampson, 1999). Note that the number of "g" responses is actually greater than the number of "d" responses

Table 2: Response totals for the visual stimuli /bɑ/, /dɑ/, and /gɑ/.

CV visual stimulus		response percentages
vowel	consonant	
ɑ	b	b 51, p 45, m 2, pl 1, y 1
	d	d 38, g 22, t 11, k 10, n 2, kr 1, r 1, ? 15
	g	g 41, d 25, k 20, t 8, ? 2, m 2, n 1, kr 1

to the silent /gɑ/ video. This may seem surprising, given the findings of Massaro (1998a) that /d/ percepts occur twice as frequently as /g/ percepts to both /ga/ and /da/ visual stimuli. However, it is inappropriate to compare the two studies in this manner: this study used an open response paradigm while Massaro's experiment used a forced choice paradigm that did not allow subjects to enter unvoiced or nasal consonants. A more appropriate comparison can be drawn by examining the perception of place of articulation in the two studies. In order to do this, the response data from this study are categorized by place of articulation in Table 3. Because the paradigm was not forced choice, there are many different types of responses, some of which are difficult to classify by place of articulation. In general, anything that does not have a labial, alveolar, or velar constriction, or that involves the formation of more than one such constriction, is classified as "other". The four categories used in this analysis are labial ("b", "p", "bh", and "m" responses), alveolar ("d", "t", "n", "s", "l", and "dh" responses), velar ("g", "k", "ng", and "gh" responses), and other (includes the responses: "ch", "r", "q", "sh", "kr", "f", "sh", "pr", "bp", "spl", "bl", "kl", "tl", "gk", "pl", "pb", "y", and "?"). Although the phoneme /n/ is sometimes classified as an interdental stop consonant (Kent and Read, 1992), it has been categorized here as an alveolar consonant following the classification scheme of Akmajian, Demers, Farmer and Harnish (1990) (see also Ladefoged, 1993). In departure from the Akmajian et al. (1990) classification scheme, the phoneme /r/ is not treated as an alveolar consonant. It has been assigned to the "other" category because there are a range of different articulations corresponding to American English /r/ (Delattre and Freeman, 1968; Ong and Stone, 1998; Guenther et al., 1998; Westbury, Hashi, and Lindstrom**, 1995). Finally, the reponses "bh", "dh", and "gh" have been assigned to the categories labial, alveolar, and velar, respectively, because questioning of subjects after the experiment revealed that these responses were intended to denote breathy utterances of /b/, /d/, and /g/.

From Table 3, it is clear that there was no bias toward an alveolar percept influencing subjects' visual perception of the /gɑ/ face. In fact, when viewing the /gɑ/ face, subjects more often perceived a velar consonant (reported 61% of the time) than an alveolar consonant (reported 34% of the time). Whether we examine the percentage of /g/ and /d/ percepts, or the percentage of velar and alveolar percepts, the results from this study are in

Table 3: Perception of place of articulation during video only test.
For each stimulus, the most frequently perceived place of articulation is shown in bold font. The percentage of /g/ and /d/ responses to the /g/ and /d/ faces are also shown in parentheses.

visual stimulus			Response percentages			
syllable type	vowel	consonant	labial	alveolar	velar	other
consonant-vowel	i	b	**85**	1	2	12
		d	0	10 (d 4)	**75** (g 55)	15
		g	0	12 (d 6)	**74** (g 40)	14
	ɑ	b	**98**	0	0	2
		d	0	**51** (d 38)	32 (g 22)	17
		g	2	34 (d 25)	**61** (g 41)	3
	u	b	**95**	0	0	5
		d	2	25 (d 15)	**46** (g 37)	27
		g	3	13 (d 8)	**71** (g 45)	13
vowel-consonant	i	b	**99**	0	0	1
		d	1	**52** (d 39)	31 (g 18)	16
		g	5	**62** (d 41)	27 (g 20)	6
	ɑ	b	**100**	0	0	0
		d	0	**57** (d 32)	42 (g 35)	1
		g	0	**50** (d 30)	**50** (g 44)	0
	u	b	**99**	1	0	0
		d	0	**77** (d 43)	22 (g 21)	1
		g	6	**58** (d 33)	34 (g 33)	2

contrast with Massaro (1998a), who found that subjects more often reported their percept of the silent video /gɑ/ to be the alveolar consonant /d/ than velar consonant /g/. There are several possible explanations for the differences in these studies. The use of an open response paradigm, and a slightly different vowel context (/ɑ/ vs /a/), may have influenced subjects' place perception in this experiment. Also, the speakers in the two studies may have different speech patterns. Finally, the video clips used in the two experiments might introduce different phonetic biases. For example, the video clips used in this experiment showed the neck and throat, which could be important for visually inducing a /g/ percept.

These results do not indicate the presence of a linguistic bias or any other bias influencing visual perception in the /ɑ/ vowel context . However, there does appear to be some bias in the /i/ and /u/ contexts. More specifically, subjects more frequently perceived a /g/ than a /d/ when viewing silent videos of the CV syllables /gi/, /di/, /gu/, or /du/. The reverse pattern was found for VC syllables. That is, subjects more frequently perceived a /d/ than a /g/ when viewing silent videos of the syllables /ig/, /id/, /ug/, /ud/. These findings also hold for the more general place of articulation categories: velar percepts were more common for CV syllables, and alveolar percepts were more common for VC syllables.

To investigate whether or not these effects were significant, a three factor, within-subjects ANOVA was performed with the factors syllable type (CV or VC), vowel context (/i/,/ɑ/ or /u/), and consonant viewed (/g/ or /d/). The dependent variable used was the difference between the number of velar responses and the number of alveolar responses. The results of this ANOVA are presented in Table 4. Syllable type was highly significant in influencing velar/alveolar perception ($p = 0.0009$), and there was a significant interaction between syllable type and vowel context ($p = 0.0122$). This is consistent with the previous observation that the /ɑ/ vowel context does not appear to share the biases which arise in the /i/ and /u/ contexts. Paired t-tests confirm that syllable type was a significant factor influencing alveolar-velar perception in the /i/ and /u/ contexts ($p \leq 0.0001$ in both contexts) and not a significant factor in the /ɑ/ context ($p = 0.5104$)[4].

