Language Learning 46:1, March 1996, pp. 3–73

Bimodal Speech Perception by Native and Nonnative Speakers of English: Factors Influencing the McGurk Effect

Debra M. Hardison
Indiana University

Two experiments explored factors affecting the influence of visual (lip-read) information on auditory speech perception, the "McGurk effect", in 120 advanced ESL learners of 4 L1s (Japanese, Korean, Spanish, and Malay) and 50 native speakers (NSs) of English. The audio and video speech signals of a female English speaker producing CV syllables with /p,f,w,r,t,k/ and /ɑ/ were combined on videotape. For nonnative speakers (NNSs), identification accuracy of /f/ and /r/ increased with matched visual cues. Visual nonlabials /t,k/ significantly influenced perception

I gratefully acknowledge the assistance of Doreen Myovich, whose voice and lip movements were recorded; David Derkacy of *Media Images*, Bloomington, Indiana for technical expertise in recording and tape editing; Mark Garton and Dan Reed for statistical consultation; and Robin Edmundson for data collection. This work has also benefited from discussions with Kathleen Bardovi-Harlig, Beverly S. Hartford, Stuart Davis, and David B. Pisoni. In addition, I thank Faridah Pawan and fellow linguists Kurumi Hiki, Jin-Seong Lee, and Beverly Ruiz for their insights on their languages of Malay, Japanese, Korean, and Venezuelan Spanish, respectively. An earlier version of this paper was presented at the 22nd Annual University of Wisconsin-Milwaukee Linguistics Symposium (Second Language Acquisition & Pedagogy), October 8, 1993.

Correspondence concerning this article should be addressed to Debra M. Hardison, Indiana University, Department of Linguistics, Memorial Hall, Bloomington, Indiana 47405 USA. Internet: hardison@indiana.edu

of auditory labials. For NSs, significant effects were noted only in noise. When stimuli involved only /p,t,k/, NSs and NNSs reported significant visual effects. Results demonstrate the influence of first language (L1) and linguistic experience on the relative information value of the cues, assumption of perceptual unity, and thus audiovisual integration. Implications for bimodal speech perception and second language instruction are discussed.

Most previous accounts of cross-language differences in speech perception have focused on auditory information and models of L2 (second language) phonetic perception. This paper reports the findings of two experiments that explored the factors affecting the integration of auditory and visual (lip-read) information in the perception of six American English (AE) consonants (/p,f,w,r,t,k/) in CV syllables (before the vowel /ɑ/) by native versus nonnative listeners of English. The nonnatives were 120 advanced learners of English as a second language (ESL) of four L1s (Japanese, Korean, Spanish, and Malay[1]).

Japanese speakers' difficulty in identifying and producing AE /r/ and /l/ has been the focus of numerous studies since a report by Goto (1971), in which he included his own problems with these sounds in conversation. He concluded that in the auditory presentation of sounds, "there was the disadvantage of not being able to read the lips of the speaker" (p. 321). Thus, of particular interest in the current investigation was L2 learners' use of visual information from lip movements to increase identification accuracy of speech sounds in an audiovisual context. These findings, together with the learners' assessments of the relative information value of the cues from each modality and the relationship between L1 and perception, have implications for a model of bimodal speech perception and audiovisual training in the perception of nonnative speech sounds. A better understanding of the acoustic-phonetic and visual inputs lays the groundwork for further studies on the process of spoken word recognition in face-to-face communication.

In addition, the native-nonnative contrast provides a new perspective on the conditions influencing audiovisual integration by highlighting not only the role of *stimulus compellingness* (degree of discordance between auditory and visual cues) in integration but also its determination on the basis of the observers' assessment of similarity between auditory and visual cues and relative weighting of the cues from each modality. These factors, in turn, are influenced by the native or nonnative observers' perceptual categories (including phonetic and visual information) to which speech input is matched for identification.

I limited the selection of consonants for this initial experiment to six, because of the length of the experimental session and possible fatigue for the participants. Of the six, /p,f,w/ exhibit visually salient labial involvement, compared to /r,t,k/, in the context of the vowel /ɑ/.[2] This selection of consonants afforded an experimental contrast in the visual discernability of consonant articulations; thus, manner and place distinctions rather than voicing were factors in the selection of consonants, as was their phonemic status in each of the L1s in this study. A previous study with native speakers of English (MacDonald & McGurk, 1978) had included the stops /p,t,k/; unlike the voiced stops, they are present phonemically in all the L1s in this study. In Korean, there are lax, aspirated, and reinforced obstruents (all voiceless). Voiced lax stops arise when a voiceless lax stop occurs in intersonorant position (Cho, 1987; Silva, 1989). Recent studies have demonstrated that their occurrence is also prosodically conditioned (Cho, 1987; Silva, 1989, 1992).[3]

I selected the three remaining consonants for the visual value of their articulations, and for problems they pose for learners. I chose the fricative /f/ for its visual quality and frequent confusion with /p/ by Korean learners (Jung, 1962; Robson, 1982). For Japanese, Vance (1987) reported a weak bilabial fricative occurring in careful speech but described it as quite different than AE [f]. Production of phones of the Spanish /f/ are less fricative than the AE counterpart and are associated with a generally lenis pronunciation of bilabial obstruents (Resnick, 1975; Stockwell &

Bowen, 1965). For speakers of Malay, /f/ has been introduced into the sound system through English loanwords (Lapoliwa, 1981).

I selected the liquid /r/ because of the absence of its articulation as an alveolar approximant in Japanese and Korean, and the problems it poses for speakers of these L1s learning English. Japanese has a voiced dental or alveolar flap occurring initially and intervocalically (Miyawaki et al., 1975; Vance, 1987), similar acoustically and articulatorily to the AE voiced dental-alveolar flap (Price, 1981), as found in medial position in the pronunciation of words such as *butter*. Korean has one phoneme that is realized as a voiced clear /l/ in word-final position and a voiced apico-alveolar flap in intervocalic position (Kim-Renaud, 1974; Koo, 1970). Since Goto's (1971) report, the AE sounds /r/ and /l/ have been the subject of many studies on L2 phoneme acquisition (for Korean, e.g., Koo, 1970, 1972; for Japanese, e.g., MacKain, Best, & Strange, 1981; Miyawaki et al., 1975; Mochizuki, 1981; Sheldon & Strange, 1982; Yamada & Tohkura, 1992) and on auditory training (Borden, Gerber, & Milsark, 1983; Gillette, 1980; Lively, Logan, & Pisoni, 1993; Lively, Pisoni, R. Yamada, Tohkura, & T. Yamada, 1994; Logan, Lively, & Pisoni, 1991). Although /w/ exists phonemically in the L1s in the present study, I selected it because of the variation in its acoustic and articulatory labial quality across languages and its potential visual similarity to /r/ from the learners' perspective. In addition, previous identification tasks of auditory stimuli on an /r/-/l/ continuum revealed that responses from the Japanese also included a /w/ category (Mochizuki, 1981; Yamada & Tohkura, 1992). Table 1 summarizes the main points concerning the realization of /p,t,k,f,w,r/ where they exist in the four L1s in this study.

