1343

Journal of Experimental Psychology:
Human Perception and Performance
2004, Vol. 30, No. 3, 445–463

Copyright 2004 by the American Psychological Association
0096-1523/04/$12.00 DOI: 10.1037/0096-1523.30.3.445

Lexical Influences in Audiovisual Speech Perception

Lawrence Brancazio
Southern Connecticut State University and Haskins Laboratories

Phoneme identification with audiovisually discrepant stimuli is influenced by information in the visual signal (the *McGurk effect*). Additionally, lexical status affects identification of auditorily presented phonemes. The present study tested for lexical influences on the McGurk effect. Participants identified phonemes in audiovisually discrepant stimuli in which lexical status of the auditory component and of a visually influenced percept was independently varied. Visually influenced (McGurk) responses were more frequent when they formed a word and when the auditory signal was a nonword (Experiment 1). Lexical effects were larger for slow than for fast responses (Experiment 2), as with auditory speech, and were replicated with stimuli matched on physical properties (Experiment 3). These results are consistent with models in which lexical processing of speech is modality independent.

A primary issue in speech perception research concerns how various sources of information contribute to phoneme identification. Two sources of information that have been studied extensively in relation to the acoustic signal are the optical signal created when a person speaks (e.g., McGurk & MacDonald, 1976; Sumby & Pollack, 1954) and higher level contextual information regarding whether a given phoneme is part of a word (e.g., Ganong, 1980; Samuel, 1981). Accordingly, there are two sets of theories that explain the respective roles of visual speech (e.g., Fowler, 1986, 1996; Green, 1998; Liberman & Mattingly, 1985; Massaro, 1987, 1998; Schwartz, Robert-Ribes, & Escudier, 1998; Summerfield, 1987) and lexical knowledge (e.g., Massaro & Oden, 1980; McClelland & Elman, 1986; Norris, McQueen, & Cutler, 2000) in speech perception. However, there has been, to date, little cross-talk between the two domains of research; visual and lexical contributions to speech perception have largely been investigated independently of one another. The present research integrates the two domains of study empirically. Specifically, the experiments reported here address the following question: Does lexical knowledge play a role in the categorization of phonemes in audiovisually presented speech? The following sections provide brief reviews of the evidence for visual and lexical contributions to phoneme identification and a discussion of how the two domains may be related.

Lawrence Brancazio, Department of Psychology, Southern Connecticut State University, and Haskins Laboratories, New Haven, Connecticut.

Some of these results were reported at the International Conference on Auditory–Visual Speech Processing, Santa Cruz, California, August 1999. This research was conducted as part of a doctoral dissertation at the University of Connecticut and was supported by National Institutes of Health (NIH) Grant HD-01994 to Haskins Laboratories and by a dissertation fellowship from the University of Connecticut. Preparation of this article was supported by NIH Grant DC-00373 to Northeastern University.

I thank my dissertation adviser, Carol Fowler, and the members of my committee, Jay Rueckl and James Green, for their advice and support on this project; Julia Irwin for videotaping the stimuli; and Tom Kies for testing the participants.

Correspondence concerning this article should be addressed to Lawrence Brancazio, Haskins Laboratories, 270 Crown Street, New Haven, CT 06511. E-mail: brancazioL1@southernct.edu

Audiovisual Integration in Speech Perception

Numerous studies have shown that seeing a person speak has a profound influence on speech perception (see, e.g., Green, 1998; Summerfield, 1987, for reviews). Speech in noise is more intelligible when a concurrent visual signal is presented (e.g., Erber, 1969; Sumby & Pollack, 1954), and visual speech can also enhance comprehension of clear auditory signals, particularly when the speech material is semantically difficult (Reisberg, McLean, & Goldfield, 1987). An even more striking demonstration of visual contributions to speech perception is the *McGurk effect*. This is a phenomenon that arises when an auditory signal specifying one phoneme is dubbed with a visual signal specifying a phoneme with a different place of articulation. Perceivers often report hearing a different phoneme than the one specified by the acoustic signal, demonstrating a visual influence on phonetic categorization. For example, listeners typically classify a clear auditory /ba/ presented with a visual signal for /ga/ as /da/, even though it is reliably identified as /ba/ when presented just auditorily (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976). This effect clearly indicates that information from the acoustic signal is perceptually integrated with information from the optical signal to arrive at a unified phonetic interpretation.

The McGurk phenomenon has been replicated and extended in numerous studies, and the visual influence has been found for a variety of audiovisual pairings. In most cases, the acoustic signal determines the perceived manner of articulation (e.g., obstruent, fricative, or nasal) and voicing, whereas the optical signal affects perceived place of articulation (MacDonald & McGurk, 1978; Summerfield, 1979). (An exception is when an acoustic /ba/ is presented with a visual /va/; the typical percept is /va/, which differs from the acoustic /ba/ in manner as well as place of articulation; Manuel, Repp, Liberman, & Studdert-Kennedy, 1983; Rosenblum & Saldaña, 1992.) The visual influence may result in a perceived consonant with the place of articulation of the visual signal (termed *visual dominance*; e.g., for auditory /ba/ + visual /da/, perceived as /da/), a place of articulation found in neither the auditory nor visual signal (an *intermediate* or *fusion* percept; e.g., for auditory /ba/ + visual /ga/, perceived as /da/, although this could be a case of dominance if /g/ and /d/ are not visually

distinguishable), or both places of articulation (a *combination*; e.g., for auditory /da/ + visual /ba/, perceived as /bda/).

One issue of debate regarding the McGurk effect concerns the point at which audiovisual integration occurs during the processing of speech (cf. Schwartz et al., 1998). Some researchers (e.g., Massaro & Stork, 1998) have argued for a relatively late stage of audiovisual integration, such that a significant amount of independent auditory-specific and visual-specific phonetic processing occurs before audiovisual integration. In keeping with this view are findings from selective adaptation experiments involving stimuli that generate a McGurk effect; such stimuli produce auditory rather than visual adaptation effects (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). However, findings of other studies suggest an earlier stage of audiovisual integration, such that integration precedes some aspects of phonetic processing (Bernstein, 1989; Green, 1998; Green & Miller, 1985). For example, visually induced changes in the perceived place of articulation of a consonant generate contextual effects on phonetic categorization that are typically generated by acoustic consonants of the same place of articulation (Fowler, Brown, & Mann, 2000; Green & Norrix, 2001), even though perceivers most likely lack explicit knowledge of the relations underlying the contextual effects. However, regardless of whether audiovisual integration occurs early or late relative to other aspects of phonetic processing, it appears clear that audiovisual integration occurs early enough to precede perceptual selection of phonetic categories; that is, phonetic decisions are not made independently for the auditory and visual inputs (even if the inputs are phonetically evaluated separately; cf. Massaro, 1998), but rather phonemes are identified on the basis of a unified, integrated percept (Green, 1998; Massaro, 1998; Schwartz et al., 1998).

On a phenomenal level, the McGurk effect is extremely compelling: There is often little awareness of the audiovisual discrepancy in the signal (Manuel et al., 1983; Saldaña & Rosenblum, 1993), and the effect occurs even when participants are made aware of the discrepancy and are told to respond solely on the basis of what they hear. However, the visual influence on consonant perception is not absolute. First, the effect does not always occur for appropriately created stimuli; there is extensive variation in the magnitude of the effect across studies (cf. Green & Kuhl, 1989; Green & Norrix, 1997; MacDonald & McGurk, 1978; Massaro, 1987). Second, the findings from the selective adaptation studies mentioned earlier (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994) illustrate that certain contextual effects produced by auditory consonants may not be produced by percepts arising from the McGurk effect. Third, even when the McGurk effect does occur, the resulting percepts apparently are not ideal exemplars of their phonetic category. Rosenblum and Saldaña (1992) found that participants selected an auditory /va/–visual /va/ token as a better match to an auditory /va/ than an auditory /ba/–visual /va/ token, even though all three tokens were identified as /va/. Additionally, Massaro and Ferguson (1993) found that listeners, using a continuous rating scale, judged auditory /da/–visual /da/ tokens to be more /da/-like than auditory /ba/–visual /da/ tokens. (In a similar vein, Green & Kuhl, 1991, and Massaro & Cohen, 1983, have found that perceivers are slower to categorize phonemes in audiovisually discrepant than congruent speech.) These latter findings indicate that even though perceivers may not always recognize an audiovisual discrepancy as such, they nonetheless detect that the visually influenced percepts are somehow distinct from percepts arising from audiovisually congruent signals.

Lexical Contributions to Speech Perception

Whereas the McGurk effect reflects contributions of visual speech on phoneme recognition, a second arena for research in phoneme recognition concerns the role of higher order lexical knowledge. A number of studies have compared categorization of phonemes in words and nonsense syllables, and there is substantial evidence that the lexical status of an utterance can affect identification of its constituent phonemes. This has been demonstrated in different paradigms, including phoneme restoration (Samuel, 1981) and, under certain conditions, phoneme monitoring (Rubin, Turvey, & van Gelder, 1976; cf. Cutler, Mehler, Norris, & Segui, 1987; Eimas, Hornstein, & Payton, 1990, Wurm & Samuel, 1997). A third empirical paradigm that provides evidence for lexical influences on phonetic categorization, which is the main focus of this article, is phoneme identification. Ganong (1980) demonstrated lexical influences on phoneme identification in a study involving stimuli from two matched acoustic continua ranging from *DASH* to *TASH* and *DASK* to *TASK* (as well as other similarly constructed stimulus sets), so that one continuum ranged from a word endpoint to a nonword endpoint, and the other continuum ranged from a nonword to a word. In an initial-consonant identification task, listeners exhibited a lexically induced boundary shift between /d/ and /t/ judgments, so that the boundary was closer to the /t/ endpoint in the *DASH–TASH* continuum than in the *DASK–TASK* continuum: Phoneme identifications that formed words were more frequent than those that formed nonwords, but only for the phonetically ambiguous stimuli in the midrange of the continuum (see also Connine & Clifton, 1987; Pitt, 1995). This lexical effect, known as the *Ganong effect*, indicates that a candidate for word recognition, if it at least partially matches the input signal, exerts an influence on phoneme identification to favor the phonemes that constitute the word. A popular current approach, incorporated in the models TRACE (McClelland & Elman, 1986) and Merge (Norris et al., 2000), is to explain this effect in terms of flow of activation between phonetic and lexical levels of representations: A prelexical phonetic processing stage activates lexical representations, and lexical activation feeds activation to phonetic representations (for the constituent phonemes of a word) to affect phonetic decisions. (Depending on the model, these phonetic-decision units are either the same phonetic units that feed activation upward to the lexical representations, as in TRACE, or instead exist at an independent stage, as in Merge; this issue is addressed in the General Discussion.) When there is competition between two phonemes at the phonetic-decision stage (owing to an ambiguous input signal), lexical activation may tilt the decision in favor of one of the phonemes (e.g., McClelland & Elman, 1986; Norris et al., 2000).

A variety of studies have explored temporal aspects of the Ganong effect, using a speeded-phoneme-identification task. Several researchers have partitioned the responses according to their reaction times and measured lexical effects separately for the fastest, slowest, and intermediate responses (e.g., Fox, 1984; Miller & Dexter, 1988; Pitt, 1995; Pitt & Samuel, 1993). Typically, the lexical shift is small (Miller & Dexter, 1988; Pitt, 1995) or nonexistent (Fox, 1984) for the fastest responses but is robust for the medium-speed and slowest responses (although McQueen,

1991, and Pitt & Samuel, 1993, found a reverse pattern for syllable-final decisions; also see Pitt & Samuel, 1993, for a meta-analysis of studies using the reaction time partition approach). Thus, lexical effects appear to develop over time, at least for stimuli with ambiguous initial consonants. One explanation for this finding is that lexical activation builds up continuously as the speech input unfolds, so that greater lexical influence on phonetic decisions may occur when the phonetic decision is made later than when it is made earlier (Connine & Clifton, 1987; Massaro & Oden, 1995; McClelland & Elman, 1986; Pitt, 1995; although see Cutler et al., 1987).

In general, the Ganong effect, along with findings of lexical influences in other paradigms, clearly demonstrates a role of lexical knowledge in the categorization of phonemes. Moreover, analyses of these lexical influences that take response time into account reveal a signature time course of the effect.

Audiovisual Speech Perception and the Lexicon

Although numerous studies have addressed the roles of optically specified and contextual information in phoneme identification, little is known about the relationship between audiovisual speech perception and lexical processing. One study demonstrated that the McGurk effect occurs for words as well as for nonsense syllables (Dekle, Fowler, & Funnell, 1992; cf. Easton & Basala, 1982), indicating that the two domains need not be considered independently of one another. A line of research more recently pursued by Bernstein and colleagues (Auer & Bernstein, 1997; Iverson, Bernstein, & Auer, 1998) involves computational modeling of the identifiability of words based on the perceptual distinctiveness of phonemes in visual-only speech (Auer & Bernstein, 1997) or in speech with visual presentation and a reduced acoustic signal (Iverson et al., 1998). The thrust of this work is therefore on the *bottom-up* processes implicated in the identification of words from an audiovisual or visual signal.

