

Changes in the McGurk Effect Across Phonetic Contexts

Michelle Hampson^{a,b}, Frank H. Guenther,^{a,c} Michael A. Cohen^a, and Alfonso Nieto-Castanon^a

^aBoston University
Department of Cognitive and Neural Systems

^bYale University
Department of Diagnostic Radiology, Fitkin B
P. O. Box 208042
New Haven, Connecticut 06520-8042
Fax Number: (203) 785-6534
Telephone Number: (203) 737-5994
Email: chell@boreas.med.yale.edu

^cMassachusetts Institute of Technology
Research Laboratory of Electronics

BOSTON UNIVERSITY TECHNICAL REPORT CAS/CNS 03-006

ABSTRACT

To investigate the processes underlying audiovisual speech perception, the McGurk illusion was examined across a range of phonetic contexts. Two major changes were found. First, the frequency of illusory /g/ fusion percepts increased relative to the frequency of illusory /d/ fusion percepts as vowel context was shifted from /i/ to /α/ to /u/. This trend could not be explained by biases present in perception of the unimodal visual stimuli. However, the change found in the McGurk fusion effect across vowel environments did correspond systematically with changes in second formant frequency patterns across contexts. Second, the order of consonants in illusory combination percepts was found to depend on syllable type. This may be due to differences occurring across syllable contexts in the timecourses of inputs from the two modalities as delaying the auditory track of a vowel-consonant stimulus resulted in a change in the order of consonants perceived. Taken together, these results suggest that the speech perception system either fuses audiovisual inputs into a visually compatible percept with a similar second formant pattern to that of the acoustic stimulus or interleaves the information from different modalities, at a phonemic or subphonemic level, based on their relative arrival times.

INTRODUCTION

Many studies have shown that vision can exert a strong influence on speech perception. For example, Sumbly and Pollack (1954) reported that the intelligibility of speech in noise was higher when listeners viewed 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 having subjects watch 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 report an auditory percept of /da/ or /ča/. In either case, the magnitude of the McGurk effect is the frequency with which subjects fail to correctly identify the acoustic stimulus (that is, the frequency with which any sort of illusory auditory percept is reported).

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 patterns are low-pass filtered. Other researchers have manipulated characteristics of the visual signal to determine how they affect perception. MacDonald, Andersen, and Bachmann (2000) found that the effect decreased as spatial resolution was degraded but did not completely disappear even at a very coarse level of spatial resolution. Temporal gating of the visual stimulus was found to decrease the McGurk effect in direct proportion to the amount of stimulus removed (Munhall & Tokhura, 1998). This may be due to the loss of dynamic visual information. Static featural 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 & 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 & Cohen, 1993; Munhall, Gribble, Sacco, & Ward, 1996), it is sensitive to mismatches in dynamics across the modalities (Munhall et al., 1996).

Despite all of this research, most studies of the McGurk effect have focussed on a limited number of phonetic contexts. Generally, the effect is studied in CV syllables in the /a/ vowel context (or the very similar /ɑ/ and /æ/ contexts). However, there are several studies that have undertaken an examination across a range of contexts. First, the perception of consonant-vowel (CV) stimuli with acoustic /b/ dubbed on visual /g/ 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 fusion effect was found to vary across these three phonetic contexts.

Secondly, a series of studies (also using CV stimuli) compared the McGurk effect in the two different vowel contexts /ɑ/ and /i/ (Green & Gerdeman, 1995; Green, Kuhl, Meltzoff, & Stevens, 1991; Green & Norrix, 1997). Unlike Green et al. (1988), these studies did not find a difference in the magnitude of the McGurk fusion effect in these different vowel contexts. However, a qualitative difference in the fusion effect was found: acoustic /b/-visual /g/ stimuli tended to produce more /d/ than /č/ percepts in the /i/ con-

text, and more /ɔ̃/ than /d/ percepts in the /α/ context. It is not clear why these studies yielded different results than the Green et al. (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 et al., 1991, Green & Gerdeman, 1995, and Green & Norrix, 1997, used the same visual stimuli).

Thirdly, Shigeno (2000) examined the McGurk effect in Japanese across the three vowel contexts /i/, /a/, and /u/ using CV syllables. Shigeno examined the overall magnitude of the McGurk effect in each vowel context by including combination and fusion illusions together. Using this approach, Shigeno found a decrease in the magnitude of the McGurk effect as vowel context was changed from /i/ to /a/ to /u/. In addition, Shigeno (2000) examined the frequency of /d/ percepts in response to acoustic /b/ - visual /g/ or /d/ stimuli and found that this specific type of illusory percept decreased as vowel context was changed from /i/ to /a/ to /u/. This is consistent with the findings of Green (1988).

One finding that is consistent across all these studies is that illusory /d/ percepts in response to acoustic /b/ - visual /g/ or /d/ stimuli were less frequent in the /α/ 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. For CV syllables, the second formant patterns for /d/ and /b/ both rise into the vowel in the /i/ context. In the /α/ 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 /ɔ̃/ is flat or slightly rising in the /α/ context, and in that sense, /ɔ̃α/ is more similar acoustically to /bα/ than /dα/ 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 /α/ context.

Finally, Jordan and Bevan (1997) examined the McGurk effect in the /i/ and /a/ vowel contexts (once again using CV stimuli). In this study, there were as many /d/ percepts reported in the /a/ context as the /i/ context in response to acoustic /b/ - visual /g/ or /d/ stimuli. This is in contrast to the studies discussed above. However, such disparity could be due to differences in acoustics of the speaker, or the precise vowel context used. For example, although the vowels /a/ and /α/ are very similar, /a/ does generally have a higher second formant frequency than /α/, 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 (and, in particular the second formant transitions) associated with those contexts (and that speaker) ? Assuming audio-visual speech stimuli are categorized based on which phonetic unit they most look and sound like, as suggested by Summerfield (1987, 1991), systematic patterns in multimodal perception across phonetic contexts should exist and should reflect the context-dependent expression (in the stimuli) of perceptually relevant sensory features.

An alternate explanation for context-dependent changes in the McGurk effect is that biases present during unimodal visual speech perception (e.g. linguistic biases) exert context-dependent influences on audio-visual speech perception. Massaro (1998) 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 /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). Different linguistic biases in different phonetic contexts might therefore produce changes in the McGurk effect across contexts.

Assuming the McGurk effect arises when normal speech perception processes are applied to unusual inputs, an examination of the different factors influencing McGurk percepts can reveal the factors involved more generally in multimodal speech perception. The roles of acoustic variability and context-dependent linguistic biases in the McGurk effect can be investigated by examining changes in the effect across phonetic contexts and comparing these with changes in unimodal speech perception across the same contexts. As discussed above, several studies have examined the McGurk effect in different phonetic contexts. However, most experiments which have tested the McGurk effect in different phonetic contexts have used only a limited range of stimuli, and few have examined context dependent changes in the perception of unimodal visual 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/ bimodal stimuli and acoustic unimodal 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 six phonetic contexts investigated include three vowel contexts, /i/, /α/, and /u/, and two different syllable types, consonant-vowel (CV) and vowel-consonant (VC). The three vowel contexts were chosen to allow comparison with the findings of previous studies by Shigeno (2000) and by Green and colleagues (Green et al., 1988; Green et al., 1991; Green & Gerdeman, 1995; Green & 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 CV syllables (see, for example, Green & Miller, 1985; MacDonald & McGurk, 1978; Massaro & Cohen, 1983; Rosenblum, Schmuckler, & Johnson, 1997). There have been a few studies involving real-word contexts (Dekle, Fowler, & Funnell, 1992; Easton & Basala, 1982; Fuster-Duran, 1996) and some investigating the effect in VCV syllables (Munhall et al., 1996; Munhall & Tohkura, 1998; Siva, Stevens, Kuhl, & Meltzoff, 1995; Smeele, Hahnen, Stevens, Kuhl, & 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.

This paper presents the findings from three experiments. In the first experiment, unimodal acoustic and visual syllables were tested in order to verify their perceptual clarity and to determine the presence of biases in unimodal perception that could affect bimodal speech perception. In the second experiment, bimodal stimuli were created by combining the acoustic and visual syllables tested in Experiment 1. These stimuli allowed an investigation of the McGurk effect across the range of phonetic contexts described above. In addition, formant tracks of the acoustic stimuli were obtained to investigate the role of second formant patterns in bimodal speech perception. Finally, to investigate the importance of the relative time-courses of auditory and visual information in determining the order of phonemes in combination percepts, a third experiment was conducted.

EXPERIMENT 1

This experiment tested the perceptual quality of the unimodal stimuli to ensure that the auditory and visual stimuli were perceptually clear, and to identify any intrinsic biases in visual perception.

METHODS

Subjects

Ten adult subjects were recruited by flyers placed around the Boston University campus. All subjects gave informed consent in accordance with a protocol reviewed and approved by the Institutional Review

Board at Boston University. 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.

Materials

A female speaker (the experimenter) was taped uttering the nine syllables /gi/, /di/, /bi/, /gɑ/, /dɑ/, /bɑ/, /gu/, /du/, and /bu/ several times each. Using Adobe Premiere, video clips of these utterances were captured for playing on the computer at 15 frames per second. Their resolution was 640 pixels wide by 480 pixels high. 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.