To investigate audiovisual integration, I based the experimental design on earlier studies (e.g., MacDonald & McGurk, 1978; McGurk & MacDonald, 1976), which demonstrated the influence of visual information (lip movements) on auditory speech perception (the "McGurk effect") by native listeners of English. These studies used dubbed videotapes to create conditions of audiovisual (A-V) discrepancy (i.e., where the visual and auditory

cues do not match, e.g., visual /ka/ combined with auditory /pa/). Other studies have used such experimentally induced intermodality discordance to investigate the processing of visual and auditory cues in speech perception (e.g., Dekle, Fowler, & Funnell, 1992; Dodd, 1977; Green & Kuhl, 1989, 1991; Green, Kuhl, & Meltzoff, 1988; Green, Kuhl, Meltzoff, & Stevens, 1991; Kuhl, Green, & Meltzoff, 1988; Sekiyama & Tohkura, 1991, 1993) and in speechreading (Easton & Basala, 1982).

In an extensive review of the literature on intersensory bias, Welch and Warren (1980) noted that the merit of investigations using such induced misperception lies in this technique's contribution to the understanding of perception in general. They proposed a model to account for intersensory discrepancy results. Supposing that one immediate response to a discrepancy between two sensory modalities is for one to bias the other, intersensory bias is the result of the perceptual system's attempt to maintain a single perceptual experience. The strength of an observer's assumption of a unitary perceptual event is presumed to be a function of the redundant properties present in the stimulus situation and the relative weighting assigned to them. In the case of speech, the properties are those of the visual and auditory cues. Welch and Warren further noted that modality bias appears to be an actual perceptual effect rather than being attributable to postperceptual decision processes or learning strategies.

The use of audiovisual discrepancy conditions reduces or eliminates redundancy between the modalities. Welch and Warren (1980) referred to this approach as the experimental means of *tagging* the available information from the modalities for the purposes of evaluating their relative contributions to the perceptual outcome.

This technique provides a valuable research tool in speech perception. In face-to-face communication, input is matched to the observers' perceptual categories (phonetic and visual) for identification. The status of the observers' perceptual categories (native or nonnative) influences the identification process which, in turn, affects the observers' assessment of the similarity be-

Table 1
L1 Consonant Descriptions

Phoneme	Language: Japanese	Korean	Spanish	Malay
/p,t,,k/	Unaspirated stops: /p/ has a weak lip activity; /t/ is lamino-alveolar.	Three phonemes at each place of articulation: lax /p,t,k/, aspirated /p^h,t^h,k^h/, and reinforced /p',t',k'/. /t/ is dental.	Unaspirated stops: /p/ is less forcefully articulated than in AE; /t/ is dental-alveolar.	/t/ is dental or dental-alveolar.
/f/	A weak bilabial fricative in careful speech; a weak labio-dental fricative may occur in less careful speech. If no audible friction, [h] may be produced.	No labial fricative.	Phones of /f/ are articulated labio-dentally and are less fricative than AE. Bilabial phones also occur associated with a generally lenis pronunciation of obstruents.	/f/ exists in loanwords from English and Dutch. Produced with weak labiodental contact; less fricative than AE.[a]
/w/	Same weak lip activity as vowel /ɯ/. Can occur only before /ɑ, ɑɪ/.	Produced with slight rounding of lips.	Labial quality less evident than in AE.	Voiced labiovelar glide.

/r/ & (/l/)	A voiced dental or alveolar flap occurs utterance-initially or intervocalically; may have some lateral articulation but no allophonic distribution.	A voiced clear /l/ occurs in syllable-final position. A voiced apico-alveolar flap occurs intervocalically. No variant in initial position.	/r/ occurs as an alveolar flap or, in initial position, as an apical trill. /l/ is a voiced alveolar lateral.	/r/ is a voiced apico-alveolar trill. /l/ is a voiced alveolar lateral.

AE=American English.

[a]In Malay, /f/ is usually realized phonetically as [p] in most loanwords, but may occur with [p] in free variation (e.g., tarip *tariff* [tarip, tarif], and poto, foto *photo* [poto, foto]. Some lexical items used mainly by educated people familiar with the source language are more commonly realized with [f] (e.g., faktor *factor* [faktor] (transcriptions from Lapoliwa, 1981).

References for each language:

Japanese: Bloch, 1950; Miyawaki et al., 1975; Price, 1981; Vance, 1987
Korean: Cho, 1987; Kim-Renaud, 1974; Koo, 1970, 1972; Martin, 1951; Silva, 1989
Spanish: Resnick, 1975; Stockwell & Bowen, 1965
Malay: Lapoliwa, 1981; F. Pawan, personal communication, May, 1995

tween the cues from each modality and thus the relative weighting given to each cue and its contribution to the eventual perceptual outcome.

The current study directed this methodology toward an investigation of the factors affecting the identification and integration of visual and auditory cues by ESL learners (in contrast with native English speakers) and the contribution of lip movements to increased accuracy in the identification of nonnative speech sounds.

Previous Work

Three principal areas of the literature are relevant to this investigation: the findings of speechreading (lipreading) studies on the visual discernability of the articulatory characteristics of different English consonants; previous studies of visual effects on native listeners' perception of speech; and experiments on nonnatives' auditory perception of English sounds.

Miller and Nicely (1955) suggested, but did not test, the use of visual cues to reduce or eliminate place of articulation confusions for consonant perception following experiments which revealed that perception of place of articulation was more affected by noise than the features of voicing and nasality were.

Speechreading / Visual Studies

Several subsequent studies investigated the visual information available from a speaker's articulatory movements. Woodward and Barber (1960) identified only four homophenous groups, that is, groups consisting of articulations not visually distinctive: bilabials, labiodentals, /w,r/, and nonlabials.[4] They found consonants with labial involvement were the most visually distinguishable.

Fisher (1968) identified five *visemes* (visually contrastive speech sounds), similar to the above groups, but with the velars /k,g/ as a separate group. In contrast, Berger (1972, 1973)

suggested 12 initial and 8 final consonant visemes, from an experiment in which consonants were placed in front of and following three different vowels and presented in a face-to-face test. Speechreading scores for consonants preceding /ɑ/ were better than for those preceding /i/ or /u/.

Binnie, Montgomery, and Jackson (1974) studied perceptual confusions of 16 consonants in CV syllables presented under audiovisual, audio-only, and visual-only conditions to normal-hearing adults, in order to evaluate overall speech intelligibility at various signal-to-noise (S/N) ratios as well as the contribution of visual information to perception.[5] From visual information only, the participants could distinguish five distinct homophenous groups: bilabials /p,b,m/, labiodentals /f,v/, interdentals /θ,ð/, palato-alveolars /ʃ,ʒ/, and the alveolar and velar sounds /t,d,n,s,z,k,g/. At all S/N ratios, visual information contributed to reduce errors, especially in the identification of stops and place of articulation. Binnie et al. suggested that acoustic cues to place information were obscured by the broad-band noise, as were consonants, such as fricatives, with wide distribution of energy in the spectrum.