A complementary approach, and the one adopted in the present research, is to examine whether there are *top-down* effects of the lexicon on audiovisual speech perception. This issue has important implications for current models of speech perception and lexical access. Most contemporary models of speech perception that provide an account of lexical effects are strictly auditory models (McClelland & Elman, 1986; Norris et al., 2000; although see Massaro, 1998). In the light of the well-established contributions of vision to speech perception, such models will ultimately need to incorporate a role for visual speech in their architecture to be complete accounts. To achieve this, it will be necessary to gain a better understanding of the interplay of cross-modal and lexical factors in speech perception.

One way to test for lexical effects in audiovisual speech perception is to test for an analogue to the auditory Ganong effect, using audiovisually discrepant stimuli that give rise to the McGurk effect. For example, consider two stimuli: an auditory token of *BEG* presented with a visual utterance of *DEG* and an auditory token of *BESK* presented with a visual utterance of *DESK*. In the former case, the auditory token is a word, but the McGurk effect (resulting in perceived initial /d/) will result in perception of a nonword; in the latter case, the auditory token is a nonword, but the McGurk effect will result in the recognition of a word. If the categorization of audiovisually discrepant speech stimuli, like that of ambiguous auditory speech stimuli, is subject to higher order influence, then the McGurk effect should occur more often with the latter stimuli than with the former.

Contrary to this expectation, however, a recent study found no influence of lexical status on the McGurk effect. Sams, Manninen, Surakka, Helin, and Kättö (1998) presented bisyllabic stimuli audiovisually, in which either the initial or medial consonant of the auditory token was /p/ and the visual token had /k/ in the same position as the auditory /p/ (but was otherwise identical to the auditory token). They selected syllables in which both /p/ and /k/ percepts formed words (in Finnish) but a /t/ percept (the most common McGurk percept for auditory /p/ and visual /k/) formed a nonword and other syllables in which only a /t/ percept, but not /p/ and /k/ percepts, formed a word. Sams et al. found no significant differences in the proportions of /p/ responses in the two conditions and found significantly fewer /t/ responses when the /t/ percept formed a word than when it formed a nonword, contrary to the lexical hypothesis.

A possible reason why lexical effects might not occur with audiovisually discrepant stimuli concerns the requirement that there be phonetic ambiguity, that is, competition between phonetic candidates, for a lexical effect. As discussed earlier, there is evidence that audiovisually discrepant stimuli do result in some degree of perceptual uncertainty (Massaro & Ferguson, 1993; Rosenblum & Saldaña, 1992), which could foster the necessary conditions for a lexical effect. However, some researchers (Burton, Baum, & Blumstein, 1989; McQueen, 1991) have argued that lexical effects on categorization may occur only under limited conditions, depending on the methods used to construct the stimuli and on their clarity (but see Burton & Blumstein, 1995; Newman, Sawusch, & Luce, 1997; Pitt & Samuel, 1993). Given such constraints, one possibility is that the sufficient conditions for a lexical effect are restricted to certain kinds of acoustically based ambiguity and exclude cross-modally induced ambiguity.[1]

Although the findings by Sams et al. (1998) suggest that lexical effects do not occur for audiovisually discrepant stimuli, this conclusion is premature, because the word and nonword stimuli in their study differed on various factors besides lexical status. First, the incongruently dubbed consonant was in the initial position in four of the six auditory words, but in only two of the four auditory nonwords. Second, the identity of the vowel following the incongruently dubbed consonant (a factor that influences the magnitude of the McGurk effect; e.g., Green, Kuhl, Meltzoff, & Stevens, 1991) varied extensively as well. For example, the vowel was /a/ for three of the four auditory nonwords in the experiment, but for only one of the six auditory words. Given the small number of stimuli used in Sams et al.'s study, it is possible that lexical effects were obscured by individual-stimulus variability in the magnitude of the McGurk effect.

The purpose of the present research was to provide another test for lexical influences on the McGurk effect while avoiding some of the potentially confounding variables in Sams et al.'s (1998) design. Stimuli in the present experiments were presented audiovisually with a cross-modal discrepancy that always occurred in the initial consonant, and the stimuli were constructed so that the auditory stimuli were either words or nonwords and the expected

[1] Note, however, that neither Burton et al. (1989) nor McQueen (1991) made explicit claims about whether lexical effects would be restricted to the auditory domain.

visually influenced percept resulted in either a word or nonword. I adopted two general strategies to uncover audiovisual lexical effects in the experiments reported here. In the first approach, adopted for Experiments 1 and 2, stimuli that differed in lexical status were matched on some, but not all, stimulus properties (e.g., all of the incongruently dubbed consonants were in initial position, and there was a restricted set of vowels that could follow the dubbed consonant). Additionally, a large number of stimuli were used, so that individual-stimulus-based variability in the McGurk effect would, on average, cancel out across the lexical conditions. The second strategy, adopted in Experiment 3, was a more conservative approach: Only a few stimulus pairs were used, but the stimuli that differed in lexical status were carefully matched on acoustic and optical properties. Thus, in Experiment 3, any changes in the McGurk effect across the lexical manipulations were unlikely to be the result of stimulus-based factors unrelated to lexical status. The primary question in all three experiments was whether the McGurk effect would be greater when it resulted in a word than when it resulted in a nonword.

In addition to testing for an audiovisual lexical effect, the experiments reported here were designed to determine the time course of such an effect. Specifically, Experiments 2 and 3 used a speeded-classification task to test whether, as in the case of the auditory Ganong effect, lexical effects are greater for slow than for fast responses (following Fox, 1984, Miller & Dexter, 1988, and others). Overall, the experiments examined whether lexical influences in the audiovisual domain would produce a similar pattern of results as comparable influences in the auditory domain.

Experiment 1

Experiment 1 had two goals. The first goal was to test for a lexical influence on the McGurk effect, using a phoneme-identification task with audiovisually discrepant stimuli (differing in place of articulation of the initial consonant) that differed in their lexical characteristics. The stimuli were devised so that lexical status of both the auditory stimulus and the expected McGurk response (with the place of articulation of the visual signal) varied independently,[2] resulting in four conditions: (a) Both the auditory stimulus and the expected McGurk percept formed (different) words, (b) only the auditory stimulus formed a word, (c) only the expected McGurk percept formed a word, and (d) neither formed a word. This design allowed separate tests of lexical effects based on lexical status of the auditory signal alone and on lexical status of the integrated percept. If both factors were important, then perceivers would make more McGurk responses (and fewer auditory responses) when the McGurk percept formed a word than when it formed a nonword and also when the auditory stimulus formed a nonword than when it formed a word.

The second goal of Experiment 1 was to determine the degree of perceptual uncertainty caused by presentation of audiovisually discrepant stimuli. As noted earlier, auditory lexical effects are typically found only for phonetically ambiguous stimuli. Given this, it is important to establish that the audiovisually discrepant stimuli used here do give rise to some degree of perceptual ambiguity (in line with the findings of Rosenblum & Saldaña, 1992, and Massaro & Ferguson, 1993). To test this, participants in Experiment 1 were asked to give explicit category goodness ratings of the consonants they perceived. Such ratings provide a reliable index of how well a particular stimulus fits a phonetic category (e.g., Miller & Volaitis, 1989). If audiovisual discrepancy results in poor category exemplars, thus providing an environment that may foster lexical effects, then ratings will be lower for the discrepant stimuli than for the congruent stimuli (in keeping with Massaro & Ferguson's, 1993, findings). In fact, the ratings should be lower for both visual (i.e., McGurk) and auditory (i.e., non-McGurk) responses to discrepant stimuli, an issue not addressed by Rosenblum and Saldaña (1992) or by Massaro and Ferguson (1993).

Note that the design used here differs from those typically used in studies of lexical influences on phonetic perception in the auditory domain (that is, studies of the Ganong effect) in two critical respects. First, it uses audiovisually discrepant and congruent tokens instead of continuum midrange and endpoint tokens to produce phonetically ambiguous and unambiguous stimuli, respectively. Second, instead of presenting multiple repetitions of tokens from a few sets of minimally distinct stimuli varying in lexical status, the present experiment used many stimuli differing in lexical status, each of which was presented only once to each participant. This approach was adopted in part because it proved difficult to generate minimally distinct lexical sets with all four possible lexical contrasts (i.e., word–word, word–nonword, nonword–word, and nonword–nonword). It was also adopted under the rationale that using many stimuli would maximize the probability that an effect would appear. The risk inherent to this approach is that the stimuli in the different lexical conditions differ on acoustic and optical properties, which could give rise to spurious differences (or obscure actual differences) between the lexical conditions in perceptual categorization. However, the items in the different conditions were globally matched on phonetic properties, and the large number of items in each cell should counter the effects of individual-item variability.

Method

Participants

Sixty-two undergraduates at the University of Connecticut, ranging in age from 18 to 25 years, participated in the experiment. All were enrolled in introductory psychology classes and received partial course credit for their participation. All of the participants were native speakers of English, and all reported having normal hearing and normal or corrected-to-normal vision. The data of 2 participants were dropped from all analyses because of a large proportion of unrecorded responses due to equipment problems. Additionally, a 3rd participant had a large proportion of unrecorded responses in the audiovisually congruent condition (see the *Results* section), but not in the audiovisually discrepant condition; consequently, the data of this participant were included in analyses of the latter condition but not of the former.

Stimuli

Stimulus selection. Sixty-four pairs of one-syllable stimuli were selected in which members of each pair were identical except for the initial consonant. The initial-consonant contrasts were /b/ and /d/, /m/ and /n/, /p/ and /t/, and /b/ and /v/, so that each pair included a bilabial consonant (/b/,

[2] For example, the expected McGurk percept for an auditory /b/ paired with a visual /d/ was /d/. In this regard, the stimuli differed from those of Sams et al. (1998), for which the expected McGurk response differed in place of articulation from the visual stimulus, such as an auditory /p/ with a visual /k/, which is typically perceived as /t/.

/m/, and /p/) and a consonant whose place of articulation was visually distinct from bilabial (/d/, /n/, and /t/ are alveolar, and /v/ is labiodental). In 16 of these pairs, both members of the pair formed words (e.g., *BELT* and *DEALT*). In another 16, the bilabial pair member (i.e., with initial /b/, /p/, or /m/) formed a word, and the other member formed a nonword (e.g., *BEG* and *DEG*). Similarly, in an additional 16 pairs, the bilabial pair member formed a nonword, and the other member formed a word (e.g., *BESK* and *DESK*). Finally, in the remaining 16 pairs, neither member formed a word (e.g., *BEDGE* and *DEDGE*). There were equal numbers of pairs with the /b/–/d/, /m/–/n/, /p/–/t/, and /b/–/v/ contrasts, respectively, in each of the four lexical sets.

All of the stimuli had a medial vowel from the set /ɪ e ɛ æ a ʌ ɚ/. Pair selection was constrained by the requirement that a change in voicing or nasality of the initial consonant of a nonword resulted in a nonword (e.g., *BEDGE*'s neighbors *MEDGE* and *PEDGE* are nonwords). Additionally, all of the words had low- to mid-frequency counts (Kučera & Francis, 1967). The frequencies (defined as the number of occurrences per million words) ranged from less than 1 to 67. The mean frequencies of the words in the four lexical sets were roughly equivalent. The stimuli and their frequencies are presented in the Appendix.

Stimulus generation. A male speaker was recorded on videotape producing multiple repetitions of the 128 syllables. Videotaping took place in a soundproof booth, with a microphone attached to the videocamera placed directly in front of the speaker (below the view of the camera). The speaker's entire face was visible in the recording.

All of the productions were digitized on an Apple Macintosh PowerMac 8500 using Adobe Premiere. Each production was digitized at 320 × 240 pixels at 30 frames per second, and the acoustic signal was simultaneously digitized using 16-bit quantization at 22 kHz (low-pass filtered at 11 kHz). Two videos from each syllable pair (e.g., one *BELT* and one *DEALT*) were edited so that their consonant releases (visually determined) occurred in the same video frame. All of the movies were edited to be exactly the same length (47 frames, or 1,567 ms), and all included the onset of the closing gesture for the initial consonant and the offset of the final consonant.