Table 4: ANOVA results for velar/alveolar perception in visual-only experiment

Source	DF	Sums of Squares	Mean Square	F-Ratio	P-value
constant	1	34.133	34.133	0.314	0.5889
subject	9	978.367	108.707		
syllable type	1	1104.130	1104.130	23.696	0.0009
sub*syl	9	419.367	46.596		
consonant viewed	1	104.533	104.533	14.941	0.0038
sub*con	9	62.967	6.996		
syl*con	1	19.200	19.200	6.830	0.0281
sub*syl*con	9	25.300	2.811		

4.A case could be made for using a pooled T-test here rather than a paired T-test, as CV syllables and their VC counterparts are very different utterances. For this reason, pooled T-tests were also run and the results were similar to those found using the paired T-tests. That is, using pooled T-test, syllable type was a significant factor in the /i/ and /u/ contexts ($p \leq 0.0001$ in both contexts), but was not significant in the /ɑ/ context ($p = 0.5907$). Results from the paired T-test are presented in the text in order to maintain consistency with other analyses done in this section for which a paired T-test was more appropriate than a pooled T-test.

Table 4: ANOVA results for velar/alveolar perception in visual-only experiment

Source	DF	Sums of Squares	Mean Square	F-Ratio	P-value
vowel	2	93.117	46.558	1.086	0.3588
sub*vow	18	771.883	42.882		
syl*vow	2	370.417	185.208	5.683	0.0122
sub*syl*vow	18	586.583	32.588		
con*vow	2	111.317	55.658	17.675	≤ 0.0001
sub*con*vow	18	56.683	3.149		
syl*con*vow	2	8.750	4.375	0.4297	0.6572
sub*syl*con*vow	18	183.250	10.181		
total	119	4895.870			

Using CV syllables in the /a/ vowel context, Hampson, Guenther, and Cohen (1998) found that the visual influences of alveolar and velar consonants on speech perception were different. Similar results were found here. The consonant viewed influenced perception ($p=0.0038$). There was, however, a strong interaction between consonant viewed and vowel context ($p \leq 0.0001$). Paired t-tests showed a significant influence of the visual consonant in the /α/ and /u/ vowel context ($p = 0.0019$ and $p \leq 0.0001$, respectively) but not in the /i/ vowel context ($p = 0.3007$). Apparently subjects were able to extract some information regarding whether the consonant viewed was velar or alveolar for consonants presented in the /α/ and /u/ vowel contexts, but not in the /i/ vowel context. It may be that the visual discriminability of /g/ and /d/ changes across vowel contexts, and that these consonants are less easily discriminable in the /i/ vowel context. Another possibility is that the availability of visual information pertinent to alveolar-velar discrimination depends ideosyncratically on the stimuli used. For example, perhaps the /gi/, /di/, /ig/ and /id/ videos used did not happen to capture distinct information specifying place of articulation, while videos of other speakers pronouncing the same syllables would provide more information about whether utterances were velar or alveolar.

Finally, an interaction was found between syllable type and the visual consonant viewed ($p = 0.0281$). This appears to be because the bias to perceive velar consonants when viewing CV syllables and alveolar consonants when viewing VC syllables was slightly more pronounced when viewing a /g/ face than a /d/ face.

One of the most noteworthy aspects of these statistical findings is that the syllable type had a stronger influence on alveolar-velar perception than the actual place of articulation of the consonant viewed. It would be of interest to see if these results replicate with different

videos, or if they are ideosyncratic to the stimuli used here. In any case, they are pertinent to the audio-visual experiment described below, which uses these same visual stimuli.

2.5 Discussion

The results from these unimodal tests establish the perceptual clarity of the acoustic and visual stimuli used for the audio-visual tests described below. They also provide information regarding unimodal perceptual biases. In particular, the visual stimuli for the syllables /gu/, /du/, /gi/, and /di/ were found to more frequently induce a /g/ percept than a /d/ percept, and their VC counterparts (/ug/, /ud/, /ig/, and /id/) were found to elicit the reverse bias. It is not clear whether these biases are linguistic in nature or not. They may, for example, arise from ideosyncracies in the video clips used. In any case, if such biases play a critical role in McGurk fusions (as posited by Massaro, 1998a), then the stimuli for the /i/ and /u/ vowel contexts should elicit different McGurk effects in CV and VC syllables. More specifically, more /g/ percepts than /d/ percepts should arise in the /i/ and /u/ vowel contexts when /bV/ syllables are dubbed onto /gV/ or /dV/ syllables. In VC syllables, the reverse is predicted: more /d/ percepts than /g/ percepts. Predictions for the /ɑ/ vowel context are simply that subjects will be biased toward a velar percept when viewing a velar utterance and toward an alveolar percept when viewing an alveolar utterance. The degree to which these predictions are satisfied in the following audio-visual experiment will be indicative of the importance of unimodal visual biases in the perception of audio-visually incongruent stimuli.

3. Experiment 2: Perception of Audio-Visual Stimuli

3.1 Stimuli

The audio-visual stimuli for this experiment were created by dubbing the auditory stimuli tested in Experiment 1 onto the silent video clips tested in Experiment 1. The burst in the acoustic syllable was always aligned with the release of the speaker's stop consonant in the visual stimulus. Temporal alignment of the audio track on the video was done by hand. Adjustments could be made with a 33 ms precision. In all audio-visual clips, the vowel and syllable type were consistent across modalities. All nine possible permutations of audio-visual consonant combinations (auditory /b/, /g/, and /d/ crossed with visual /b/, /g/, and /d/) were created for each vowel context and each syllable type. This resulted in a total of twenty-seven consonant-vowel video clips, and twenty-seven vowel-consonant video clips.

3.2 Subjects

Sixteen adult subjects who had not participated in Experiment 1 were recruited by flyers placed around the Boston University campus. They all had English as their first language, and normal or corrected-to-normal vision. None of the subjects reported any history of a speech or hearing disorder.

3.3 Procedure

Subjects were tested alone in a dimly-lit room. They were seated approximately 2 feet in front of a computer monitor with speakers on either side, and a keyboard in front of them. The experiment was response-paced, with a new video being played immediately following the previous response.

The CV syllables and VC syllables were played in separate blocks. Half the subjects were exposed to the CV syllable set first, and the other half were exposed to the VC syllable set first. Within each block, the twenty-seven different video clips were played in a random order. Each clip appeared three times within the block. Subjects were directed to report what consonant sounds they heard after each clip was played, by typing in the letter, or the set of letters, which best represented their percept, and pressing the "enter" key. It was emphasized that subjects should watch the video clips throughout, but should always report what they heard.