Summerfield (1979) asked listeners with normal hearing to write down isolated sentences spoken in British English and presented at a −12 dB S/N ratio, the noise in this case being a segment of continuous prose. Summerfield scored the number of main words correctly written as a percentage, and compared results with those from audiovisual conditions. In two of these conditions, one showing the speaker's face and the other the lips only, the accuracy of responses significantly improved—by 43% and 31%, respectively. Listeners commented that the absence of teeth and tongue movement in the latter condition led to the poorer performance.

Visual Effects in Discrepant Conditions

McGurk and MacDonald (1976) further investigated the role of visual information in speech perception by NSs of British

English, using an audiovisual discrepancy technique. They limited the stimulus materials to vocalizations of the English syllables /bɑ/, /pɑ/, /gɑ/, /kɑ/, dubbed onto filmed lip movements of these syllables so as to create combinations where the audiovisual stimuli were matched and mismatched, for example, /kɑ/ lips-/pɑ/ voice. Participants were to watch the film and repeat what they had heard. In an audio-alone condition, identification accuracy averaged 99% for the four syllable combinations. Responses in the A-V condition fell into four principal categories: accurate responses that matched the auditory cue; those that matched the visual cue; *fused* responses (i.e., those representing new elements not actually presented either visually or auditorily); and *combination* responses (i.e., those composed of both the visual and auditory stimuli). For combinations of a visual velar and an auditory bilabial consonant (e.g., /kɑ/-/pɑ/) the majority of responses (81%) were /tɑ/ (fused). On the other hand, visual bilabial and auditory velar pairs (e.g., /pɑ/-/kɑ/) produced primarily combination responses (44%) (e.g., /pɑkɑ/) as well as those matching the visual cue /pɑ/ (37%). This pattern was also typical of the voiced stimulus pairs.

The McGurk effect also occurs with sentences (McGurk, 1981). When McGurk dubbed the utterance *My bab pope me poo brive* onto lip movements for *My gag koke me koo grive*, the reported percept was *My dad taught me to drive*. Notice, however, that the sounds which differ between auditory and visual modalities are the bilabial and velar stops (e.g., /b/ in *bab* matched with /g/ in *gag*). These findings are consistent with the previous study using only the stops.

Dodd (1977) asked normal-hearing adolescents to repeat CVC words presented to them in white noise. Results indicated that under substantial white noise, where the auditory and visual inputs differed, more information from the visual modality was reported. Fewer errors were made for bilabial, labiodental, and interdental consonants than for other more posteriorly articulated consonants.

Subsequent studies, however, have revealed that the magni-

tude of the McGurk effect varies widely. In fact, findings from an experiment by MacDonald and McGurk (1978) showed a different pattern of responses than those of their first study (McGurk & MacDonald, 1976). Stimuli in their second experiment included voiced and voiceless stops, as well as the nasals /m,n/, in CV syllables with the vowel /ɑ/, and were dubbed as in their earlier study to provide sequences of matched and mismatched audiovisual pairs. If we separate the consonants into two groups, the labials /p,b,m/ and nonlabials /t,d,n,k,g/, then input combinations where the visual and auditory stimuli were taken from the same group revealed no significant perceptual errors. The authors reported an accuracy rate of 80% or less as significantly different from chance ($p<.05$). In this study, the combination of visual /kɑ/ and auditory /pɑ/ produced primarily correct /pɑ/ responses (70%) rather than fused /tɑ/ responses as in the earlier experiment. The stimulus pair of visual /pɑ/ and auditory /kɑ/ in the second study revealed 82% correct /kɑ/ responses and only 9% each of /pɑ/ and the combination response /pkɑ/, although these latter two were the predominant responses in the first experiment. Similar discrepancies between the two studies occurred with the voiced stops.

Summerfield (1979) used natural video recordings synchronized with synthetic speech syllables along continua ranging among /ɑbɑ/, /ɑdɑ/, and /ɑgɑ/. For the /ɑbɑ/-/ɑdɑ/ continuum, visual /ɑdɑ/ increased the number of responses reflecting the visual cue relative to the audio-alone condition. For the /ɑgɑ/-/ɑbɑ/ continuum, visual /ɑgɑ/ increased the number of /ɑgɑ/ responses. However, visual /ɑbɑ/ only increased the number of /ɑbdɑ/ responses when paired with stimuli along the /ɑbɑ/-/ɑdɑ/ continuum. Although highly visible, the bilabial articulation did not dominate place of articulation in the responses but appeared to contribute only additional information to the percept.

Variable influence of the visual cue also appears to depend on the speech sound. Massaro (1987) reported on a study by Repp, Manuel, Liberman, and Studdert-Kennedy (1983) that found less influence of discrepant visual information on the perception of

auditory /d/ compared to /b/. There was no velar stop in the stimulus set. A closer look at MacDonald & McGurk's (1978) findings shows that visual /p/ had a greater influence on auditory /t/ (only 58% correct responses) compared with auditory /k/ (82% correct responses). The same pattern of results was noted for the voiced stops.

Kuhl, Green, and Meltzoff (1988) found that the auditory level during presentation of the syllables (visual /gɑ/ and auditory /bɑ/) was an important factor in the occurrence of visual effects. Kuhl et al. suggested the results may seem somewhat counterintuitive; they found that the louder the auditory signal, the greater the visual effects: that is, the number of illusory /dɑ/ responses increased significantly with increases in auditory signal levels from "soft" (45 dB) to "moderate" (58 dB), and from "moderate" to "loud" (66 dB).

Green et al. (1988) found that the McGurk effect varied according to the vowel used to create the CV syllable with /g/ as the visual stimulus and /b/ as the auditory stimulus. The number of illusory /d/ responses was highest when the consonant was placed before the vowel /i/, moderate for /ɑ/, and "almost nonexistent" for /u/. From these results, it appears that visual influence varies according to the salience of the articulatory gesture for a specific consonant in the context of a given vowel. Rounded vowels obscure the information value of visual cues that distinguish consonants.

Easton and Basala (1982) found no influence from visual articulatory information in the auditory perception of CVC syllables and spondaic words. Their experiments focused primarily on speechreading ability under visual-only and dubbed visual-auditory discrepancy conditions. With competing cues, although visual information had little effect on auditory speech recognition and no combination errors occurred, discrepant auditory information did substantially influence speechreading accuracy, depending on such factors as the observer's general ability to read lips and the phonetic relation between lipped and dubbed words. Where visual lip information and auditory dub were completely different

(i.e., where bimodal discordance was the greatest), the biasing effects of audition on lipreading ability were reduced.

In studies of spatial perception, important factors in the degree to which one perceptual system may bias another include (a) the compellingness of the stimulus situation (i.e., the degree of discordance), and (b) the observers' awareness of the intersensory discrepancy, and their assumption of the task as a single or unitary event. The latter may be influenced by the task instructions or the presence of discordance in the materials (Warren, Welch, & McCarthy, 1981; Welch & Warren, 1980). Thus, in a second experiment, Easton and Basala (1982) included a low compelling condition in which a man's voice was dubbed over the videotape of a speaking woman. The least amount of auditory bias on lipreading ability (highest lipreading accuracy) occurred in the low compelling condition, which involved discrepancies between the initial and final visual and auditory sounds of monosyllabic words.