Procedure

Subjects 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 speaker's face in the videos was approximately 3" wide by 4" high. Directions were given verbally prior to initiation of the session, after which the experimenter 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; see 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 "?".

RESULTS

Auditory only tests

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 I. 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.

Visual only tests

The totals across all subjects for consonant-vowel syllables in the /ɑ/ vowel context are shown in Table II. Note that the number of "g" responses is actually greater than the number of "d" responses to the silent /gɑ/ video. This may seem surprising, given the findings of Massaro (1998) 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.

Table I: Results from Auditory-Only Test

Auditory Syllable	Response Percentages
bi	b 93, p 6, d 1
di	d 98, t 2
gi	g 96, k 3, gk 1
bα	b 98, p 2
dα	d 100
gα	g 95, k 5
bu	b 91, p 4, bl 3, g 1, v 1
du	d 100
gu	g 100
ib	b 100
id	d 100
ig	g 100
αb	b 100
αd	d 90, n 10
αg	g 100
ub	b 100
ud	d 100
ug	g 100

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 III. 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”, “pr”, “bp”, “spl”, “bl”, “kl”, “tl”, “gk”, “pl”, “pb”, “y”, and “?”). Although the phoneme /n/ is sometimes classified as an interdental

stop consonant (Kent & 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 & Freeman, 1968; Guenther et al., 1999; Ong & Stone, 1998; Westbury, Hashi, & Lindstrom, 1995). Finally, the responses “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/.

Table II: Response Totals for the Visual Stimuli /bɑ/, /dɑ/, and /gɑ/.

Visual Stimulus	Response Percentages
bɑ	b 51, p 45, m 2, pl 1, y 1
dɑ	d 38, g 22, t 11, k 10, n 2, kr 1, r 1, ? 15
gɑ	g 41, d 25, k 20, t 8, ? 2, m 2, n 1, kr 1

From Table III, 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 contrast with Massaro (1998), who found that subjects more often reported their percept of the silent video /gɑ/ to be the alveolar consonant /d/ than velar consonant /g/. These results are, however, consistent with the study of Shigeno (2000) in which subjects more often reported a /ga/ percept than a /da/ percept in response to a /ga/ face. There are several possible explanations for the differences in these studies. The type of response paradigm used (i.e. forced choice or open response) may have influenced subjects’ place perception. Also, the speakers in the three studies may have had different speech patterns. Finally, the video clips used in the 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/, and /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 was the difference between the number of velar responses and the number of alveolar responses. Syllable type was highly significant in influencing velar/alveolar perception ($F(1,9) = 23.70, p = 0.0009$), and there was a significant interaction between syllable type and vowel context ($F(2,18) = 5.68, p = 0.0122$). 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 ($t(19)$

= -6.82, $p \leq 0.0001$ and $t(19) = -5.89$, $p \leq 0.0001$, respectively) and not a significant factor in the /ɑ/ context ($t(19) = -0.67$, $p = 0.5104$).

Table III: Perception of Place of Articulation During Video-Only Test.

Visual Syllable	Response Percentages			
	Labial	Alveolar	Velar	Other
bi	85	1	2	12
di	0	10 (d 4)	75 (g 55)	15
gi	0	12 (d 6)	74 (g 40)	14
bɑ	98	0	0	2
dɑ	0	51 (d 38)	32 (g 22)	17
gɑ	2	34 (d 25)	61 (g 41)	3
bu	95	0	0	5
du	2	25 (d 15)	46 (g 37)	27
gu	3	13 (d 8)	71 (g 45)	13
ib	99	0	0	1
id	1	52 (d 39)	31 (g 18)	16
ig	5	62 (d 41)	27 (g 20)	6
ɑb	100	0	0	0
ɑd	0	57 (d 32)	42 (g 35)	1
ɑg	0	50 (d 30)	50 (g 44)	0
ub	99	1	0	0
ud	0	77 (d 43)	22 (g 21)	1
ug	6	58 (d 33)	34 (g 33)	2

Using acoustic stimuli along a /ga/-/da/ continuum, Hampson, Guenther, and Cohen (1998) found that dubbing the acoustic syllables on a video of a speaker uttering /da/ induced more /da/ percepts and fewer /ga/ percepts than the dubbing the syllables on a video of the same speaker uttering /ga/. That is, for CV syllables in the /a/ vowel context, visual influences of alveolar and velar consonants on speech perception were found to be different. This challenges the notion that alveolar and velar visual tokens are perceptually equivalent (i.e. it suggests that they are not visemes in a strict sense). The perception of unimodal visual stimuli reported here yields similar results in that alveolar and velar consonants were perceived differently ($F(1,9) = 14.94$, $p=0.0038$). There was, however, a strong interaction between consonant viewed and vowel context ($F(2,18) = 17.68$, $p \leq 0.0001$). Paired t-tests showed a significant influence of the visual conso-

nant in the / α / and /u/ vowel context ($t(19) = 3.60$, $p = 0.0019$ and $t(19) = 4.73$, $p \leq 0.0001$, respectively) but not in the /i/ vowel context ($t(19) = -1.06$, $p = 0.3007$). 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 (perhaps because the high front location of the tongue for /i/ obscures tongue movements to and from this vowel). 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 ($F(1,9) = 6.83$, $p = 0.0281$). 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. There was no main effect of vowel context, and no significant three-way interaction between vowel context, syllable type, and consonant viewed.

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.

DISCUSSION

The results from these unimodal tests establish the perceptual clarity of the acoustic stimuli used in the audio-visual experiment, and provide information regarding unimodal visual 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, 1998), 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 since the unimodal visual stimuli in those contexts induced more /g/ than /d/ percepts. In VC syllables in the /i/ and /u/ contexts, the reverse is predicted: more /d/ fusion percepts than /g/ fusion percepts. Predictions for the / α / vowel context are that subjects will be more likely to report a velar percept when viewing a velar utterance than when viewing an alveolar utterance, and conversely, more likely to report an alveolar percept when viewing an alveolar utterance than when viewing a velar 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.

EXPERIMENT 2

METHODS

Subjects

Fifteen adult subjects who had not participated in Experiment 1 were recruited in the New Haven area, both by flyers placed around the Yale University campus and by word of mouth. All subjects gave

informed consent in accordance with a protocol reviewed and approved by the Human Investigations Committee of the Yale University School of Medicine. 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.

Materials

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 video capturing introduced a delay between the auditory track and the video, so that the dubbed acoustic syllable could not be aligned based on the location of the original acoustic syllable. Therefore, the acoustic syllable was aligned with the visual syllable by hand. The spectral envelope was examined and aligned so that sharp increases (decreases) in amplitude at the start (end) of the syllables were aligned with movements into (out of) vocal tract closure for CV (VC) syllables. In addition, information regarding the release of the speaker's stop consonant, as determined visually in the video track, allowed adjustment of the location of the acoustic syllable such that the acoustic burst was aligned with the vocal tract release. Adjustments could be made with a 33 ms precision, corresponding to half the presentation time of each video frame. 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 CV video clips, and twenty-seven VC video clips.

Procedure

Subjects 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 speaker's face in the videos was approximately 3" wide by 4" high.

The CV syllables and VC syllables were played in separate blocks. The order of these blocks was randomized. 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.

Analysis of Acoustic Stimuli: Formant Tracking

Formant tracks were found for each of the digitized (44 kHz) acoustic syllables. First, the data was pre-emphasized (first order linear filter, $a_1 = -0.95$) and windowed using a 20 ms moving Hamming window. Linear predictive coding was then performed using a model with 58 poles. This number of poles was chosen because it revealed the formant pattern clearly across all 18 acoustic syllables.

RESULTS

Illusory fusion percepts

Acoustic /b/-visual /d/ and acoustic /b/-visual /g/ stimuli induced illusory auditory percepts of nonlabial consonants across the different vowel contexts. A summary of responses to the CV stimuli is provided in Table IV². Results pertinent to the following discussion are highlighted in bold font.

In Panel a of Figure 1, the percentage of the time that subjects responded /d/ to stimuli in the CV syllable set is plotted as a function of vowel context. The solid line shows the percentage of /d/ responses when

Table IV: Percentage responses made to the CV audio-visual stimuli.

