LANGUAGE AND SPEECH, 2006, 49 (1), 21–53 21

Visual Influences on Perception of Speech and Nonspeech Vocal-Tract Events

Lawrence Brancazio[1], **Catherine T. Best**[2] **Carol A. Fowler**[3]

[1] *Southern Connecticut State University, New Haven, CT and Haskins Laboratories, New Haven, CT*

[2] *MARCS Auditory Laboratories, University of Western Sydney, Penrith NSW, Australia and Haskins Laboratories, New Haven, CT*

[3] *University of Connecticut, Storrs, CT, Haskins Laboratories, New Haven, CT, and Yale University, New Haven, CT*

Key words

audiovisual speech perception

clicks

McGurk effect

nonspeech

Abstract

We report four experiments designed to determine whether visual information affects judgments of acoustically-specified nonspeech events as well as speech events (the "McGurk effect"). Previous findings have shown only weak McGurk effects for nonspeech stimuli, whereas strong effects are found for consonants. We used click sounds that serve as consonants in some African languages, but that are perceived as nonspeech by American English listeners. We found a significant McGurk effect for clicks presented in isolation that was much smaller than that found for stop-consonant-vowel syllables. In subsequent experiments, we found strong McGurk effects, comparable to those found for English syllables, for click-vowel syllables, and weak effects, comparable to those found for isolated clicks, for excised release bursts of stop consonants presented in isolation. We interpret these findings as evidence that the potential contributions of speech-specific processes on the McGurk effect are limited, and discuss the results in relation to current explanations for the McGurk effect.

1 Introduction

In 1976, McGurk and MacDonald reported an effect on speech perception of dubbing an acoustic signal corresponding to one consonant-vowel sequence (e.g., /ba/-/ba/) onto a videotaped event in which a speaker mouthed a different sequence (e.g., /ga/-/ga/). Listeners reported hearing a consonant reflecting the integration of phonetic information acquired optically and acoustically (e.g., /da/-/da/). Subsequent research

Acknowledgments: This research was supported by NIH Grant, HD-01994 and DC-00403 to Haskins Laboratories. We thank Julia Irwin for help with videotaping and digitizing the movies and Robert Dietz for running subjects.

Address for correspondence: L. Brancazio, Department of Psychology, 501 Crescent St., Southern Connecticut State University, New Haven, CT 06515;
e-mail: <brancazioL1@southernct.edu>.

(e.g., MacDonald & McGurk, 1978) demonstrated that certain dubbings will result in perception of the optically-specified consonant (e.g., /da/ when auditory /ba/ is dubbed with visual /da/) or a consonant that incorporates features from each modality (e.g., /na/, an alveolar nasal, when an auditory bilabial nasal /ma/ is dubbed with a visual alveolar plosive /da/). This phenomenon, which we will refer to as the "McGurk effect," following popular practice (e.g., Green, Kuhl, Meltzoff, & Stevens, 1991; Rosenblum & Saldaña, 1992), has excited considerable interest and has stimulated much research. A reason for this interest in the McGurk effect, aside from its phenomenal vividness, is its demonstration of cross-modal contributions to a unified perceptual experience. In particular, it provides compelling evidence that speech perception is not only an auditory process, but involves the extraction of a phonetic message across modalities.

The McGurk effect has been reported for a variety of audio-visually incongruent speech stimuli and is often phenomenally quite compelling (Manuel, Repp, Liberman, & Studdert-Kennedy, 1983). However, it varies in magnitude and even fails to occur in some dubbings. For example, an auditory syllable /da/ presented with a visual /ga/ is perceived as /da/ despite the audiovisual incongruency (MacDonald & McGurk, 1978), presumably because the visual distinction between /da/ and /ga/ is subtle. McGurk effects for vowels, although reliable, are small (Massaro & Cohen, 1993; Summerfield & McGrath, 1984). Such findings indicate that there are constraints on the range of audiovisual dubbings for which there is a visual influence on perception.

To date, the limiting conditions for the McGurk effect have not been clearly established. One question regarding the scope of the McGurk effect concerns whether it is restricted to phonologically relevant stimuli (e.g., speech sounds), or whether it might occur for appropriately constructed nonspeech stimuli. This question highlights a long-running debate among speech researchers, namely whether speech perception is accomplished by specialized brain mechanisms that are solely devoted to the task of perceiving speech, or whether it is accomplished by means of general perceptual systems.

The most prominent theory associated with the specialized speech mechanism view is the motor theory (Liberman & Mattingly, 1985). According to motor theorists, speech perception is achieved by a specialized phonetic "module" (see Fodor, 1983) that retrieves gestures of the vocal tract, whereas nonspeech acoustic inputs undergo general auditory system analysis. By this view, the McGurk effect is a consequence of processing gestural information in both the acoustic and optical signals by the phonetic module. Thus, McGurk effects should be limited to stimuli that engage the phonetic module or any other brain module that makes use both of optical and of acoustic information of a specific type (such as, perhaps, one responsible for sound localization, a domain in which visual effects are well-known; Driver, 1996; Radeau, 1994). Possibly, McGurk-like effects would occur for other stimuli as well by means of some higher-level cognitive process, but these effects would "not portray the striking power of the McGurk effect" (Saldaña & Rosenblum, 1993, p.407).

In contrast, other theories do not maintain a distinction between the perception of speech and nonspeech stimuli. For example, proponents of the direct-realist theory of speech perception (Best, 1984, 1995; Fowler, 1986, 1994; Rosenblum, 1987), while

agreeing with motor theorists that the objects of speech perception are gestures, propose that perception of speech is fundamentally like perception of other events: both nonspeech and speech events create structure in acoustic and optical (and haptic; Fowler & Dekle, 1991) signals that specify their source, and perceptual systems detect this structure as information about the event (Gibson, 1966; 1979). According to this view, the McGurk effect occurs because information in the dubbed audio-visual signal corresponds to an event that contrasts with the event specified by the acoustic signal alone, regardless of whether the stimuli are speech. To date, proponents of this view have not explored what the critical factors for a strong McGurk effect might be, but from a direct-realist perspective, the critical factors should involve aspects of the underlying kinematic properties of the dubbed events and how they are specified by the acoustic and optical signals (see Rosenblum & Saldaña, 1996, for a discussion of the role of kinematic primitives in audio-visual integration).

Another theory that rejects the view that the McGurk effect is specific to speech is the Fuzzy Logical Model of Perception (or FLMP, Massaro, 1987, 1998), According to the FLMP, categorization in both speech and nonspeech involves extracting features from input signals and using them to evaluate stored prototypes. The McGurk effect arises for audio-visually incongruent stimuli because of the contributions of visual features in the selection of a prototype. By this view, the McGurk effect can occur for both speech and nonspeech stimuli, provided that the relevant prototypes include both auditory and visual features. However, because prototype descriptions develop from experience, the effect will occur only if there are sufficient built-up associations of auditory and visual features (see also Diehl & Kluender, 1989; but see Fowler & Dekle, 1991, for findings challenging this assumption). Additionally, according to the theory, the magnitude of the McGurk effect will be related to the relative ambiguity of the auditory and visual cues.

To date, few studies have offered evidence for a strong nonspeech McGurk effect involving visual influences on auditory event identification (excluding visual influences on auditory *location*; cf. Rosenblum, 1994). Rosenblum and Fowler (1991) presented a model clapping his hands at different levels of visible effort and auditory loudness, with the auditory and visual levels of effort cross-dubbed. When listeners were instructed to rate the loudness of the claps, based only on what they heard, the effects of video effort on loudness judgments were small and reached significance only at one of the four levels of auditory effort. Saldaña and Rosenblum (1993) tested for a McGurk effect using dubbed cello bow and pluck sounds: they created a continuum of cello sounds ranging from a bow sound to a pluck, which were dubbed onto video presentations of a person bowing or plucking a cello string. There were significant effects of the video display on ratings of the sounds along a pluck-bow continuum, but the effect lacked the phenomenal vividness of the McGurk effect for speech; notably, the dubbed videos failed to turn a clear pluck sound into a bow sound or vice versa, unlike the effect with dubbed consonants. More recently, de Gelder and Vroomen (2000; see also Massaro & Egan, 1996) found a visual influence of static faces with happy and fearful expressions on judgment of emotion expressed vocally, but as with the findings for hand-claps and plucks and bows, this effect was weaker and qualitatively different than the effect found for dubbed consonants. Thus, the

visual influences on auditory judgments of nonspeech events observed to date have been considerably weaker than the effect found for consonants.

These weak McGurk effects for nonspeech events appear to be consistent with the view that phonological significance is required for a strong McGurk effect. However, this is a premature conclusion, because these stimuli differed from consonants in other ways than the speech/nonspeech distinction. For example, as pointed out by Saldaña and Rosenblum (1993), the McGurk effect might be stronger for categorically perceived stimuli (such as stop consonants) than for continuously perceived stimuli (such as vowels and plucks and bows). The effect might also be stronger for stimuli with less robust auditory cues, as predicted by the FLMP (Massaro, 1987); auditory cues for stop consonants are arguably more transient than the ones that distinguish, for example, plucks from bows. Stop consonants differ along very different acoustic dimensions than the nonspeech stimuli that have thus far been examined: acoustically, stop consonants differ primarily in frequency at their onsets (e.g., Fant, 1973); in contrast, Saldaña and Rosenblum's plucks and bows and Rosenblum and Fowler's loud and soft hand-claps differed from one another only in variations in amplitude, and sentences varying in emotion differ primarily in F0. It is possible that these distinctions in physical (or psychophysical) properties, rather than the presence versus absence of phonological significance, underlie the variation in the magnitude of the McGurk effect between stop consonants and the nonspeech stimuli tested to date. In other words, perhaps a strong McGurk effect occurs for stimuli that have certain physical properties, and stop consonants happen to possess those properties.

The purpose of the present research was to investigate the importance of phonological status in relation to nonlinguistic stimulus properties in fostering a strong McGurk effect. In doing so, we attempted to address the question of whether speech-specific mechanisms are responsible for the McGurk effect.

Because there is a broad range of potentially critical physical or event properties, we opted to take an incremental approach to the problem: We tested the McGurk effect for stimuli that share many properties of stop consonants, but that would not be identified by our listeners as speech. Specifically, we used variants of the consonantal clicks that serve as phones in some languages of Africa. These have kinematic properties that are similar to those of consonants of spoken English: they are produced by making complete constrictions somewhere in the oral cavity and then releasing the constrictions, resulting in a distinct pattern of change in the acoustic frequencies at release, as in English stop production. The clicks we selected for use are also visibly distinct from one another, thus providing an appropriate environment for the McGurk effect to emerge. Moreover, previous work has found that for native speakers of English, clicks are perceived as nonspeech (Best, McRoberts, & Sithole, 1988); notably, Best and Avery (1999) found that native speakers of English do not exhibit a right-ear advantage when discriminating clicks although they do for native English consonant contrasts, whereas native speakers of a click language (Zulu) do show a right-ear advantage for the same click stimuli.