3.4 Results

Table 5: Percentage responses made to the CV audio-visual stimuli
Numbers are rounded off to the nearest percent.

acoustic stimulus		visual stimulus		
vowel	consonant	b	d	g
i	b	b 100	**d 88, g 6**, gd 2 ? 4	**d 85, g 4, b 4** t 2, f 2, ? 2
	d	d 48, bd 17 b 31, db 2 pd 2	d 100	d 96, gd 4
	g	g 42, bg 23 b 15, gb 2 p 8, bkg 4 k 2,? 2, pb 2	g 98, ? 2	g 100
α	b	b 100	**d 25, g 19, b 40** v 4, **th 2**, f 2 dg 2, bf 2, ? 4	**d 15, g 46 b 25, th 2** gd 2, f 2, ? 8
	d	d 50, bd 25 b 19, db 6	d 100	d 100
	g	g 79, bg 17 b 2, gb 2	g 100	g 100
u	b	b 100	**d 6, g 77, b 10** ? 6	**d 2, g 79 b 17**, ? 2
	d	d 54, bd 25 b 19, db 2	d 100	d 100
	g	g 65, bg 29 b 4, gb 2	g 100	g 100

A summary of the results from the CV syllable set is provided in Table 5[5] (individual data are available in Appendix C of Hampson, 1999). Results pertinent to the following discussion are highlighted in bold font.

This paper is concerned with McGurk fusions, and will therefore focus on the subjects' perceptions of acoustic /b/-visual /d/ and acoustic /b/-visual /g/ stimuli across the different

5.The responses "gh", "dh", and "bh" were treated as "g", "d", and "b" responses, respectively, since they were reported by subjects to represent breathy examples of the stop consonants /g/, /d/, and /b/.

vowel contexts (for discussion of combination percepts, see Part II). In Panel a of Figure 1 the percentage of the time that subjects responded /d/ to the stimuli in the CV syllable set is plotted as a function of vowel context. In agreement with the previous findings of Green and colleagues (Green, Kuhl and Meltzoff, 1988; Green, Kuhl, Meltzoff and Stevens, 1991; Green and Gerdeman, 1995; Green and Norrix, 1997), acoustic /b/-visual /g/ stimuli elicited a decreasing number of /d/ percepts as vowel context was changed from /i/ to /ɑ/ to /u/. Additionally, the acoustic /b/- visual /d/ stimuli elicited a similar response pattern. That is, the frequency of /d/ responses to acoustic /b/-visual /d/ stimuli decreased as vowel context was shifted from /i/ to /ɑ/ to /u/, as shown by the dashed line in Panel a.

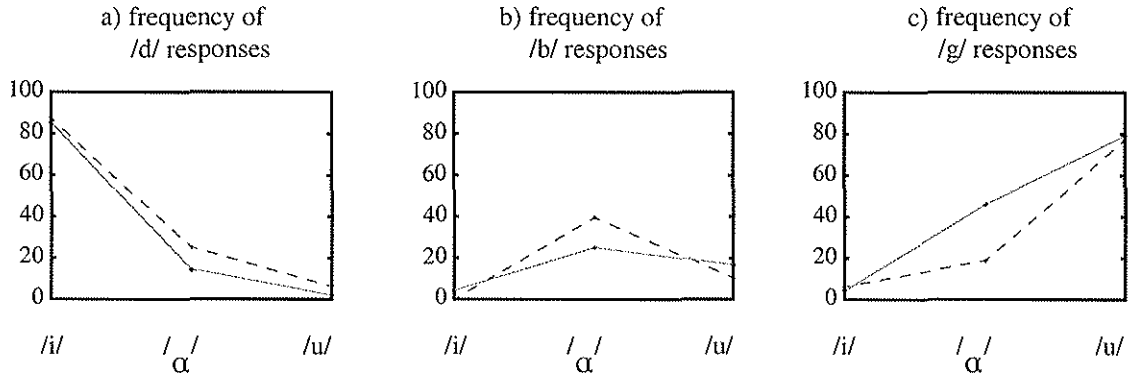


Figure 1. Response patterns for the CV syllable set. These graphs show how responses to acoustic /b/-visual /g/ stimuli (solid lines) and acoustic /b/-visual /d/ stimuli (dashed lines) change across vowel contexts. Panel a illustrates the frequency of /d/ responses across vowel contexts, Panel b illustrates the frequency of /b/ responses across vowel contexts, and Panel c illustrates the frequency of /g/ responses across vowel contexts.

However, in contrast with the findings of Green, Kuhl, Meltzoff and Stevens (1991), Green and Gerdeman (1995), and Green and Norrix (1997), the decrease in /d/ percepts as vowel context was changed from /i/ to /ɑ/ was not accompanied by a notable increase in /ð/ percepts. In fact, "th" responses in this study never amounted to more than 2% of the responses in any context. For this reason, the frequency of /ð/ percepts is not plotted in Figure 1. The findings of this study also differ from those of Green, Kuhl and Meltzoff (1988), who found a decrease in the magnitude of the McGurk effect as the vowel context was changed from /i/ to /ɑ/ to /u/. That is, their stimuli induced a steadily increasing frequency of /b/ percepts as the vowel context was shifted from /i/ to /ɑ/ to /u/ (see Green, 1996, for more details). Panel b of Figure 1 illustrates the frequency of /b/ percepts induced by the CV stimuli of this study. Although the frequency of /b/ percepts did increase in the /ɑ/ context relative to the /i/ context, it decreased again in the /u/ context. Rather than a decrease in the magnitude of the McGurk effect or an increase in the frequency of /ð/ percepts as vowel context was shifted from /i/ to /ɑ/ to /u/, the dominant trend found in this study was an increase in /g/ percepts. This is illustrated in Panel c of Figure 1.

Although this increase in /g/ percepts is different from previous findings of Green and colleagues (Green, Kuhl and Meltzoff, 1988; Green, Kuhl, Meltzoff and Stevens, 1991;

Green and Gerdeman, 1995; Green and Norrix, 1997), it is not necessarily incompatible with their findings from a theoretical perspective. Green (1996) suggested that the findings of Green, Kuhl, Meltzoff and Stevens (1991) and Green and Gerdeman (1995) were a result of the different second formant (F2) patterns of /d/, /b/ and /ð/ in the different vowel contexts. He noted that F2 is rising for both /d/ and /b/ in the /i/ vowel context which may allow these consonants to be easily confused in that context. In the /ɑ/ vowel context, however, the F2 transition is falling for /d/ and rising for /b/. Therefore, the acoustic stimulus /bɑ/ may be sufficiently different from /dɑ/ to prevent an acoustic /bɑ/-visual /gɑ/ stimulus from being mistaken for /dɑ/. However, the second formant pattern for /ðɑ/ is generally more similar to that of /bɑ/ (than the second formant pattern of /dɑ/ is) in that it has a flat or even somewhat rising transition. Green(1996) suggested that this may have resulted in a higher frequency of /ðɑ/ than /dɑ/ fusion percepts in the studies of Green, Kuhl, Meltzoff and Stevens (1991) and Green and Gerdeman (1995). A similar explanation, based on second formant patterns, may be applied to our findings.