In contrast, Green et al. (1991) dubbed a male voice onto the video image of a female talker, and vice versa, using the CV syllables /bɑ/ and /gɑ/. A discrepancy in the gender of the speaker between the audio and video signals did not significantly reduce the McGurk effect.

In addition, Dekle et al. (1992) reported a strong McGurk effect when monosyllabic words presented auditorily were discrepant with observed words in consonantal place of articulation. They combined auditory stimuli beginning with /b/ with visual ones beginning with /v/, and auditory stimuli beginning with /m/ with those beginning with either /d/ or /g/. They found only a weak reverse effect of dubbed words on judgments of lipread words.

In their discussion of the materials used by Easton and Basala (1982), Dekle et al. (1992) noted that some of the pairs of words did not provide visually discernible discrepant place information, for example, *word* and *whirl*. In addition, Easton and Basala had instructed their participants to identify the "word" in the identification task; however, Dekle et al. noted that the predicted McGurk percept for some of the pairs would not have been a real word.

In summary, from these studies with native English listeners, information from one modality may influence the other in speech perception, subject to factors such as the adjacent vowel, auditory signal level, compellingness of the stimulus situation, auditory intelligibility, presence of noise, and cognitive factors such as the observers' awareness of bimodal discordance.

To investigate further the influence of noise in the McGurk effect, and the strength of the effect beyond English listeners, Sekiyama and Tohkura (1991) conducted experiments with Japanese speakers listening to 10 Japanese syllables in noise-free and noise-added conditions. An earlier confusion analysis study using multidimensional scaling by Sekiyama, Joe, and Umeda (1988) had revealed that untrained Japanese observers could easily discriminate labials from nonlabials with an accuracy rate of 92% in lipreading Japanese syllables; therefore, Sekiyama and Tohkura expected the McGurk effect. Their stimuli included voiced and voiceless stops, /w/, /r/, and the nasals /m,n/. In the absence of noise, they found visual effects mostly with auditory /p/, /w/, /t/, stimuli for which audio-alone intelligibility had been slightly less than 100%. For the auditory labials /p/ and /w/, the visual effect was significant for visual nonlabials. In noise, the McGurk effect was stronger and more widespread. Perception of auditory nonlabials was significantly reduced by visual labials sharing manner of articulation. Similarly, auditory labial stimuli, when combined with visual nonlabials, produced significant visual effects and a predominance of nonlabial responses. Sekiyama and Tohkura had also asked participants to report any perceptual discrepancy between audio and visual stimuli; however, the authors did not discuss this point further. For the Japanese, listening to Japanese syllables, the occurrence of the McGurk effect was contingent upon auditory intelligibility and the presence of noise.

Later, the same authors explored the McGurk effect with Japanese speakers listening to 10 AE syllables and AE speakers listening to 10 Japanese syllables (Sekiyama & Tohkura, 1993). The Americans, tested in the U.S., had neither studied Japanese

nor lived in Japan. Although the Japanese (university students) had studied English in junior and senior high schools, with an emphasis on reading and grammar, they had not lived in an English-speaking country. For visual /gɑ/ and /bɑ/, there were more visual effects when participants were presented with stimuli from the nonnative rather than the native language. For AE speakers, visual /g/ reduced identification accuracy of auditory bilabials /b,p,m/ and visual /b/ generally influenced identification of auditory nonlabials, but to a lesser extent for velar stops. For the Japanese listening to AE syllables, visual /g/ had significant effects on auditory labials, for example, when combined with auditory /p/, it produced 72% /t/ responses. Visual /b/ had a significant effect on auditory /t/ but not /d/, and no effect at all on the velar stops. Of particular interest to the present study is the auditory intelligibility of AE /r/ at 67% with no improvement when the visual cue was added. Sekiyama and Tohkura argued that perceptual processing by the Japanese was "vision-independent", despite their apparent lipreading ability, and thus differed from that of AE speakers. They proposed an Auditory Intelligibility Hypothesis to account for the differences. This states that the Japanese incorporate visual cues only when auditory information is insufficient for speech perception.

In contrast, experiments by Massaro, Cohen, Gesi, Heredia, and Tsuzaki (1993) found that speakers of Japanese, Spanish, and AE demonstrated the McGurk effect. The authors presented participants with stimuli synthesized along auditory and visual dimensions, using a /bɑ/-/dɑ/ continuum and corresponding articulatory gestures on an animated face. They argued that, given cross-linguistic processing similarities in the results, the underlying mechanisms for speech perception are similar also.

Apart from these last two studies on the McGurk effect, which considered different language and cultural factors, other cross-language studies of speech perception using auditory stimuli alone have revealed several factors affecting the ability to perceive and produce nonnative contrasts.

Nonnative Speech Perception Studies

Previous studies have reported on the development of new phonetic categories to accommodate VOT (voice-onset time) differences between the L1 and L2 (e.g., Flege, 1989; Flege & Eefting, 1987). Strange and Jenkins (1978) have argued that contrasts based on temporal characteristics (e.g., VOT, as found in the distinction between /p/ and /b/) are easier to acquire than those based on spectral cues (e.g., /r/-/l/). Native English speakers have been able to identify a new VOT category after a single training session (Pisoni, Aslin, Perey, & Hennessy, 1982). In contrast, training studies involving the /r/-/l/ contrast have spanned several weeks (e.g., Lively et al., 1993). The AE /r/-/l/ contrast may be more difficult to acquire than VOT distinctions.

The AE /r/-/l/ contrast has dominated the perception literature, in particular with regard to the problems it poses for Japanese listeners. Mann (1986) found that, although unable to categorize /r/ and /l/ as separate phonemes, Japanese listeners demonstrated sensitivity to the acoustic effects of coarticulation involving these sounds. When presented with synthetic tokens along a /dɑ/-/gɑ/ continuum preceded by natural tokens of /ɑl/ and /ɑr/, the Japanese, like the English speakers, responded to the contrasting effects each of these has on the location of the /d/-/g/ category boundary by indicating more /g/ percepts with /ɑl/ than /ɑr/.

Mochizuki (1981) investigated the perception of natural speech tokens of /rɑ/ and /lɑ/ by Japanese listeners. When /r/ and /l/ were presented in different positions, as well as in initial and final clusters in words, accuracy of identification varied according to the position—from 64% for /r/ in initial consonant clusters to 98% for /l/ in final position. In general, accuracy of identification was lower for the liquids in clusters and intervocalic position. Performance on labeling tasks using a series of synthetic /rɑ/-/lɑ/ tokens was poorer (average 84%) than that for natural speech. The Japanese responses also included the identification of /w/ along this continuum, with the peak more than 40% higher than

that for the Americans. However, in other studies, Japanese with more conversational English experience showed better performance on identification and discrimination tasks involving synthetic /r/ and /l/ tokens (MacKain et al., 1981; Miyawaki et al., 1975).