Numbers are rounded off to the nearest percent. 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.

acoustic stimulus		visual stimulus		
vowel	consonant	b	d	g
i	b	b 98, bt 2	d 91, g 4, dt 4	d 87, g 4, b 2, q 2, dt 2, ? 2
	d	<i>bd 40, d 36,</i> <i>b 18, md 2,</i> <i>th 2, bg 2</i>	d 96, g 4	d 89, g 4, dg 4, ? 2
	g	<i>bg 42, g 42,</i> <i>mg 7, b 2,</i> <i>p 2, d 2, k 2</i>	g 98, d 2	g 98, ? 2
α	b	b 100	d 40, g 40, b 11, th 2, bg 2, t 2, gd 2	d 31, g 56, b 4, th 2, dg 2, dt 2, ? 2
	d	<i>bd 56, d 24,</i> <i>b 7, md 4,</i> <i>g 7, bg 2</i>	d 91, g 7, dg 2	d 76, g 18, dt 2, ? 4
	g	<i>bg 44, g 44,</i> <i>mg 9, bgf 2</i>	g 96, bg 2, gs 2	g 98, bg 2
u	b	b 100	d 13, g 69, b 9, ? 9	d 7, g 62, b 13, bg 2, dg 2, k 2, ? 11
	d	<i>bd 49, d 44,</i> <i>md 4, b 2</i>	d 96, g 4	d 96, g 2, gd 2
	g	<i>g 44, bg 36,</i> <i>mg 9, b 7,</i> <i>dg 2, ? 2</i>	g 100	g 98, dg 2

/b/ is dubbed on a /g/ face. In agreement with the previous findings of Shigeno (2000) and Green and colleagues (Green et al., 1988; Green et al., 1991; Green & Gerdeman, 1995; Green & Norrix, 1997), acoustic /b/-visual /g/ stimuli elicited a decreasing number of /d/ percepts as vowel context was changed from /i/ to /α/ to /u/. Additionally, as shown by the dashed line in Panel a of Figure 1, 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 /α/ to /u/. To evaluate the significance of this change in the frequency of /d/ perception across vowel contexts, a two factor within-subjects ANOVA was performed. The factors were vowel context (/i/, /α/ or /u/) and visual consonant (/g/ or /d/) and the dependent variable was the number of /d/ percepts. The effect of vowel context was significant ($F(2,14) =$

41.97, $p \leq 0.0001$). The effect of the visual consonant was approaching significance as well ($F(1,14) = 3.8571$, $p=0.07$), suggesting that subjects were more likely to perceive /d/ when /b/ was dubbed on a /d/ face than when it was dubbed on a /g/ face. However, the influence of the visual consonant was much smaller than the influence of vowel context as clearly illustrated in Figure 1a. There was no significant interaction between the factors. In summary, the change in frequency of /d/ fusion percepts across vowel contexts reported by Shigeno (2000) and Green and colleagues was reproduced in the CV data set of this study.

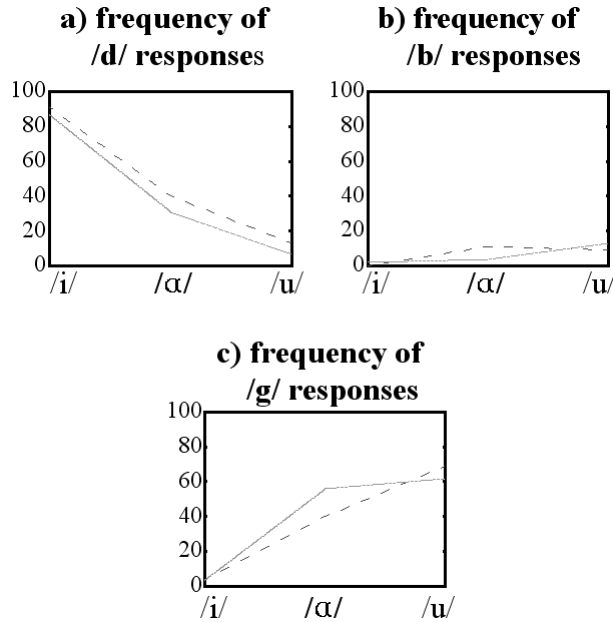


Figure 1. Response patterns for the CV syllable set. These graphs show how responses to acoustic /b/-visual /g/ stimuli (solid lines) and acoustic /b/-visual /d/ stimuli (dashed lines) change across vowel contexts. (a) Frequency of /d/ responses across vowel contexts. (b) Frequency of /b/ responses across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

However, in contrast with the findings of Green et al. (1991), Green and Gerdeman (1995), and Green and Norris (1997), the decrease in /d/ percepts as vowel context was changed from /i/ to /α/ 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 CV 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 et al. (1988) and Shigeno (2000), who found a large decrease in the magnitude of the McGurk effect as the vowel context was changed from /i/ to /α/ to /u/. That is, their stimuli induced a steadily increasing frequency of /b/ percepts as the vowel context was shifted from /i/ to /α/ 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 slightly in the /α/ and /u/ contexts relative to the /i/ context, this change was not comparable to that reported by Green et al. (1988) or Shigeno (2000). To evaluate the significance of this change in the frequency of /b/ percepts across vowel contexts, a two-factor, within-subject ANOVA was performed. The factors were vowel context (/i/, /α/ or /u/) and visual consonant(/g/ or /d/) and the dependent variable was the number of /b/ percepts. Neither of the variables, nor their interaction was significant at the $p < 0.05$ level. Therefore, there was no significant change in the magnitude of the McGurk fusion effect across vowel contexts in the CV data set of this study.

Rather than a decrease in the magnitude of the McGurk effect or an increase in the frequency of /ǝ/ 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. A two-factor, within-subject ANOVA was performed to evaluate the change in frequency of /g/ percepts across vowel context. The factors were vowel context (/i/, /ɑ/ or /u/) and visual consonant (/g/ or /d/) and the dependent variable was the number of /g/ percepts. Vowel context was found to be significant at the $p < 0.05$ level ($F(2,14) = 18.082, p \leq 0.0001$). Neither the visual consonant nor the interaction was significant. Although this increase in /g/ percepts is different from previous findings of Green and colleagues (Green et al., 1988; Green et al., 1991; Green & Gerdeman, 1995; Green & Norrix, 1997), it is not necessarily incompatible with their findings from a theoretical perspective. Green (1996) suggested that the findings of Green et al. (1991) and Green and Gerdeman (1995) were a result of the different second formant (F2) patterns of /d/, /b/ and /ǝ/ 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 /ǝɑ/ 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 /ǝɑ/ than /dɑ/ fusion percepts in the studies of Green et al., (1991) and Green and Gerdeman (1995). A similar explanation, based on second formant patterns, may be applied to our findings.

The formant tracks of the acoustic CV syllables used in this study are shown in Figure 2. The top panel of Figure 2 shows the results of the LPC analysis. The first two formant patterns of each syllable were traced by hand; these tracings are displayed in the lower panel. Estimates of the slopes of the second formant transitions were formed by evaluating the slopes of the initial segments of these tracings. In the /i/ vowel context, the second formant patterns for /b/ and /d/ are qualitatively similar in that they are rising transitions (the slopes of these F2 transitions are approximately 28 Hz/ms and 6 Hz/ms, respectively) while the second formant pattern for /g/ is falling (F2 slope of approximately -4 Hz/ms). Subjects who were exposed to an acoustic /bi/ stimulus dubbed onto a face enunciating /gi/ or /di/ may have perceived /di/ because it is acoustically similar to /bi/ and visually similar to the syllable viewed. In contrast, in the /u/ vowel context, the syllable /bu/ has a flat second formant pattern (zero slope) which more closely resembles the slightly falling second formant pattern of /gu/ (F2 slope $\cong -3$ Hz/ms) than the somewhat more steeply falling formant pattern of /du/ (F2 slope $\cong -5$ Hz/ms). This may explain why the acoustic /bu/-visual /gu/ or /du/ stimuli were more often perceived as /gu/ than /du/. Finally, in the /ɑ/ vowel context, the rising second formant transition of /b/ (F2 slope $\cong 25$ Hz/ms) is qualitatively different from the falling transitions of both /d/ and /g/ (slopes of approximately -12 Hz and -10 Hz, respectively), and in this context, illusory /g/ and /d/ percepts occurred with a similar frequency.

Interestingly, the data of Shigeno (2000) also shows an increase in /g/ percepts in response to acoustic /b/ - visual /g/ or /d/ stimuli as vowel context was changed from /i/ to /a/ to /u/. Unfortunately, this effect was not evaluated for significance, and the formant patterns of the speaker in that study were not described. However, the qualitative similarity in the data of Shigeno (2000) is encouraging in that it suggests this pattern of increasing /g/ percepts (in response to acoustic /b/ - visual /g/ or /d/ stimuli) across the vowel con-

texts /i/, /ɑ/ (or /a/), and /u/ is not an anomaly of these stimuli, but a more general feature of those vowel contexts.