In using these clicks, we narrowed our scope from the general question of whether certain nonlinguistic properties are required for a strong McGurk effect, to the more specific question of whether the McGurk effect is always strong for a

particular type of event, namely those which involve a rapid release of a vocal tract constriction. In doing so, we followed the recommendation of Saldaña and Rosenblum (1993) that "future research should be designed to implement nonspeech sounds that have characteristics of consonants ... in order to demonstrate a more phenomenally striking nonspeech McGurk effect" (p.415).

In the following experiments, we examined the magnitude of the McGurk effect for voiceless stop consonants in consonant-vowel syllables (in Experiment 1) and for isolated clicks (in Experiment 2). A finding of McGurk effects of comparable magnitudes in the two experiments would indicate that the McGurk effect is strong for stimuli that involve a rapid release of vocal tract constrictions, regardless of whether or not they are perceived as speech. In contrast, a finding of a weaker McGurk effect for clicks could be attributed either to their lack of phonological significance or to the physical differences that exist between the stop consonant syllables and the isolated clicks. Experiments 3 and 4 provided a systematic examination of the relative contributions of these physical differences; they tested the McGurk effect for clicks coarticulated with a following vowel and for stop bursts presented without a following vowel, respectively. Thus, these experiments were designed to examine the relative contributions of both physical factors and phonological significance, the latter of which we expected to differ across the stimuli of the four experiments.

2 Experiment 1

The purpose of the first experiment was to establish the McGurk effect using English voiceless stop consonant syllables, namely, /pa/, /ta/, and /ka/. We used voiceless stops, as opposed to the more commonly used voiced stops, because in a later experiment we only presented the release bursts of the stops, and the bursts of voiced consonants would not have been appropriate for this use. Typically, McGurk experiments are conducted by instructing participants to identify a critical phoneme in a syllable or word. However, when the stimuli are unfamiliar, such as the clicks we used in subsequent experiments, participants may not have consistent labels for them, and an identification task may provide unreliable results. Therefore, we used an AXB categorization task in which participants were required to compare two stimuli (A and B) to an anchor stimulus (X) and to choose the one that was the better match. Although an AXB task has been used before to test for a McGurk effect (Rosenblum & Saldaña, 1992), the format we adopted was unique.

In our test, X was presented only auditorily, whereas A and B (henceforth the 'test tokens') were presented audio-visually. In this way, participants could not base their matching decisions on the visual similarity of the A and B tokens to the X token. The auditory component of either A or B matched X in place of articulation, while the other had a different auditory place of articulation. We had three kinds of trials. First, in "Auditory-Alone" trials, the auditory tokens were presented without videos, in order to establish the overall categorizability of the auditory stimuli. In the "Match" trials, A and B were both audio-visually congruent, whereas in the "Mismatch" trials, the visual displays were switched so that both were audio-visually incongruent. Table 1 shows examples of the Match and Mismatch Trials. The effect of incongruent dubbings on Mismatch trials was determined by comparing performance on these

trials to otherwise identical trials with congruent dubbings (Match trials). If, on a Mismatch trial, the test token that acoustically differs from X in place of articulation (e.g., B in Table 1) sounds like a better category match to X than does the test token that shares auditory place of articulation with X (A in Table 1), then participants will choose B as the better match to X. Thus, the test requires a visually-induced change in the perceived category of at least one of the test tokens. If there is no visual effect, then the audio-visually incongruent token that is an auditory match to X should be selected consistently in both the Match and Mismatch conditions.

Table 1

Examples of the Match and Mismatch AXB trial types, using stimuli from Experiment 1

Stimulus Modality	*Match*			*Mismatch*		
	A	X	B	A	X	B
Auditory	/pa/	/pa/	/ta/	/pa/	/pa/	/ta/
Visual	/pa/		/ta/	/ta/		/pa/

2.1 *Method*

2.1.1 Participants

The participants were 25 undergraduates at the University of Connecticut. All received course credit for their participation. All reported normal hearing and normal or corrected-to-normal eyesight, and all were native speakers of English. Data from one participant were dropped from the analyses for failure to meet our performance criteria (see Procedure below).

2.1.2 Materials

2.1.2.1 *Visual stimuli*

Visual stimuli were recorded in a room with bright lighting and a plain backdrop. An adult female (the second author) was videotaped producing the syllables /ba/, /da/, and /ga/ in randomized sequences. The tokens were digitized at 30 frames/second at 320-by-240 pixels using a Radius board mounted in a Macintosh computer, and edited using Adobe Premiere (San Jose, CA). One token of each syllable was selected from these productions, based on the clarity of the articulatory movements and visual similarity (in duration, head position, etc.) across the tokens of different consonants. Movies of each syllable were truncated so that they were the same duration (1630ms), had at least two frames preceding the onset of the consonant gesture, and included the full duration of the vowel opening and closing gestures.

2.1.2.2 *Auditory stimuli*

To obtain noise-free recordings, the auditory tokens were recorded in a separate session, in a soundproof booth. The same talker produced multiple tokens of the syllables /pa/, /ta/, and /ka/ in randomized sequences, which were recorded onto a digital audio tape recorder (DAT) and then input to the Haskins VAX computer system in pulse code modulation (PCM) format, where acoustic analyses were performed. The duration and fundamental frequency (F0) of each syllable were measured in the acoustic analysis program HADES (Rubin, 1995), and three tokens of each syllable were selected that were roughly matched in duration and F0. The stimuli were then digitally amplified so that they were matched in peak RMS amplitude as well. Of the three tokens of each category, one was selected to be dubbed (for A and B tokens of the AXB test), while the other two were dubbed to a continuous black video screen (for X). The acoustic A and B tokens were selected because they were the most similar to one another in duration and F0 across the three categories. In that way, we made the acoustic differences between the test (dubbed) tokens and the comparison (X) tokens as uniform as possible.

2.1.2.3 *Audio-visual dubbing*

The selected auditory tokens were converted into AIFF format, and then imported into Adobe Premiere, where those selected to be dubbed were paired with the digitized visual token. Dubbing was accomplished by aligning the acoustic release burst with the video frame in which the consonant release was first visible. Auditory /pa/ was paired with visual /ba/, /da/, and /ga/ (which, because they are visually indistinguishable from their voiceless counterparts, will henceforth be called /pa/, /ta/, and /ka/, respectively); auditory /ta/ was paired with visual /pa/ and /ta/; and auditory /ka/ was paired with visual /pa/ and /ka/. (The /ta/-/ka/ and /ka/-/ta/ combinations were not used because the alveolar and velar places of articulation are not sufficiently distinct to give rise to a McGurk effect; e.g., MacDonald & McGurk, 1978). Each of the dubbed audiovisual tokens was saved as an individual movie file. Additionally, movies were made of each auditory token paired with a black screen.

2.1.2.4 *Test sequences*

AXB trials were constructed by concatenating a dubbed movie, an X token movie (a black screen), and another dubbed movie, followed by a 3.5s long silent movie of a black screen. The interstimulus intervals (ISI), measured from the acoustic offset of one token to the acoustic onset of the next token within a trial, were approximately 1050ms. Sixteen AXB trials were constructed for each of the three trial types (Audio-Alone, Match and Mismatch), along with two other trial types that will not be reported here, for a total of 80 trials.

In the Auditory-Only, Match, and Mismatch trials, the consonant contrast between the A and B stimuli was either /p/-/t/ or /p/-/k/; the X token was either /pa/ or the alternative consonant in the contrast (/ta/ or /ka/); and the assignment of /pa/ and /ta/ or /ka/ to either the A or B position of the triad was balanced. Additionally, two X tokens of each syllable were used. The three classes of trial differed in the visual components: in Auditory-Only trials, the visual component was a black screen for A, X, and B; in Match trials, A and B had congruent visual tokens (e.g., auditory

/pa/-visual /pa/; auditory /ta/-visual /ta/); in the Mismatch trials, the visual components of the A and B stimuli in the Match trials were switched.

We created a randomized sequence of the 16 Audio-Alone AXB trials, and a separate randomized sequence of the 64 Audio-Visual AXB trials (with a full randomization of all trial types). Additionally, practice sets were created for the Audio-Alone and Audio-Visual sequences: each contained five AXB trials, and in the Audio-Visual practice sequence they were all trials from the Match condition. The sequences were output directly to videotape using Adobe Premiere. All participants received the same Audio-Alone sequence and the same Audio-Visual sequence.

2.1.3 Procedure

Participants were given answer sheets that offered 'A' and 'B' as response options for each trial. In the Audio-Alone portion, they were instructed that they would hear sets of three English syllables and that they were to write on their answer sheet which of the first or the third syllables sounded more like the second. In the Audio-Visual portion, they were told that the first and third syllables in the sequence, but not the second, would be accompanied by a visual presentation of a face saying a syllable. Participants were informed that the syllables were dubbed, and that the visual presentation would not necessarily match the acoustic signal. Participants were instructed to base their decisions only on how similar the syllables *sounded*. It was also stressed that participants should, nonetheless, watch the video screen at all times except when they were marking their answer sheet. The participants were shown the practice sequence for each portion of the experiment immediately preceding the respective test sequence, and were given feedback if requested.

In this and all of the other experiments, the order of the presentation of the Audio-Alone and Audio-Visual tests was counterbalanced so that half of the participants performed the Audio-Alone test first, and half the Audio-Visual test.

The sequences were presented visually to the participants on a color TV monitor (20 inch screen) with full-screen video, and played through a VCR. The sound was fed from the VCR output through an amplifier, and played through an eight-inch speaker mounted on top of the monitor. Participants were seated approximately eight feet from the monitor, and were run in groups of one to four.

2.2 *Results*

There were two screening measures. Participants had to make correct responses in at least 70% of trials in the Audio-Alone condition and in at least 70% of trials in the Audio-Visual Match condition. The data of one participant failed to meet either of these criteria and were excluded from all analyses. The remaining participants averaged 98.2% correct responses on the Audio-Alone trials.