A schematic diagram of typical second formant patterns of /bV/, /dV/ and /gV/ syllables across vowel contexts is shown in Figure 2. In the /i/ vowel context, the second formant patterns for /b/ and /d/ are qualitatively similar (rising transitions) while the second formant pattern for /g/ is very different (a falling transition). Subjects who are exposed to an acoustic /bi/ stimulus dubbed onto a face enunciating /gi/ or /di/ may tend to perceive /di/ because it is acoustically similar to /bi/ and visually similar to the syllable viewed. In the /u/ vowel context, however, /b/ generally has a second formant pattern more similar to /g/ than to /d/[6]. In the /u/ vowel context, therefore, it is not surprising that an acoustic /b/-visual /g/ or /d/ stimulus would be more likely to be perceived as /g/ (which has an F2 pattern somewhat similar to that of the acoustic stimulus) than /d/ (which, at least in terms of second formant patterns, is extremely different from the acoustic stimulus). The situation in the /ɑ/ vowel context is somewhere between these two. It appears that /b/ typically has a second formant pattern more similar to /d/ than /g/ in the /ɑ/ context. However, it is important to note that for the very similar vowel /ɔ/ the opposite relationship holds, the second formant pattern for /bɔ/ is more similar to /gɔ/ than to /dɔ/. It appears that the relationships between the F2 transitions of these three consonants are changing rapidly around the /ɑ/ vowel context. While several studies have found that acoustic /bɑ/-visual /dɑ/ or /gɑ/ stimuli elicit mostly /dɑ/ percepts (e.g., Green and Norrix, 1997), some studies, including this one, have found a higher frequency of /gɑ/ percepts in response to these stimuli (see also Munhall, Gribble, Sacco and Ward, 1996)[7].

6. Both /gu/ and /du/ have falling F2 transitions, in contrast to the rising transition of /bu/, but the transition for /gu/ is very shallow and in that respect is similar to the transition for /bu/. The F2 transition for /du/, on the other hand, falls very sharply and is thus dramatically different from the slightly rising transition of /bu/.

7. Note the distinction between the /a/ vowel context which has been used in many McGurk experiments including the original study by McGurk and MacDonald (1976) and which generally elicits a high frequency of /d/ fusion percepts, and the /ɑ/ vowel context which was used in this experiment, and which has been used in several other North American studies including those of Green and Norrix (1997) and Munhall et al. (1996). The reason for using the /ɑ/ vowel context here, rather than using the /a/ context chosen by McGurk and MacDonald (1976) is that the vowel /a/ does not exist in most English dialects, including the dialect of the region where this study was conducted.

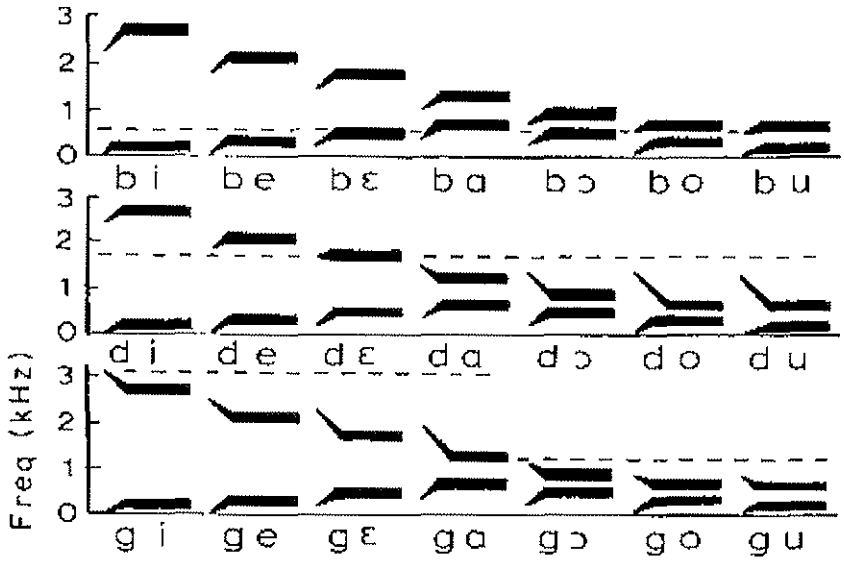


Figure 2. First and second formant patterns from Delattre, Liberman, and Cooper (1955). This figure was adapted by Kent and Read (1992) and has been reprinted here with permission by Singular Publishing Group, Inc.

Assuming that the changes in fusion percepts across vowel context can be attributed to differences in the second formant patterns in the different contexts, there remains the issue of why some studies (Green, Kuhl, Meltzoff and Stevens, 1991; Green and Gerdeman, 1995; Green and Norrix, 1997) have found an increase in /ð/ percepts as the vowel context was changed from /i/ to /ɑ/ (or /a/), while other studies (including the current study and that of Green, Kuhl, and Meltzoff, 1988) have not. Perhaps the magnitude of the release burst of the acoustic /bɑ/ stimulus plays a role. The acoustic stimuli in this study were chosen for their perceptual clarity, and as a result, they had pronounced release bursts. This may have prevented subjects from confusing the acoustic /b/ stimulus with the fricative /ð/, regardless of how similar the second formant patterns of these consonants were.

If changes in the McGurk fusion phenomenon across vowel contexts arise from changes in the second formant patterns across contexts, what sort of response patterns should occur in the VC syllable set? Given that the formant patterns for VC syllables are generally similar to those of the corresponding CV syllables reversed in time (Olive, Greenwood and Coleman, 1993), we would expect similar fusion percepts to occur across syllable sets. In other words, based on results from the CV syllable set, it is expected that acoustic /Vb/-visual /Vg/ or /Vd/ stimuli will produce a decreasing frequency of /d/ responses and an increasing frequency of /g/ responses as vowel context is changed from /i/ to /ɑ/ to /u/.