Using natural speech, Sheldon and Strange (1982) also found that difficulties in perceiving the /r/-/l/ contrast varied with position in the word (i.e., prevocalic liquids in consonant clusters produced more errors whereas the best performance was obtained with word-final postvocalic liquids).

In a further investigation into the process of acquiring the /r/-/l/ contrast, Yamada and Tohkura (1992) reported that 14 of 15 Japanese listeners perceived a /w/ category between /r/ and /l/ when presented with the initial CV portion of a synthesized *right-light* continuum. The authors noted that the Japanese /w/ can precede the vowels (/ɑɪ/ and /ɑ/) used in this and in Mochizuki's (1981) study; this may have contributed to the perceptual assimilation of the nonnative stimulus into a native phoneme category (Best, McRoberts, & Sithole, 1988), in this case, the Japanese /w/. Best et al. proposed that the native phoneme category to which a nonnative sound would assimilate should be predictable on the basis of the similarity in phonetic-articulatory features between the nonnative and the native category. Of the three categories, AE /r/, /l/, and /w/, only the latter is similar to the realization of a Japanese phoneme (see also Best & Strange, 1992).

In summary, these cross-language studies on the perception of /r/ and /l/ by Japanese listeners show that numerous factors affect such perception, including experience with English, the phonological context of the liquids, their word position, and the nature of the stimuli (i.e., natural vs. synthetic).

In addition to these studies investigating the variables influencing Japanese speakers' ability to identify and discriminate /r/ and /l/, several studies have had the objective of training them to identify these sounds (Lively et al., 1993, 1994; Logan et al., 1991). Japanese learners of ESL received three weeks of training using forced-choice identification tasks that emphasized

natural exemplars from multiple talkers with /r/ and /l/ in a variety of phonetic environments in minimal pairs (Lively et al., 1993, 1994). Results showed a significant improvement in identification accuracy as well as generalization of training to novel words spoken by a familiar talker and by a new talker.

Borden et al. (1983) investigated the influence of training for 10 Korean learners of ESL in the perception and production of /r/ and /l/. Perception results indicated an improvement in the categorization of /r/ and /l/ on a synthesized /rɑ/-/lɑ/ continuum after training that involved identification of the contrast in nonsense syllables and words in initial position and initial consonant clusters. Pretest and posttest production scores did not differ significantly (cf. Sheldon, 1985).

One additional study provides evidence for the use of visual cues by Japanese speakers as a strategy in face-to-face communication in English. Hattori (1987) surveyed Japanese students who had lived in the U.S. for at least two years before returning to Japan, and had used English to communicate with their friends. Results indicated that, following their experience in the U.S., they felt both that they made eye contact more than their friends who had lived only in Japan and that those who made eye contact were more friendly. Hattori suggested that this practice had developed early in their stay in the U.S., not only because of the American custom of looking at the face of one's interlocutor but also out of the necessity for these nonnative speakers to "catch as much information as possible from their interlocutor" (p. 115) to compensate for difficulty in understanding English.

The aforementioned studies have dealt primarily with the auditory perception of nonnative contrasts. In conversational contexts, however, visual information from a speaker's lip movements is also available. The findings of Sekiyama and Tohkura (1993) revealed no benefit of visual /r/ to the Japanese speakers in Japan in the identification of auditory AE /r/. Given Hattori's (1987) results, this may be attributable to the lack of English conversational input to participants who had received only classroom instruction in Japan.

The first experiment reported here addresses the issue of the relative influence of visual and auditory cues on the perception of three visually salient labial consonants /p,f,w/ and three nonlabial articulations /r,t,k/ by 90 advanced ESL learners in the U.S. of four L1s: Japanese, Korean, Spanish, and Malay, plus a control group of 25 native speakers. Videotapes of audiovisual and audio-alone conditions with and without noise were used. This research was motivated by the following questions:

1. Do the learners' responses reveal any influence of visual cues on their perception of the consonants? How do L1 categories affect audiovisual perception? From the findings of the studies summarized earlier, I hypothesized that of the consonants used in this study, /f/ and /r/ would pose the most identification problems, and that in those instances where the auditory and visual cues matched, accuracy of identification would increase.

2. If visual effects are found, what do they indicate about the processing of nonnative visual and auditory cues? I hypothesized that the labial sounds would be more visually salient and thus might contribute more to perception, especially with auditory cues of relatively low intelligibility.

3. Is there any difference in the use of visual information by nonnative versus native speakers of English? Do visual effects vary in the presence of noise? Sekiyama and Tohkura (1991) found that for NSs of Japanese listening to Japanese syllables, the occurrence of the McGurk effect depended upon auditory intelligibility in the audio-alone condition and the presence of noise. The addition of visual cues for nonnatives might facilitate identification accuracy of nonnative sounds in the presence and absence of noise.

Experiment 1

Method

Participants

Experiment 1 involved 115 participants (25 NSs and 90 NNSs of English). Nonnatives, ranging in age from 18–25 years, were speakers of the following L1s: Japanese (25), Korean (20), Spanish (20), and Malay (25). These participants were advanced ESL learners in an English program at Indiana University, and participated as part of their listening skills class. They had not received any specific training in the discrimination or articulation of English sounds. Level of proficiency in this program is assessed by a battery of skill tests, including listening comprehension. Their TOEFL scores were over 500.[6] In an experiment such as this, involving the perception of CV syllables rather than the aural comprehension of language in context, it is difficult to determine an appropriate independent proficiency measure. I selected overall level placement because it represented individuals who, within each L1 group, shared similar experience with English; that is, exposure to the language and native speaker contact. On a Language Experience Questionnaire, the Japanese, Korean, and Malaysian learners reported 8–12 years of English study, and the Spanish reported 6–8 years. Participants across language groups reported having spent 2–6 months in the U.S. prior to this study. They also reported frequent conversations with native speakers outside of class.

I selected those with considerable English language experience for this experiment to assess the extent to which the use of visual information may have developed as part of an effective listening strategy in English conversational situations. The selection of advanced learners also ensured comprehension of task instructions, and a relatively good sound-grapheme correspondence in English, because I instructed them to write their responses on answer sheets. I considered a closed-set identifica-

tion task but discarded it as too confining to accommodate all possible responses.

In addition, I paid 20 undergraduates, without prior knowledge of the experiment, $5 for their participation as a NS control group. All participants had corrected-to-normal vision with no reported hearing problems.

Materials

I used six English consonants in this experiment: three (/p,f,w/) with marked labial involvement, and three (/r,t,k/) with comparatively less visually salient articulatory movements. Selection of these consonants was guided by several factors: (a) their phonemic status in the participants' L1s as described earlier, (b) their visual value in English, (c) observed perceptual difficulties of ESL learners, and (d) length of the experimental session. Taking into consideration the fatigue that the repetitive nature of this type of experimental design can cause, and the fact that participants were available for only one session, I limited the number of consonants to six.