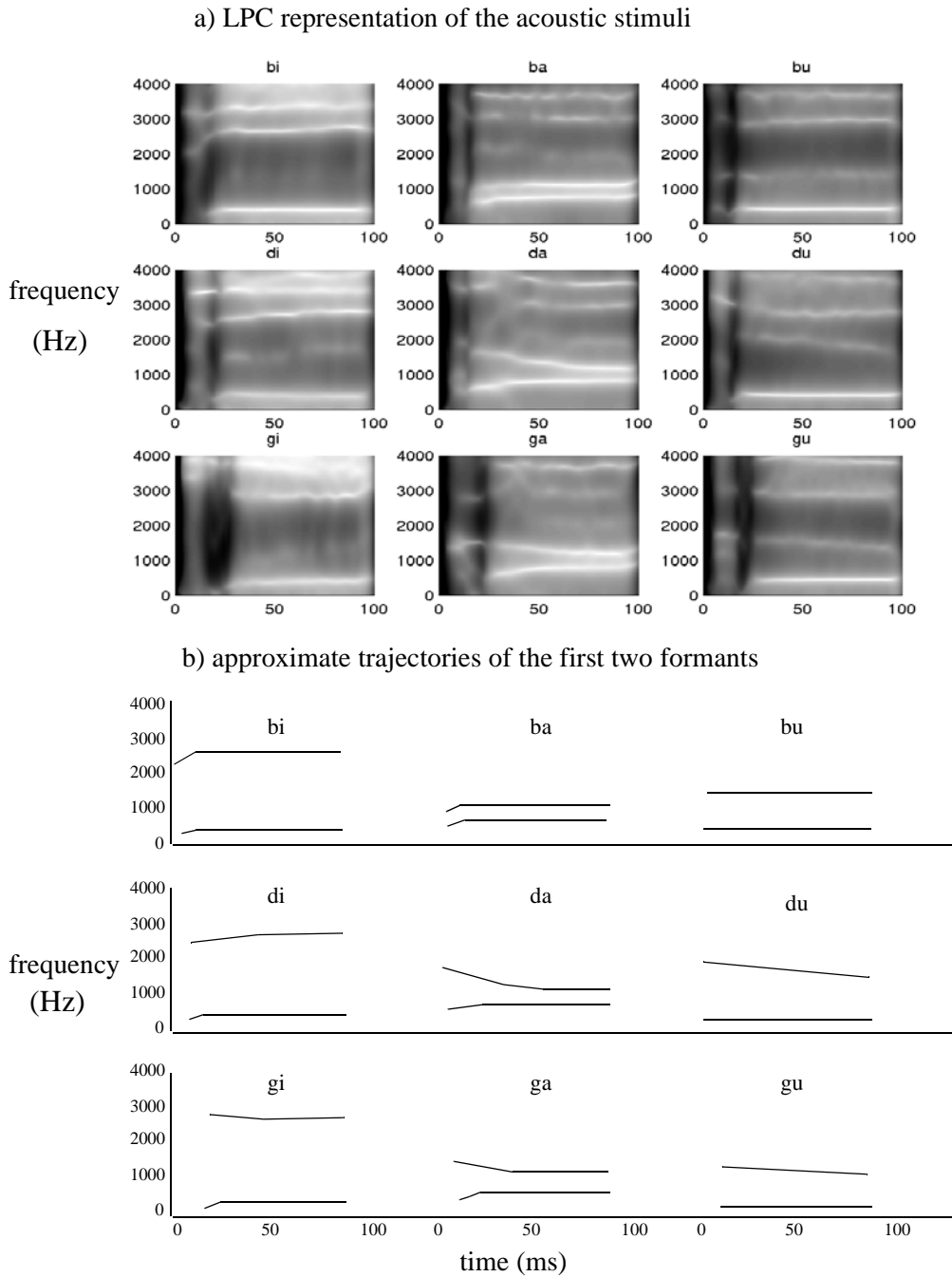


Figure 2. (a) Formant tracks of the CV syllables produced by LPC analysis. Peaks in the spectrum appear in white and low energy areas are darker. (b) Hand-sketched estimates of the trajectories of the first two formants (piece-wise straight line tracings of the LPC peaks).

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 et al., 1991; Green & Gerdeman, 1995; Green & Norrix, 1997) have found an increase in /č/ percepts as the vowel context was changed from /i/ to /ɑ/ (or /a/), while other studies (including the current study and that

of Green et al., 1988) have not. Perhaps the magnitude of the release burst of the acoustic /b α / 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 / β /, regardless of how similar the second formant patterns of these consonants were.

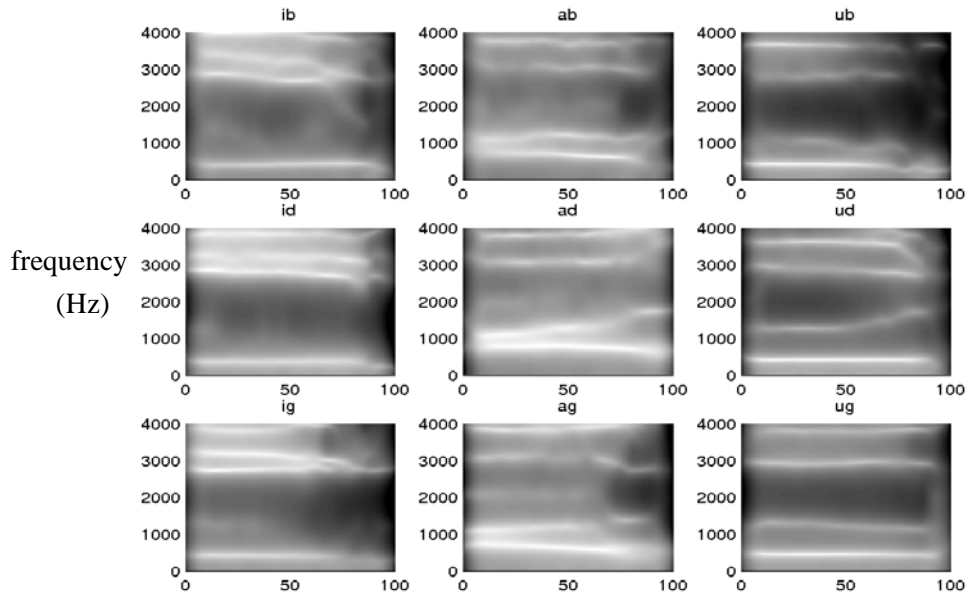
If changes in the McGurk fusion phenomenon across vowel contexts arise from changes in the second formant patterns across contexts, what sort of response patterns should occur in the VC syllable set? The formant patterns of the VC syllables are shown in Figure 3. These patterns are similar to the CV formant tracks reversed in time. As a result the response patterns to VC stimuli are expected to be similar to the corresponding CV stimuli. Estimates of the slopes of the second formant frequency transitions for these syllables were formed by evaluating the slope of the last segment of the tracings shown in the bottom panel (except in a couple cases where sharp jogs occurred in the tracings near the end of the syllable, in which case slopes were averaged over the last couple of line segments).

Examination of Figure 3 reveals that the syllable /ib/ has a steeply falling second formant transition (slope $\cong -43$ Hz/ms) which is much more similar to the falling /id/ transition (slope $\cong -27$ Hz/ms) than the rising /ig/ transition (slope $\cong 5$ Hz/ms). Therefore more /id/ than /ig/ fusion percepts are predicted in response to acoustic /ib/-visual /id/ or /ig/ stimuli. Second, in the / α / context, / α b/ has a flat second formant pattern which is qualitatively different from the rising transitions of both / α d/ and / α g/ (19 and 8 Hz/ms respectively), but which is somewhat more similar to / α g/. In this case, / α g/ percepts are expected to occur more frequently than / α d/ percepts in response to acoustic / α b/-visual / α g/ or / α d/ stimuli. Finally, in the /u/ context, the second formant patterns for both /ub/ and /ug/ are approximately flat (-3 and -2 Hz/ms), while the second formant pattern for /ud/ rises (with a slope of 13 Hz/ms). Acoustic /ub/-visual /ug/ or /ud/ stimuli are thus expected to produce many more /ug/ than /ud/ percepts. In summary, it is predicted that acoustic /Vb/-visual /Vg/ or /Vd/ stimuli will produce a decreasing frequency of illusory /d/ percepts and an increasing frequency of illusory /g/ percepts as vowel context is changed from /i/ to / α / to /u/.

The results for the VC syllable set are provided in Table V. Results pertinent to the following discussion are highlighted in bold font. 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 4. 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 / α / to /u/ (shown in Panel a of Figure 4). A two-factor (vowel context and visual stimulus), within-subject ANOVA confirmed the significance of this trend. Vowel context was found to have a significant effect on the number of /d/ fusion percepts in the VC data set ($F(2,14) = 40.38$, $p \leq 0.0001$), while neither the visual stimulus nor its interaction with vowel context were significant.

The increase in /g/ percepts found in the CV syllable set is also reproduced in the VC syllable set (see Panel c of Figure 4) in that there are many more /g/ percepts in the / α / and /u/ contexts than in the /i/ context. However, the number of /g/ percepts in the / α / context of the VC syllable set is markedly greater than found in the / α / context of the CV syllable set. This may be because (for the tokens of this speaker) /g/ is more similar acoustically to /b/ than the competing /d/ percept in / α C/ syllables (but not in /C α / syllables). That is, the slope of the second formant pattern for / α g/ is more similar to that of / α b/ than the second formant slope of / α d/ whereas the second formant pattern for /g α / and /d α / are about equally different from that of /b α /. Despite differences in the precise pattern of /g/ perception across fusion contexts in VC as compared to CV syllables, the existence of a dramatic increase in /g/ perception across vowel contexts was reproduced. A two-factor (vowel context and visual stimulus), within-subject ANOVA confirmed the significance of this change in frequency of /g/ fusion percepts across vowel contexts ($F(2,14) = 18.44$, $p \leq 0.0001$). Neither the visual stimulus nor the interaction between visual stimulus and vowel context was found to be significant..