Audiovisual dubbing was accomplished in Adobe Premiere. For a given stimulus pair, one acoustic token of each syllable (taken from different utterances than the ones used for the videos) was temporally aligned at the point of consonant release with each of the two videos. The point of consonant release in the acoustic tokens was determined to be the point of burst release for voiced (/b/, /d/) and voiceless (/p/, /t/) stops, the offset of frication for /v/, and the transition from nasal murmur to vowel for the nasal stops (/m/, /n/). The dubbing process resulted in the creation of two audiovisually congruent and two discrepant stimuli for each stimulus pair. For example, the *BELT–DEALT* pair produced auditory *BELT* + visual *BELT* (henceforth, a*BELT*-v*BELT*), a*DEALT*-v*DEALT*, a*BELT*-v*DEALT*, and a*DEALT*-v*BELT* stimuli. As a consequence, equal numbers (32 each) of audiovisually discrepant stimuli were created in which both the auditory and visual tokens were words (henceforth, aW-vW), in which only the auditory token, and not the visual token, was a word (aW-vNW), in which only the visual token was a word (aNW-vW), and in which neither was a word (aNW-vNW). Note that a word–nonword stimulus pair such as *BEG–DEG* generated one aW-vNW stimulus (a*BEG*-v*DEG*) and one aNW-vW stimulus (a*DEG*-v*BEG*), as did a nonword–word pair such as *BESK–DESK*. In contrast, word–word pairs generated only aW-vW stimuli, and nonword–nonword pairs generated only aNW-vNW stimuli.

Procedure

The stimuli were presented on a 17-in. (43.18-cm) monitor connected to an Apple Macintosh PowerMac 9600. The video image filled approximately half of the screen, was approximately 4.6 in. high × 6.2 in. wide, and was surrounded by a black screen. (Although the image size was somewhat smaller than that commonly used in audiovisual speech research, Jordan & Sergeant, 1998, found that the McGurk effect was highly resistant to variation in image size. Notably, they found no reduction in the magnitude of the McGurk effect for images far smaller than the ones used here.) The sound was presented through speakers placed on each side of the computer monitor at a comfortable listening level. Participants were tested individually. They were seated in front of the computer at a normal viewing distance (approximately 18 in. [45.72 cm]), thus creating a viewing angle of approximately 15°.

The participants were instructed to watch the computer monitor, listen to the speakers, and type on the computer keyboard the letter or letters corresponding to the initial consonant (or combination of consonants, such as "bd," "mn," or "pt") that they heard. They were informed that the syllables were dubbed and that they should identify the initial consonant(s) that they heard regardless of the visual stimulus. Participants were also instructed to rate the goodness of the consonant(s) that they heard (regardless of what they saw) as an example of its phonetic category, on a scale ranging from 1 (*poor*) to 5 (*excellent*). Note that the goodness instructions specifically referred to the heard consonant, which, unbeknownst to the participants, could be a visually influenced percept. The experimenter sat in the room to confirm that the participants watched the monitor.

Stimulus presentation was controlled by PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). Each trial began with a beep and a text that warned participants to orient to the monitor. After a 300-ms interval, an audiovisual token was presented, followed by a prompt to type in the identification response. After that response, a second prompt requested a goodness rating. The intertrial interval was 1,800 ms. The experiment began with four audiovisually congruent trials that were not included in the analyses.

Each participant was randomly assigned to one of four groups and was presented a different set of 64 of the 256 audiovisual stimuli according to the group. Each of these four stimulus sets (presented to different groups) contained only one of the four possible audiovisual pairings for a given stimulus pair (e.g., one set included a*BELT*-v*DEALT*, and another included a*BELT*-v*BELT*). The sets were balanced so that each contained one stimulus of each audiovisual pairing (audio bilabial or nonbilabial, video bilabial or nonbilabial) by lexical status of the auditory (W vs. NW) and visual (W vs. NW) tokens by the consonant contrast (/b/ vs. /d/, /m/ vs. /n/, /p/ vs. /t/, or /b/ vs. /v/). Each participant received a different randomized sequence of the stimuli.

Results

Two sets of analyses are reported here. The first examined the goodness ratings, to determine whether the McGurk effect resulted in poor category exemplars. This analysis compared ratings of audiovisually congruent and discrepant stimuli, looking separately at visual and auditory responses to the latter set of stimuli. The second set of analyses examined whether lexical status had an influence on patterns of consonant identification, with separate analyses for trials involving audiovisually congruent stimuli and those involving discrepant stimuli. In all of the analyses reported here, a response was counted as an *auditory* response only if it matched the phoneme specified in the auditory signal. A response was counted as a *visual* response only if it matched the visual signal in place of articulation and the auditory signal in voicing and manner.[3] For example, for an auditory /b/–visual /d/, "d" counted as a visual response, but "t," "n," and "th" (for /ð/) did not, because lexical status for these responses was not controlled. The only exception to this rule was with the /b/–/v/ contrast, in which

[3] The labels *auditory* and *visual* for the two classes of responses were chosen for ease of exposition. Despite the implication of the terminology, visual responses did not reflect pure visually determined percepts. Rather, they were the consequence of audiovisual integration because they incorporated visual information for place of articulation with acoustically specified information about voicing and manner.

the visual response did not match the auditory signal in manner; the visual response for auditory /b/–visual /v/ was "v," and for auditory /v/–visual /b/ it was "b."

Overall, accuracy was very high in the congruent trials, with correct identifications of the initial phoneme on 98.7% of the trials. In the discrepant trials, 49.7% of the responses were visual, and 30.8% were auditory. (The remaining 19.5% of the responses consisted of fusions, combinations, and miscellaneous other responses.)

Goodness Ratings

The mean goodness ratings (computed by participant, collapsing across all other stimulus factors) were compared for three response categories: correct responses to audiovisually congruent tokens (*congruent–correct*) and visual and auditory responses to audiovisually discrepant tokens (*discrepant–visual* and *discrepant–auditory*, respectively). The mean ratings in each of these three categories are presented in Table 1. The table reveals that ratings were highest for the congruent–correct responses, intermediate for the discrepant–visual responses, and lowest for the discrepant–auditory responses. A one-way repeated measures analysis of variance (ANOVA) with three levels (i.e., congruent–correct response, discrepant–visual response, and discrepant–auditory response) was computed by participants only. (The items analysis provided unreliable results owing to the vastly unequal proportions of auditory and visual responses for individual tokens.) The ANOVA revealed a significant difference among the means, $F_1(2, 118) = 146.38$, $p < .0001$. Specific comparisons revealed significant differences among all three means: The smallest $t_1(59)$ of the three comparisons was 5.57, $p < .0001$, for the contrast between discrepant–auditory and discrepant–visual. These results indicated that the McGurk percepts were considered poorer category exemplars than percepts of audiovisually congruent tokens and that the auditorily determined percepts from discrepant stimuli were rated as especially poor exemplars of their phonetic category.

Identification Results

Audiovisually congruent trials. A 2 (lexicality: word vs. nonword) × 2 (neighbor lexicality: word vs. nonword, corresponding to the stimulus pair member in the discrepant trials; e.g., a*BESK*-v*BESK* has the word neighbor *DESK*) × 2 (place of articulation of the initial consonant: bilabial, including /b/, /m/, and /p/, vs. nonbilabial, including /d/, /n/, /t/, and /v/) ANOVA on the percentage of correct responses was conducted by participants (repeated measures on all factors) and by items (between items on all factors). There was a main effect of lexicality by participants only, $F_1(1, 58) = 5.14$, $p < .05$; $F_2(1, 120) = 1.78$, *ns*; with greater accuracy for words (99.3% correct) than for nonwords (98.0% correct). There was also an effect of place of articulation that was also significant by participants, $F_1(1, 58) = 7.22$, $p < .01$, but not by items, $F_2(1, 120) = 2.49$, *ns*, with greater accuracy for bilabials (99.4% correct) than for nonbilabials (97.9% correct). There was no effect of neighbor lexicality, and there were no interactions among the factors. Thus, there were minimal differences in accuracy across the conditions in the congruent trials.

Table 1
Goodness Ratings in Experiment 1

Audiovisual condition	Response	Goodness rating	
		M	*SD*
Congruent	Correct	4.25	0.48
Discrepant	Visual	3.28	0.57
Discrepant	Auditory	2.81	0.81

Note. Ratings ranged from 1 (*poor*) to 5 (*excellent*).

Audiovisually discrepant trials. The mean percentages of visual responses are shown in Figure 1, separated by the lexical status of the auditory response (matching the auditory token's initial consonant) and the visual response (matching the visually presented initial consonant). Figure 1 reveals that the incidence of visual responses varied considerably depending on the lexical status of both the visual and auditory percepts, with a decreasing incidence of visual responses across the aNW-vW, aW-vW, aNW-vNW, and aW-vNW conditions. The comparison of greatest lexical disparity, between the aW-vNW and aNW-vW cases, revealed a 33.1% difference.

The percentages of visual responses were analyzed in a 2 (auditory lexicality: auditory–word vs. auditory–nonword) × 2 (visual lexicality: visual–word vs. visual–nonword) × 2 (auditory place of articulation: auditory bilabial vs. nonbilabial, with visual nonbilabials and bilabials, respectively) ANOVA, conducted by participants (repeated measures on all factors) and by items (between items on all factors). There was a significant main effect of both auditory lexicality, $F_1(1, 59) = 55.68$, $p < .0001$; $F_2(1, 120) = 5.60$, $p < .05$, with more visual responses in the auditory–nonword (55.3%) than in the auditory–word (44.0%) conditions, and visual lexicality, $F_1(1, 59) = 117.86$, $p < .0001$; $F_2(1, 120) = 20.65$, $p < .0001$, with more visual responses in the visual–word (60.6%) than in the visual–nonword (38.8%) conditions. There was no interaction between the two factors. There was also a significant main effect of auditory place of articulation, $F_1(1, 59) = 25.44$, $p < .0001$; $F_2(1, 120) = 12.09$, $p < .001$, with a greater McGurk effect for auditory bilabials (58.2% visual responses) than for auditory nonbilabials (41.2%). However, this asymmetry is not of theoretical concern for the present experiment, because auditory place of articulation did not interact with either of the lexical factors.

Because the response options were unrestricted in the experiment, participants made visually influenced responses other than the expected visual responses (e.g., fusions and combinations). Therefore, it is possible that the lexical influences on the percentage of visual responses did not reflect an effect on the overall incidence of the McGurk effect (defined as any response that deviates from the one consistent with the acoustic signal) but only on the relative incidences of different types of visually influenced responses. In fact, the pattern of auditory (non-McGurk) responses was a near inversion of the pattern of visual responses: Auditory responses occurred most often in the aW-vNW condition (46.3%), to an intermediate degree in the aW-vW (25.8%) and aNW-vNW (30.0%) conditions, and least often in the aNW-vW condition (20.9%). Statistical analyses confirmed that these lexical effects on auditory responses were significant. Thus, there were lexical influences on the overall incidence of the McGurk effect, as well as on the incidence of the specifically defined visual responses.

In summary, there was a pronounced visual influence on categorization (the McGurk effect) when there was an audiovisual

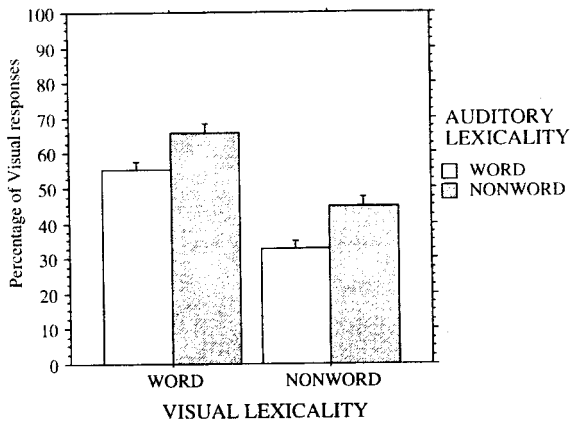

Figure 1. Percentages of visual responses in Experiment 1 as a function of the lexical status of the syllables formed by auditory and visual responses. Error bars represent standard errors.

discrepancy. Additionally, there was a strong lexical influence on categorization of the discrepant stimuli, in both visually influenced and auditorily based responses and with both the auditory and visual lexicality factors playing significant roles.

Discussion

The primary finding of Experiment 1 was that lexical factors exerted a strong influence on the McGurk effect: There were significantly fewer responses consistent with the acoustic signal (auditory responses), and more responses that incorporated the visually specified place of articulation (visual responses), for vW-aNW stimuli than in the opposite case. Thus, the lexical factors affected both the overall incidence of the McGurk effect and the incidence of particular visually influenced percepts.

An interesting aspect of the results is that both auditorily based and visually influenced lexicality had independent significant effects on phoneme identification. This finding is noteworthy because it indicates that the incidence of both visual and auditory responses were affected not only by the lexical status of the syllable formed by the selected phoneme but also by the lexical status of the syllable that would have been formed by a phoneme that was not selected. For example, there were more visual responses in the aNW-vW condition (e.g., aBESK-vDESK) than in the aW-vW condition (e.g., aBELT-vDEALT). This suggests that when two word candidates are potential matches to an audiovisually discrepant signal, they exert opposing influences on phoneme decisions, in keeping with the notion of competition between partially activated lexical candidates that is central to several theories of lexical access (e.g., Luce & Pisoni, 1998; Marslen-Wilson, 1990; McClelland & Elman, 1986; Norris, 1994).