I paired each consonant in a CV syllable with the low back vowel /ɑ/. All of the L1s in this study have a similar vowel. There is evidence that the visual discernability of consonants is enhanced when adjacent to vowels produced with the mouth wide open (Berger, 1972), avoiding such influences on the consonant's articulation as anticipatory rounding when preceding a rounded vowel. In addition, /ɑ/ occurs in open syllables in English without a diphthongization that might be confused visually with another phoneme.

Experimental Design

I recorded the audio and video signals as a female native English speaker pronounced each of the six CV syllables: /pɑ/, /fɑ/, /wɑ/, /rɑ/, /tɑ/, and /kɑ/. She enunciated each syllable clearly although articulation was quite natural. The signals were re-

corded with a Panasonic WV 6000 video camera and Panasonic AG 7400 audio equipment. From these original signals, I carried out dubbing to produce three repetitions of all possible combinations of stimuli on videotape in a randomized order, that is, each auditory stimulus matched with each visual one so as to produce both matched and mismatched combinations. The speaker's entire head appeared on the videotape. I performed the editing with the assistance of a professional audiovisual specialist, using the Panasonic AG 7400 and the industrial editing deck Panasonic AG 1950 with the editing controller AG A96 to control dubbing timing in order to synchronize the consonant releases of the audio and video signals. For each syllable's visual articulation, a new audio signal replaced the original one, and then was examined frame by frame (each frame is about 33 ms.) to ensure the most compelling condition. For example, dubbing was guided by the closure release for the consonants /p,t,k/.

I investigated potential visual effects in the presence and absence of noise. To measure auditory intelligibility of the six syllables, I included audio-alone conditions, one without and one with noise. A broad bandwidth noise with relatively equal intensity over a range of 100–7,000 Hz was added to the audio signal on the tape. The signal-to-noise ratio was 0 dB, and was set at the peak of the syllable (i.e., during the vowel). To test visual influences, I used two audiovisual conditions (with and without noise). All groups participated in four conditions in the following order: nA-V (audiovisual stimuli in noise), nA (audio-alone in noise), A-V (audiovisual without noise), and A (audio-alone without noise). In all conditions, a warning tone preceded the stimulus, which was then followed by several seconds allowing participants to write down their responses. Each presentation (from one tone to the next) comprised a 9-second unit. In the nA and A conditions, the screen remained black; however, the time course was the same. Following Massaro and Cohen (1983) and Sekiyama and Tohkura (1991), the audiovisual presentations preceded the audio-alone ones, and the presentations in noise preceded those in the clear to avoid the influence of hearing clear audio-alone

signals on the conditions in noise and with visual cues. For each participant, the total number of trials was as follows: 216 in the nA-V and A-V conditions (3 repetitions×36 syllable combinations ×2 conditions) and 36 in the nA and A conditions (3 repetitions× 6 syllables×2 conditions) for 252 trials. Each experimental session lasted less than 1 hour, including instructions, questions, and a break at the mid point.

Procedure

Stimuli were presented at a comfortable 70 dB SPL (measured at the participants' heads) via the speakers of a 24-inch color TV monitor in a sound-attenuated room. Testing was done with small groups, with an average of five per group. To generalize the findings of the current study to natural communicative situations as much as possible, participants were instructed to imagine themselves in face-to-face communication with the speaker whose entire head appeared life-sized on the TV monitor. They were asked to sit at a distance from the TV screen that was comfortable for them, approximately 4 feet from the screen. Observation of each session revealed that both NSs and NNSs followed the instructions and looked at the screen at the sound of each tone. They were told they would hear a sequence of English sounds, not real words, and were given numbered sheets on which to write down what they heard, using standard spelling letters. They were instructed to respond to every stimulus. The response sheets were collected at the conclusion of each condition. The participants were not informed of the presence of audiovisual discrepancies to discourage them from allocating more attention to the visual modality than would be expected in a speech situation, where bimodal input would be assumed by a listener to represent a single perceptual event (Easton & Basala, 1982).

Several practice trials (including matched and mismatched stimuli) preceded each experimental session. As a result of pilot testing, this was found to be helpful for the NNSs who were unaccustomed to writing sequences of letters representing sounds

(e.g., *ta*) rather than actual words. No difficulties, however, arose during these trials. Inspection of responses revealed that participants had understood the instructions. The procedure was exactly the same for the NS group.

Results and Discussion

I tabulated the types and frequencies of responses in the A-alone and A-V conditions, with and without noise, separately for each L1 group so as to isolate any L1 influence. Participants provided a single response to each stimulus, and all responded to every stimulus. As the consonants are the focus here, the consonant portions of the stimuli and responses appear in each of the following tables. The consonants presented visually are horizontally across the top, and those presented auditorily appear vertically on the left. The responses and their frequencies appear for each A-V combination. The first vertical column of responses represents the percentage correct in the A-alone condition (auditory intelligibility). I have converted the frequencies to percentages in order to permit comparison of patterns across the groups that did not contain the same number of participants. I compared the frequency of correct responses in each cell in the A-V condition with the frequency of correct responses for the auditory syllable in the A-alone condition. Symbols indicate significant differences based on the chi-square test.[7] Discussion of the results begins with the NS data, followed by the results from the Japanese, Korean, Spanish, and Malaysian groups, and then a discussion of the relationship between L1 and the distribution of responses. For all groups, comparison of the results in the noise-free and noise-added conditions revealed that the addition of noise had a rather limited influence, that is, only for certain A-V combinations did visual effects reach significance only in noise and not in the clear. Thus, information within the tables represents the results of the noise-free conditions for each L1 group. In those instances where A-V responses showed a significant difference from the A-alone presentation in noise only, symbols refer to the responses in

noise given at the bottom of the table with the appropriate level of statistical significance. The discussion of the results includes any notable findings.

Native Speaker Results

As shown in Table 2, NSs were able to identify correctly all consonant stimuli in the audio-alone condition. In all instances of concordance between visual and auditory cues (the diagonal cells), responses were 100% accurate; however, where these cues did not match, as shown in the off-diagonal cells, some erroneous responses occurred, although not to a statistically significant degree in the absence of noise.

In the A-V condition, auditory /k/ was accurately perceived across the board regardless of the visual cue. In fact, the right lower quadrant of the table reveals no confusion at all when both visual and auditory cues were taken from /w,r,t,k/. Auditory /t/, /r/, and /w/ were almost as resistant to visual influence as /k/, showing only a slight reduction in identification accuracy in combination with visual labials, especially /p,f/. Mismatched stimuli combinations of the bilabial stop /p/ and labiodental fricative /f/, close in place of articulation although different in manner, showed some evidence of confusion, with only 79% accurate identification of auditory /f/; the remainder of the responses included the visual cue /p/. The influence of visual nonlabials /t,k/ on auditory labial /p/ was slight for native speakers listening to clear audio signals.