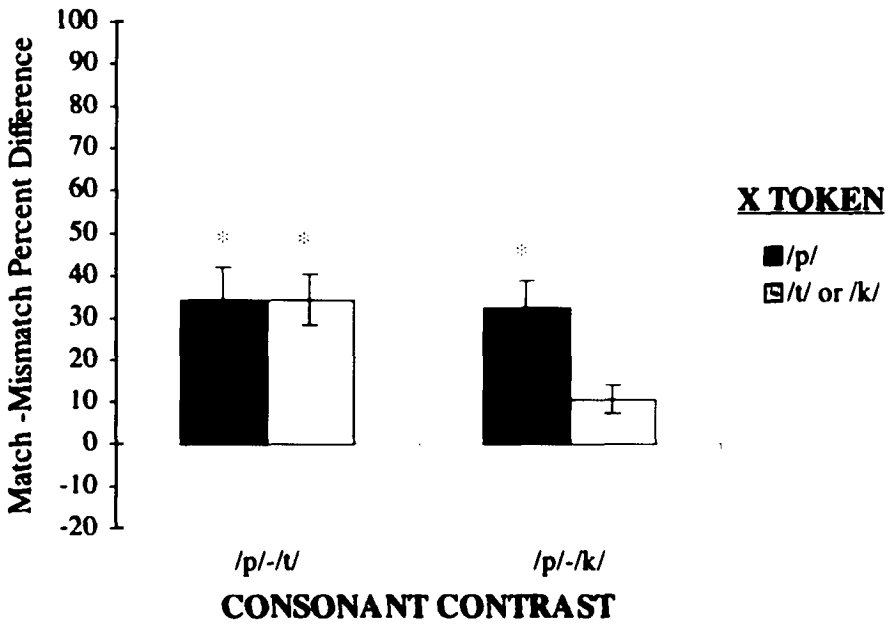
On Match and Mismatch trials, we computed the percentage of trials on which participants selected the audio-visual syllable with the same auditory place of articulation as X as the better match to X. This percentage should be low on Mismatch trials to the extent that the incongruent dubbings alter the perception of the A and

B tokens, such that the syllable with an auditory match X sounds less like X, and the other syllable, whose visual (but not auditory) component matches auditory X in place of articulation, sounds more like X. Match trials, on which the percentage should be high, provided a baseline for assessment of performance on Mismatch trials. Thus, our dependent measure was the Match minus Mismatch difference in the percentage selection of syllables with an auditory match to X. The consonant contrast on each trial was either /pa/-/ta/ or /pa/-/ka/. The X token was either /pa/ or the nonlabial alternative (/ta/ or /ka/).

The overall difference between the Match and Mismatch conditions was 27.9%, a value significantly greater than zero, $t(95) = 8.88$, $p < .0001$. Analogous one-sample t-tests performed on each cell mean revealed that each one differed from zero at $p < .005$ or less, smallest $t(23) = 3.12$. An ANOVA on the difference scores revealed significant main effects of Consonant Contrast, $F(1, 23) = 6.59$, $p < .05$, and of X token, $F(1, 23) = 5.38$, $p < .05$, as well as a significant interaction, $F(1, 23) = 6.34$, $p < .05$, shown in Figure 1. The Match-Mismatch difference score was nearly identical in three of the cells; it was much smaller when X was /ka/ in /pa/-/ka/ contrasts than in the other conditions.

Figure 1

Difference between Match and Mismatch trials in the percentage of selections of the audio-visual token with an auditory place-of-articulation match to X, for English stop consonant syllables in Experiment 1. Means are grouped according to the Consonant Contrast (/p/-/t/ or /p/-/k/) and the identity of the X token. Asterisks specify difference scores that are significantly greater than zero



2.3 Discussion

The results demonstrate a visual influence on speech perception, and further revealed differences in the strength of the effect depending on the phonological contrasts and the particular audiovisual pairings; these differences are broadly consistent with the typical phonetic classification for these audiovisual pairings. Although we will not do so for the later experiments, where our primary focus is the overall effect magnitude, here we will discuss the results in some depth in order to provide an understanding of how results in the AXB tasks correspond to typical phonetic classification results.

We found significant mismatch effects for all of our conditions, but the effect was much weaker in one condition, namely when the contrast was between /pa/ and /ka/, and X was /ka/, than in the others. The pattern of results can be interpreted in light of the typical responses for the incongruent stimuli. Table 2 presents, for the four test conditions in the Mismatch condition, the typical visually-influenced percepts for the incongruent tokens in the AXB trials. Across the four conditions, the typical McGurk percept for the incongruent token with an auditory match to X is either a different consonant than X (rows 1 and 3 of Table 2) or a combination of consonants that includes X (rows 2 and 4). Likewise, the typical McGurk percept for the incongruent token with a different auditory place of articulation than X is usually either the same consonant as X (row 2) or a combination that includes X (rows 1 and 3). The exception to this is the condition (row 4) in which X is /ka/ and the incongruent token is $A_{/pa/}$-$V_{/ka/}$, which is typically perceived as /ta/. Table 2 demonstrates that in the first three conditions the expected McGurk percept for the incongruent token with a different auditory place of articulation than X provides a better match to X. In contrast, in the fourth condition the McGurk percept for the incongruent token with the auditory match to X provides a better match, consistent with the smaller mismatch effect.

Table 2

Design of Mismatch trials in Experiment 1, along with typical visually-influenced responses to the incongruent A and B stimuli, and expected responses if there is no visual influence ('Auditory Choice') and if there is a McGurk effect ('McGurk Choice')

Consonant Contrast	*X Token*	*Sample AXB Trial*			*Auditory Choice*	*Expected McGurk Percepts*			*McGurk Choice*
		A	X	B		A	X	B	
/pa/-/ta/	/pa/	$A_{/pa/}$-$V_{/ta/}$	$A_{/pa/}$	$A_{/ta/}$-$V_{/pa/}$	A	/ta/	/pa/	/pta/	B
/pa/-/ta/	/ta/	$A_{/ta/}$-$V_{/pa/}$	$A_{/ta/}$	$A_{/pa/}$-$V_{/ta/}$	A	/pta/	/ta/	/ta/	B
/pa/-/ka/	/pa/	$A_{/pa/}$-$V_{/ka/}$	$A_{/pa/}$	$A_{/ka/}$-$V_{/pa/}$	A	/ta/	/pa/	/pka/	B
/pa/-/ka/	/ka/	$A_{/ka/}$-$V_{/pa/}$	$A_{/ka/}$	$A_{/pa/}$-$V_{/ka/}$	A	/pka/	/ka/	/ta/	A

Overall, Experiment 1 showed that our AXB task captures the typically-reported perceptual effects of audiovisual incongruities in speech. We turn now to our test for

a McGurk effect on stimuli that lack phonological significance for our participants, namely vocal-tract clicks.

3 Experiment 2

In our attempt to find a nonspeech McGurk effect, we used a bilabial click (which is rare among click languages, used only by speakers of southern Khosian languages such as !Xõõ (Ladefoged & Maddieson, 1996; Ladefoged & Traill, 1994) along with a dental click and a lateral click, used in a number of other African languages (e.g., Zulu, Xhosa). Clicks are produced with two locations of complete closure in the oral cavity permitting a suction to be formed between them when the primary articulator is drawn down for release of the anterior closure. The posterior closure (velar, except for bilabial clicks, in which case it is dental) is released only after the anterior constriction release. Thus the initial release produces a suction release noise. Informally, a bilabial click resembles a "kissing" sound; dentals can be described as a "tsk" sound, and laterals (which have a visual asymmetry of the jaw and tongue during the click production) produce a sharp "cluck" sound (similar to a "giddyap"). The bilabial and dental clicks are relatively similar acoustically, both in frequency and duration, and are both fairly distinct from the lateral (Ladefoged & Traill, 1994). Velar clicks do not exist in any language, and would in fact be articulatorily impossible given the requirement that clicks have two releases, normally with a velar secondary release. The stimuli for Experiment 2 were clicks produced in isolation; that is, they were produced without a following vowel.

These clicks are similar to English stop-vowel syllables of Experiment 1 in that they involve a full closure of the vocal tract followed by an abrupt release. Additionally, the places of articulation of two of the clicks we used, namely the bilabial and dental, are similar to those of /p/ (a bilabial) and /t/ (an alveolar), two of the stimuli in Experiment 1. However, the clicks differ from the stimuli of Experiment 1 in phonological status, as they are not native speech sounds for English-speaking listeners, who have been found to hear clicks as nonspeech sounds (Best et al., 1988). The clicks also differ from the stimuli of Experiment 1 in physical properties; they lack a following vowel and vocalic transitions from the consonant release to the vowel steady-state portion, and the click releases differ from English stop consonant bursts both aerodynamically and in their acoustic makeup (Ladefoged & Traill, 1994). We will return to these physical differences in the discussion.

We tested for effects of incongruent dubbings on perception of the clicks, using the procedures of Experiment 1. This enabled us to test the hypothesis that stimuli that are similar to stop consonants in that they involve a rapid release of a vocal tract constriction will exhibit a strong McGurk effect, regardless of their phonological status.

3.1 *Method*

3.1.1 Participants

Participants were 25 undergraduates at the University of Connecticut. All received course credit for their participation. They all reported normal hearing and normal

or corrected-to-normal eyesight, and all were native speakers of English. An additional 10 undergraduates served as participants in a follow-up test of the perceived phonological significance of the stimuli.

3.1.2 Materials

3.1.2.1 *Visual stimuli*

The visual stimuli were recorded in the same session, by the same speaker, as in Experiment 1. The speaker is a native speaker of English, but has phonetic training and can produce clicks, both in isolation and coarticulated with a vowel. The talker produced six repetitions of bilabial, dental, and lateral clicks in isolation, intermixed with other productions in randomized sequences. Of these, one bilabial, one dental, and one lateral production were selected to provide videos for the experiment. All digitizing, selecting, and editing procedures were the same as in Experiment 1. The visual stimuli were all 1400ms in duration.

3.1.2.2 *Auditory stimuli*

The auditory tokens were recorded separately, in the same session (with the same recording conditions) as in Experiment 1. The speaker produced up to 20 tokens each of the bilabial, dental, and lateral clicks in isolation (in randomized sequences); these were recorded to DAT and input into the Haskins VAX system in PCM format, where acoustic measurements were performed. We measured click durations and centroid frequency at the click midpoint. Three tokens from each click category were selected on the basis of similarity among these measurements. The stimuli were digitally amplified, where necessary, so that the clicks were matched in RMS amplitude. As in the previous experiments, one of the three tokens served as the dubbed tokens; the others served as X tokens.

3.1.2.3 *Audiovisual dubbing*

The procedure for dubbing was identical to that used in Experiment 1, except that here all possible dubbing pairs were used (i.e., bilabial-dental, bilabial-lateral, and dental-lateral).[1]

3.1.2.4 *Test sequences*

The procedure was identical to that of Experiment 1. However, because more contrasts were used in this experiment, the Audio-Alone test sequence had 24 trials (instead of 16), and the Audio-Visual test sequence had 96 trials (rather than 64). The acoustic ISIs were approximately 1375ms. The difference in ISI from Experiment 1 is primarily due to the acoustic signals (isolated clicks) used here being much shorter than the

[1] Although our tests included dental-lateral contrasts, we will not present data from them. This contrast involves two nonlabial places of articulation, and tests with stop consonants do not typically find evidence for a McGurk effect with such contrasts (e.g., MacDonald & McGurk, 1978). In line with those findings, we found no McGurk influences in these cross dubbings.

syllables in Experiment 1 while the visual stimuli in the two experiments were more similar in duration.