Goodness ratings were collected to determine the extent to which audiovisually discrepant signals give rise to perceptual uncertainty. The findings showed that audiovisual discrepancy results in less robust phonetic percepts than does audiovisual congruency, in accordance with the discrimination tests of Rosenblum and Saldaña (1992) and the continuous rating measures of Massaro and Ferguson (1993; cf. Massaro, 1987). These goodness-rating results are consistent with the assumption that audiovisual discrepancy gives rise to perceptual ambiguity and therefore that discrepant stimuli will provide an appropriate context for lexical effects to emerge, as they did.

The goodness-rating results also provide new insight into the nature of the different types of responses given to audiovisually discrepant stimuli. Auditory-based categorization of these stimuli resulted in even poorer category exemplars than did visually influenced categorization. The fact that auditory responses had low ratings is significant because it eliminates one explanation for the failure to obtain a McGurk effect on some trials: Participants were not simply tuning out the visual display of the face (either by not looking at the screen or by shifting attentional resources away from vision). To the contrary, participants were sensitive to the audiovisual discrepancy when they made non-McGurk responses.

This outcome is consistent with the possibility that when perceivers are presented with audiovisual speech stimuli, they automatically integrate the information across modalities, even in instances when the McGurk effect does not occur. Perceptual integration of an audiovisually discrepant signal may yield partial support for more than one phoneme candidate (e.g., a perceptually integrated auditory /b/–visual /d/ stimulus provides support for both /b/ and /d/), one of which is ultimately selected (e.g., Massaro, 1987). The selected candidate may be consistent with the auditory signal (for auditory /b/–visual /d/, /b/), but it is not necessarily an auditory percept per se; the selection process may be blind to the modality of the information that supported the phoneme. The fact that auditory percepts received lower goodness ratings than visual percepts attests to the powerful influence of visual information for place of articulation in the present experiment.

In summary, Experiment 1 provides strong evidence for lexical influences on the McGurk effect. Experiments 2 and 3 extend this finding by examining the time course of the audiovisual lexical effect and by replicating it under more stringent constraints.

Experiment 2

Experiment 2 was a speeded-classification variant of Experiment 1, using the same stimuli and overall design. Most tests of lexical effects in the auditory domain have involved a speeded-classification task (e.g., Connine & Clifton, 1987; Fox, 1984; Miller & Dexter, 1988; Pitt, 1995; Pitt & Samuel, 1993), and it is possible that the magnitude of lexical effects could become inflated when responses are untimed. If participants are allowed extra time to reflect on their responses, it could lead to greater awareness that certain responses form words and therefore result in a conscious response bias toward phonemes that form words. Accordingly, Experiment 2 tested whether the lexical effect found in Experiment 1 would replicate under speeded conditions. Furthermore, it provided a test of whether, as in the auditory domain (Fox, 1984; Miller & Dexter, 1988; Pitt, 1995; Pitt & Samuel, 1993), lexical effects would be larger for slower responses.

Method

Participants

Sixty-four undergraduates at the University of Connecticut, ranging in age from 18 to 25 years, participated in the experiment. All were enrolled in introductory psychology classes and received partial course credit for their participation. All of the participants were native speakers of English, and all reported having normal hearing and normal or corrected-to-normal vision.

Stimuli

The stimuli were slightly modified versions of the movies used in Experiment 1. First, each movie was truncated from the onset so that the point of consonant release occurred at the eighth frame, or 267 ms from the movie onset. In most cases, the truncated movies began with articulator movements into the initial-consonant closure. Second, the visual and auditory portions of the movies were saved as separate files (because of hardware storage constraints) so that audiovisual dubbing would be conducted online by PsyScope.

Additionally, six nonwords with the initial consonants of the four contrastive sets (/b/–/d/, /m/–/n/, /p/–/t/, and /b/–/v/) were generated to serve as practice items. The 24 stimuli were recorded on audiotape and were digitized at 22 kHz (low-pass filtered at 11 kHz). These auditory stimuli were not paired with visual tokens and, other than their initial consonants, were phonetically unrelated to the test stimuli.

Procedure

The experiment was conducted with the same equipment as in Experiment 1. Participants were instructed to watch the computer monitor as the stimuli were presented and to identify the initial consonant that they heard, regardless of what they saw, as quickly as possible. Responses were made by pressing one of two keys on a button box.

The stimuli were blocked according to the initial-consonant contrast: The four blocks each consisted of 16 stimuli from one of the /b/–/d/, /m/–/n/, /p/–/t/, or /b/–/v/ contrasts, respectively. The order of the blocks was counterbalanced across participants using a Latin square design. Assignment of the responses to the response keys (e.g., *B*:left and *D*:right or *M*:left and *N*:right) was fixed by block. The bilabial responses in all four blocks were always assigned to the same key; the mapping of bilabial responses to the left or right key was counterbalanced across participants.

Stimulus presentation was controlled by PsyScope. Each participant was randomly assigned to one of four groups and was presented a different set of 64 of the 256 possible audiovisual stimuli according to the group. The four sets of 64 stimuli were the same as in Experiment 1.[4] Each trial began with a beep and a text that warned the participants to orient to the monitor, along with a display of the response alternatives (letters corresponding to the phoneme choices were displayed on the monitor above colored rectangles, with the positions and colors of the rectangles corresponding to the respective response keys). After a 750-ms interval, a movie and a sound file were simultaneously presented. Recording of the button-box response began at the onset of stimulus presentation. A trial ended with a button-box response unless no response was given in 3,000 ms; there was an intertrial interval of approximately 2,500 ms.

At the beginning of each block, participants were informed of the response alternatives. They were then given 6 auditory-only practice trials at the beginning of each set. Thus, each block consisted of 6 practice and 16 test trials, for a total of 24 practice and 64 test trials.

Results

Analyses of the data are reported in two sections. The first examines the overall identification results. The second involves a partitioning of the responses into reaction time bins, to examine the magnitude of the lexical influences on identification separately for each partition.

Overall Identification

Audiovisually congruent trials. As in Experiment 1, participants' accuracy in identifying the initial consonants in the audiovisually congruent trials was analyzed using a 2 (lexicality: word vs. nonword) × 2 (neighbor lexicality: word vs. nonword) × 2 (place of articulation: bilabial vs. nonbilabial) ANOVA on the percentage of correct responses, conducted by participants (repeated measures on all factors) and by items (between items on all factors). Overall, accuracy was very high (98.2% correct). There were no effects of lexicality or neighbor lexicality. There was a marginal interaction of the factors, $F_1(1, 63) = 5.75$, $p < .05$; $F_2(1, 120) = 3.10$, $p = .08$, but the differences among the cells were minimal (ranging from 97.5% for nonwords with word neighbors to 99.0% for words with word neighbors). There was also a marginal interaction of place of articulation and neighbor lexicality, $F_1(1, 63) = 5.27$, $p < .05$; $F_2(1, 120) = 3.10$, $p = .08$, but with minimal variation among the cell means (97.2%–99.2%).

Audiovisually discrepant trials. Because responses were forced choice, only the percentage of visual responses was analyzed. Overall, participants made visual responses 65.6% of the time. The pattern of results across the lexical conditions was similar to that found in Experiment 1: Visual responses occurred most often in the aNW-vW condition (73.1%), intermediately often in the aW-vW (68.2%) and aNW-vNW (63.8%) conditions, and least often in the aW-vNW condition (57.2%). The overall effects were somewhat smaller than in Experiment 1; the disparity between the lexical extremes (aW-vNW and aNW-vW) was 15.8%, in contrast to the 33.1% difference found in Experiment 1.

A 2 (auditory lexicality: auditory–word vs. auditory–nonword) × 2 (visual lexicality: visual–word vs. visual–nonword) × 2 (auditory place of articulation: auditory bilabial vs. nonbilabial) ANOVA was performed by participants (repeated measures) and by items (between items) on the percentages of visual responses. The main effect of auditory lexicality was significant by participants, $F_1(1, 63) = 14.28$, $p < .001$, but not by items, $F_2(1, 120) = 2.22$, *ns*, and the main effect of visual lexicality was significant by both subjects and items, $F_1(1, 63) = 33.10$, $p < .0001$; $F_2(1, 120) = 6.94$, $p < .0001$. There was no interaction between the two factors. There was a significant main effect of auditory place of articulation, $F_1(1, 63) = 76.61$, $p < .0001$; $F_2(1, 120) = 40.18$, $p < .0001$, with more visual responses for auditory bilabials than for auditory nonbilabials, as in Experiment 1. There was also an interaction between auditory lexicality and auditory place of articulation that was significant by participants, $F_1(1, 63) = 5.31$, $p < .05$, but not by items, $F_2(1, 120) = 1.07$, *ns*, with the auditory lexicality effect occurring only for the auditory bilabial tokens.

Reaction Time Partition Analyses

Each participant's responses were sorted by reaction time and were assigned to equal-size fast, medium, and slow partitions, following Miller and Dexter (1988). The mean boundaries (averaged across the boundaries that were computed individually for each participant) between the fast and medium partitions and between the medium and slow partitions were 622 ms and 812 ms, respectively. Three 2 (auditory lexicality) × 2 (visual lexicality) repeated measures ANOVAs (by participants only, because the partitions were determined separately for each participant) were performed on the percentages of visual responses in the fast,

[4] Owing to an error in assigning participants to conditions, the three manipulations of group, button–response pairing, and block order were not fully counterbalanced with respect to one another; however, each was fully counterbalanced when the other two manipulations were collapsed.

medium, and slow partitions.[5] The means are presented in the top, middle, and bottom panels, respectively, of Figure 2.

In the fast partition, 73.5% of the responses were visual. The top panel of Figure 2 shows no consistent pattern of differences among the lexical conditions. Accordingly, the ANOVA revealed that neither main effect of auditory lexicality or visual lexicality, nor their interaction, was significant, $F_1(1, 32) < 1$, in contrast to the analysis including all responses. Moreover, there was no difference between the cells representing the lexically extreme conditions, namely aW-vNW versus aNW-vW (1.8% difference, *ns*).

In the medium partition, 68.5% of the responses were visual. The middle panel of Figure 2 contains a pattern of results that is globally similar to that found in the overall analysis, with more visual responses in aNW-vW than in aW-vNW and with the aW-vW and aNW-vNW conditions intermediate. The ANOVA revealed that the 6.0% difference between the auditory–word and auditory–nonword conditions of auditory lexicality was not significant, $F_1(1, 51) = 2.25$, *ns*. However, there was a main effect of visual lexicality, $F_1(1, 51) = 12.38$, $p < .001$, with more visual responses in the visual–word (75.2%) conditions than in the visual–nonword (62%) conditions. There was no interaction between the factors. Despite the absence of a significant auditory lexicality effect, the comparison of the lexically extreme conditions (aNW-vW and aW-vNW) produced a significant 19.3% difference, $t_1(52) = 3.88$, $p < .0005$.

Finally, in the slow partition, 58.1% of the responses were visual. As with the medium partition, the bottom panel of Figure 2 contains a globally similar pattern of results to the overall analysis. An ANOVA revealed main effects of both auditory lexicality, $F_1(1, 63) = 15.45$, $p < .005$, with more visual responses in the auditory–nonword (63.7%) conditions than in the auditory–word (52.5%) conditions, and visual lexicality, $F_1(1, 63) = 6.99$, $p = .01$, with more visual responses in the visual–word (63.5%) conditions than in the visual–nonword (53.7%) conditions. There was also a significant interaction of the two factors, $F_1(1, 63) = 4.45$, $p < .05$, reflecting an auditory lexicality effect only for the visual–word conditions and likewise a visual lexicality effect only for the auditory–nonword conditions. This interaction can be observed in the bottom panel of Figure 2, which shows that visual responses were more frequent in the aNW-vW condition than in the other three conditions, which all had similar results. As in the medium partition, there was a significant 20.6% difference, $t_1(63) = 4.80$, $p < .0001$, between the lexically extreme conditions (aNW-vW and aW-vNW).

In summary, the reaction time partition analysis revealed no lexical effects for fast responses but significant lexical influences on the McGurk effect for medium-speed and slow responses. Although the specific patterns of lexical influences differed between the medium and slow response categories (in particular, in the relative magnitudes of the auditory and visual lexicality effects), the overall effects of lexical status were highly robust at both response speeds. An additional finding was a decrease in the proportion of visual responses from the fast to the slow partition.