With noise, accuracy of identification was 100% in three distinct areas: the A-alone condition, instances of visual-auditory concordance, and again the visual and auditory nonlabials in the right lower quadrant. Two remarkable differences, however, occurred. The visual effect of /k/ on auditory /p/ was significant, as was visual /p/ on auditory /f/ ($p<.001$, $df=1$, $n=150$). When the audio signal was disrupted, native listeners appeared to weigh the relative strengths of the available cues. In cases where the auditory and visual cues shared features of articulation, as with

Table 2
Response Types & Frequencies (%)
Native Speaker of English, Audio Alone, and Audiovisual Conditions

	Visual Cue													
Auditory Cue	None		p		f		w		r		t		k	
p	p	100	p	100	p	91	p	98	p	100	p	93	p	91*
					f	7	pw	2			t	5	k	9
					pl	2					k	2		
f	f	100	f	79#	f	100	f	92	f	97	f	100	f	100
			p	14			fw	8	p	3				
			pl	5										
			pw	2										
w	w	100	w	93	w	92	w	100	w	100	w	100	w	100
			bw	7	vw	8								
r	r	100	r	97	r	97	r	100	r	100	r	100	r	100
			br	3	vr	3								
t	t	100	t	88	t	100	t	100	t	100	t	100	t	100
			pt	12										
k	k	100	k	100	k	100	k	100	k	100	k	100	k	100

In noise: *p 60•, #f 61•. •$p<.001$ ($df=1$, $n=150$).

the voiceless stops /p,t,k/, the possibility of competition was greater. The acoustic cue signaling the place feature of the bilabial stop /p/ did not appear to contribute sufficient information in noise for some native listeners (60% accuracy) in order to override the strength and clarity of the visual cue for /k/. Most of the erroneous responses to this combination were /k/, although /h/, which is visually indistinguishable from /k/, was reported by three participants. Both Binnie et al. (1974) and Miller and Nicely (1955) had found that cues to place of articulation were substantially affected by noise. The major concentration of energy during the noise burst following the voiceless aspirated stops is lowest in frequency and intensity for /p/ (Ladefoged, 1982; Miller & Nicely, 1955). There was also a significant effect of visual /p/ on auditory /f/ reducing accuracy of identification to 61% ($p<.001$). The random noise pattern of [f] is scattered in the middle and lower parts of the frequency range; thus, the added noise, also spread across frequencies, may have allowed the visual bilabial /p/, close in place of articulation, to be relatively more informative.

Nonnative Results

The NNS results will be discussed by L1 group in the following order: Japanese, Korean, Spanish, and Malay (see Tables 3–6). For each group, discussion begins with the audio-alone responses followed by responses to matched A-V cues and then mismatched cues. This discussion will focus on the patterns that emerged from the noise-free and noise-added conditions. I analyzed the data per L1 group in order to control for any L1 influence; however, I noted similarities across these groups. The next section will present a discussion of the relationship between L1 and distribution of responses.

Japanese. The audio-alone responses in the leftmost column of Table 3 show that the Japanese were able to identify the stops /p,t,k/ with 100% accuracy, and /w/ with 97% accuracy. This was also the case in the presence of noise. As one might expect from the findings of previous studies, identification of /r/ was much

lower at 65%. Erroneous responses for auditory /r/ were all /l/. Participants were not aware that /l/ was not among the consonants used in this experiment. Recall that all repetitions of /r/ in audio-alone and audiovisual conditions were dubbed from one original utterance, and thus, were identical. Although advanced in many aspects of their English proficiency, these learners had difficulty attending to the acoustic cues that define the AE /r/ category. As advanced learners, however, they were aware that /r/ and /l/ pose difficulties for some L2 learners; consequently, some of these responses may represent instances of hypercorrection. There was a significant increase ($p<.05$, $df=1$, $n=150$) in the identification of /r/ with visual cues /w/ and /r/, which were apparently perceived as similar in articulation. For some speechreading researchers, /w/ and /r/ belong to the same viseme or visual category (Fisher, 1968; Woodward & Barber, 1960). Identification of /f/ in the A-alone condition was also low. Most of the incorrect responses were /p/. According to Vance (1987), a weak voiceless bilabial fricative [ϕ] occurs in Japanese in careful speech. In less careful pronunciation, the lower lip and upper teeth may be approximated, resulting in a weak labiodental fricative. These learners did not fully distinguish the bilabial and labiodental phonetic and visual categories in English, despite the difference in manner of articulation (stop vs. fricative). Perception of /f/ improved with the addition of visual /f/; however, this difference was not statistically significant. Perception of /f/ was significantly reduced by visual /p/ ($p<.01$), /t/, and /k/ ($p<.001$), even without noise, attesting to the ambiguous nature of this sound for them. Perception of the stops /p,t,k/, and of /w/ was 100% accurate where visual and auditory cues matched, as shown by the diagonal cells. Although /p/ was identified with 100% accuracy in audio-alone conditions and with visual /p/, visual /f/ significantly affected perception of the stop, resulting in responses of /t/ or /f/. Of particular interest was the dominance of visual /r,t,k/ on auditory labial /p/ producing primarily nonlabial responses. The magnitude of the visual effect of /k/ with and without noise was greatest on auditory /p/ (12% accuracy without noise, Table 3).

Table 3

Response Types and Frequencies (%): Japanese, Audio-Alone, and Audiovisual Conditions

Auditory Cue	Visual Cue: None		p		f		w		r		t		k	
p	p	100	p	100	p	67•	p	93	p	69■	p	16•	p	12•
					f	8	f	7	f	16				
					t	25			t	12	t	82	t	50
									k	3	k	2	k	38
f	f	56	f	33■	f	70	f	53*	f	65	f	27•	f	27•
	p	40	p	50	p	20	p	27	p	20	p	9	p	16
	pl	4	pl/pr	10	pl	10	pl/pr	10	pl/pr	13	pl/pr	12	pl/pr	8
			pw	7			pw	8	t	2	t	46	t	31
							w	2			s	3	s	3
											st	3	st	3
													k	12
w	w	97	w	97	w	87	w	100	w	100	w	75▲	w	67■
	l	2	p	3	l	8					l	14	l	25
	r	1			r	5					r	11	r	8
r	r	65	r	70	r	63	r	85▲	r	87▲	r	60	r	52
	l	35	l	30	l	37	l	15	l	13	l	40	l	48
t	t	100	t	99	t	100	t	99	t	100	t	100	t	100
			p	1			tw	1						
k	k	100	k	100	k	100	k	100	k	100	k	100	k	100

▲$p<.05$. ■$p<.01$. •$p<.001$ ($df=1$, $n=150$). In noise: *f 44▲ (Audio-Alone, $f=63$)

However, visual /k/ also significantly influenced perception when the auditory cues were /f/ or /w/, pointing to its information value relative to the auditory cues.