The results for the VC syllable set are provided in Table 6 (individual data are available in Hampson, 1999). The frequencies of /b/, /d/, and /g/ percepts in response to acoustic /Vb/-visual /Vg/ or /Vd/ stimuli in the three different vowel contexts are plotted in Figure 3. As seen in the CV syllable set, the frequency of /d/ fusion percepts to the VC stimuli drops dramatically as the vowel context is changed from /i/ to /ɑ/ to /u/ (shown in Panel a of Figure 3). However, the trend of increasing /g/ percepts found in the CV syllable set is not reproduced in the VC syllable set (see Panel c of Figure 3). In particular, in the /u/

Table 6: Percentage responses made to the VC audio-visual stimuli
Numbers are rounded off to the nearest percent.

acoustic stimulus		visual stimulus		
vowel	consonant	b	d	g
i	b	b 90, db 8 bd 2	**d 54, b 44,** db 2	**d 48, g 2,b 44,** bd 2, dg 2, ? 2
	d	d 40, db 40 b 12, bd 8	d 100	d 100
	g	g 46, gb 31 b 4, bg 15 gdb 2, ? 2	g 96, gd 2, b 2	g 100
ɑ	b	b 98, db 2	**d 25, g 19, b 44** gd 6, dg 4, ? 2	**d 13, g 42** **b 31**,dg 4,gb 2 gd 2, ? 6
	d	b 38, db 31 d 21, bd 8 n 2	d 98, g 2	d 96, g 2, n 2
	g	g 46, gb 38 b 8, bg 6 gd 2	g 100	g 100
u	b	b 100	**d 4, g 17, b 69** gb 6, bg 2, v 2	**d 2, g 17** **b 79**, gb 2
	d	d 50, db 42 b 4, bd 2 dp 2	d 100	d 100
	g	g 48, gb 40 b 8, bg 4	g 100	g 100

vowel context, the number of /g/ percepts is very low (17%). This can be attributed to the weak overall McGurk effect which occurred in this context. As shown in Panel b of Figure 3, the frequency of /b/ percepts was very high in the /u/ context. In fact, subjects' percepts were dominated by the acoustic input an average of 74% of the time when they were exposed to the acoustic /ub/-visual /ug/ and acoustic /ub/ - visual /ud/ stimuli, implying a very weak McGurk effect to these stimuli. It is not clear why the McGurk effect in the /uC/ context was so weak, while the McGurk effect in all other contexts was substantial. It may be particularly difficult to induce a McGurk effect in the /u/ vowel context. Although the /Cu/ stimuli of this experiment did produce a significant McGurk effect, Green, Kuhl, and Meltzoff (1988) reported almost no McGurk effect in response to /Cu/ stimuli. One problem with inducing a McGurk effect in the /u/ vowel context may be that the lip-rounding in this context is visually similar to a labial closure and the visual stimulus is thus somewhat compatible with the consonant /b/.

Regardless of the magnitude of the effect across contexts, it is of interest for the purposes of this discussion to examine the quality of the McGurk effect, and how that changes

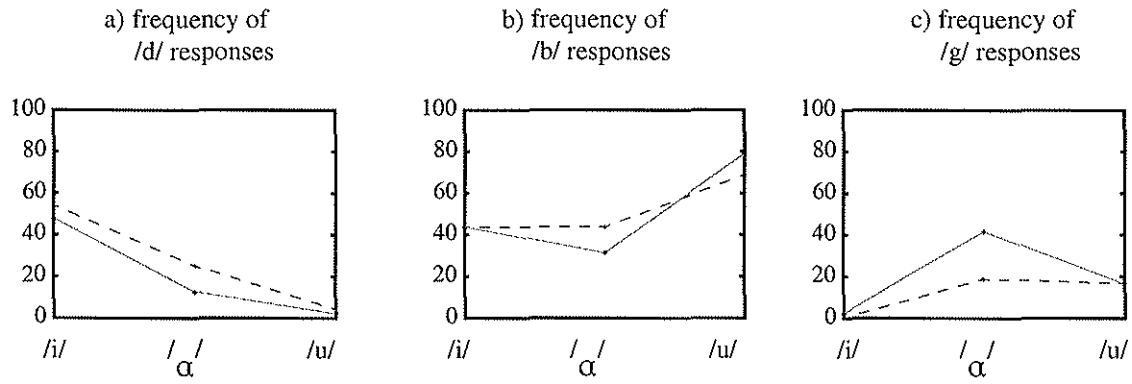
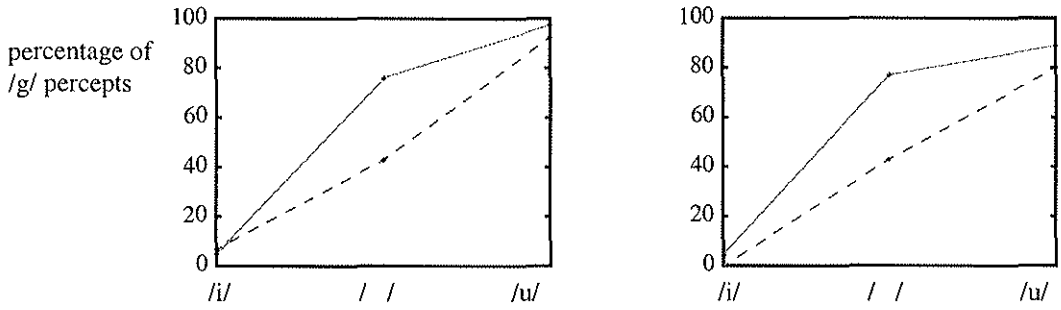


Figure 3. Response patterns for the VC syllable set. These graphs show how responses to acoustic /b/-visual /g/ stimuli (solid lines) and acoustic /b/-visual /d/ stimuli (dashed lines) change across vowel contexts. Panel a illustrates the frequency of /d/ responses across vowel contexts, Panel b illustrates the frequency of /b/ responses across vowel contexts, and Panel c illustrates the frequency of /g/ responses across vowel contexts.

across vowel context. In order to examine the qualitative nature of the effect across contexts, in a manner which is independent of changes in the magnitude of the effect, the data have been normalized based on the total number of /d/ and /g/ illusory responses occurring in each context (note that no other illusory percept, including /ð/, ever occurred more than 6% of the time in any context). The normalized data are presented in Figure 4. This figure depicts the frequency of /g/ responses relative to the frequency of /d/ responses across vowel contexts. For example, for the auditory /bi/ - visual /di/ stimulus, the percentage of /g/ responses (6.25 - the data in Table 5 is rounded to the nearest percent) is divided by the percentage of responses which were either a /d/ or a /g/ (87.5 + 6.25 = 93.75) to yield the relative percentage of /g/ responses ($6.\dot{6}$ %), which is the starting position of the dashed line in panel a. The percentage of /d/ percepts relative to /g/ percepts in each context is implicitly represented Figure 4. For example, the relative percentage of /d/ percepts in response to auditory /bi/ -visual /di/ stimuli is $100.0 - 6.\dot{6} = 93.\dot{3}$ % .

Examining Figure 4, it is apparent that the *relative* frequencies of /g/ and /d/ percepts changed dramatically across the different vowel contexts in a systematic manner. The percentage of illusory percepts which were /g/ increased steadily from /i/ to /α/ to /u/ (and the percentage of illusory percepts which were /d/ decreased steadily). Importantly, this trend held for both syllable types (panels a and b), and for both visual conditions (denoted by the solid and dashed lines) as vowel context was changed from /i/ to /a/ to /u/. In fact, these changes led to a qualitatively different McGurk "fusion" in the /u/ vowel context. The traditional McGurk fusion involves acoustic /b/-visual /g/ stimuli inducing /d/ percepts. This study found, in the /u/ vowel context, that acoustic /b/-visual /d/ stimuli gave rise to /g/ percepts.