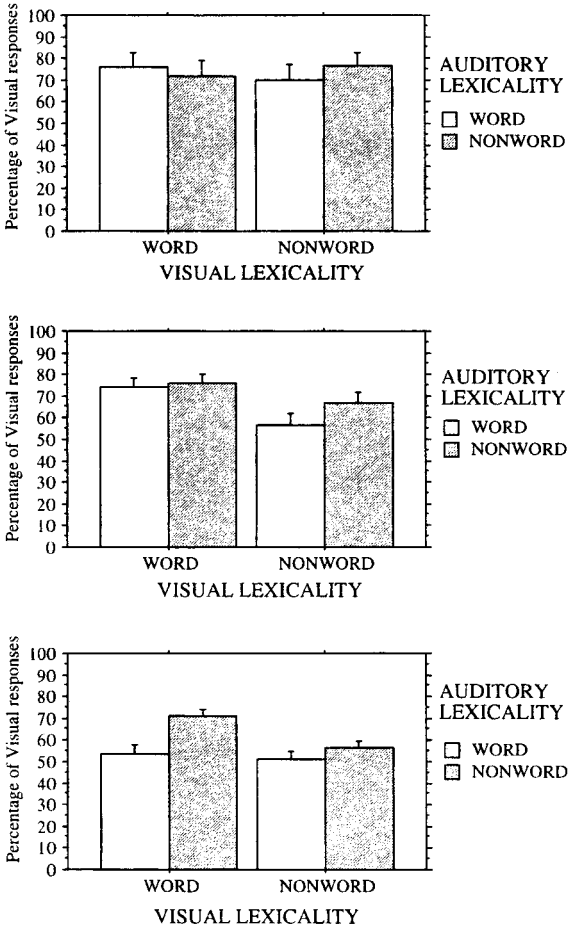

Figure 2. Percentages of visual responses in Experiment 2 as a function of the lexical status of the syllables formed by auditory and visual responses and as a function of response speed. Results for fast, medium, and slow responses are shown in the top, middle, and bottom panels, respectively. Error bars represent standard errors.

Discussion

Experiment 2 revealed robust lexical influences on the McGurk effect, replicating the results of Experiment 1 in a forced-choice, speeded-response task. Although the magnitude of the effect was somewhat reduced in comparison with the effect in Experiment 1, it remained robust. Moreover, an analysis of the time course of the lexical effect revealed a pattern similar to findings with auditory stimuli: Lexical effects were robust for medium-speed and slow

[5] Because there were only eight discrepant trials within each lexical condition (e.g., aW-vW or aNW-vW) and these trials could fall into any of the three partitions, several participants had cells with no responses in one or more of the partitions. Because congruent-trial responses were much faster than discrepant-trial responses (in keeping with previous findings by Green & Kuhl, 1991, and Massaro & Cohen, 1983), there were substantially fewer discrepant trials in the fast partition than in the medium and slow partitions. Consequently, most of the missing cells in the analysis of the discrepant stimuli occurred in the fast partition. (The auditory place of articulation factor was not included because its inclusion reduced the number of items in each condition to four, resulting in a drastic increase in the incidence of empty cells in the partitions.) Each partition analysis included all of the participants that had no missing cells for that partition; therefore, different but overlapping subsets of participants were used in each of the three analyses. In all, 33 participants were included in the fast-partition analysis, 52 were included in the medium-partition analysis, and all 64 were included in the slow-partition analysis.

responses but not for fast responses. This latter finding could reflect, as has been argued in the case of auditory lexical effects, a gradual buildup of lexical activation over time (Connine & Clifton, 1987; Massaro & Oden, 1995; McClelland & Elman, 1986; Pitt, 1995), although it could also have simply been a consequence of fast responses being initiated before the end of the stimuli and hence, before recognition of the word (a concern that also applies to findings in the auditory domain; cf. Pitt & Samuel, 1993).

Two aspects of the results merit additional mention. First, a notable outcome of this experiment was that visual responses were most frequent in the fast partition and least frequent in the slow partition. This suggests that visual responses overall were made more quickly than auditory responses, and in fact, a direct comparison of response times revealed that visual responses were, on average, 150 ms faster than auditory responses. There are at least two possible reasons for this. One possibility is that the reaction time difference is a reflection of the difference in category goodness between the two classes of responses reported in Experiment 1: Auditory responses were poorer category exemplars than visual responses, and accordingly, those responses were made more slowly. Another possibility is that visual responses were made more quickly because visual information for place of articulation was available earlier than auditory information, owing to anticipatory lip and tongue movements before the acoustic onset (cf. Munhall & Tohkura, 1998).

A second notable finding is that although the lexical effects in the medium and slow partitions were robust at a global level of analysis, the specific patterns of results differed for the two sets of responses. Specifically, the auditory lexicality factor produced a significant effect only in the slow partition, whereas the visual lexicality factor appeared to have a greater influence in the medium partition than in the slow partition. The reason for this difference is not clear. Possibly, it reflects an earlier time course of visually presented information relative to auditory information. However, it may be an artifact of stimulus factors unrelated to lexical status: The stimuli in the aW-vW condition were considerably shorter, on average, than the stimuli in the other conditions, and the changes in the lexical effects from the medium to the slow partition were primarily due to a large drop-off in the percentage of visual responses in the aW-vW condition (see Figure 2). Future studies, in which the temporal properties of the stimuli in the different lexical conditions are controlled, are required to assess whether auditory and visual lexical influences have independent time courses.

To summarize, Experiments 1 and 2 revealed significant lexical influences on the McGurk effect. Experiment 2 also revealed that the time course of the lexical effect is typical of lexical effects found in the traditional Ganong paradigm. However, one weakness of these studies is that the stimuli in the different lexical conditions were not matched on all stimulus properties, leaving open the possibility that the apparent lexical effects were due to uncontrolled stimulus variability in the magnitude of the McGurk effect. The fact that the pattern of results was quite orderly in both experiments, with the two lexically neutral conditions (aW-vW and aNW-vNW) providing intermediate outcomes to the two lexical extreme conditions (aW-vNW and aNW-vW), makes this alternative interpretation less plausible. Nonetheless, to address this concern, Experiment 3 provides a more conservative test for lexical influences on the McGurk effect, using stimuli that are physically matched. It also provides a test of whether the time course of the lexical effect observed in Experiment 2 will replicate with these matched stimuli.

Experiment 3

As noted earlier, experiments exploring lexical influences in an auditory phoneme-identification task, following Ganong (1980), typically involve stimuli that have been constructed so that pairs of syllables differing in lexical status are physically identical except for their final phonemes. That strategy was adopted for Experiment 3, in a test of whether lexical influences on the McGurk effect would emerge under more stringent stimulus constraints than were used in Experiments 1 and 2.

In Experiment 3, auditory words and nonwords, differing only in their final consonant (e.g., *PACE* and *PAKE*), were paired with incongruent visual stimuli (e.g., with an initial /t/). These stimuli were selected so that the visual response for the auditory word (*PACE*) formed a nonword (*TACE*) and the visual response for the auditory nonword (*PAKE*) formed a word (*TAKE*). Critically, the auditory word and nonword in a pair were physically identical except for the final consonant, and both were paired with the same discrepant visual stimulus. Thus, any difference between the two stimuli in the magnitude of the McGurk effect can be attributed to lexical factors. Note that in contrast to Experiments 1 and 2, this experiment only involved two lexical conditions: Either the auditory response formed a nonword and the visual response formed a word, or the auditory response formed a word and the visual response formed a nonword.

As in the preceding experiment, participants in Experiment 3 made speeded, forced-choice responses. Therefore, Experiment 3 was designed to address the same questions raised in Experiment 2. First, would the incidence of visual responses be influenced by lexical status of the response, under speeded conditions? Second, would lexical effects emerge for medium-speed and slow responses, but not for fast responses?

Method

Participants

Sixty-two members of the University of Connecticut community participated in the experiment, including 53 undergraduates enrolled in introductory psychology classes who received partial course credit for their participation and 9 psychology graduate students. Their ages ranged from 18 to approximately 35 years. The data of 2 participants were not included in the analyses: One was not a native speaker of English, and the other did not attend to the computer monitor throughout the experiment. All of the remaining participants were native speakers of English and reported having normal hearing and normal or corrected-to-normal vision.

Stimuli

Stimulus selection. Three sets of matched word–nonword minimal-syllable pairs were selected. Within each syllable pair, the two syllables differed only by a minimal change in place of articulation in the initial consonant (e.g., *PACE–TACE*). Each pair was matched to a second minimal pair that differed from the first pair only in the final consonant (e.g., *PAKE–TAKE*). The change in place of articulation within a pair was always between a bilabial and an alveolar consonant and always preserved voicing and manner; the three pairs of initial consonants were /b/–/d/, /m/–/n/, and /p/–/t/. The change in the final consonant always caused a reversal in the

lexical status of the pair members, so that one pair within a set had a bilabial-initial word and an alveolar-initial nonword, and the other pair had a bilabial-initial nonword and an alveolar-initial word. The three sets were *BENCH–DENCH–BENSE–DENSE*, *MESS–NESS–MECK–NECK*, and *PACE–TACE–PAKE–TAKE*.

Stimulus generation. A male speaker was recorded on videotape producing multiple repetitions of each of the 12 syllables (4 syllables in each lexical set) in a similar recording environment as that used to record the stimuli of Experiments 1 and 2. The same speaker was also recorded on audiotape in a separate session that took place in a quiet room, again producing multiple repetitions of each syllable.

Digitization of the videos was accomplished in the same manner as for the stimuli of Experiment 1. The video tracks of one token of each of the 12 items were saved as individual movies. All of the movies were edited so that the visible point of initial consonant release occurred in the 8th frame of the movie (or after 267 ms). Additionally, the movies were truncated after the 12th frame (400 ms), so that participants only viewed the initial consonant–vowel (CV) portion; this was accomplished online using PsyScope (see the *Procedure* section). The visual truncation was performed so that a single visual token could be presented with both members of a pair of auditory stimuli that differed in lexical status, without generating an audiovisual mismatch on the final consonant for either one. (For example, a visual *DENSE* could be paired with an auditory *BENCH* without a final-position discrepancy between the auditory /č/ and visual /s/). As a result, there were two tokens of each of 6 visible CV syllables, two in each set. For example, in the *BENCH* set, there were two visual tokens of /bɛ/, one originating from an utterance of /bɛns/ (*BENSE*) and one from an utterance of /bɛnč/ (*BENCH*), and two visual tokens of /dɛ/, one originating from an utterance of /dɛns/ and one from an utterance of /dɛnč/.[6]

The auditory speech tokens were digitized on an Apple Macintosh PowerMac 7100 using SoundScope (GW Instruments, Somerville, MA) with 16-bit quantization and a sampling rate of 22 kHz and were low-pass filtered at 11 kHz. Six tokens of each syllable were selected. Silence was added to the beginning of each syllable so that the acoustic consonant release would be synchronized with the consonant release in the visual token. (The point of consonant release was identified with waveform and spectrographic displays as the release burst for /b/, /d/, /p/, and /t/ and as the nasal murmur offset for /m/ and /n/.)

An acoustic cross-splicing technique was adopted to eliminate psychoacoustic differences between the words and nonwords. Each stimulus was truncated at a zero crossing at the onset of the final consonant, leaving only the initial CV portion (or CVn, in the *BENCH* set). This resulted in 12 tokens of each initial portion (e.g., there were 12 tokens of /bɛn/, 6 originating from utterances of /bɛns/ and 6 originating from utterances of /bɛnč/). Two final consonants excised from other tokens were appended to each initial portion to create two new stimuli; thus, each token of /bɛn/ was paired with a final /č/ and with a final /s/ to produce tokens perceived as *BENCH* and *BENSE*, respectively. Therefore, within each set, there were equal numbers of words and nonwords that were matched pairwise on acoustic properties, with a total of 48 acoustic tokens in each set. For example, in the *BENCH* set, there were 12 acoustic tokens each of *BENCH*, *BENSE*, *DENCH*, and *DENSE*.

Audiovisual dubbing was conducted online by PsyScope to create audiovisually discrepant and congruent stimuli. To create audiovisually discrepant stimuli, both members of each pair of acoustically matched syllables (e.g., an acoustic token of *BENSE* and an acoustic token of *BENCH*, originating from a single acoustic /bɛn/ stem) were presented with a single discrepant visual token (e.g., a visual /dɛ/). (One half of the acoustically matched syllable pairs within a particular set were presented with one truncated discrepant visual token, such as visual /dɛ/ from /dɛns/, whereas the other half of the pairs were presented with a different truncated discrepant visual token, such as visual /dɛ/ from /dɛnč/.) As a result, the pairs of lexically contrastive audiovisually discrepant stimuli were matched on auditory and visual properties, so that only the acoustic final consonant differed between word and nonword syllables.

One half of the acoustic tokens were also selected for creation of audiovisually congruent stimuli. There were equal numbers of words and nonwords among these stimuli. These acoustic tokens were paired with visual tokens with the same initial consonant (e.g., acoustic /bɛns/ paired with visual /bɛ/); the visual token's final consonant, which was not viewed, matched the acoustic final consonant for half of the congruent stimuli (e.g., half of the acoustic /bɛns/ tokens used in congruent trials were paired with a visual /bɛ/ from /bɛns/, and the other half were paired with a visual /bɛ/ from /bɛnč/).

Overall, each auditory token was presented with a discrepant visual token, and half of the auditory tokens were also presented with a congruent visual token. Thus, there were 48 discrepant stimuli (24 auditory words and 24 auditory nonwords) and 24 congruent stimuli (12 words and 12 nonwords) in each of the three sets.