3.1.3 Procedure

The instructions were similar to those used in Experiment 1, except that the clicks were described as "sounds." Thus, in both the Audio-Alone and Audio-Visual portions of the test, participants were told that they would hear a sequence of three short sounds and that they were to indicate on their answer sheets whether the first or third sounded most like the middle one; in the Audio-Visual portion, they were told that the first and third sounds would be accompanied by a video display of a woman opening her mouth. No reference was made to the fact that the sounds themselves were produced by a vocal tract. In other regards, the experiment was conducted in exactly the same manner as Experiment 1.

3.1.3.1 *Phonological significance test*

We conducted a follow-up test to confirm that the click stimuli were not perceived as speech. Participants in this phonological significance test received the same instructions as the participants in the regular experiment. However, they were only presented with the Audio-Visual portion of the experiment, and only the first 24 trials were shown. They were then given a questionnaire in which they were asked to describe the stimuli. The first item of the questionnaire asked the participants, in an open-ended manner, to describe the qualities of the sounds that had been presented. Subsequent questions asked specifically whether the sounds resembled (1) any particular environmental events, (2) any nonspeech mouth sounds, and (3) any speech sounds. In each case that participants responded "yes," they were asked to describe what the stimuli resembled.

3.2 Results

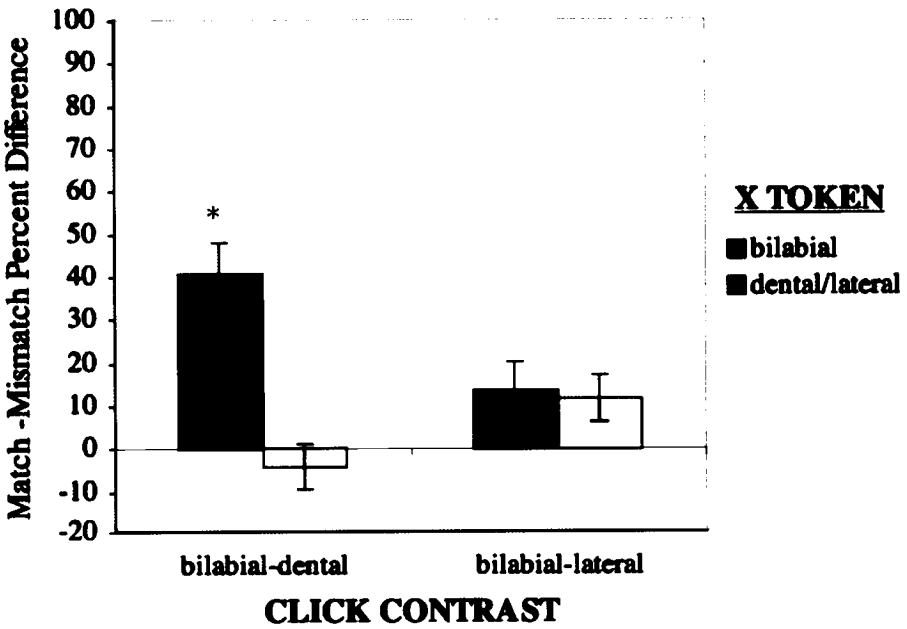
The data of one participant failed to meet our performance criteria (see Experiment 1) and were excluded from all analyses. In the Audio-Alone part of the experiment, the remaining participants overall selected the test token that matched X 87.5% of the time.

As in Experiment 1, we examined the difference between Match and Mismatch conditions in the percentage of selections of the token that matched X in auditory place of articulation. The click contrast was either bilabial-dental or bilabial-lateral, and the X token was either bilabial or the alternative (dental or lateral).

Overall, participants selected the X token's auditory match 15.4% less often in the Mismatch than in the Match condition; this overall difference score is significantly greater than zero, $t(95) = 4.39$, $p < .0001$. One-sample t-tests performed on each cell mean of the Click Contrast by X-token crossing (see Fig. 2) found a significant difference from zero at the $p < .01$ level only for the bilabial-dental comparison when X was bilabial, $t(23) = 5.42$, $p < .0001$. The Match-Mismatch difference score was 40.6% in that condition. Thus, there was a reliable effect of audiovisual discrepancy on matching judgments only with particular audiovisual pairings.

Figure 2

Difference between Match and Mismatch trials in the percentage of selections of the audiovisual token with an auditory place-of-articulation match to X, for isolated clicks in Experiment 2. Means are grouped according to the Click Contrast (bilabial-dental or bilabial-lateral) and the identity of the X token. Asterisks specify difference scores that are significantly greater than zero



An ANOVA on the difference scores revealed a significant main effect of X token, $F(1, 23) = 7.62$, $p < .05$, with the difference score being greater when X was bilabial than when it was dental or lateral. There was also a significant interaction of Click Contrast and X token, $F(1, 23) = 39.33$ $p < .0001$. Simple-effects analyses on this interaction revealed that the X token effect was significant only for the bilabial-dental comparison, $F(1, 23) = 23.96$, $p < .0001$. The interaction is displayed in Figure 2, which demonstrates that the Match-Mismatch difference for the bilabial-dental comparison when X is bilabial is much greater than in any other condition.

3.2.1 Phonological significance test

In answer to the first, open-ended, question, all 10 of the participants in the phonological significance test reported that they perceived the stimuli primarily as nonspeech events. In the subsequent specific questions, five participants responded that some of the stimuli resembled speech sounds; however, most of these speech identifications involved /k/, which most likely corresponds to the lateral click. In contrast, nine of the 10 participants reported that the stimuli sounded like nonspeech mouth sounds (such as kissing). Thus, it appears that the stimuli were not, for the most part, treated as speech stimuli by our participants.

3.3
Discussion

The purpose of Experiment 2 was to test for a McGurk effect using nonspeech stimuli that are physically similar to English stop consonant syllables, namely vocal-tract clicks. We did find a significant effect of audiovisual incongruency for these clicks. However, the visual effect was limited in its scope; there was an effect only for the bilabial-dental contrast, not the bilabial-lateral, and only when the comparison token was bilabial, not dental. The difference between the bilabial-dental and bilabial-lateral contrasts may reflect the overall greater acoustic similarity between the bilabial and dental clicks than between the bilabial and lateral clicks: The bilabial and dental clicks we used both had gradual amplitude contours, whereas the lateral clicks had very abrupt and high-amplitude onsets. The finding that the effect was found for the bilabial-dental contrast only when the X token was bilabial is difficult to interpret because we do not know exactly how the incongruent stimuli were perceived, but we note that the finding has a parallel in the Experiment 1 results with the /p/-/k/ contrast. One possible explanation for this finding, in line with our explanation from Experiment 1, is that the A_{den} -$V_{bil.}$ token was perceived as a combination of a bilabial and a dental, and was therefore compatible with both the bilabial and dental X tokens, whereas the $A_{bil.}$-V_{den} token was perceived as something intermediate between a bilabial and dental (but closer to the bilabial X token). This would result in the A_{den} -$V_{bil.}$ token being consistently selected as the better match to the dental X token, and often selected as the better match to the bilabial X token, which is what we found.

Our phonological significance test indicates that participants in our study primarily perceived the stimuli as nonspeech vocal-tract events. Therefore, the experiment provides evidence for a nonspeech McGurk effect. This finding demonstrates that phonological significance is not a necessary condition for a McGurk effect to occur, and thus concurs with findings of effects of audiovisual incongruencies for nonspeech events reported by Rosenblum and Fowler (1991) and Saldaña and Rosenblum (1993).

Although we did find a McGurk effect for clicks, there were clear differences between the results for stop consonants in Experiment 1 and the clicks in Experiment 2. Overall, the McGurk effect for isolated clicks was much less robust than the effect for English stop-consonant syllables. There was a 27.9% difference between the Match and Mismatch trials in Experiment 1 but only a 15.4% difference in Experiment 2; moreover, the effect only occurred in one of the four conditions in Experiment 2, but three of the four in Experiment 1.

The weaker McGurk effect for clicks is consistent with the view that the McGurk effect is stronger for stimuli with phonological significance. However, the weaker effect could also be the consequence of physical differences between the stop consonants and the clicks, rather than the clicks' nonspeech status. As mentioned earlier, there are two primary ways in which the stimuli of Experiments 1 and 2 differ physically. The first is syllable structure, namely the presence versus absence of vocalic transitions and a following vowel. One of the hallmarks of the production of stop consonants is that they are coarticulated with a vowel, either preceding or following. The consequences of consonant-vowel coarticulation on the acoustic signal, manifested in

formant transitions, are well-known to provide rich information for stop consonant identification (e.g., Liberman, Delattre, Cooper, & Gerstman, 1954). Accordingly, vocalic transitions might play an important role in the processes that underlie the McGurk effect (see Green & Norrix, 1997). The second physical difference is the consonant release. Although stop consonants and clicks are similar in that they involve the sudden onset of airflow from a full closure of the vocal tract, they do differ aerodynamically; click releases produce an initial influx of air, unlike stop consonants. As a result, click and corresponding stop consonant releases are not acoustically equivalent (Ladefoged & Traill, 1994).

We examine the importance of these physical differences for the McGurk effect in Experiments 3 and 4. In Experiment 3, we test for a McGurk effect using clicks that are coarticulated with a following vowel; in Experiment 4, we test for an effect using English stop consonant release bursts without a following vowel. Therefore, Experiments 1 through 4 represent the four cells of a crossing of the variables stimulus type (stop consonant vs. click) and syllabic structure (release + vowel vs. release-only). If the smaller effects for isolated clicks in Experiment 2 than for the stop-consonant syllables in Experiment 1 were due purely to differences in syllable structure (i.e., absence of a following transitions and vowel), then there should be a strong McGurk effect for the click-vowel syllables of Experiment 3, but a relatively weak effect for the stop release bursts in Experiment 4. On the other hand, if physical differences between the click and stop-consonant releases were the determining factor, then there should also be relatively weak effects for click-vowel syllables in Experiment 3, but strong effects for the stop release bursts in Experiment 4.

Experiments 3 and 4 may also provide a basis for examining the role of phonological significance for a strong McGurk effect, depending on whether the click-vowel syllables are perceived as having phonological significance or not. We will return to this issue when we discuss the results of Experiment 3.

4 Experiment 3

The click categories used in Experiment 2, namely bilabials, dentals, and laterals, were again used here. However, the clicks used in the present experiment, like the stop consonants of Experiment 1, were coarticulated with the following vowel /a/. Thus, the stimuli match those of Experiment 1 in syllable structure and match those of Experiment 2 in physical properties of the click release.

We used the same tests and conditions in Experiment 3 as we did in Experiment 2. Additionally, as in Experiment 2, we ran an additional group of participants to determine whether the stimuli were perceived as speech.

4.1 *Method*

4.1.1 Participants

Participants were 24 undergraduates at the University of Connecticut. All received course credit for their participation. They all reported normal hearing and normal

or corrected-to-normal eyesight, and all were native speakers of English. Data from one participant were excluded from the analyses for failure to meet our performance criteria. An additional 10 undergraduates served as participants in a follow-up test of the phonological significance of the stimuli.