To verify that vowel context had a significant effect on the qualitative nature of the McGurk fusion effect, a three factor ANOVA was performed with the factors vowel context (/i/, /α/, or /u/), syllable type (CV or VC), and visual stimulus (/g/ face or /d/ face). The dependent variable used was the percentage of total /d/ and /g/ responses which were



(a) CV syllable set (b) VC syllable set

Figure 4. Percentage of total /g/ or /d/ responses which were /g/ in the CV syllable set (panel a) and VC syllable set (panel b). Responses to acoustic /b/-visual /g/ stimuli are indicated with solid lines and responses to acoustic /b/-visual /d/ stimuli are indicated by dashed lines.

/g/[8]. The results of the ANOVA are provided in Table 7. Vowel context is easily significant at the $p = 0.05$ level ($p = 0.0002$).

Table 7: ANOVA results for fusion data

Source	DF	Sums of Squares	Mean Square	F-Ratio	P-value
constant	1	31294.500	31294.500	18986.00	≤ 0.0001
syllable	1	61.369	61.369	37.232	0.0258
vowel	2	15174.700	7587.350	4603.100	0.0002
syl*vow	2	62.442	31.221	18.941	0.0501
visual	1	574.731	574.731	348.680	0.0029
syl*vis	1	10.446	10.446	6.338	0.1282
vow*vis	2	598.832	299.416	181.65	0.0055
error	2	3.297	1.648		
total	11	16485.800			

Syllable type was also a significant factor ($p=0.0258$), reflecting the fact that /g/ percepts were less frequent in the VC syllable set than in the CV syllable set (even after normalizing for the different magnitudes of the effect in different contexts). This decrease in /g/ percepts in the VC syllable set is compatible with the response biases found in the unimodal visual data (see Section 2). That is, when viewing silent videos of /gV/ and /dV/ utter-

8. This dependent variable was used, rather than the dependent variable used in analyzing data from the visual-only experiment (% alveolar responses - % velar responses), because it normalizes for differences in the magnitude of the effect across contexts.

ances, people tended to have velar percepts more often in the CV syllable set, and alveolar percepts more often in the VC syllable set. This effect was dependent on vowel context (it was limited to the /i/ and /u/ vowel contexts, with no significant effect of syllable type in the /ɑ/ context). Similarly, in the audio-visual experiment, the influence of syllable type across vowel contexts is approaching significance (the interaction between syllable type and vowel has a significance level p=0.0501). Unfortunately, it is not possible to do T-tests to rigorously establish the nature of this interaction, as we do not have enough data per subject in this case. However, there does appear to be a pattern similar to that seen in the data from the unimodal visual tests, although much less pronounced. That is, in both the /i/ and /u/ vowel contexts, there is a slightly lower frequency of /g/ percepts in the VC syllable set (2% and 84% respectively) than in the CV syllable set (6% and 95% respectively), while the response frequencies in the /ɑ/ vowel context are similar for the two syllable types (60% and 59% in the VC and CV syllable sets, respectively). It must be noted that these small differences across syllable sets are overlaid on a much larger pattern of change in response which is similar across syllable sets (i.e. the increase in /g/ responses relative to /d/ responses as vowel context is changed from /i/ to /a/ to /u/).

Another aspect of this fusion data which reflects response biases found in the unimodal visual tests is the influence of the visually presented consonant. This factor is significant ($p = 0.0029$) and there is a significant interaction between the visually presented consonant and the vowel context ($p = 0.0055$). Examination of Figure 4 reveals that the pattern of this interaction is similar to that seen in the unimodal visual tests. That is, the influence of the visual consonant (as indicated by the difference between the dashed and solid lines in Figure 4) seems to be stronger in the /ɑ/ and /u/ contexts than the /i/ context. However, in the audio-visual experiments, the influence of the visual consonant is much stronger in the /ɑ/ context than the /u/ context, which was not seen in the unimodal visual results (where both contexts showed a strong influence of visual consonant on alveolar-velar perception). With this one exception, response patterns to the audio-visual stimuli reflect response patterns found in the unimodal visual tests, albeit to a very diminished degree. In summary, it appears that visual biases do play a role in the audio-visual fusion results, as suggested by Massaro (1998a). However, their influence is small in comparison to the large differences found in the effect across vowel contexts.

As noted earlier, differences in the nature of McGurk fusions across vowel context are qualitatively similar for both types of visual stimuli, in both the CV and VC syllable sets. Although there is not enough data per subject to analyze vowel effects independently in each syllable set and for each of the visual conditions, the trend of increasing /g/ fusion percepts (relative to /d/ percepts) as vowel context was changed from /i/ to /a/ to /u/ was pronounced in all four cases.

3.5 Discussion

This experiment did not replicate the findings of Green, Kuhl, and Meltzoff (1988) as the magnitude of the McGurk effect in the CV syllable set did not decrease substantially as vowel context was changed from /i/ to /a/ to /u/. However, a trend similar to that reported by Green, Kuhl, and Meltzoff (1988) was found in the VC syllable set. That is, a decreasing number of McGurk fusions were found to occur in response to VC stimuli as the vowel

context was changed from /i/ to /a/ to /u/. It appears that there are several factors involved in determining the magnitude of the McGurk fusion effect. Perhaps vowel context does play a role, but other variables (such as the particular visual stimuli used, for example) must also be important.

Although this study found differences in the magnitude of the McGurk effect across syllable types, similar qualitative changes in the effect appeared in both the CV and VC syllable sets. That is, of the illusory percepts which did occur, a decreasing proportion of them were /d/ percepts and an increasing proportion of them were /g/ percepts as vowel context was changed from /i/ to /a/ to /u/. Not only did this pattern hold for both syllable types, but it held regardless of whether the visual stimulus used was /g/ or /d/, and it was very pronounced in all cases (see Figure 4).

4. General Discussion

The major finding of this paper was that the nature of McGurk "fusions" varied systematically with vowel context. The number of /g/ responses relative to the number of /d/ responses increased dramatically as vowel context was changed from /i/ to /a/ to /u/. In fact, the acoustic /bu/-visual /du/ stimulus induced a /gu/ percept 77% of the time. This finding is troublesome for theories of motor-based speech perception, as the syllable /gu/ is not an articulatory compromise between /du/ and /bu/. The gestures of all three articulations are very different, both in terms of musculature involved and in terms of the locations of the major constrictions in the vocal tract. In fact, the velar constriction formed when pronouncing /gu/ is even farther back in the vocal tract than the alveolar constriction formed when pronouncing /du/. In this sense, /gu/ is less similar to /bu/ (which in volves a labial constriction) than /du/ is. The tendency for subjects to perceive /gu/ when presented with an acoustic /bu/-visual /du/ stimulus is thus very difficult to explain from a motoric perspective.