Procedure

The experiment was conducted with the same equipment as in the previous experiments. Participants were randomly assigned to one of three groups, each of which was presented with the stimuli from one of the three sets (the *BENCH*, *MESS*, or *PACE* sets). Twenty participants ran in each of the three conditions. As in Experiment 2, participants were instructed to watch the computer monitor as the stimuli were presented and to identify as quickly as possible the initial consonant that they heard regardless of what they saw; responses were made by pressing one of two keys on a button box. Within each group, the assignment of responses to buttons (e.g., *B*:left and *D*:right vs. *D*:left and *B*:right) was counterbalanced across participants.

As in the previous experiments, stimulus presentation was controlled by PsyScope. Each trial began with a warning beep and a printed message reminding the participant to watch the computer monitor, along with a display of the response alternatives and their correspondence to the response keys. After a 500-ms interval, a movie and a sound file were simultaneously presented; the movie was truncated after 400 ms (12 frames). Recording of the button-box response began at the onset of stimulus presentation. A trial ended with a button-box response unless no response was given in 3,000 ms; there was an intertrial interval of approximately 2,500 ms.

The experimental block began with 6 practice trials with audiovisually congruent tokens, followed by 8 unrecorded trials and then by the 72 test trials in a different randomized sequence for each participant. Before the experimental block, the participants were run in a block of trials that involved presentation of audiovisually congruent and discrepant nonword stimuli, which is not discussed here.

Results

Analyses of the data are reported in two sections, in the same manner as Experiment 2. The first section examines the overall identification results, and the second section examines identification within reaction time partitions.

[6] It is likely that the truncated visual signals preserve some coarticulatory information in the vowel for the final consonant, which could produce a subtle influence on perception of the acoustic final consonant. However, because equal numbers of visual tokens with each final consonant in a given stimulus set were used in the experiment, these coarticulatory effects would not affect the overall pattern of results. A similar concern about the acoustic cross-splicing technique (described in the following paragraphs) is resolved in the same manner.

Overall Identification

Audiovisually congruent trials. Accuracy in the congruent trials was very high (97.8% correct). The mean percentages of correct responses for word and nonword stimuli were compared in a 2 (lexicality: word vs. nonword) × 2 (place of articulation of the initial consonant: bilabial, including /b/, /m/, and /p/, vs. alveolar, including /d/, /n/, and /t/) × 3 (consonant pair: /b/–/d/ vs. /m/–/n/ vs. /p/–/t/, corresponding to the three stimulus sets) ANOVA conducted by participants (repeated measures on lexicality and place of articulation and between subjects on consonant pair) and by items (between items on all factors). There was an effect of lexicality that was significant by participants only, $F_1(1, 57) = 4.46$, $p < .05$; $F_2(1, 60) = 2.76$, $p = .10$, with slightly higher accuracy for nonwords (98.4%) than for words (97.2%). There was also an effect of consonant pair that was significant by items, $F_1(2, 57) = 2.81$, $p = .07$; $F_2(2, 60) = 8.71$, $p < .001$, reflecting slightly lower accuracy for the /m/–/n/ set (95.6%) than for the /b/–/d/ and /p/–/t/ sets (99.0% for both). Thus, as in the previous experiments, performance on the congruent trials was highly accurate in all of the conditions, despite small differences among the conditions.

Audiovisually discrepant trials. In the discrepant trials, participants made visual (McGurk) responses 37.7% of the time, a lower incidence than in the previous experiments. (A pilot test suggested that the much smaller magnitude of the McGurk effect was related to the visual truncation procedure; when the full visual tokens were presented to participants, the magnitude of the McGurk effect was much greater.) The mean percentages of visual responses are presented in Figure 3 separately for conditions differing in lexical status of the responses (either the auditory response formed a word and the visual response formed a nonword, or the auditory response formed a nonword and the visual response formed a word), in auditory place of articulation of the initial consonant, and in the initial-consonant pair. As Figure 3 demonstrates, visual responses were more prevalent when they formed a word than when they formed a nonword in all of the individual comparisons, but the magnitude of this difference varied across the conditions. Overall, visual responses were made 40.7% of the time when they formed a word and 34.7% of the time when they formed a nonword.

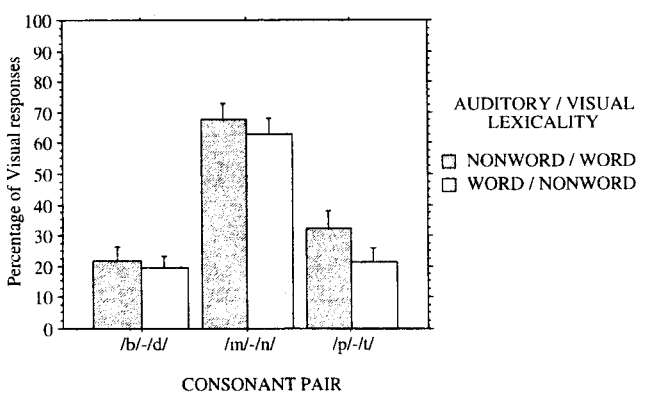

Figure 3. Percentages of visual responses in Experiment 3 as a function of lexical status and the initial-consonant contrast (/b/–/d/, /m/–/n/, or /p/–/t/). NONWORD/WORD refers to stimuli for which the auditory stimulus is a nonword and the visual response forms a word; WORD/NONWORD refers to stimuli for which the auditory stimulus is a word and the visual response forms a nonword. Error bars represent standard errors.

These apparent differences were tested in a 2 (lexicality: aW-vNW vs. aNW-vW) × 2 (auditory place of articulation: auditory bilabial vs. nonbilabial) × 3 (consonant pair: /b/–/d/ vs. /m/–/n/ vs. /p/–/t/) ANOVA performed by participants (repeated measures on lexicality and place of articulation and between subjects on consonant pair) and by items (between items on place of articulation and consonant pair and repeated measures on lexicality, because pairs of lexically contrastive stimuli shared acoustic and optical properties). The overall effect of lexicality was significant, $F_1(1, 57) = 19.86$, $p < .0001$; $F_2(1, 66) = 18.70$, $p < .0001$. There was also a significant interaction between lexicality and consonant pair, $F_1(2, 57) = 3.65$, $p < .05$; $F_2(2, 66) = 3.49$, $p < .05$, and a three-way interaction that was significant by participants, $F_1(2, 57) = 3.14$, $p = .05$, but not by items, $F_2(2, 66) = 2.36$, $p > .10$. To unpack these interactions, separate ANOVAs were conducted for each of the three lexical sets. In the /b/–/d/ set, the effect of lexicality was not significant, $F_1(1, 19) < 1$; $F_2(1, 22) < 1$, nor was the interaction with auditory place of articulation, $F_1(1, 19) < 1$; $F_2(1, 22) < 1$. In the /m/–/n/ set, there was a significant effect of lexicality, $F_1(1, 19) = 6.88$, $p < .05$; $F_2(1, 22) = 4.16$, $p = .05$, and an interaction of lexicality and auditory place of articulation that was significant by participants only, $F_1(1, 19) = 4.19$, $p = .05$; $F_2(1, 22) = 2.18$, *ns*, reflecting a larger lexical effect for auditory bilabials. Finally, in the /p/–/t/ set, there was a significant lexicality effect, $F_1(1, 19) = 17.27$, $p < .001$; $F_2(1, 22) = 25.02$, $p < .0001$, and no interaction with auditory place of articulation, $F_1(1, 19) = 2.84$, *ns*; $F_2(1, 22) = 3.22$, *ns*. Overall, the effect of lexical status on the magnitude of the McGurk effect was significant in both the /p/–/t/ and /m/–/n/ sets (although in the latter case, the extent of the lexical influence depended on whether the auditory initial consonant was /m/ or /n/) but was not significant for the /b/–/d/ set.[7]

Reaction Time Partition Analyses

Each participant's responses were sorted by reaction time and were assigned to fast, medium, and slow partitions, using the same method as in Experiment 2. The mean boundaries between the fast and medium partitions and between the medium and slow partitions (averaged across individual participants' boundaries) were 589 ms and 749 ms, respectively. Three 2 (lexicality: aW-vNW vs. aNW-vW) × 3 (consonant pair: /b/–/d/ vs. /m/–/n/ vs. /p/–/t/) repeated measures ANOVAs (by participants) were performed on the percentages of visual responses in the fast, medium, and slow partitions. The means for each consonant pair in each of the partitions are presented in Figure 4.

In the fast partition, visual responses were given on 44.6% of the trials. As the top panel of Figure 4 demonstrates, there were

[7] Independent of the lexical manipulation, there was considerable variation in the magnitude of the McGurk effect across the stimuli. There was a significant main effect of consonant pair, $F_1(2, 57) = 20.71$, $p < .0001$; $F_2(2, 66) = 311.76$, $p < .0001$, with more visual responses for the /m/–/n/ pair (65.3%) than for the /b/–/d/ (20.8%) and /p/–/t/ (27.0%) pairs, and a significant interaction of consonant pair and auditory place of articulation, $F_1(2, 57) = 3.90$, $p < .05$; $F_2(2, 66) = 30.37$, $p < .0001$, reflecting a greater McGurk effect for auditory bilabials than for alveolars for the /p/–/t/ set and an opposite pattern for the /b/–/d/ and /p/–/t/ sets. However, these differences are not relevant to the question of whether there are lexical influences on the McGurk effect.

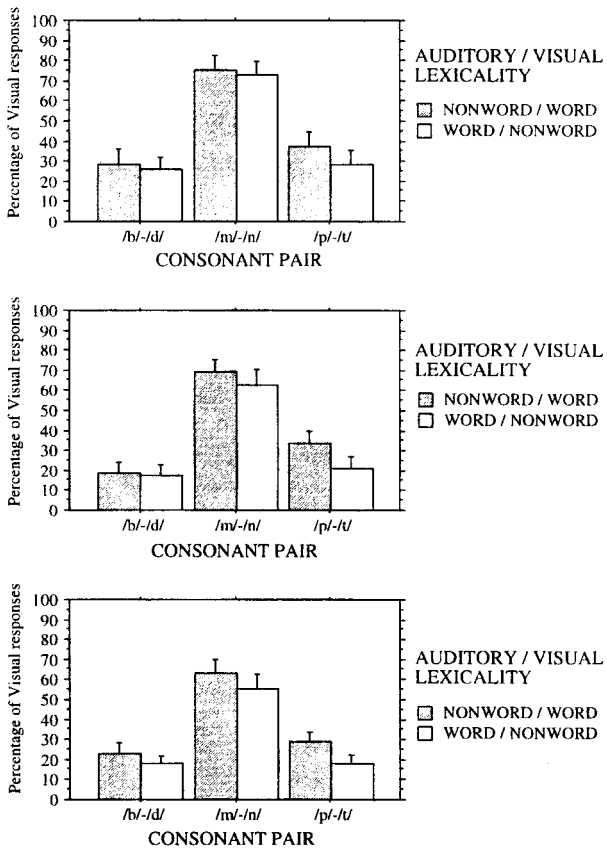

Figure 4. Percentages of visual responses in Experiment 3 as a function of lexical status, the initial-consonant contrast (/b/–/d/, /m/–/n/, or /p/–/t/), and response speed. NONWORD/WORD refers to stimuli for which the auditory stimulus is a nonword and the visual response forms a word; WORD/NONWORD refers to stimuli for which the auditory stimulus is a word and the visual response forms a nonword. Results for fast, medium, and slow responses are shown in the top, middle, and bottom panels, respectively. Error bars represent standard errors.

slightly more visual responses in the aNW-vW condition than in the aW-vNW condition for all three consonant sets. However, the overall 4.6% difference between the lexical conditions was not significant, $F_1(1, 57) = 2.32$, *ns*, nor did lexicality interact with consonant pair, $F_1(1, 57) < 1$.

In the medium partition (see middle panel of Figure 4), visual responses were given on 37.2% of the trials. There was a significant effect of lexicality, with 6.8% more visual responses in the aNW-vW condition than in the aW-vNW condition across the sets, $F_1(1, 57) = 7.66$, $p < .01$. There was no interaction of lexicality and consonant pair, $F_1(1, 57) = 1.60$, *ns*.

Finally, in the slow partition (see bottom panel of Figure 4), visual responses were given on 34.4% of trials. As in the medium partition, there was a significant effect of lexicality, with 7.8% more visual responses in the aNW-vW condition than in the aW-vNW condition across the sets, $F_1(1, 57) = 9.33$, $p < .005$, and no interaction of lexicality and consonant pair, $F_1(1, 57) < 1$.

Overall, the results were similar to those found in Experiment 2. There were significant lexical effects in the medium and slow partitions but not in the fast partition.

Discussion

Experiment 3 provided a replication of the lexical effects on identification of audiovisually discrepant stimuli reported in Experiments 1 and 2, in the absence of any potential confound between lexical status and acoustic or optical stimulus properties that might affect the incidence of visual responses. As in the previous experiments, participants in Experiment 3 made more visual responses to audiovisually discrepant stimuli when the auditory stimulus was a nonword and the visual response resulted in a word than when the auditory stimulus was a word and the visual response resulted in a nonword. Additionally, as in Experiment 2, the magnitude of this lexical effect was inversely related to response speed, because it was significant for medium and slow responses but not for fast responses.