4.1.2 Materials

4.1.2.1 *Visual stimuli*

The visual stimuli were recorded in the same session, by the same speaker as in Experiments 1 and 2. She produced six repetitions of bilabial, dental, and lateral clicks preceding /a/, intermixed with other productions in randomized sequences. Of these, one bilabial, one dental, and one lateral production were selected to provide videos for the experiment. All digitizing, selecting, and editing procedures were the same as in previous experiments. The visual stimuli were all 1465ms in duration.

4.1.2.2 *Auditory stimuli*

Auditory tokens of bilabial, dental, and lateral click-/a/ syllables were recorded separately in the same session and under the same recording conditions as the auditory tokens of Experiments 1 and 2. Three tokens from each click category were selected on the basis of similar click releases, and, as much as possible, on vowel qualities. However, it proved impossible to select three bilabial tokens that were good matches on their clicks and on their vowels. Therefore, we selected three bilabials on the basis of click similarity, and three others on the basis of vowel similarity. The three similar clicks were then each spliced onto the vocalic segment of one of the tokens selected on the basis of vowel qualities. The splice point was immediately after the acoustic offset of the click noise portion, and prior to the onset of voicing. This procedure was not necessary for the dentals and laterals, because it was possible to select three tokens of each that met all of the selection conditions. For all of the clicks, the amplitudes of the click and vocalic segments were independently modified (where necessary) so that they would be roughly equivalent for all nine tokens. As in the previous experiments, one token from each of the three click categories was chosen to be the dubbed token, while the others were used as X tokens.

4.1.2.3 *Audio-Visual dubbing*

The procedure for dubbing was identical to that of Experiment 2. As in Experiment 2, the experiment included dental-lateral pairings, but we will not report the results for trials involving those stimuli.

4.1.2.4 *Test sequences*

The test sequences corresponded to those of Experiment 2. Because the acoustic stimuli in this experiment were much longer than those used in the preceding experiment while the visual stimuli were similar in length, the acoustic ISIs for this experiment were consequently shorter in this experiment, approximately 850ms.

4.1.3 Procedure

The instructions were similar to those used in Experiments 1 and 2, except that the clicks were referred to as "sounds" that are immediately followed by the vowel "ah." Thus, the instructions suggested that the click and vowel portions of each acoustic signal constituted separate events. Participants were instructed to base their judgments only on the initial sound, and not on the vowel portion. In the Audio-Visual portion, participants were told that the first and third sounds would be accompanied by a video display of a woman opening her mouth; the instructions did not indicate that the clicks were produced by a vocal-tract, although it was clear that the vowel portions of the acoustic signals were. In other regards, the experiment was conducted in exactly the same manner as Experiments 1 and 2.

4.1.3.1 *Phonological significance test*

We conducted an additional test to assess whether or not the clicks were perceived as speech. As in Experiment 2, participants in the phonological significance test received the same instructions as the participants in the regular experiment, but they were only presented with the Audio-Visual portion of the experiment, and only the first 24 trials were shown. They were then given a questionnaire in which they were asked to describe the stimuli. The questionnaire was nearly identical to that used in Experiment 2, with an initial open-ended question and subsequent specific questions asking about nonspeech-like or speech-like interpretations of the sounds; the only difference was that the participants were asked to answer the items with particular reference to the click sounds and not to the following vowel.

4.2 *Results*

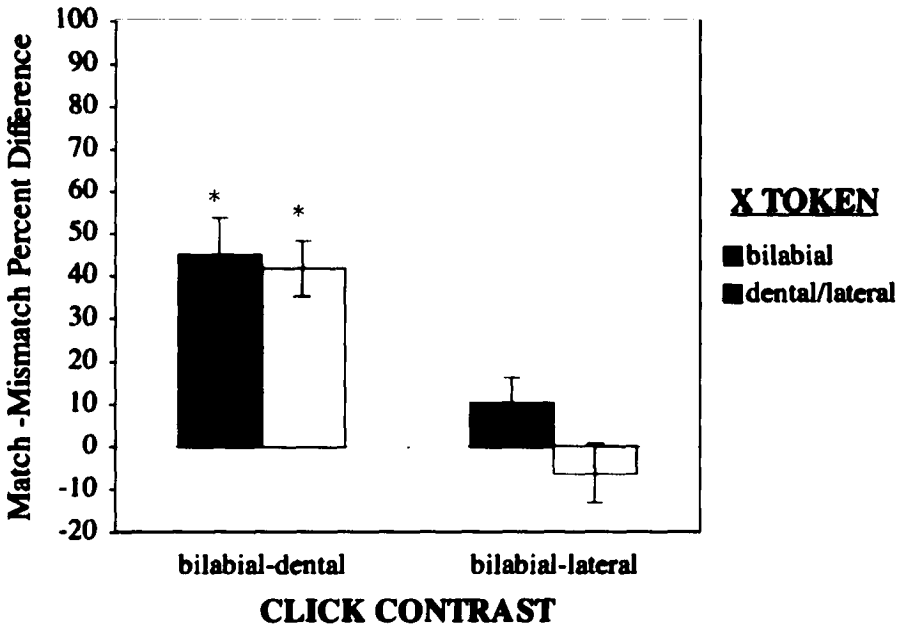
In the Audio-Alone portion of the experiment, participants overall selected the test token that matched X on 96.6% of the trials.

We analyzed the difference between responses on the Match and Mismatch trials in the same way as for the previous experiments, computing the difference in selections of the token that matched X in auditory place of articulation, and using the same factors of Click Contrast (bilabial-dental vs. bilabial-lateral) and X token (bilabial or other, i.e., dental or lateral) as in Experiment 2. Overall, participants selected the test token with an auditory place of articulation matching that of X 22.6% less often in the Mismatch condition than in the Match condition; this difference score is significantly greater than zero, $t(95) = 5.52$, $p < .0001$. One sample t-tests revealed that the means for both bilabial-dental contrasts were significantly greater than zero (smaller $t(23) = 5.18$, $p < .0001$), but that neither of the bilabial-lateral contrasts' means were significantly different from zero.

The ANOVA on the difference scores revealed only a significant main effect of Click Contrast, $F(1, 23) = 43.04$, $p < .0001$, with a larger difference score for the bilabial-dental comparison than for the bilabial-lateral comparison. Neither the main effect of X token nor the interaction reached significance. Figure 3 shows that an effect of the visible articulations is only evident in the bilabial-dental comparisons.

Figure 3

Difference between Match and Mismatch trials in the percentage of selections of the audio-visual token with an auditory place-of-articulation match to X, for click-vowel syllables in Experiment 3. Means are grouped according to the Click Contrast (bilabial-dental or bilabial-lateral) and the identity of the X token. Asterisks specify difference scores that are significantly greater than zero



4.2.1

Phonological significance test

In answer to the first, open-ended, question, nine of the 10 participants reported that they perceived the click releases primarily as consonants, and the 10th participant reported in a follow-up question that the stimuli resembled consonants. Seven of the participants reported hearing consonants with either a bilabial (/b/, /p/, or /m/) or alveolar (/d/, /t/, or /n/) place of articulation, indicating that the bilabial and dental clicks were assimilated to native phonological categories. Additionally, nine of the participants reported hearing /k/, which probably corresponds to the lateral click. Six participants did report that the clicks also resembled nonspeech events (either of environmental or vocal-tract origins). Overall, though, it appears likely that the participants did attend to phonological properties of the stimuli in performing the AXB task.

4.3 *Discussion*

In Experiments 1 and 2, we found a strong McGurk effect for English stop-consonant syllables and a weaker effect for isolated clicks. The purpose of Experiment 3 was

to determine whether the inclusion of vocalic transitions and a vowel following the clicks would increase the magnitude of the McGurk effect to the level found for the English stop-consonant syllables.

Overall, we found evidence for a strong McGurk effect for the bilabial-dental contrast (for both the bilabial and alveolar X tokens), but the McGurk effect was not significant for the bilabial-lateral contrast. As with the results of Experiment 2, the overall stronger effects for the bilabial-dental contrast may reflect the greater acoustic similarity of the bilabial click release to that of the dental, relative to the lateral.

A comparison of the results of Experiments 2 and 3 reveals a stronger effect of audiovisual incongruency in the latter experiment, as there was a 22.6% difference between the Match and Mismatch conditions in Experiment 3 compared to the 15.4% difference in Experiment 2. Thus, the McGurk effect is stronger for clicks that are coarticulated with a following vowel than for clicks produced in isolation. In contrast, the McGurk effect for the click syllables was closer in magnitude to those found for English stop-consonant syllables in Experiment 1 (27.9%). Although the McGurk effect for clicks was weak for the bilabial-lateral contrast, the effect for the bilabial-dental contrast was actually larger than the effects for the /p/-/t/ and /p/-/k/ contrasts in Experiment 1 (see Figures 1 and 3).

As we noted earlier, the English stop-consonant syllables of Experiment 1 and the isolated clicks of Experiment 2 differ physically in two regards, namely the consonant release (stop consonant vs. click) and the syllable structure (following vowel vs. no vowel). They also differ in their phonological significance. The stimuli of Experiment 3 share the consonant release properties of the clicks of Experiment 2, but share the syllable structure of the stop-consonant syllables of Experiment 1. (We cannot be absolutely certain that the click release properties of the isolated clicks and the coarticulated clicks are identical; it is possible that coarticulating the clicks with a following vowel changed the click release in some way that fostered a stronger McGurk effect. However, we have replicated the findings of Experiment 2 for the bilabial-dental contrast using truncated versions of the coarticulated clicks of Experiment 3; i.e., we removed the vocalic transitions and vowels following the click release. The results were virtually identical to the findings of Experiment 2.) The results of our phonological significance test in Experiment 3 suggest that our English-speaking participants perceptually assimilated the clicks to their native stop-consonant phonetic categories, and performed the AXB tasks using the assimilated phonetic characteristics (Best, 1994; Best et al., 1988),[2] meaning that the present

[2] This outcome stands in marked contrast to previous reports that native English listeners do not assimilate click syllables to their native phonological categories (Best & Avery, 1999; Best et al., 1988). This discrepancy may be due to the fact that, whereas the clicks in previous studies were produced by a native speaker of a click language, the click syllables used in the present study were produced by a native speaker of English; perhaps the speaker's click productions were influenced by her native phonological inventory. The clicks in the present study were also presented in conjunction with an articulating face, which could have increased the salience of the stimuli's relationships to native speech utterances. However, we note that participants in an auditory-only version of our phonological significance test also reported that the clicks resembled English consonants.

stimuli apparently share the phonological significance of the stimuli of Experiment 1. Overall, the finding in Experiment 3 of a strong McGurk effect for click syllables indicates that the weaker effect for isolated clicks than for English stop-consonant syllables was not the result of physical differences between click releases and English stop consonant releases, but it could have been due either to the absence of a vowel or to the lack of phonological significance in the isolated clicks.