a) LPC representation of the acoustic stimuli



b) approximate trajectories of the first two formants

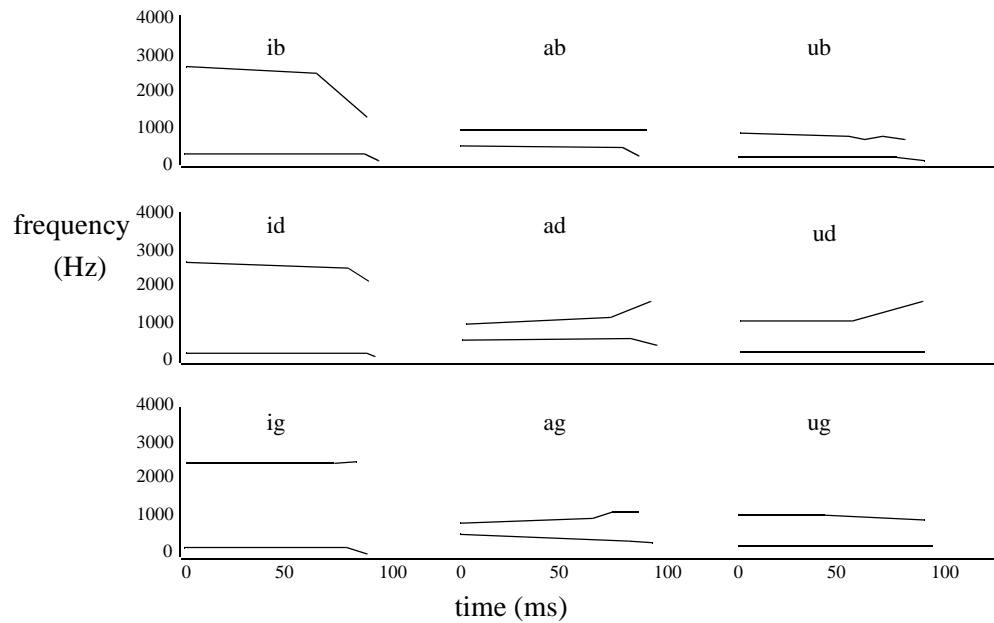


Figure 3. (a) Formant tracks of the VC syllables produced by LPC analysis. Peaks in the spectrum appear in white and low energy areas are darker. (b) Hand-sketched estimates of the trajectories of the first two formants (piece-wise straight line tracings of the LPC peaks).

Table V: Percentage responses made to the VC audio-visual stimuli

Numbers are rounded off to the nearest percent. 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.

acoustic stimulus		visual stimulus		
vowel	consonant	b	d	g
i	b	b 91, bd 9	d 71, b 20 , bd 4, dth 4	d 82, b 13 , dth 4
	d	<i>bd 47, b 24, db 16, d 13</i>	d 100	d 100
	g	<i>bg 51, g 24, gb 13, b 4, gd 2, dbg 2, bk 2</i>	g 98, gk 2	g 98, gk 2
α	b	b 96, bd 2, bp 2	d 20, g 56, b 13 bg 4, dg 4, gd 2	d 18, g 58, b 13 , db 2, bg 2, gd 4, ? 2
	d	<i>bd 42, b 31, db 11, d 7, gb 7, bg 2</i>	d 89, g 7, bd 2, gd 2	d 84, g 9, gd 7
	g	<i>bg 49, gb 27, g 16, b 7, gd 2</i>	g 100	g 100
u	b	b 100	d 4, g 53, b 31 v 4 , bg 4, gb 2	g 51, b 29 , v 7 , gb 4, db 2, bg 2, dg 2, bp 2
	d	<i>bd 42, db 31, d 27</i>	d 100	d 98, dth 2
	g	<i>bg 31, gb 31, g 31, b 4, dgd 2</i>	g 100	g 100

To allow comparison of contextual effects in the McGurk fusion illusion with changes in the perception of the unimodal visual stimuli, a three factor, within-subjects ANOVA was performed on responses to acoustic /b/ - visual /g/ or /d/ stimuli. The factors were syllable type (CV or VC), vowel context (/i/, /α/ or /u/), and visual consonant (/g/ or /d/). The dependent variable used was the difference between the number of alveolar (including /d/ and /t/) responses and the number of velar (including /g/ and /k/) responses. This dependent variable was selected to be consistent with the unimodal analysis, although very few velar or alveolar responses other than /g/ or /d/ were reported in the bimodal experiment. The only significant factor or interaction found was a main effect of vowel context ($F(2,14) = 61.218, p \leq 0.0001$). This is in stark contrast to the findings from the unimodal data set in which the significant main effects were syllable type and consonant viewed, and in which vowel context was not found to be significant. Therefore, response biases in the unimodal data set cannot explain the pattern of responses in the bimodal experiment. In particular, the shift in illusory auditory perceptions of acoustic /b/ - visual /g/ or /d/ stimuli from alveolar to velar percepts as vowel context was shifted from /i/ to /α/ to /u/ cannot be explained by biases found in the unimodal visual tests.

In summary, both the CV and VC syllable sets yielded similar context-dependent patterns of fusion effects. In the CV syllable set, a decrease in the frequency of illusory /d/ percepts in response to acoustic /b/-visual /g/ or /d/ stimuli was found as vowel context was shifted from /i/ to /α/ to /u/, and a corresponding increase in the frequency of illusory /g/ percepts occurred. In the VC syllable set, a similar decrease in the frequency of illusory /d/ percepts in response to acoustic /b/-visual /g/ or /d/ stimuli was found as vowel context was shifted from /i/ to /α/ to /u/. In addition, an increase in the frequency of /g/ percepts occurred as vowel context was changed from /i/ to /α/ or /u/. These changes could not be explained by biases found in the unimodal visual tests but were consistent with changes found in the second formant frequency patterns of the stimuli in the different vowel contexts. There was no significant change in the magnitude of the McGurk effect found across vowel contexts or in response to the different visual stimuli used (i.e the /d/ face or the /g/ face). In general, there was no effect of visual stimulus viewed on the illusory percept produced, although an increase in /d/ percepts in response to a /d/ face (as compared to a /g/ face) was approaching significance in the CV data set.

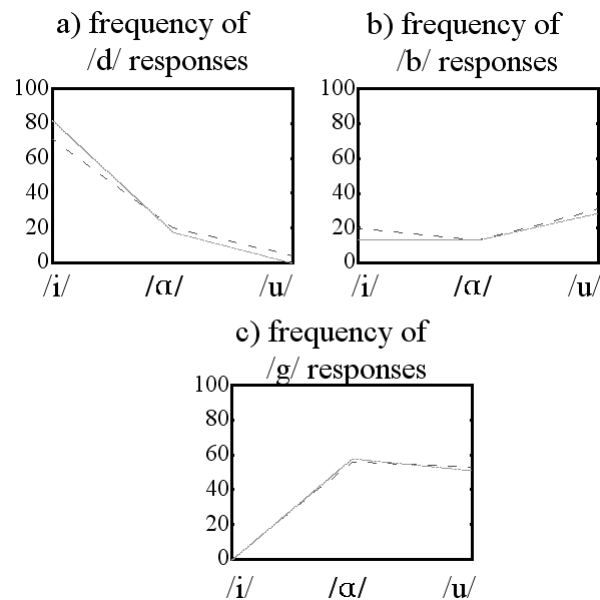


Figure 4. Response patterns for the VC syllable set. These graphs show how responses to acoustic /b/-visual /g/ stimuli (solid lines) and acoustic /b/-visual /d/ stimuli (dashed lines) change across vowel contexts. (a) Frequency of /d/ responses across vowel contexts. (b) Frequency of /b/ responses across vowel contexts. (c) Frequency of /g/ responses across vowel contexts.

Illusory combination percepts

Illusory auditory percepts of consonant combinations were reported in response to stimuli with an acoustic /g/ or /d/ dubbed on a visual /b/ utterance. The combination percepts reported by subjects in response to acoustic /b/-visual /g/ or /d/ stimuli are italicized in Tables 4 and 5. For this discussion, percepts combining the acoustically presented consonant with any labial consonant are considered combination percepts, as the visual stimuli are not expected to convey much information regarding manner of articulation.

In prior studies of the McGurk effect, involving CV syllables, combination percepts almost always had the labial consonant leading the nonlabial (e.g. MacDonald and McGurk, 1978; Massaro and Cohen, 1993). The results for CV syllables in this experiment are consistent with previous findings. That is, many combination percepts did occur when /gV/ and /dV/ syllables were dubbed on a face saying the corresponding /bV/ syllable, and these percepts always involved consonant combinations of the form ‘bg’ or

'bd' rather than 'gb' or 'db' (see italicized entries in Table IV). Combination responses were also common for VC syllables. However, in this context, subjects reported many percepts in which the nonlabial consonant led the labial. That is, many combination percepts reported in response to the VC stimuli were of the form 'gb' or 'db' rather than 'bg' or 'bd' (see italicized entries in Table V).

DISCUSSION

Illusory fusion percepts

This experiment did not replicate the findings of Green et al. (1988) or Shigeno (2000) as the magnitude of the McGurk effect in the CV syllable set did not decrease substantially as vowel context was changed from /i/ to /α/ to /u/. Rather, a change in the type of illusory percept was found to occur across vowel contexts. 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 /α/ to /u/. A similar pattern was found in the VC data set. In both cases, the trends were consistent with changes found in the second formant patterns of the syllables across vowel contexts.

Illusory combination percepts

One possible reason why the order in which consonants are perceived in combination percepts might change across syllable types is that the perceived order of consonants is influenced by the timing between visual and auditory inputs, which changes with syllable type. When the acoustic burst is aligned with lip-opening, CV audio-visual clips have the lip-closing portion of the /b/ gesture before any sound is heard, while VC clips have the lip-opening portion of the /b/ gesture after the acoustic syllable has ended. It is possible that these parts of the articulation are ignored. However, the speech perception system appears to utilize visual information from throughout articulation rather than gaining information only from specific points or periods during an utterance (Munhall and Tokhura, 1998). In the case of CV syllables, some information regarding the visual consonant is arriving prior to any information regarding the acoustic consonant since the movement of the speaker's face begins before the onset of the acoustic speech signal. For VC syllables, information regarding the visual consonant is still arriving after the completion of the acoustic syllable. The temporal relationship between the period of unimodal visual input and the occurrence of the acoustic syllable may therefore influence the temporal relationship between the perceptions of the visual and acoustic consonants.