One clear difference between the results for Experiment 3 and those from the previous experiments was in the small lexical effects. In fact, the effects were robust only for certain stimulus sets (the *PACE* set and to a lesser extent the *MESS* set), although all of the sets exhibited numerical trends in the expected direction. (Note that lexical effects do not emerge for all appropriately constructed stimuli in the auditory domain, and the reasons for this variability in the emergence of the effect are not clear; e.g., Pitt & Samuel, 1993.) Overall, the 6.0% difference between the aNW-vW and aW-vNW conditions in Experiment 3 was considerably smaller than the 33.1% and 15.8% differences in Experiments 1 and 2. There are a variety of reasons why this reduction in the magnitude of the effect might have occurred. One is that the effects in the earlier experiments might have been inflated due to the failure to control for nonlexical stimulus differences between the stimuli in the different lexical conditions. Another possible source of the difference stems from the fact that in Experiments 1 and 2, participants were presented with 64 lexically unique stimuli (32 of which were audiovisually discrepant), whereas in Experiment 3, participants were presented with multiple tokens of stimuli from a limited lexical set of four syllables. Task monotony has been found to reduce lexical effects in phoneme-monitoring tasks (Cutler et al., 1987), and perhaps the repetitiveness of the stimulus presentation in Experiment 3 reduced the effect. A third difference between experiments that may have played a role is the quality of the auditory stimuli. The auditory and visual stimuli for Experiments 1 and 2 were taken from the same recording session, and the videocamera generated some background noise in the auditory stimuli. In contrast, in Experiment 3, the auditory stimuli were recorded in a separate session than the visual stimuli and were overall of a higher quality. Perhaps the relatively high quality of the stimuli in Experiment 3 reduced the effectiveness of the lexical manipulations, in line with similar effects of stimulus quality in the auditory domain (e.g., Burton et al., 1989; McQueen, 1991).

Finally, a fourth possibility for the small lexical effect is that the magnitude of the lexical effect is somehow tied to the magnitude of the McGurk effect. The overall incidence of visual responses was considerably lower than in Experiment 2, and this appears to be at least partly related to the visual truncation procedure. (As noted earlier, pilot testing with untruncated versions of the stimuli produced a more robust McGurk effect.) Thus, it is plausible that factors that reduced the magnitude of the McGurk effect (including the truncation procedure) also caused a reduction in the lexical effect, although the available data do not permit a direct test of this possibility. Note, however, that when comparing across the three

stimulus sets in Experiment 3, there was not a correlation between the magnitude of the McGurk effect and that of the lexical effect; the *PACE* set exhibited a weak McGurk effect but exhibited the largest lexical effect of the three sets.

Overall, it is not clear, on the basis of the present findings, the extent to which any of these factors are responsible for the reduction of the lexical effect in Experiment 3. Future research could be designed to examine the effects of task monotony, acoustic stimulus quality, and visual truncation directly in the audiovisual phoneme-identification task.

A final issue to note regards the relative speed of auditory and visual responses. In Experiment 2, visual responses were faster than auditory responses, which could have been a consequence either of the difference in relative goodness of the percepts or of the earlier availability of visual than auditory information. In contrast, there was no reliable difference in reaction times between visual and auditory responses in Experiment 3 (on average, auditory responses were 11 ms faster). There was a tendency for faster visual responses (by 83 ms) for the *MESS* set, which had a strong McGurk effect (see Figure 3), but not for the *BENCH* and *PACE* sets, which had much weaker McGurk effects. These outcomes suggest that the relative speed of visual and auditory responses may be more a function of the relative goodness of the percepts than of the relative availability of the information from each modality.

In summary, Experiment 3 demonstrated that lexical knowledge influences identification of audiovisually discrepant stimuli and that these lexical influences emerge over time, even when the stimuli that differ in lexical status are closely matched on stimulus properties. Thus, the experiment provides a replication of the main findings of Experiments 1 and 2.

General Discussion

The goal of this study was to test whether lexical knowledge affects categorization of audiovisually discrepant speech stimuli. Overall, the three experiments reported here provided a highly consistent picture: The McGurk effect (i.e., visually influenced categorization) was more prevalent when it resulted in identification of a phoneme that formed a word than one that formed a nonword. This finding is analogous to the Ganong effect, in which an acoustic segment that is ambiguous between two phonemes is identified more often as a phoneme that forms a word (Ganong, 1980). The findings are at odds with those of Sams et al. (1998), who failed to observe a higher incidence of the McGurk effect when the effect formed a word (and the auditory signal was a nonword) than when it did not (and the auditory signal was a word). However, as noted in the introduction, Sams et al. did not control for syllable position of the audiovisual mismatch (initial, medial) or vowel context in their stimuli, both of which could have affected the magnitude of the McGurk effect, and their lexical conditions differed extensively along these dimensions. In contrast, in all of the present experiments, the audiovisual mismatch was always in the initial position, and vowel context was similar across the lexical conditions in Experiments 1 and 2 and was held constant across conditions in Experiment 3. Under these conditions, a robust lexical effect was obtained. Given these findings, the most likely explanation for Sams et al.'s failure to find a lexical effect is that the lexical influence was obscured by influences of other, nonlexical, stimulus variables.

The results reported here clearly demonstrate that lexical influences on phonetic categorization are not restricted to the auditory domain; lexical knowledge affects phoneme identification for phonetically ambiguous stimuli regardless of whether the ambiguity is acoustically or audiovisually based. Moreover, lexical effects on initial phonemes show similar temporal constraints in the auditory and audiovisual domains; in both cases, lexical effects are weak or nonexistent when a response is made quickly and are stronger for slower responses (e.g., Fox, 1984; Miller & Dexter, 1988).

A question that arises from the present findings is how a model might account for lexical influences on the McGurk effect, in conjunction with the well-established findings of auditory lexical effects. Two related issues are addressed here: whether current models of lexical effects in auditory speech perception can be extended to account for the present results and when audiovisual integration occurs relative to other aspects of processing, including lexical access.

As noted in the introduction, speech perception models such as TRACE (McClelland & Elman, 1986) and Merge (Norris et al., 2000) incorporate a process by which activation from a prelexical phonetic stage of processing flows bottom up to a lexical stage, and activation from lexical representations then affects the activation of phonetic units involved in phonetic decisions. These models diverge, however, in their assumptions about whether this process is interactive, with bidirectional flow of activation between high-level and low-level processing stages (McClelland & Elman, 1986; see also Samuel, 1981), or strictly autonomous and bottom up, with no influence of higher order processing on low-level processes (e.g., Norris et al., 2000; see also Cutler et al., 1987). In the interactive model TRACE, activation flows bidirectionally between units at feature, phoneme, and lexical levels; lexical effects on phoneme identification (which occurs at the phoneme stage) emerge when activation from a word provides top-down support for one of its constituent phonemes at the phoneme level, particularly when there are competing partially activated phonemes. In contrast, the autonomous model Merge permits only a unidirectional flow of activation, so that lexical representations do not affect processing in the prelexical representations that feed up to them. In Merge, the bottom-up speech signal is processed at a prelexical stage, which then activates representations at a higher lexical stage, similar to TRACE. However, because there can be no feedback from the lexical stage to the prelexical stage, the model includes a separate phoneme decision stage, to which both the prelexical and lexical stages provide input; thus, lexical activation can affect phoneme decisions.

Where would a visual speech input fit into these architectures? As discussed in the introduction, there is a debate over the extent to which phonetic processing occurs before audiovisual integration or subsequent to it. However, as noted earlier, although some evidence suggests that certain aspects of phonetic processing might occur before audiovisual integration (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994), other findings indicate that audiovisual integration may precede or be contemporaneous with other aspects of phonetic processing (Fowler, et al., 2000; Green & Norrix, 2001). This latter set of findings suggests that in the context of the TRACE and Merge models, audiovisual integration would occur at the featural or phonetic stages in TRACE and at the prelexical stage in Merge. In TRACE, this could be accommodated by the addition of visual features at the feature stage, possibly (but not necessarily) with an additional audiovisual

integration stage before the phoneme stage. In Merge (which does not specify separate feature and phoneme stages), an audiovisual integration mechanism would be incorporated into the prelexical stage. In either model, therefore, audiovisual integration would occur at a stage before activation of lexical representations.

In the audiovisual instantiations of both TRACE and Merge, the results of the present study can be explained as follows, using the stimuli from the *PACE* series in Experiment 3 as an example. When an auditory /pe/ and a visual /te/ are presented together, so that there is auditory information for a voiceless stop consonant (/p/, /t/, or /k/) and conflicting auditory and visual information for bilabial (/b/, /p/, or /m/) and alveolar (/d/, /t/, or /n/) places of articulation, audiovisual integration results in partial activation for both /p/ and /t/. When the final consonant in the auditory signal is /s/, activation of /s/, in conjunction with the partial activation for /p/, results in partial activation for the word *PACE* at the lexical level; although there is also partial activation of /t/, there is no word *TACE* to be activated at the lexical level. The activation from *PACE* feeds back to the phonetic-decision stage (the phoneme level in TRACE and the separate phonetic-decision stage in Merge), where /p/ may be selected over /t/ as a result of the top-down lexical support. Additionally, in TRACE, a feedback loop could emerge in which top-down lexical activation feeds back to the phoneme level and then to the feature level, which would include auditory- and visual-specific features; the activation feeding up from auditory and visual features to phoneme units would then be accordingly modulated. In either model, the lexical effect would be more pronounced when the phonetic decision is made slowly, as the delay in making a decision enables lexical activation to build up and exert a greater influence.

As the preceding account implies, there is an important distinction between lexical influences on the McGurk effect (a behavioral measure) and lexical influences on the process of audiovisual integration itself. Note that in both of the models just described, lexical effects emerge as a result of lexical influences on the selection of a phoneme from multiple phonetic candidates (which were activated on the basis of audiovisual integration) at a stage subsequent to audiovisual integration. In the audiovisual Merge, lexical effects can only emerge during the phoneme selection process; the process of audiovisual integration is encapsulated (along with other prelexical processes) from higher order influences. In the audiovisual TRACE model, lexical effects would result (as in Merge) from lexical influences on phoneme selection (due to top-down influences on phoneme units) but could additionally result from top-down lexical influences on the audiovisual integration process itself (by virtue of top-down activation feeding back to auditory-specific and visual-specific feature units, as noted earlier). However, the emergence of lexical effects in TRACE would not require this latter mechanism. Thus, although a finding of lexical influences on the McGurk effect might be regarded as evidence for top-down effects on the audiovisual integration mechanism (cf. Sams et al., 1998), this is not a necessary conclusion.

Clearly, the present results do not provide a basis for distinguishing the TRACE- and Merge-based explanations for lexical influences on the McGurk effect (and, in fact, it has proven difficult to distinguish TRACE and Merge empirically in the auditory domain; cf. Norris et al., 2000, and following commentaries), but the current study represents a starting point for future empirical tests to distinguish the models. As noted above, the audiovisual TRACE and Merge models differ on whether the process of audiovisual integration, as well as the phonetic processing stages that precede audiovisual integration (e.g., auditory- and visual-specific processing stages), can be affected by higher level lexical processing. Thus, more sophisticated empirical designs involving audiovisual speech might offer a way to tap into early stages of processing, either before audiovisual integration (i.e., involving auditory- and visual-specific processing stages) or at the point of integration, and test for lexical influences at those stages (cf. Brancazio & Fowler, 2000; Fowler & Brancazio, 2000).

Despite their differences, the audiovisual versions of TRACE and Merge outlined here share an assumption that audiovisual integration precedes lexical access and the locus of lexical effects. Stated slightly differently, the process of lexical access (which, in turn, gives rise to lexical influences on phoneme identification) in these models is *modality independent*. That is, there is a single, modality-independent pathway to the lexicon (by virtue of the fact that activation of lexical representations follows audiovisual integration), and lexical representations themselves are therefore neither auditory nor visual.[8] This assumption clearly is consistent with early-integration models of audiovisual integration. However, although on first consideration, the modality-independence assumption might appear to be inconsistent with late-integration models, this is not necessarily the case: It is possible for there to be extensive phonetic processing before audiovisual integration (e.g., late integration), and nonetheless for audiovisual integration to occur before lexical activation.

An alternative to the modality-independent view is that lexical activation occurs before audiovisual integration rather than vice versa. In this vein, a number of models could be proposed that embody a *modality-specific* lexical architecture. For example, one possibility is that there are independent auditory and visual streams of processing that each affect phoneme identification and that only the auditory stream, and not the visual stream, is affected by top-down lexical knowledge.[9] (Because anticipatory visual information is available earlier than auditory information, the visual route will generally be faster, meaning that visual responses should predominate in fast responses and that lexical effects, arising from the auditory-lexical route, should be larger in slow responses; both outcomes were observed in the present experiments.) However, the results of Experiments 1 and 2 pose a problem for such a model, because there were lexical effects that were independent of the lexical status of the auditory signal. Specifically, visual responses were more common with stimuli such as a*BESK*-v*DESK* than with stimuli such as a*BEDGE*-v*DEDGE*, despite the common nonword status of the auditory signals. Such findings indicate that if there are separate auditory and visual processing routes, then both must be influenced by lexical factors.