The phenomenon is more easily explained by a sensory-based theory of speech perception, since in terms of an important sensory dimension, the slope of the second formant frequency transition, /gu/ is generally midway between /bu/ and /du/. That is, /gu/ is more similar visually to the stimulus /du/ than /bu/ is, and /gu/ is *generally* more similar acoustically to the stimulus /bu/ than /du/ is, at least in terms of the second formant transitions[9]. Therefore, the percept /gu/ is usually a good compromise between what is being seen and what is being heard when an acoustic /bu/-visual /gu/ or /du/ stimulus is presented. In general, the changes found in the McGurk effect across vowel contexts are compatible with a sensory model of speech perception in which the second formant pattern plays an important role.

An alternative explanation for context-dependent changes in the McGurk effect is that the effect is caused by linguistic biases that change across vowel contexts. The data from the visual-only test (Experiment 1) do reveal certain biases in the visual perception of these

9. The word generally is emphasized here because the second formant relationships between /b/, /d/ and /g/ are not always the same as those shown in Figure 2.

stimuli, which could reflect linguistic expectations; however, these biases cannot explain the response patterns seen in the audio-visual test. For example, silent videos of /gi/ and /di/ elicited /d/ percepts only 5% of the time and alveolar percepts of any sort only 11% of the time. The responses to these unimodal visual stimuli were strongly biased towards velar percepts in general (75% of the responses), and /g/ percepts in particular (47% of all responses). In contrast, when these videos were dubbed with the sound /bi/ in Experiment 2, they nearly always elicited /d/ percepts (87% of the time), and rarely resulted in /g/ percepts (only 5% of the time). In general, the biases seen in the visual-only experiment cannot explain the frequency of different audio-visual "fusion" percepts. This does not imply that they do not play a role in audio-visual perception, only that they are not the major determinants of /g/-/d/ response frequencies.

The data from Experiments 1 and 2 do suggest that biases present in unimodal visual perception (which could be linguistic in nature, but are not necessarily) may play a secondary role in audio-visual illusions. The tendency for subjects to respond /g/ to the silent videos of /gi/, /di/, /gu/, and /du/, and to respond /d/ to the silent videos of /ig/, /id/, /ug/, and /ud/ was also seen in the data from the audio-visual experiment (although this tendency was much less pronounced in the latter case). In Experiment 2 /d/ responses increased relative to /g/ responses as the syllable type was changed from CV to VC. In the /i/ vowel context, the relative percentage of /d/ responses increased from 94% to 98%, and in the /u/ vowel context it increased from 5% to 16%. In the /ɑ/ vowel context, syllable type did not affect /g/-/d/ perception (approximately 40% of /g/ or /d/ responses were /d/s in both the CV and VC syllable sets). In this context, subjects were more likely to respond with /d/ than /g/ when they were viewing the /d/ face, and more likely to respond with /g/ than /d/ when viewing the /g/ face, which was also consistent with the response patterns from the visual-only experiment.

Although an emphasis on second formant patterns was maintained throughout this analysis (for the sake of simplicity), there are clearly other acoustic features, such as the third formant frequency, which are important in unimodal auditory perception. Kewley-Port (1982) found that the second formant onset frequencies of the three phonemes /b/, /d/, and /g/ were not statistically separable in all vowel contexts. However, when both the second and third formant onset frequencies were considered, these phonemes could be distinguished within a given vowel context. Figure 5 below was taken from Kewley-Port (1982). This image illustrates the relationships of these three phonemes in the different vowel contexts, in an acoustic space defined by the second and third formant onset frequencies. This two-dimensional acoustic space makes similar predictions to the simple one-dimensional space considered in this paper (which is based on the slope of the second formant transition) regarding perception of consonants in the /u/, /ɑ/ and /i/ vowel contexts. In the /i/ vowel context, /d/ is more similar to /b/ than /g/ is, and in the /u/ vowel context, it is /g/ which is more similar to /b/. The /ɑ/ vowel context is more complex, and relationships between the consonants will depend on the tokens chosen. Given two /bɑ/ tokens, one with a high F3 onset frequency and one with a lower F3 onset frequency, will the token with lower F3 onset produce more /g/ percepts when these acoustic tokens are dubbed on videos of /gɑ/ and /dɑ/ utterances? If so, this would provide evidence for the importance of F3 in the perception of these consonants. If audio-visual speech is a form of sensory pattern recognition, as the results of Experiment 2 seem to imply, then studies

of bimodal speech perception should provide information regarding unimodal speech perception, and studies of unimodal speech perception should provide information regarding bimodal speech perception.

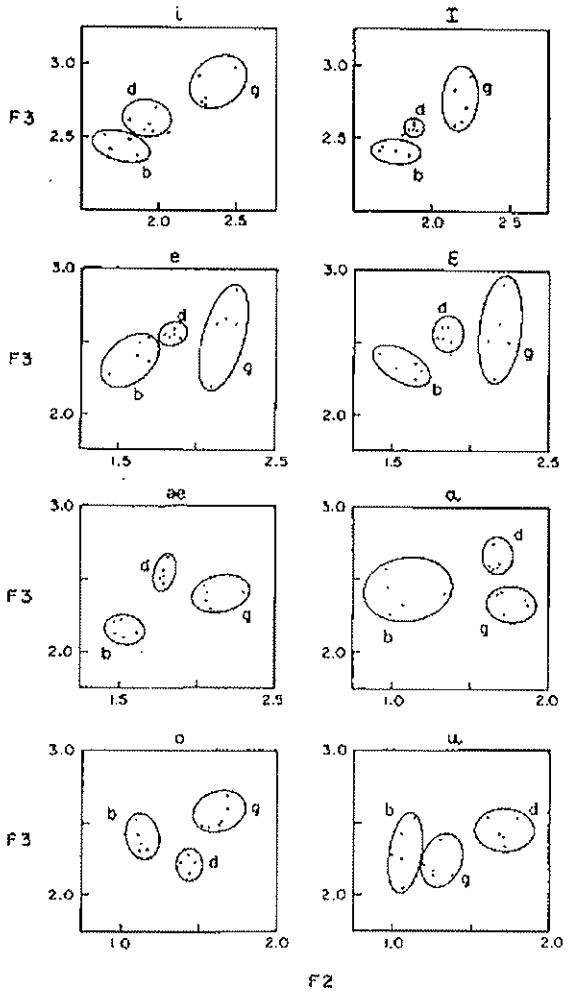


Figure 5. Distributions of /d/, /g/ and /b/ utterances in a space defined by the onset frequencies of the second and third formants. Each panel shows utterances in one vowel context (as indicated above the panel). Frequency is in kHz. This diagram was reprinted with permission from Kewley-Port, 1982.