We examine the influence of syllable structure and phonological status on the magnitude of the McGurk effect in Experiment 4 by testing the McGurk effect in English stop-consonant release bursts presented without a following vowel. Because these stimuli have phonological significance, the comparison of stop-consonant releases and stop-consonant syllables, unlike the comparison of isolated clicks and click syllables, does not confound phonological status and syllable structure. Therefore, we can make the following predictions. If the presence of vocalic transitions and a following vowel foster a strong McGurk effect, then the effect of isolated bursts should be considerably weaker than that of their syllabic counterparts, and possibly comparable in magnitude to that of the isolated clicks. If, instead, phonological significance is the critical factor, then the effect of the English bursts should be strong and comparable in magnitude to the full stop-consonant syllables.

5 Experiment 4

We excised the release bursts from productions of /pa/, /ta/, and /ka/ similar to those used in Experiment 1 from the remainder of the syllables. These excised bursts were paired with the intact full-syllable visual stimuli used in Experiment 1. Otherwise, the experiment was identical to Experiment 1. We did not conduct a phonological significance test for these stimuli, because we determined that the acoustic stimuli were clearly identifiable as segments of the stop consonants /p/, /t/, and /k/, especially when listeners were previously informed of the stimuli's phonetic character. To ensure that the participants in the experiment were aware of the stimuli's phonological significance, we identified them for the participants at the outset of the experiment.[3]

5.1 *Method*

5.1.1 Participants

The participants were 26 undergraduates at the University of Connecticut, who received course credit for their participation. All reported normal hearing and normal or corrected-to-normal eyesight, and all were native speakers of English.

[3] We also ran a condition, which we do not report here, in which participants were not informed of the speech nature of the stimuli. We found that changes in the instructions had minimal consequences for the overall results.

5.1.2 Materials

5.1.2.1 *Visual stimuli*

The visual stimuli were those used in Experiment 1.

5.1.2.2 *Auditory stimuli*

The stimuli were selected from the productions of /pa/, /ta/, and /ka/ recorded and digitized for Experiment 1. However, while tokens were selected for Experiment 1 on the basis of their similarity in duration and F0, for this experiment three tokens of each syllable were selected based on the clarity of their release bursts, played in isolation from the remainder of the syllable. All of the selected stimuli were truncated at a zero-crossing in the waveform so that they were all 27 ms in length; this duration was selected so that none of the bursts themselves would be removed, but that only aspiration and vocalic portions were eliminated. Because the bursts differ in duration, for some tokens (in particular, the /p/ productions), some aspiration remained along with the burst. The stimuli were digitally amplified so that they were matched in peak RMS amplitude as well. As in previous experiments, of the three tokens of each category, one was selected to be dubbed (for A and B tokens of the AXB test), while the other two were undubbed (for X).

5.1.2.3 *Audio-visual dubbing*

The dubbing procedure was identical to that used in Experiment 1.

5.1.2.4 *Test sequences*

The test sequences corresponded to those of Experiment 1. The acoustic ISIs were approximately 1575 ms, and thus longer than the ISIs in Experiment 1. This difference is due to the fact that the same videos were used in Experiments 1 and 4, but the auditory tokens used in Experiment 4 were much shorter than those used in Experiment 1. A sequence of six of the trials was accidentally repeated in the videotape (Trials 43 to 48); thus, there were 70 trials instead of 64. The second repetition of this series of trials (Trials 49 to 54) was dropped from the analysis.

5.1.3 Procedure

The instructions were similar to those used in Experiment 1, except that participants were explicitly told that the auditorily presented stimuli were the beginnings of the syllables /pa/, /ta/, and /ka/, artificially truncated. In other regards, the experiment was conducted in exactly the same manner as Experiment 1.

5.2 Results

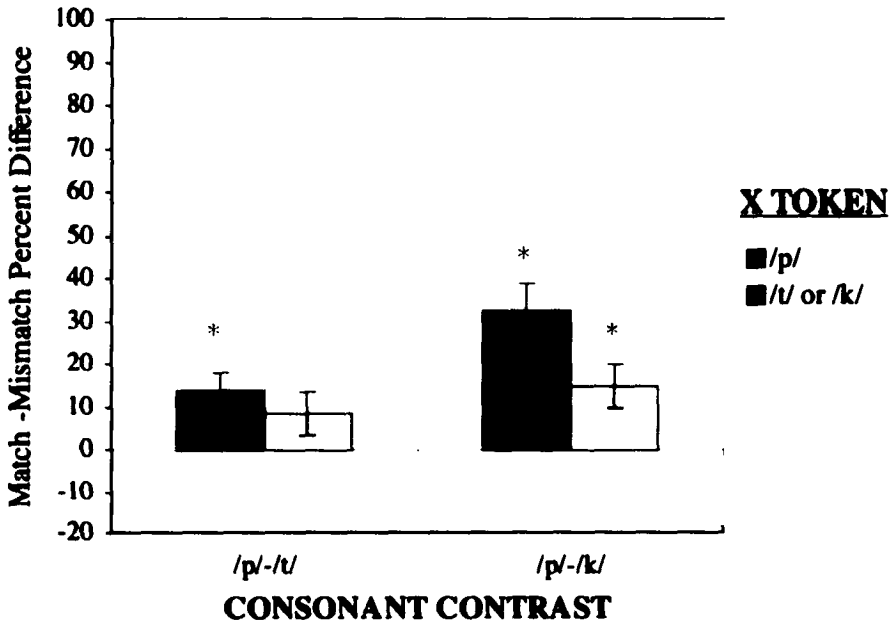
Data from two participants were excluded from the analyses for failure to meet our performance criteria of 70% accuracy in the Audio-Alone trials and in the Match trials. In the Audio-Alone portion of the experiment, participants overall selected the test token that matched X on 94.0% of the trials.

As in previous experiments, we computed the difference between the Match and Mismatch conditions in selections of the audiovisual token matching auditory X in place of articulation, using the same factors as in Experiment 1, namely Consonant Contrast (/p/-/t/ vs. /p/-/k/) and X token (/p/ vs. the nonlabial alternative, either /t/ or /k/).

Overall, participants selected the test token that was an auditory match to X 17.2% less often in the Mismatch condition than in the Match condition. This difference is significantly greater than zero, $t(95) = 6.30$, $p < .0001$. The mean difference scores for the crossing of Consonant Contrast and X token are presented in Figure 4. One-sample t-tests for each condition revealed difference scores that were significantly greater than zero for three of the conditions (smallest, $t(23) = 2.93$, $p < .01$), with the exception being the /p/-/t/ contrast when the X token was /t/.

Figure 4

Difference between Match and Mismatch trials in the percentage of selections of the audiovisual token with an auditory place-of-articulation match to X, for isolated English stop consonant release bursts in Experiment 4. Means are grouped according to the Consonant Contrast (/p/-/t/ or /p/-/k/) and the identity of the X token. Asterisks specify difference scores that are significantly greater than zero



An ANOVA on the difference scores revealed significant main effects of Consonant Contrast, $F(1, 23) = 8.90$, $p < .01$, with a greater difference score for the /p/-/k/ comparison than for the /p/-/t/ comparison, and of X token, $F(1, 23) = 7.05$, $p < .05$, with a greater difference score when X was /p/ than when it was the alternative (/t/ or /k/). There was also a marginal interaction of Consonant Contrast and

X token, $F(1, 23) = 3.83$, $p = .06$, reflecting the much greater difference score for the /p/-/k/ contrast when the X token was /p/ than for the other conditions.

5.3 Discussion

In Experiment 4, we found a significant McGurk effect for excised stop bursts from the English syllables /pa/, /ta/, and /ka/. However, the effects were overall weak in magnitude. Although there were significant effects of audiovisual discrepancy in three of the four conditions, only one of the effects (for the /p/-/k/ contrast when X was /ka/) was comparable in magnitude to the significant effects found in the previous three experiments.

Overall, the McGurk effect for the excised stop bursts (17.2% difference between the Match and Mismatch conditions; $d = .89$[4]) was much weaker than the effects found for both full English syllables (27.9%; $d = 1.29$) and click syllables (22.6%; $d = 1.22$). In contrast, the effect was similar in magnitude to that observed for isolated clicks in Experiment 2 (15.4%; $d = .87$). To evaluate these differences, we conducted a cross-experiment analysis of variance on the Match-Mismatch differences, including the factors Release Type (English stop vs. Click) and Syllable Structure (Vowel vs. Release-only) and collapsing over the other factors (Consonant/Click Contrast and X token). There was a significant main effect of Syllable Structure, $F(1, 92) = 5.15$, $p < .05$, with vowel-less stimuli in both languages showing a marked reduction in the magnitude of the McGurk effect (clicks: 7% difference; stops: 11% difference), but no effect of Release Type, $F < 1$, and no interaction of Release Type and Syllable Structure, $F < 1$. Thus, comparison of the results of the four experiments indicates that variation in syllable structure appeared to have similar consequences for clicks and stop consonants: for both, the McGurk effect was stronger when the click/stop consonant was followed by a vowel than when it was presented in isolation. There was no reliable overall difference in the McGurk effect between clicks and English stop consonants.

We note that there were also differences in performance on the auditory baseline task among the four experiments. Auditory discriminability was lower for the isolated clicks (87.5%) than for the other conditions, and discriminability for the excised stop bursts (94.0%) was also somewhat lower than that for the click syllables (96.6%) and English syllables (98.2%). A cross-experiment analysis of variance revealed significant effects of both Syllable Structure (Vowel vs. Release-only), $F(1, 92) = 22.28$, $p < .0001$, and Release Type (English stop vs. Click), $F(1, 92) = 8.23$, $p < .01$; the interaction was marginal, $F(1, 92) = 3.09$, $p = .08$. The fact that the isolated clicks (and, to a lesser extent, the excised stop bursts) were less discriminable than the stimuli with vowels (both clicks and English stops) is relevant to the interpretation of the McGurk results. In general, visual influences in speech perception (including the McGurk effect) are

[4] To calculate Cohen's d (the ratio of mean difference to SD, a measure of effect size), we used the Match-Mismatch score averaged across trial types (i.e., different consonant/click contrasts and different X tokens) for each participant. Typically, a value of d greater than 0.8 indicates a large effect.

stronger for less clear auditory stimuli. Accordingly, the auditory discriminability results suggest that, setting aside effects of syllable structure, the Match-Mismatch difference should have been *greater* for the isolated clicks and excised stop bursts than for the other stimuli. Therefore, our results may have underestimated the effect of syllable structure on the magnitude of the McGurk effect.[5]

Earlier, we listed three factors that might influence the magnitude of the McGurk effect for our stimuli, stemming from our finding of substantially smaller effects for isolated clicks than for English stop consonant syllables. Two were physical factors, namely syllable structure and differences between English stop consonant and click releases, and the third was the linguistic factor of phonological significance. The strong McGurk effect for click syllables, found in Experiment 3, suggested that the difference between click and stop consonant releases was not the critical factor underlying the weak effect for isolated clicks. However, because the click syllables resembled English syllables, the strong effect for these stimuli could have been the consequence of either phonological significance or the presence of a following vowel and transitions.