---

[8] Note that the suggestion here is that the mechanisms of lexical access are modality independent, which is not meant to imply that the modality of the input has zero consequence for lexical access. To the contrary, Auer and Bernstein (1997) and Iverson et al. (1998) have demonstrated that the differences between the confusability of phonemes in auditory speech and in visual (and audiovisual) speech have significant implications for word recognition in each domain. However, these modality-dependent effects on word recognition may reflect differences in the information provided to the lexical access system rather than in the organization of the lexical access system itself.

[9] I thank a reviewer for the suggestion of this possibility.

Accordingly, an alternative modality-specific model could include separate auditory and visual processing streams that have independent routes to the lexicon, with lexical feedback occurring on a modality-specific basis. (Following this logic, there could be separate auditory and visual lexicons as well.) According to such a model, when the auditory signal forms a word, the word's representation is activated and provides top-down support for the initial phoneme of the word, and likewise when the visual signal forms a word, that word's representation provides top-down support for that initial phoneme. When the auditory and visual initial phonemes conflict, whichever phoneme has greater top-down support is more likely to be selected. By this account, the effect of visual lexicality (e.g., more visual responses for stimuli such as a*BESK*-v*DESK* than for stimuli such as a*BEDGE*-v*DEDGE*) observed in Experiments 1 and 2 arose from recognition of the visually presented word from speech reading, independent of the auditory signal. However, a problem for this account arises from the relative impoverishment of the visual signal. Without auditory information about manner and voicing, the initial phoneme in a visual presentation of a word such as *DESK* could plausibly be /d/, /t/, or /n/, because these phonemes are all members of a viseme grouping (e.g., Jackson, 1983), and the overall word is consistent with numerous phonetic interpretations (see Iverson et al., 1998). In contrast, words presented in the auditory signal (if analyzed independently of the conflicting visual signal) are relatively unambiguous. Thus, to the extent that visual lexical effects occurred independent of auditory processing, they would be considerably weaker than auditory lexical effects. Although direct comparison of the auditory and visual lexicality effects in Experiments 1 and 2 is hampered by the fact that the physical properties of the stimuli were not fully controlled across lexical conditions, note that the effect of visual lexicality was numerically larger than the effect of auditory lexicality. There was no indication that the lexical status of the auditory signal had a greater impact on phoneme identification than that of the visual stimulus.

Thus, although the present results do not rule out the possibility of independent auditory and visual routes to the lexicon, they are more consistent with the view that audiovisual integration precedes lexical access. Further studies may provide a better basis for elucidating the exact mechanisms of lexical access for audiovisual speech.

Although the preceding discussion focuses on possible extensions of existing models, there is also a current model of speech perception (which is not explicitly modality specific or modality independent with respect to lexical access) that might also account for the present findings, namely, the fuzzy logical model of perception (or FLMP; Massaro, 1989; Massaro & Oden, 1980, 1995). In the FLMP, phonetic features, which are either auditory or visual, are independently evaluated and assigned values based on the degree to which the input matches stored prototype representations, and these feature values are integrated, leading to perceptual selection of the best-matched prototype (Massaro, 1987, 1998; Massaro & Cohen, 1983). Because auditory and visual features are evaluated independently, the FLMP is considered a late-integration model (Schwartz et al., 1998). In the FLMP, lexical context is treated as another cue and is integrated in the same manner as auditory and visual cues. However, descriptions of the FLMP (e.g., Massaro, 1987, 1998) have not specified exactly when evaluation of lexical context occurs in relation to audiovisual integration. If there is a common integration stage for auditory, visual, and lexical cues, then the FLMP could explain the current findings in the following manner. In the case of an auditory /pek/ syllable with a visual /te/, auditory and visual features will collectively partially match both the /p/ and /t/ prototypes; however, the /ek/ lexical context will favor the /t/ prototype because it forms the word *TAKE*, leading to more /t/ responses than if the context was /es/.[10]

Finally, the issue of whether there are modality-independent or modality-specific routes to the lexicon is relevant for recent exemplar-based models of the lexicon (e.g., Goldinger, 1998). On the basis of findings suggesting that listeners store detailed information about the surface properties (e.g., voice characteristics) of individual utterances of words (e.g., Palmeri, Goldinger, & Pisoni, 1993), Goldinger (1998) has proposed that the lexicon is composed of detailed representations of individual utterances rather than abstract lexical entries. In the light of the growing evidence for the contributions of visual speech in lexical access (including the present findings as well as those of Auer & Bernstein, 1997, and Iverson et al., 1998), it will be important for such accounts to specify the role of visual information in an exemplar lexicon. For example, stored exemplars for words could separately encode acoustic and optical details, in keeping with a modality-specific framework, or could instead encode details in a metric that is modality independent, such as articulatory properties (cf. Goldinger, 1998; Sheffert & Fowler, 1995).

In summary, the experiments reported here demonstrate lexical influences on the McGurk effect that parallel previously reported influences on perceptually ambiguous acoustic stimuli. These findings require models of lexical influences in speech perception to incorporate visual input into their architecture. Most current theories (including TRACE and Merge) can be extended in ways that enable them to account for the present findings, if they do not already do so (e.g., FLMP). However, the audiovisual implementations of these models may be challenged by more subtle manipulations of audiovisual speech than the ones used here. Moreover, the present findings raise the question of whether lexical processes are modality independent, occurring postintegration, or instead involve modality-specific representations. Although modality-independent models arguably better accommodate the results, future studies modeled on the present one can be designed to address this issue more directly. Thus, in sum, it is hoped that the present research can provide an impetus for further investigations into the nature of lexical influences in audiovisual speech perception and for the refinement of models of audiovisual lexical access.

[10] This explanation of the results does not require a resolution of whether lexical access is modality-specific or modality-independent in the FLMP. In its explanation of lexical effects, the FLMP treats the context (e.g., the syllable rime) as independent from the initial consonant, which is where the audiovisual discrepancy occurs in the present experiments. In the example used here, the context could be evaluated either on the basis of the auditory signal (/es/) or audiovisually (auditory /es/ with visual /e/), and the outcome would be the same in either case (/es/).

References

Auer, E. T., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *Journal of the Acoustical Society of America, 102,* 3704–3710.

Bernstein, L. E. (1989). Independent or dependent feature evaluation: A

question of stimulus characteristics. *Behavioral and Brain Sciences, 12,* 756–757.

Brancazio, L., & Fowler, C. A. (2000). Merging auditory and visual phonetic information: A critical test for feedback? *Behavioral and Brain Sciences, 23,* 327–328.

Burton, M. W., Baum, S. R., & Blumstein, S. E. (1989). Lexical effects on the phonetic categorization of speech: The role of acoustic structure. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 567–575.

Burton, M. W., & Blumstein, S. E. (1995). Lexical effects on phonetic categorization: The role of stimulus naturalness and stimulus quality. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1230–1235.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 25,* 257–271.

Connine, C. M., & Clifton, C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 291–299.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology, 19,* 141–177.

Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics, 51,* 355–362.

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics, 32,* 562–570.

Eimas, P. D., Hornstein, S. B. M., & Payton, P. (1990). Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language, 29,* 160–180.

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

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. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99,* 1730–1741.

Fowler, C. A., & Brancazio, L. (2000). Feedback in audiovisual speech perception. In A. Cutler, J. M. McQueen, & R. Zondervan (Eds.), *Proceedings of the Workshop on Spoken Word Access Processes* (pp. 87–90). Nijmegen, the Netherlands: Max Planck Institute for Psycholinguistics.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 877–888.

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 526–540.

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

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 3–26). Hove, England: Psychology Press.

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics, 45,* 34–42.

Green, K. P., & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 278–288.

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 & Psychophysics, 50,* 524–536.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics, 38,* 269–276.

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 and Hearing Research, 40,* 646–665.

Green, K. P., & Norrix, L. W. (2001). Perception of /r/ and /l/ in a stop cluster: Evidence of cross-modal context effects. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 166–177.

Iverson, P., Bernstein, L. E., & Auer, E. T. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication, 26,* 45–63.

Jackson, P. L. (1983). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90,* 99–115.

Jordan, T. R., & Sergeant, P. C. (1998). Effects of facial image size on visual and audio–visual speech recognition. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 155–176). Hove, England: Psychology Press.

Kučera, H., & Francis, W. N. (1967). *A computational analysis of present-day American English.* Providence, RI: Brown University Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1–36.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24,* 253–257.

Manuel, S. Y., Repp, B. H., Liberman, A. M., & Studdert-Kennedy, M. (1983, November). *Exploring the "McGurk effect."* Paper presented at the 24th Annual Meeting of the Psychonomic Society, San Diego, CA.

Marslen-Wilson, W. D. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148–172). Cambridge, MA: MIT Press.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for scientific inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of perception. *Cognitive Psychology, 21,* 398–421.

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

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

Massaro, D. W., & Ferguson, E. L. (1993). Cognitive style and perception: The relationship between category width and speech perception, categorization, and discrimination. *American Journal of Psychology, 106,* 25–49.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (pp. 129–165). New York: Academic Press.

Massaro, D. W., & Oden, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1053–1064.

Massaro, D. W., & Stork, D. G. (1998, May–June). Speech recognition and sensory integration. *American Scientist, 86,* 236–244.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McGurk, H., & MacDonald, J. (1976, December 4). Hearing lips and seeing voices. *Nature, 264,* 746–748.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 433–443.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 369–378.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46,* 505–512.

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

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 873–889.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23,* 299–370.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309–328.

Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1037–1052.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 699–725.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 97–114). Hillsdale, NJ: Erlbaum.

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics, 30,* 309–314.

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

Rubin, P., Turvey, M. T., & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics, 19,* 394–398.

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

Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *Journal of the Acoustical Society of America, 95,* 3658–3661.

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication, 26,* 75–87.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110,* 474–494.

Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio–visual fusion in speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 85–108). Hove, England: Psychology Press.

Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language, 34,* 665–685.

Sumby, W. H., & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Summerfield, A. Q. (1979). Use of visual information for speech perception. *Phonetica, 36,* 314–331.

Summerfield, A. Q. (1987). Preliminaries to a comprehensive account of audio–visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 3–52). Hillsdale, NJ: Erlbaum.

Wurm, L. H., & Samuel, A. G. (1997). Lexical inhibition and attentional allocation during speech perception: Evidence from phoneme monitoring. *Journal of Memory and Language, 36,* 165–187.

Appendix

Stimuli Used in Experiments 1 and 2

Word–word pairs		Word–nonword pairs		Nonword–word pairs		Nonword–nonword pairs	
Bilabial	Nonbilabial	Bilabial	Nonbilabial	Bilabial	Nonbilabial	Bilabial	Nonbilabial
BELT (29)	DEALT (22)	BAND (53)	DAND	BEAF	DEAF (12)	BARVE	DARVE
BIN (9)	DIN (1)	BEG (11)	DEG	BESK	DESK (65)	BEDGE	DEDGE
BUCK (20)	DUCK (9)	BENCH (35)	DENCH	BISH	DISH (16)	BELF	DELF
BUG (4)	DUG (15)	BUNT (3)	DUNT	BODGE	DODGE (11)	BILN	DILN
BAN (7)	VAN (32)	BARGE (7)	VARGE	BAGUE	VAGUE (25)	BAIF	VAIF
BAIL (12)	VEIL (15)	BATCH (5)	VATCH	BALVE	VALVE (3)	BALP	VALP
BENT (34)	VENT (10)	BATHE (4)	VATHE	BAHS	VASE (4)	BANCH	VANCH
BET (20)	VET (1)	BIRD (31)	VIRD	BERB	VERB (4)	BUSP	VUSP
MAP (13)	NAP (4)	MASK (9)	NASK	MAB	NAB (1)	MAFT	NAFT
MELT (4)	KNELT (8)	MIST (14)	NIST	MERVE	NERVE (12)	MEMP	NEMP
MILL (11)	NIL (1)	MUNCH (1)	NUNCH	MIP	NIP (3)	MERB	NERB
MOCK (8)	KNOCK (15)	MYTH (35)	NITH	MURSE	NURSE (17)	MILCH	NILCH
PACK (25)	TACK (4)	PAGE (66)	TAGE	PAB	TAB (1)	PAHS	TAHS
PAR (13)	TAR (12)	PINCH (8)	TINCH	PEXT	TEXT (60)	PAV	TAV
PIP (23)	TIP (22)	POND (25)	TOND	PUNG	TONGUE (35)	PILTH	TILTH
PUN (1)	TON (13)	PUP (2)	TUP	PUSK	TUSK (0)	PULF	TULF

Note. Kučera–Francis (1967) frequencies of word stimuli are in parentheses.

Received December 4, 2000
Revision received June 2, 2003
Accepted October 27, 2003 ■