Massaro and Cohen (1993) performed an experiment which has bearing on this issue. They varied the time-alignment of auditory /da/-visual /ba/ syllables, and had viewers report their resulting auditory percepts. As the auditory input was shifted earlier relative to the video, the number of /dba/ responses increased very slightly, but probably insignificantly, relative to the number of /bda/ responses. Even at the greatest auditory lead time tested (200 ms), there were still many more /bda/ percepts than /dba/ percepts. It is difficult to draw any conclusions from this finding as the stimulus with an auditory lead time of 200 ms could still have had visual information for the bilabial preceding the onset of the acoustic syllable (i.e. the lips could have begun moving towards labial closure prior to the start of the acoustic stimulus). Thus it is not clear from these results whether temporal relationships between the auditory and visual inputs determine the temporal relationships of consonants in the final percept, but it does appear to be a reasonable possibility.

Another possible explanation why the consonants perceived in combination illusions may often occur in a different order in VC syllables than they do in CV syllables is related to the context dependent ability of labial consonants to mask the visual expression of alveolar consonants. That is, perhaps the utterance /bda/ is visually more similar to /ba/ than the utterance /dba/, while /adb/ is more visually similar to /ab/ than /abd/. This might be the case, for example, if adjacent labial consonants more easily mask alveolar

consonants that are embedded within the syllable (between the labial consonant and the vowel), than those on the edge of the syllable.

Experiment 3 was designed to test the hypothesis that the order of consonants in combination percepts is influenced by the temporal relationship between the acoustic and visual stimuli. If this hypothesis is verified, it may explain why many combination percepts in the VC data set had consonant order reversed as compared to the typical combination percepts occurring in CV syllables. However, if this hypothesis is not supported some other explanation must be considered, such as compatibility between the consonant combination perceived and the visual stimulus.

EXPERIMENT 3

The goal of this experiment was to test whether misalignment of the acoustic and visual syllables can alter the order of consonants in the final percept. One approach would be to use consonant-vowel stimuli and to shift the acoustic signal (the nonlabial consonant) earlier relative to the visual stimulus (the labial consonant). If this produced a change in combination illusions from the typical labial-preceding-nonlabial percepts to percepts in which the nonlabial consonant leads the labial, it would provide evidence that the temporal relationship between visual and acoustic information affects the order of consonants in combination percepts. As discussed above, Massaro and Cohen (1993) performed an experiment of this nature, except that the maximum auditory lead time tested was 200 ms. As lip closure during the production of syllable-initial labial consonants often exceeds 200 ms, the videos used by Massaro and Cohen (1993) may have all had the visual gesture leading into labial closure preceding any auditory information regarding the nonlabial consonant. A larger auditory lead time may be necessary to induce a change in the order of consonants perceived. Unfortunately, although the McGurk effect is fairly robust to audio-visual asynchronies, large misalignment of the auditory and visual information tends to disrupt audio-visual integration. For example, Munhall, Gribble, Sacco and Ward (1996) found the magnitude of the McGurk fusion effect diminished with increasing audio-visual misalignment. Similarly, the number of combination illusions in the Massaro and Cohen (1993) study decreased as the auditory lead time was increased. Therefore, it may not be possible to examine illusory combination percepts at the auditory lead times necessary to ensure auditory information in CV syllables preceded the labial closure.

An alternate approach is to manipulate the alignment of the auditory track of VC syllables rather than manipulating the alignment of the auditory track of CV syllables. There are two aspects of this approach which make it less likely to suffer from a disrupted McGurk effect due to large audio-visual misalignment. First, as the number of 'bd' combination percepts is not already saturated at 100% in the VC data set, the audio track can be shifted later, to see if more 'bd' percepts can be induced, as well as earlier (to see if more 'db' percepts can be induced). This allows us to examine variation in two directions and thereby increases the probability that a significant change in perception can be observed within the window in which integration occurs. Second, audio-visual integration is more robust to misalignment when the auditory stimulus lags the visual stimulus rather than leading it (Dixon and Spitz, 1980). This phenomenon is apparent in the results of Massaro and Cohen (1993). The magnitude of the McGurk combination illusion obtained in that study was much greater when the auditory information lagged the visual information by 200 ms than when the auditory information preceded the visual information by 200 ms. Therefore, when introducing auditory delays, a greater misalignment may be used without disrupting audiovisual integration than when the auditory stimulus is shifted earlier relative to the visual stimulus.

Therefore, the approach adopted in this experiment was to manipulate the audiovisual alignment in VC stimuli rather than introducing auditory leads in CV stimuli. If the timing between incoming visual and acoustic information influences the order of consonants in combination percepts, then changes in the alignment of the auditory track of a VC stimulus should influence the order of consonants perceived. That is, auditory delays should result in significantly more combination percepts in which the labial (visual) conso-

nant precedes the nonlabial (acoustic) consonant, and should result in less of the nonlabial-preceding-labial combination percepts found to occur during normal alignment in Experiment 2.

METHODS

Subjects

Eight adult subjects were recruited in the New Haven area, both by flyers placed around the Yale University campus and by word of mouth. All subjects gave informed consent in accordance with a protocol reviewed and approved by the Human Investigations Committee of the Yale University School of Medicine. 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.

Stimuli

The audio-visually congruent / α b/ syllable, the audio-visually congruent / α d/ syllable, and the incongruent acoustic / α d/-visual / α b/ syllable from Experiment 2 were used in Experiment 3. The inclusion of the two congruent syllables was expected to increase subjects' perception that the viewed stimuli were natural utterances and thus to increase their tendency to have a unified percept in response to the incongruent stimuli. In addition, three different misalignments of the incongruent syllable were produced in which the acoustic signal was shifted to lag the visual signal by 333 ms, 167 ms, and -167 ms (i.e. a lead of 167 ms).

Procedure

Subjects were seated alone in a room approximately 2 feet in front of a computer monitor with speakers on either side, and a mouse in front of them. The experiment was response-paced, with a new video being played immediately following the previous response. The speaker's face in the videos was approximately 3" wide by 4" high. Directions were given verbally prior to initiation of the session. In particular, subjects were instructed to indicate their auditory percept after each video clip was played by clicking on the correct button.

A six-alternative forced-choice paradigm was used. Subjects were instructed to watch the videos, but to select the response option closest to what they heard. The available choices were 'Vb', 'Vd', 'Vbd', 'Vdb', 'VbVd', and 'VdVb'. Subjects were told that 'V' was a placeholder for any vowel sound. The last two options were included to determine whether consonants in combination percepts are always interleaved at a phonemic level, or whether, with sufficient acoustic delay, the entire acoustic syllable is perceived as being heard after the visual syllable. In either case, the percept will be illusory as the visual stimulus will be affecting the auditory percept (that is, subjects will be reporting *hearing* both consonants, although only one is present acoustically). The experiment was divided into two blocks to allow subjects a break midway through the experiment. Within each block, the audio-visual clips were presented in a randomized order such that they each appeared 5 times (for a total of 30 clips per block).

RESULTS

Subjects responded 'd' to the audio-visually congruent / α d/ stimulus 100% of the time, and 'b' to the audio-visually congruent / α b/ stimulus 98% of the time. Responses to the acoustic / α d/-visual / α b/ stimuli, at the four different alignments tested, are summarized in Table VI. The percentage of 'Vd' responses is the number of 'correct' auditory percepts (i.e. those which were not corrupted by the incongruent visual stimulus). A large percentage of 'Vd' responses implies a weak McGurk effect, and a small percentage of 'Vd' responses implies a strong McGurk effect. The weakest McGurk effect occurred in the stimulus with the largest misalignment. There were never more than 7% VbVd or VdVb responses, therefore most combination illusions were single syllable percepts.

Table VI: Percentage responses made to the four different audiovisual alignments of acoustic / α d/-visual / α b/.

Numbers are rounded off to the nearest percent.

response option	Delay of auditory stimulus (ms)			
	-167	0	167	333
Vd	8	8	8	23
Vb	3	6	9	6
Vdb	54	33	8	18
Vbd	29	51	70	49
VdVb	6	1	3	1
VbVd	1	1	4	4

The frequency of the single syllable combination percepts ‘Vdb’ and ‘Vbd’ are indicated by the black and white bars in Figure 5, respectively. The stimulus with an acoustic lead time of 167 ms resulted in more ‘Vdb’ responses than ‘Vbd’ responses. The audio-visually aligned stimulus produced more ‘Vbd’ than ‘Vdb’ responses, and the stimulus with an acoustic delay of 167 ms resulted in even more ‘Vbd’ and fewer ‘Vdb’ responses. However, this trend did not continue into the stimulus with the greatest acoustic delay, as the number of ‘Vbd’ percepts dropped off for the stimulus with the 333 ms acoustic delay (perhaps due to a weakening of the McGurk effect at this large misalignment), and the number of ‘Vdb’ percepts increased (albeit not significantly).