The results of Experiment 4, in relation to the previous experiments' results, point to an overall effect of syllable structure but not of phonological significance. Two aspects of the results are noteworthy. First, the McGurk effect was similar in magnitude for the isolated clicks in Experiment 2 and the syllable structure-matched excised stop consonant bursts in Experiment 4, even though they differ in phonological status. Second, the comparison between the McGurk effect for stop consonant syllables and their isolated bursts closely parallels the comparison between click syllables and isolated clicks, even though the latter comparison involves a change in phonological status and the former does not. Thus, the magnitude of the McGurk effect was not affected by a change in phonological status with no change in syllable structure, and it was affected by a change in syllable structure regardless of whether there was a change in phonological status or not. Therefore, the results indicate that phonological significance does not necessarily increase the magnitude of the McGurk effect.

6 General Discussion

Our experiments were designed to explore the extent to which a strong McGurk effect arises from speech-specific processing and the extent to which it depends on

[5] We also have converging evidence for the effect of syllable structure on the McGurk effect for the clicks and English syllables from other AXB tasks that were included in our experiments but that we did not report here. In these tasks, the A and B tokens consisted of the same auditory token paired with either matched or mismatched videos, and the McGurk effect was indexed by discrimination that differed from chance (i.e., if the matched and mismatched videos did not affect perception, performance on these tasks would be 50%). Critically, because the A and B auditory tokens were identical, auditory discriminability of the tokens is irrelevant to performance on these trials. Nonetheless, the results on those trials was broadly consistent with the results on the Match-Mismatch trials, with greater visual influence observed for English and click syllables than for isolated clicks and stops.

physical properties of the dubbed stimuli independent of their linguistic significance. We examined these possibilities using stimuli that were similar to the English stop consonant-vowel syllables that do give rise to a robust McGurk effect, both in the events themselves and in their acoustic and optical consequences. The clicks we chose provide a close parallel to stop consonant contrasts such as /p/-/t/ and /p/-/k/ in two ways: first, they involve the rapid release of vocal tract constrictions and second, they differ from one another both visually and in acoustic frequency at their onsets. Therefore, as a first attempt to gain leverage on the question of what are the limiting factors for the McGurk effect, we tested whether dubbing events that involve the rapid release of vocal tract constrictions is sufficient to produce a strong McGurk effect, even if the stimuli lack phonological significance.

In Experiment 2, we found evidence for a McGurk effect for the nonspeech clicks. However, the effect was much weaker in magnitude than that found for stop consonant-vowel syllables. Our results therefore do not support the hypothesis that a strong McGurk effect will occur for any dubbed events involving the rapid release of vocal tract constrictions. The question that follows concerns what factors determine the magnitude of the effect for vocal-tract events. As we pointed out earlier, one potential critical factor is phonological significance, in line with speech-specific explanations of the McGurk effect such as the motor theory (Liberman & Mattingly, 1985); other potential factors stem from physical differences between the isolated clicks and stop consonant-vowel syllables. We addressed these possibilities by testing for a McGurk effect using click-vowel syllables and stop consonant release bursts in addition to stop consonant syllables and isolated clicks. We found strong McGurk effects of similar magnitude for English stop consonants and clicks presented in syllables, both of which were perceived as having phonological significance, and weak effects of similar magnitude for stop consonant and click releases when presented in isolation, the former of which had phonological significance and the latter of which did not. The decreases in the McGurk effect from the vowel to no-vowel conditions were of similar magnitudes for clicks and stop consonants. We interpret these findings as evidence that syllable structure is a critical factor for a strong McGurk effect, and that phonological significance does not necessarily increase the magnitude of the effect.

This latter conclusion requires some qualification, however, because our findings do not rule out the possibility that phonological significance does contribute to the McGurk effect under certain conditions; in particular, it is possible that the McGurk effect would have been weaker for our click-vowel stimuli had they lacked phonological significance. As a consequence, one possibility consistent with our results is that phonological significance increases the magnitude of the McGurk effect for stimuli that have a vowel, even though (as indicated by the results of Experiments 1 and 4) it apparently does not when they include only an isolated consonant release. In line with this possibility, one explanation for our results from the perspective that speech-specific mechanisms underlie the McGurk effect could be that such mechanisms are engaged to a greater extent when a vowel is present than when it is not, because speech normally involves coarticulation between consonants and vowels. By this view, the strong McGurk effects for the English and click syllables are the consequence of the stimuli being perceived as speech, whereas the effect is weaker for

the isolated bursts because they lack a vowel and engage the speech-specific processes to a lesser extent. The significant but weak McGurk effect for the isolated clicks, according to this account, must arise from some other, nonlinguistic audiovisual integration process.

It is also possible to account for our findings without reference to speech-specific processes. In the introduction, we discussed two theories that disagree with the motor-theory prediction that the McGurk effect is specific to speech, namely the FLMP (Massaro, 1987; 1998) and direct realism (Best, 1984, 1995; Fowler, 1986; 1994). Here we will discuss how each theory could account for our findings.

As we discussed earlier, the magnitude of the McGurk effect depends on a variety of factors in the FLMP framework. One critical factor is the ambiguity of the cues: the McGurk effect will be stronger for stimuli with ambiguous auditory cues than those with more robust auditory cues. Another factor is whether the perceiver has prototype descriptions for the stimuli that include both auditory and visual cues; the McGurk effect will be weaker for stimuli with which the perceiver has had little experience and thus no developed prototype (Massaro, 1987). With regard to the present findings, the former factor cannot account for our findings, although the latter might. Consider first the nature of the acoustic signals in our experiments. Isolated click and stop release bursts are acoustically less robust than click and stop consonant syllables; in the former, only the release distinguishes the stimuli, whereas in the latter, both the release and the formant transitions distinguish them. Consequently, we found weaker McGurk effects for the more transient acoustic signals, contrary to the expectations of the FLMP. On the other hand, experience with the stimuli might account for the differences in the McGurk effect across our experiments. The clicks were likely to be less familiar to our participants than were the stop consonant syllables (even though the clicks do closely resemble commonly used nonverbal mouth actions such as a kiss, a "tsk," and a "cluck"), and perhaps the McGurk effect was weaker for that reason. (However, see Fowler & Dekle, 1991, for evidence against experience-based explanations of the McGurk effect.) By this account, the effect was presumably strong for the click syllables because they matched the participants' stop consonant-vowel syllable prototypes. Finally, the weak effects for the stop bursts might be accounted for by the FLMP because, according to the theory, prototypes for speech are for consonant-vowel syllables rather than isolated consonants; thus, the isolated bursts provided poor matches to the prototypes. To summarize, the FLMP might account for our findings, without assuming speech-specific reasons for the McGurk effect, by appealing to explanations that involve learning and the use of syllabic prototypes for speech stimuli, but apparently not to explanations involving psychophysical aspects of the stimuli.

The direct-realist perspective, like the FLMP, predicts that the magnitude of the McGurk effect does not depend on phonological significance of the stimuli. According to the direct-realist framework, the critical factors that determine the magnitude of the McGurk effect should involve the particular ways in which the acoustic and optical signals specify the underlying kinematic aspects of the dubbed events. There are at least two physical differences that may have contributed to a strong McGurk effect in our stimuli: the presence of acoustic formant transitions, and the presence of a vowel.

One possible reason for the weaker McGurk effect for the isolated clicks and stop bursts is that the acoustic signals lacked formant transitions. Formant transitions might be important for the McGurk effect because they constitute a dynamic signal that provides information about kinematic properties of the movement of the vocal-tract articulators. (Click releases and stop bursts are also dynamic signals, in that their frequency components do change over time, but these changes in frequency occur over a very short time frame relative to transitions; Fant, 1973; Ladefoged & Traill, 1994). Because the optical signals for the stimuli provided dynamic information about the underlying kinematic properties of the event (Rosenblum & Saldaña, 1996), an acoustic signal that also specifies the underlying kinematic properties will be commensurate with the optical signals. By this view, the effects for isolated clicks and release bursts are weaker than the respective effects for click and stop consonant syllables because the acoustic signals provide less information about speech dynamics.

A second possibility is that the presence of the vowel itself (in addition to the transitions into the vowel) is critical for a McGurk effect, for reasons independent of phonological significance. The presence of the vowel may contribute by effectively conveying that the acoustic signal specifies a vocal-tract event, and not some other kind of event. A given acoustic signal might, when presented out of context, be confusable as different kinds of events (which share certain kinematic properties, such as the striking of one object against another; Gaver, 1993). Notably, Best et al. (1988) reported that clicks are often categorized by native English listeners as nonvocal tract events such as fingers snapping (and we found some comparable responses in our phonological significance test for the isolated clicks). From the perspective of an English listener, the acoustic signal of a click release could be due to a vocal-tract event, but it could also be due to some other type of event that involves rapid release of a contact between two objects. The same may be true of the excised stop bursts as well, despite the fact that they can be categorized phonetically. On the other hand, a vowel may not plausibly be produced by anything but a vocal tract. Because the McGurk effect entails perception of a unitary event in spite of a cross-modal discrepancy in the signal, perhaps the effect, and audio-visual integration in general, is stronger to the extent that the acoustic and optical signals unambiguously and uniquely specify the same *type* of event (although necessarily different tokens of the type, because otherwise there would be no discrepancy to be resolved). Conversely, the McGurk effect may be weaker when either the acoustic signal or the optical signal is consistent with more than one type of event. Stated in general terms, a strong McGurk effect should occur when acoustic and optical signals uniquely specify nonvocal tract events that have similar underlying kinematic properties to one another. In line with this "unique specification" hypothesis, perhaps the presence of a following vowel renders the source of a click or stop burst unambiguous, and audiovisual integration is stronger for this reason, regardless of whether the initial release is perceived as speech or not.[6]

[6] Findings that a robust McGurk effect occurs even when there is a gender mismatch between the acoustic and optical signals (Green et al., 1991) and when only a point-light display of the face is presented (Rosenblum & Saldaña, 1996) indicate that "unique specification" must refer to the underlying kinematic properties of a particular event rather than its surface

The preceding two explanations for variation in the McGurk effect have in common an emphasis on aspects of how the acoustic and optical signals specify their source events, and neither invokes speech-specific mechanisms. It is possible that both constraints play a role in determining the magnitude of the McGurk effect. However, these accounts may not provide a complete explanation for why the McGurk effect varies in its magnitude for different stimuli. For example, the dynamic-signal hypothesis explains why the effect is weak for steady-state vowels, as both the acoustic and optical signals are relatively static; it might also account for the weak effects for de Gelder and Vroomen's (2000) emotional expression stimuli, which involved static faces. However, it is not clear whether it can explain the weak effects for Saldaña and Rosenblum's (1993) pluck-bow stimuli and Rosenblum and Fowler's (1991) hand-clap stimuli; the acoustic signals for those stimuli are dynamic in the sense that they involve changes in amplitude, but they do not have changes in frequency. The unique-specification hypothesis can account for Saldaña and Rosenblum's (1993) and Rosenblum and Fowler's (1991) results by attributing the weak McGurk effects to the acoustic stimuli being consistent with other events besides a finger or bow vibrating a string and clapping hands, respectively. It does not, however, explain why the McGurk effect is weaker for vowels than for consonants (e.g., Massaro & Cohen, 1993; Summerfield & McGrath, 1984), because both consonants and vowels are uniquely vocal-tract events. Thus, each hypothesis can provide only partial constraints on audiovisual integration; they would necessarily act in conjunction with a variety of other constraints, such as that the effect might be greater in magnitude for stimuli that are categorically perceived (Saldaña and Rosenblum, 1993). (Because we do not know whether the stimuli of Experiments 2 through 4 would be categorically perceived, we cannot evaluate this hypothesis with respect to our results.)