References

[1] Akmajian, A., Demers, R.A., Farmer, A.K., and Harnish, R.M. (1990). Linguistics, an introduction to language and communication, third edition. The MIT Press, Cambridge, Massachusetts.

[2] Dekle, D.J., Fowler, C.A., and Funnell, M.G. (1992). Audiovisual integration in perception of real words. Perception and Psychophysics, 51(4):355–362.

[3] Delattre, P. and Freeman, D.C. (1968). A dialect study of American r's by x-ray motion picture. Linguistics, 44:29–68.

[4] Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America, 27(4):769–773.

[5] Dodd, B. (1977). The role of vision in the perception of speech. *Perception*, 6:31–40.

[6] Easton, R.D. and Basala, M. (1982). Perceptual dominance during lipreading. Perception and Psychophysics, 32(6):562–570.

[7] Erber, N.P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech and Hearing Research, 12:423–425.

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

[9] Fowler, C.A. and Dekle, D.J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. Journal of Experimental Psychology: Human Perception and Performance, 17(3):816–828.

[10] Fuster-Duran, A. (1996). Perception of conflicting audio-visual speech: An examination across Spanish and German. In Stork, D.G. and Hennecke, M.E., editors, Speechreading by Humans and Machines, pages 135–151. Springer-Verlag.

[11] Green, K. and Norrix, L.W. (1997). Acoustic cues to place of articulation and the McGurk effect: The role of release bursts, aspiration and formant transitions. Journal of Speech, Language, and Hearing Research, 40:646–665.

[12] Green, K.P. (1996). The use of auditory and visual information in phonetic perception. In Stork, D.G. and Hennecke, M.E., editors, Speechreading by Humans and Machines, pages 55–77. Springer-Verlag.

[13] Green, K.P. and Gerdeman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. Journal of Experimental Psychology: Human Perception and Performance, 21(6):1409–1426.

[14] Green, K.P., Kuhl, P.K., and Meltzoff, A.N. (1988). Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. Journal of the Acoustical Society of America, 84:S155.

[15] Green, K.P., Kuhl, P.K., Meltzoff, A.N., and Stevens, E.B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. Perception and Psychophysics, 50(6):524–536.

[16] Green, K.P. and Miller, J.L. (1985). On the role of visual rate information in phonetic perception. Perception and Psychophysics, 38(3):269–276.

[17] Guenther, F.H., Espy-Wilson, C.Y., Boyce, S.E., Matthies, M.L., Zandipour, M., and Perkell, J.S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. Journal of the Acoustical Society of America, in press.

[18] Hampson, M. (199). An investigation of speech reference frames: Modelling and psychophysics. Unpublished doctoral dissertation, Boston University, Boston.

[19] Hampson, M., Guenther, F., and Cohen, M. (1998). Visual influences on the perception of alveolar/velar place discrimination. Journal of the Acoustical Society of America, 104(3),Pt.2:1854.

[20] Jordan, T.R. and Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. Journal of Experimental Psychology: Human Perception and Performance, 23(2):388–403.

[21] Kent, R.D. and Read, C. (1992). The Acoustic Analysis of Speech. Singular Publishing Group, Inc.

[22] Keppel, G. (1991). Design and Analysis: A Researcher's Handbook. Prentice Hall.

[23] Kewley-Port, D. (1982). Measurements of formant transitions in naturally produced stop consonant-vowel syllables. Journal of the Acoustical Society of America, 72(2):379–389.

[24] Ladefoged, P. (1993). A Course in Phonetics. Harcourt Brace College Publishers, third edition.

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

[26] Liberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech revisited. Cognition, 21:1–36.

[27] MacDonald, J. and McGurk, H. (1978). Visual influences on speech perception processes. Perception and Psychophysics, 24(3):253–257.

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

[29] Massaro, D.W. (1998). Illusions and issues in bimodal speech perception. Proceedings paper: Auditory-Visual Speech Processing. Terrigal, New South Wales, Australia.

[30] Massaro, D.W. and Cohen, M.M. (1983). Evaluation and integration of visual information in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 9(5):753–771.

[31] Massaro, D.W. and Cohen, M.M. (1990). Perception of synthesized audible and visible speech. Psychological Science, 1(1):55–63.

[32] Massaro, D.W. and Cohen, M.M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. Speech Communication, 13:127–134.

[33] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. Nature, pages

746–748.

[34] Munhall, K.G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. Perception and Psychophysics, 58(3):351–362.

[35] Munhall, K.G. and Tokhura, Y. (1998). Audiovisual gating and the time course of speech perception. Journal of the Acoustical Society of America, 104(1):530–539.

[36] Olive, J.P., Greenwood, A., and Coleman, J. (1993). Acoustics of American English Speech. Springer.

[37] Ong, D. and Stone, M. (1998). Three-dimensional vocal tract shapes in /r/ and /l/: A study of MRI, ultrasound, electropalatography, and acoustics. Phonoscope, 1:1–13.

[38] Reisberg, D., McLean, J., and Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In Dodd, B. and Campbell, R., editors, Hearing by Eye: The Psychology of Lip-reading, chapter4, pages 97–113. Lawrence Erlbaum Associates Ltd.

[39] Rosenblum, L.D. and Saldana, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology: Human Perception and Performance, 22(2):318–331.

[40] Rosenblum, L.D., Schmuckler, M.A., and Johnson, J.A. (1997). The McGurk effect in infants. Perception and Psychophysics, 59(3):347–357.

[41] Siva, N., Stevens, E.B., and Kuhl, P.K. (1995). A comparison between cerebral-palsied and normal adults in the perception of auditory-visual illusions. Journal of the Acoustical Society of America, 98(5):2983.

[42] Smeele, P. M.T., Hahnlen, L.D., Stevens, E.B., and Kuhl, P.K. (1995). Investigating the role of specific facial information audio-visual speech perception. Journal of the Acoustical Society of America, 98(5),Pt.2:2983.

[43] Sumby, W.H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, 26(2):212–215.

[44] Walker, S., Bruce, V., and O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. Perception and Psychophysics, 57(8):1124–1133.

[45] Westbury, J.R., Hashi, M., and Lindstrom, M.J. (1995). Differences among speakers in articulation of American English /r/: An x-ray microbeam study. In Elenius, K. and Branderud, P., editors, Proceedings of the XIIIth International Congress of Phonetic Sciences, volume 4, pages 50–57. Stockholm, Sweden: Kungliga Tekniska Hogskolan and Stockholm University.