To investigate whether the order of consonants in combination percepts was significantly different with different audio-visual alignments, a logistic regression was fit via maximum likelihood to the subjects’ combination responses. An analysis of deviance was used to evaluate the hypothesis that the slope of the regression was zero (Venables and Ripley, 1997, p 227). The null hypothesis was rejected at the $p \leq 0.05$ level. That is, the change in the order of consonants in combination percepts was found to be significant for the pooled group data ($F(1,246) = 26.21, p < 0.0001$).

DISCUSSION

Very few two-syllable illusory percepts (VbVd or VdVb) were reported in response to visual / α b/ - acoustic / α d/ stimuli. The order in which consonants were perceived to be heard in one-syllable combination percepts was found to depend on the temporal relationship between the acoustic and visual information in the stimulus. It therefore appears that consonant percepts are interleaved at a phonemic (or subphonemic) level during combination illusions, and that their order in the final percept is determined, at least in part, by the relative timing of intermodal information. Alternatively, information from the two modalities may be integrated continuously (incorporating at this point the arrival time of the information in the two modalities) prior to any phonetic classification.

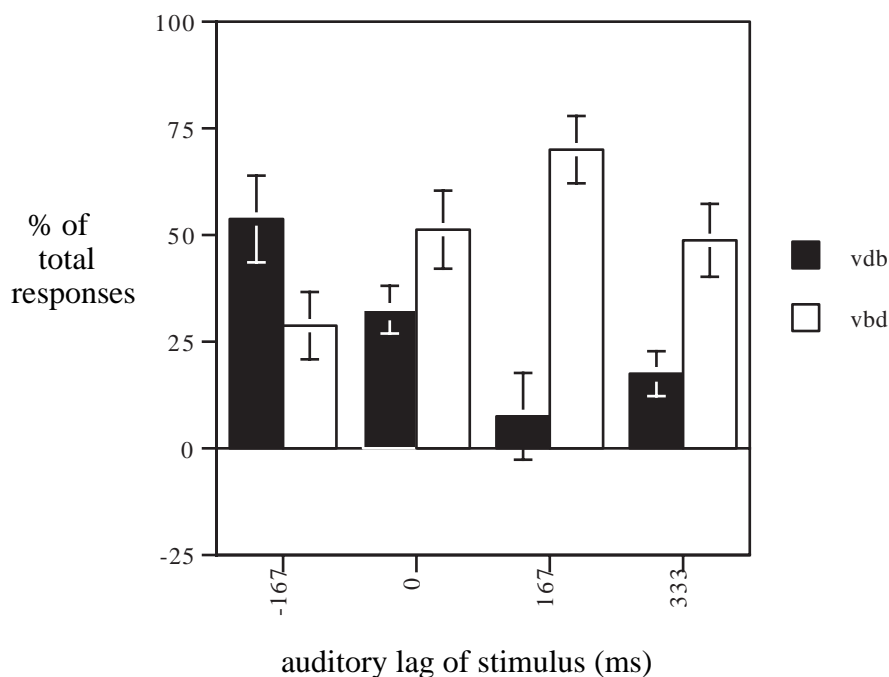


Figure 5. Percentage of responses which were ‘Vdb’ (in black) and ‘Vbd’ (in white) shown for each of the four alignments of the acoustic /αd/-visual /αb/ stimuli. Standard error is indicated by error bars .

GENERAL DISCUSSION

A review of current models of audio-visual fusion in speech perception is provided by Schwartz, Robert-Rimes, and Escudier (1998). The findings of the current paper should help constrain these theories of bimodal speech perception. For example, Schwartz et al. (1998) describe the following motor recoding model of audio-visual fusion. A “vocal tract configuration can be derived independently for each modality: the resultant configurations are fused and then presented to the phonetic classifier” (pp. 92). This model provides a simple explanation for the well-known McGurk fusion effect in which acoustic /ba/-visual /ga/ stimuli are perceived as /da/ since the alveolar constriction of /d/ is midway in the vocal tract between the labial constriction of /b/ and the velar constriction of /g/. However, the findings reported in the current paper are less congruent with this model. In particular, the finding that acoustic /bu/-visual /du/ stimuli resulted in illusory /gu/ percepts 77% of the time is difficult to explain. That is, /g/ is formed by a constriction further back in the vocal tract than either /b/ or /d/, so it seems unlikely that a /g/ percept would arise from a fusion of the vocal tract configurations for /b/ and /d/. However, other motor-based models of bimodal speech perception may be compatible with this data. For example, models that suggest that sensory information is fused across modalities prior to conversion into a motoric code may be able to explain these findings depending on the reference frame fusion is assumed to occur in. In any case, the results of these experiments should help constrain current theories of speech perception by elucidating the pattern of variation in audio-visual illusions across syllable type, visual image, and vowel context. In particular, two major context dependent effects have been reported here: changes in illusory fusion percepts across vowel contexts and changes in the order of consonants in illusory combination percepts with shifts in audiovisual alignment.

Illusory fusion percepts in different vowel contexts

Examination of the formant patterns of the acoustic stimuli revealed that the systematic changes found in fusion percepts across vowel contexts may be explained by changes in acoustics across contexts. More specifically, the increase in /g/ responses relative to /d/ responses as vowel context was changed from /i/ to /α/ to /u/ reflects changes in second formant patterns across these vowel contexts. For example, in terms of the slope of the second formant frequency patterns, /gu/ is midway between /bu/ and /du/. Therefore, /gu/ is more similar visually to the stimulus /du/ than /bu/ is, and /gu/ is more similar acoustically to the stimulus /bu/ than /du/ is, at least in terms of the second formant patterns. That is, the percept /gu/ is a good compromise between what is being seen and what is being heard when an acoustic /bu/-visual /gu/ or /du/ stimulus is presented. This finding thus supports the view of Summerfield (1987): “The ‘illusory’ percept that is most likely to occur is the consonant that is most easily confused auditorily with the acoustical consonant and which is most visually compatible with the visible consonant” (p. 26). The fuzzy logical model of speech perception (Massaro, 1987a,1987b) is also compatible with the findings reported here. In this model, fuzzy values represent the acoustic and visual similarities of prototypical syllables to the stimulus, and the integration of these values across modalities determines which prototype best matches the incoming sensory input. The changes found in the McGurk effect across vowel contexts provide support for these theories of speech perception, and more specifically, highlight the role of second formant frequency patterns in multimodal speech perception. Other acoustic features that are important in unimodal auditory perception, such as the third formant frequency pattern, may also play an important role in multimodal speech perception. This could be investigated further via the use of controlled synthetic acoustic stimuli.

An alternate explanation for context-dependent changes in the McGurk fusion effect is that the effect is caused by linguistic biases that change across vowel contexts (Massaro, 1998). The data from the visual-only test (Experiment 1) do reveal certain biases in the visual perception of these 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 (89% of the time), and rarely resulted in /g/ percepts (only 4% of the time). In general, the biases seen in the visual-only experiment cannot explain the frequency of different audio-visual “fusion” percepts.

Combination illusions and the relative timing of auditory and visual information

Many studies report that audiovisual integration of speech information is robust to a range of misalignments between the modalities. Although people can detect audiovisual asynchrony in brief nonspeech stimuli when the intermodal misalignments are as small as 20 ms (Hirsch and Sherrick, 1961), people are generally less capable of detecting audiovisual misalignments in speech stimuli (perhaps due to the greater complexity of the latter). For example, McGrath and Summerfield (1985) tested subjects’ abilities to detect the alignment of buzzing noises with the opening of liplike images, and concluded that subjects were unlikely to detect auditory lags in speech stimuli of less than 40 ms. Furthermore, detection of asynchrony in sentences is much greater. Dixon and Spitz (1980) reported that subjects were unable to detect auditory lags up to 257 ms and auditory leads of up to 131 ms in recorded sentences.

It is possible that the degree of misalignment required for detection of audiovisual asynchrony may be greater than the degree of misalignment necessary to disrupt subconscious integration across modalities. However, evidence from a range of studies suggests that the interaction between different modalities during bimodal speech perception is generally robust to audiovisual misalignment. For example, using pulse trains to capture the pattern of glottal pulses in the acoustic signal, McGrath and Summerfield (1985)

found that the presence of such acoustic information improved lipreading capability. This improvement in lipreading performance was robust to the introduction of auditory lags, showing little influence of lags up to 80 ms. In addition, many studies have shown that vision can influence auditory speech perception even when large misalignments (on the order of hundreds of milliseconds) are introduced between the modalities (e.g. Munhall, Gribble, Sacco and Ward, 1996; Massaro and Cohen, 1993).

Generally, studies of the integration of misaligned speech information across modalities have examined the strength of cross-modal influences (e.g. Munhall, Gribble, Sacco, and Ward, 1996; Massaro, Cohen, and Smeele, 1996). However, using incongruent audiovisual stimuli, it is possible to investigate the nature, as well as the existence, of audiovisual integration. In particular, when integration does occur, it is possible to examine whether the differences in time of arrival of information across modalities are used by the perceptual system to order the resulting percepts, and if so, at what level (e.g. syllable, phoneme, phonetic feature) such ordering occurs.