To summarize, our study was intended to narrow the gap between previously tested nonspeech events, which produce weak McGurk effects, and stop consonants, which produce strong effects, by testing the effect using nonspeech vocal-tract events that are similar to stop consonants. Although we did not find a strong nonspeech McGurk effect, our results did reveal a consistent pattern across our manipulations of syllable structure and type of consonant release, and we interpret our findings as revealing limitations on the possible role of phonological significance in determining the magnitude of the McGurk effect. Nonetheless, our findings can be accounted for both by views that assume speech-specific mechanisms and by those that do not.

---

properties (in other words, acoustic and optical signals may both specify the kinematics of a vocal-tract event, even if they are from different vocal tracts). We note that our "unique-specification" hypothesis bears some resemblance to Welch and Warren's (1980) account of cross-modal integration effects, namely that intersensory interactions are stronger to the extent that the perceiver's "unity assumption," or the degree to which the perceiver believes that the signals from two modalities stem from a single source event, is strong. However, our "unique-specification account" differs from their account in that they emphasized the perceiver's awareness of and beliefs concerning the intermodal relationship, a point which has been challenged (Green et al., 1991; Radeau, 1994), whereas our account emphasizes the properties of the signals themselves, in particular focusing on how well they specify the particular kind of event for the perceiver.

However, neuroimaging studies involving the stimuli we used in these studies may provide an avenue for distinguishing these accounts. Recent fMRI evidence suggests that portions of the left superior temporal sulcus (STS) and Broca's area may be involved in the audio-visual integration of speech (e.g., Calvert, 2001; Calvert & Campbell, 2003; Calvert, Campbell, & Bremmer, 2000; Paulesu, Perani, Blasi, Silani, Borghese, de Giovanni, Sensolo, & Fazio, 2003; see also Möttönen, Krause, Tiippana, & Sams, 2002, but see Jones & Callan, 2003; Olson, Gattenby, & Gore, 2002). These regions may also be involved in the auditory and visual processing of nonlinguistic biological motion (e.g., Grossman, Donnelly, Price, Pickens, Morgan, Neighbor, & Blake, 2000; Puce, Allison, Bentin, Gore & McCarthy, 1998; Rizzolatti, Fogassi, & Galese, 2002). However, Campbell, Macsweeney, Surguladze, Calvert, McGuire, Suckling, Brammer, & David (2001; see also Calvert, Bullmore, Brammer, Campbell, Williams, McGuire, Woodruff, Iversen, & David, 1997) reported a dissociation between the presentation of linguistic and nonlinguistic mouth movements (gurning movements for the latter) in activation of the left STS and Broca's area, and suggested that the STS may play a special role for speech. Thus, the possibility remains that areas of the brain responsible for audio-visual integration of speech (and thus for the McGurk effect) are specifically tuned to speech signals. Our finding of a weak but significant McGurk effect for clicks (which involve nonlinguistic biological movement similar to speech) is especially interesting in light of these neuroimaging findings.

Of particular interest is whether our audio-visual nonspeech click stimuli, as well as our click-vowel stimuli and isolated English stop bursts, would activate common brain regions as our English stop-consonant stimuli. Comparison of activation patterns in the left STS and Broca's area when these stimuli are presented auditorily, audio-visually with a congruent visual token, and audiovisually with an incongruent visual token might shed some light on the question of whether the different magnitudes of the McGurk effect for the different stimuli are associated with different levels of activation of common brain regions, or instead in activation of different regions (in line with the speech-specific mechanism view). Of particular concern would be the changes in pattern of activation when switching from English stop consonant syllables to stop bursts (thus, with a change in syllable structure but not phonological status), in contrast to the changes when switching to the isolated clicks (thus, with a change in both syllable structure and phonological status). If speech-specific mechanisms are responsible for the audio-visual integration of speech, giving rise to the McGurk effect, then the first contrast should involve a change in overall level of activation of common brain regions, but the second should involve a change in the locations of the activated regions. In conclusion, the empirical approach we have adopted here, namely of testing for a McGurk effect with nonspeech stimuli that are similar to speech, might offer new avenues for neuroimaging tests that could offer important insights into the processes that give rise to the McGurk effect.

References

BEST, C. T. (1984). Discovering messages in the medium: Speech and the prelinguistic infant. In H. E. Fitzgerald, B. Lester, & M. Yongman (Eds.), *Advances in pediatric psychology*, Vol.2 (pp. 97–145). New York: Plenum.

BEST, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge: MIT Press.

BEST, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Cross-language speech perception* (pp. 171–204). Baltimore: York Press.

BEST, C. T., & AVERY, R. A. (1999). Left hemisphere advantage for click consonants is determined by linguistic significance. *Psychological Science*, **10**, 65–70.

BEST, C. T., McROBERTS, G. W., & SITHOLE, N. M. (1988). Examination of perceptual reorganization for non-native speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 345–360.

CALVERT, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, **11**, 1110–1123.

CALVERT, G. A., BULLMORE, E. T., BRAMMER, M. J., CAMPBELL, R., WILLIAMS, S. C. R., McGUIRE, P. K., WOODRUFF, P. W. R., IVERSEN, S. D., & DAVID, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, **276**, 593–596.

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

CALVERT, G. A., CAMPBELL, R., & BREMMER, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, **10**, 649–657.

CAMPBELL, R., MacSWEENEY, M., SURGULADZE, S., CALVERT, G., McGUIRE, P., SUCKLING, J., BRAMMER, M. J., & DAVID, A. S. (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research*, **12**, 233–243.

De GELDER, B., & VROOMEN, J. (2000). The perception of emotions by ear and eye. *Cognition and Emotion*, **14**, 289–311.

DIEHL, R. L., & KLUENDER, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, **1**, 121–144.

DRIVER, J. (1996). Enhancement of selective listening of illusory mislocation of speech sounds due to lip-reading. *Nature*, **381**, 66–68.

FANT, G. (1973). *Speech sounds and features*. Cambridge, MA: MIT Press.

FODOR, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

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

FOWLER, C. A. (1994). Speech perception: Direct realist theory. In *Encyclopedia of language and linguistics* (pp. 4199–4203). Oxford: Pergamon Press.

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

GAVER, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*, **5**, 1–29.

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

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

GREEN, K. P., KUHL, P. K., MELTZOFF, A. N., & STEVENS, E. B. (1991). Integrating speech information across talkers, genders and sensory modality: Female faces and male voices in the McGurk effect. *Perception and Psychophysics*, **50**, 524–536.

GREEN, K. P., & NORRIX, L. W. (1997). Acoustic cues to place of articulation and the McGurk effect: The role of release bursts, aspiration, and formant transitions. *Journal of Speech, Language, and Hearing Research*, **40**, 646–665.

GROSSMAN, E., DONNELLY, M., PRICE, R., PICKENS, D., MORGAN, V., NEIGHBOR, G., & BLAKE, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience*, **12**, 711–720.

JONES, J. A., & CALLAN, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport*, **14**, 1129–1133.

LADEFOGED, P., & MADDIESON, I. (1996). *The sounds of the world's languages*. Oxford, U.K.: Blackwell Publishers.

LADEFOGED, P., & TRAILL, A. (1994). Clicks and their accompaniments. *Journal of Phonetics*, **22**, 33–64.

LIBERMAN, A. M., DELATTRE, P. C., COOPER, F. S., & GERSTMAN, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, **68**, 1–13.

LIBERMAN, A. M., & MATTINGLY, I. G. (1985). The motor theory revised. *Cognition*, **21**, 1–36.

MacDONALD, J., & McGURK., H. (1978). Visual influences on speech perception. *Perception and Psychophysics*, **24**, 253–257.

MANUEL, S. Y., REPP, B. H., LIBERMAN, A. M., & STUDDERT-KENNEDY, M. (1983) Exploring the "McGurk effect." Talk presented at the 24th meeting of the Psychonomic Society, November, 1983.

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

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

MASSARO, D. W., & COHEN, M. M. (1993) Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, **13**, 127–134.

MASSARO, D. W., & EGAN, P. B. (1996). Perceiving affect from the voice and face. *Psychonomic Bulletin and Review*, **3**, 215–221.

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

MÖTTÖNEN, R., KRAUSE, C. M., TIIPPANA, K., & SAMS, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research*, **13**, 417–425.

OLSON, I. R., GATENBY, J. C., & GORE, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research*, **14**, 129–138.

PAULESU, E., PERANI, D., BLASI, V., SILANI, G., BORGHESE, N. A., de GIOVANNI, U., SENSOLO, S., & FAZIO, F. (2003). A functional-anatomical model for lipreading. *Journal of Neurophysiology*, **90**, 2005–2013.

PUCE, A., ALLISON, T., BENTIN, S., GORE, J. C., & McCARTHY, G. (1998). Temporal cortext activation in humans viewing eye and mouth movements. *Journal of Neuroscience*, **18**, 2188–2199.

RADEAU, M. (1994). Auditory-visual spatial interactions and modularity. *Cahiers de Psychologie Cognitive*, **13**, 3–51.

RIZZOLATTI, G., FOGASSI, L., & GALLESE, V. (2002). Motor and cognitive functions of the ventral premotor cortex, *Current Opinion in Neurobiology*, **12**, 149–154.

ROSENBLUM, L. D. (1987). Towards an ecological alternative to the motor theory. *Perceiving-Acting Workshop*, **2**, 25–28.

ROSENBLUM, L. D. (1994). How special is audiovisual speech integration? *Cahiers de Psychologie Cognitive*, **13**, 110–116.

ROSENBLUM, L. D., & FOWLER, C. A. (1991). An audio-visual investigation of the loudness/effort effect for speech and nonspeech events. *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 976–985.

ROSENBLUM, L. D., & SALDAÑA, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception and Psychophysics*, **52**, 461–473.

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

RUBIN, P. E. (1995). HADES: A case study on the development of a signal analysis system. In A. Syrdal, R. Bennett, & S. Greenspan (Ed.), *Applied signal technology*. (pp. 501–520) Florida: CRC Press.

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

SUMMERFIELD, Q., & McGRATH, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, **36A**, 51–74.

WELCH, R. B., & WARREN, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, **88**, 638–667.