The experiments presented here used incongruent multimodal stimuli to examine how visual and acoustic speech inputs are combined by the speech perception system. The order of consonants in illusory (single syllable) combination percepts occurring in response to visual /b/ - acoustic /d/ stimuli was dependent on audio-visual alignment. This indicates that the perceptual system can interleave multimodal information at a phonemic or other sub-syllabic level based on intermodal timing of the incoming sensory input. This may be achieved by the ordering of phonemic or subphonemic percepts after phonetic categorization or by the integration of auditory and visual speech information as continuous variables prior to phonetic categorization (e.g., Summerfield, 1992; Braida, 1991).

Although phonetic context and audio-visual alignment influenced the magnitude of the combination illusion and the order in which the consonants were perceived, the nature of the illusion was robust. That is, the illusory percept always involved a combination of the two consonants presented, rather than some third consonant which was distinct from both the acoustic consonant and the visual consonant. When the modalities are reversed, however, so subjects are viewing a nonlabial consonant and listening to a labial consonant, a third consonant, which has been hypothesized to be a fusion of the visual and acoustic stimuli, is often perceived (MacDonald and McGurk, 1978). These two different effects (combination vs. fusion) illustrate that the modality in which phonemes are presented affects how they are integrated with other phonemic inputs. This is probably because the set of features defining phonemes depends upon modality, and these features vary in saliency. Summerfield (1987) proposed that the resolution of audio-visual conflicts depends on the visual confusability of the visual stimulus and the auditory confusability of the acoustic stimulus. This is similar to proposing that the saliency of features in the different modalities may determine which features are preserved in multimodal percepts. For example, both /b/ and /g/ are more likely to be perceived intact when a visual /b/ is presented in synchrony with an acoustic /g/ (combination percepts often occur to such stimuli), than when an acoustic /b/ is dubbed on a visual /g/ (fusion percepts often occur to such stimuli). This is probably because the phoneme /b/ has a very visually distinct labial closure and the phoneme /g/ has a prominent acoustic burst. These phonetic features may be so salient and unique that they cannot be ignored or fused perceptually with other features. The more subtle cues which distinguish acoustic /b/ or visual /g/ from other consonants may be more easily ignored or blended with other inputs by the perceptual system, thus allowing (for example) an acoustic /b/-visual /g/ stimulus to be perceived as /d/.

This hypothesis is consistent with the findings of Green and Norrix (1997). They reported that editing out the /g/ burst of acoustic /gV/-visual /bV/ stimuli resulted in a decrease in the number of combination illusions, and an increase in the number of /b/ percepts (except for the male talker in the /a/ vowel context). Interestingly, close examination of the data of Green and Norrix (1997) also reveals that /d/ and /č/ percepts arose after removal of the /g/ burst (although the increase in these responses was not tested for significance). This is consistent with the hypothesis that the distinctive burst of /g/ plays an important role in inducing combination percepts. Similar research may help further elucidate why some conflicting sensory inputs are fused at the phonemic level while others are concatenated.

NOTES

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, & Meltzoff, 1988; Green, Kuhl, Meltzoff & Stevens, 1991, Green & Gerdeman, 1995, and Green & 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 /α/ and the formant patterns of the stimuli (see Table 2 of Green & Norrix, 1997) are typical of the vowel /α/. As several studies of Green and colleagues shared stimuli (see Green & Gerdeman, 1995 and Green & 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 /α/ 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 /α/.
2. 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/.

ACKNOWLEDGEMENTS

This research was supported by NIH grant R01 DC02852 to F. Guenther.

REFERENCES

- Akmajian, A., Demers, R.A., Farmer, A.K., & Harnish, R.M. (1990). *Linguistics, an introduction to language and communication, third edition*. The MIT Press, Cambridge, Massachusetts.
- Braida, L. 1991. Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology*, 1991, **43A** (3), 647-677.
- Dekle, D.J., Fowler, C.A., & Funnell, M.G. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics*, **51**(4), 355-362.
- Delattre, P. & Freeman, D.C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics*, **44**, 29-68.
- Dixon, N.F. and Spitz, L. 1980. The detection of auditory and visual desynchrony. *Perception*, **9**, pp. 719-721.
- Dodd, B. (1977). The role of vision in the perception of speech. *Perception*, **6**, 31-40.
- Easton, R.D. & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics*, **32**(6), 562-570.
- Erber, N.P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech & Hearing Research*, **12**, 423-425.
- Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective.

- Journal of Phonetics*, **13**, 3-28.
- 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.
- Green, K.P. (1996). The use of auditory and visual information in phonetic perception. In Stork, D.G. & Hennecke, M.E., editors, *Speechreading by Humans and Machines*, pages 55–77. Springer-Verlag.
- 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, pp. 1409–1426.
- Green, K.P., Kuhl, P.K., & 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.
- 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, pp. 524–536.
- Green, K.P. & Miller, J.L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, **38**(3), 269–276.
- 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**, pp. 646–665.
- Guenther, F.H., Espy-Wilson, C.Y., Boyce, S.E., Matthies, M.L., Zandipour, M., & Perkell, J.S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *Journal of the Acoustical Society of America*, **105**(5), 2854–65.
- Hampson, M., Guenther, F., & 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.
- 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, pp. 388–403.
- Kent, R.D. & Read, C. (1992). *The Acoustic Analysis of Speech*. Singular Publishing Group, Inc.
- Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook*. Prentice Hall.
- 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.
- Ladefoged, P. (1993). *A Course in Phonetics*. Harcourt Brace College Publishers, third edition.
- Lieberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychological Review*, **74**:431–461.
- Lieberman, A.M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.
- MacDonald, J., Andersen, S., and Bachmann, T. 2000. Hearing by eye: How much spatial degradation can be tolerated? *Perception*, **29**(10), 1155-1168.
- MacDonald, J. and McGurk, H. 1978. Visual influences on speech perception processes. *Perception and Psychophysics*, **24** 3, pp. 253–257.
- Massaro, D.W. (1987a). Speech perception by ear and eye. In Dodd, B. & Campbell, R., editors, *Hearing by Eye: The Psychology of Lip-Reading*, pages 53–83. Hillsdale, N.J.: Lawrence Erlbaum Associ-

ates.

- Massaro, D.W. (1987b). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.
- Massaro, D.W. (1998). Illusions and issues in bimodal speech perception. Proceedings paper: Auditory-Visual Speech Processing. Terrigal, New South Wales, Australia.
- Massaro, D.W. & Cohen, M.M. (1983). Evaluation and integration of visual information in speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **9**(5), 753–771.
- Massaro, D.W. and Cohen, M.M. 1993. Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, **13**, pp. 127–134.
- Massaro, D.W., Cohen, M., and Smeele, P.M.T. 1996. Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, **100**(3), 1777-1786.
- McGrath, M. and Summerfield, Q. 1985. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, **77**(2), 678-685.
- McGurk, H. and MacDonald, J. 1976. Hearing lips and seeing voices. *Nature*, **264**, pp. 746–748.
- Munhall, K.G., Gribble, P., Sacco, L., and Ward, M. 1996. Temporal constraints on the McGurk effect. *Perception and Psychophysics*, **58** 3, pp. 351–362.
- 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, pp. 530–539.
- Ong, D. & Stone, M. (1998). Three-dimensional vocal tract shapes in /r/ and /l/: A study of MRI, ultrasound, electropalatography, and acoustics. *Phonoscope*, **1**, 1–13.
- Reisberg, D., McLean, J., & 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.
- Rosenblum, L.D. & Saldana, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **22**(2), 318–331.
- Rosenblum, L.D., Schmuckler, M.A., & Johnson, J.A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, **59**(3), 347–357.
- Schwartz, J., Robert-Rimes, J., & Escudier, P. (1998). Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. In Campbell, R., & Dodd, B., editors, *Hearing by Eye II: Advances in the psychology of speechreading and audio-visual speech*, pages 85-108. Hove, UK: Psychology Press/Erlbaum (Uk) Taylor & Francis.
- Shigeno, Sumi. 2000. Influence of vowel context on the audio-visual speech perception of voiced stop consonants. *Japanese Psychological Research*, **42**(3), 155-167.
- 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, pp. 2983.
- Smeele, P. M.T., Hahnlen, L.D., Stevens, E.B., and Kuhl, P.K. 1995. Investigating the role of specific facial information in audio-visual speech perception. *Journal of the Acoustical Society of America*, **98** 5,Pt.2, pp. 2983.
- Sumbly, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**(2), 212–215.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech percep-

- tion. In Dodd, B. & Campbell, R., editors, *Hearing by Eye: The Psychology of Lip-Reading*, pages 3–52. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Summerfield, Q. (1991). Visual perception of phonetic gestures. In Mattingly, I.G. & Studdert-Kennedy, M., editors, *Modularity and the Motor Theory of Speech Perception*, pages 117–138. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Summerfield, Q. 1992. Lipreading and audio-visual speech perception. *Phil. Trans. R. Soc. Lond. B.*, **335**, 71-78
- Venables, W.N. and Ripley, B.D. 1997. *Modern Applied Statistics with S-Plus, 2nd Edition*. Springer-Verlag NY, Inc.
- Westbury, J.R., Hashi, M., & Lindstrom, M.J. (1995). Differences among speakers in articulation of American English /r/: An x-ray microbeam study. In Elenius, K. & 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.