Korean. As shown in Table 4, responses from the Korean group in the A-alone condition revealed 100% accuracy in the perception of the stops /p,t,k/ and of /w/. The only change in the identification of these consonants with the addition of noise was the perception of /p/, which was reduced slightly to 96%. The 80% correct identification of /r/ without noise (78% with noise) exceeded that of the Japanese. Identification of /f/ showed the greatest accuracy of all the nonnative groups (92% without noise; 87% with noise). The addition of visual information enhanced identification of /f/ and /r/ to 97% and 95%; however, as the A-alone figures were already high, the difference did not reach statistical significance. Where auditory and visual cues were mismatched, significant visual effects on the auditory labials were found in the noise-free and noise-added conditions. The visual nonlabials /t,k/ and /r/ influenced the perception of /p/ ($p<.001$, $df=1$, $n=120$). Visual /k/ also significantly reduced perception of auditory /f/. For this L1 group, there was a significant reduction in the perception of auditory /t/ with and without noise in the presence of visual /p/ ($p<.01$). Visual /p/ also reduced the perceptual accuracy of auditory /f/ ($p<.001$) and /w/ in both A-V conditions; however, the percentage of correct responses for /w/ decreased more in noise to 34% ($p<.001$). The incorrect responses in this case were largely bilabial as a result of the bilabial articulation of the visual cue. The accuracy of identification of /p/ also decreased significantly ($p<.001$) in the presence of visual /f/.

Spanish. Responses from the Spanish, shown in Table 5, indicate that identification of the stops /p,t,k/ as well as the approximants /r,w/ in both the noise-free and noise-added audio-alone conditions was 100% accurate. However, perception of /f/ was confused with /p/. Unlike the Japanese and Korean learners, the Spanish did not have difficulty with the audio-alone identification of English /r/, with or without noise.[8] The addition of visual /f/ provided only a slight percentage increase in accurate percep-

tion of auditory /f/. The majority of erroneous responses included the stop /p/. The remainder of the diagonal cells (matched A-V stimuli) revealed 100% accuracy in both A-V conditions. In cases of mismatched A-V stimuli, with and without noise, the visual nonlabials /t/ and /k/ significantly reduced the perception of the auditory labial /p/ ($p<.001$) to 30% and 10%, respectively, without noise, and 20% and 2% with noise. Erroneous responses reflected the visual cues /t,k/, which also influenced accurate identification of auditory /f/ and /w/, the latter especially in noise where identification of /w/ was reduced to only about 76% accuracy. In both the noise-free and noise-added conditions, identification of auditory /f/ was significantly reduced to only 7% accuracy when combined with visual /p/ ($p<.001$), and visual /f/ reduced identification of auditory /p/ to 67% ($p<.01$). Visual /w/ also influenced perception of auditory /f/, which was reduced to 35% accuracy in the noise-free condition ($p<.01$). Visual /p/ also reduced identification of /r/ to 52% in noise ($p<.001$). Many erroneous responses included the auditory cue with a bilabial stop, but voiced (e.g., /br/). Visual /f/ significantly reduced perception of /w/ in noise (72%) ($p<.05$). The majority of incorrect responses maintained the voicing feature of the auditory cue combined with the labiodental place cue to produce /v/. Combinations of visual and auditory nonlabials in the right lower quadrant showed accurate identification.

Malay. Like the Spanish, the Malaysian learners' responses showed 100% accuracy in the perception of /p,t,k/ and /w,r/ in audio-alone conditions with and without noise, as noted in Table 6. Accurate identification of /f/ without noise improved from 75% in the A-alone condition to 82% with the addition of the visual /f/ cue, but only to a significant degree in noise where identification increased from 68% to 93% ($p<.01$, $df=1$, $n=150$). In cases where the auditory and visual cues were mismatched, the Malaysians also showed a significant visual effect of the nonlabial stops /t,k/ when the auditory cue was the bilabial stop /p/ ($p<.001$) with and without noise. Perception of /r/ and /t/ was influenced by visual labials but only to a significant degree with noise ($p<.05$), where identification of /r/ was reduced in the presence of visual /p/ to 77%,

Table 4

Response Types and Frequencies (%): Korean, Audio-Alone & Audiovisual Conditions

Auditory Cue	Visual Cue: None		p		f		w		r		t		k	
p	p	100	p	100	p	40•	p	58•	p	45•	p	18•	p	12•
					f	42	f	32	f	45				
					t	18	t	2	t	10	t	82	t	67
							w	8					k	21
f	f	92	f	45•	f	97	f	88	f	92	f	70*	f	53■
	p	8	p	55	p	3	p	2	p	6	r	6	p	4
							wf	8	fw	2			l	8
							v	2					fr/fl	4
											t	22	t	20
											k	2	k	11
w	w	100	w	67■	w	67■	w	100	w	87	w	73▲	w	72▲
			v	15	v	15								
			m	10										
			pr/pl	4	r	9			rw/lw	3			r	5
			b	2										
			l	2	l	9			l	10	l	27	l	23

r	r	80	r	67	r	87	r	83	r	95	r	77	r	72
	l	20	l	25	l	9	l	6	l	5	l	23	l	28
			br	8	vr	4	w	6						
							rw	5						
t	t	100	t	67▪	t	100	t	87	t	93	t	100	t	100
			p	33			f	7	tw	7				
							tw	6						
k	k	100	k	85	k	100	k	100	k	100	k	100	k	100
			p	15										

▲$p<.05$. ▪$p<.01$. •$p<.001$ ($df=1$, $n=120$).
In noise: *f 57▪ (Audio-Alone, f=88)

Table 5

Response Types and Frequencies (%): Spanish, Audio-Alone, and Audiovisual Conditions

	Visual Cue													
Auditory Cue	None		p		f		w		r		t		k	
p	p	100	p	100	p	67■	p	90	p	93	p	30•	p	10•
					f	13	w	3						
					t	10	pw	7	t	7	t	62	t	10
					k	10					k	8	k	80
f	f	62	f	7•	f	64	f	35■	f	57	f	27•	f	25•
	p	31	p	86	p	25	p	46	p	20	p	2	p	23
	pl	7	pl	7	pl	7	pw	8	pw	7	pl	5	pl	7
					v	4	v	7	v	7	t	64	t	13
							w	4	w	2	k	2	k	23
									b	7			b	9
w	w	100	w	87	w	80*	w	100	w	100	w	80	w	95
			v	6	v	20					v	13		
			b	7							l	7	l	5
r	r	100	r	64■	r	89	r	100	r	100	r	100	r	100
			br	32	br	9								
			b	2	vr	2								
			p	2										

t	t	100	t p	77 23	t	100	t p tw	91 2 7	t	100	t	100	t	100
k	k	100	k p	95 5	k	100	k kw	98 2	k	100	k	100	k	100

▪$p<.05$. ▪$p<.01$. •$p<.001$ ($df=1$, $n=120$).
In noise: *w 72▪

Table 6

Response Types and Frequencies (%): Malay, Audio-Alone, and Audiovisual Conditions

	Visual Cue													
Auditory Cue	None		p		f		w		r		t		k	
p	p	100	p	100	p	77▲	p	84*	p	80#	p	28•	p	16•
					f	23	f	12	f	16				
							pw	4	pw	2				
									t	2	t	72	t	8
													k	76
f	f	75	f	48■	f	82	f	65	f	80	f	71	f	60
	p	15	p	52	p	18	p	23	p	12	p	8	p	28
	pl	10					w	5	w	2	t	13	t	4
							pw	3	pw	3	pl	5		
							fw	4	pl	3	kl	3	k	8
w	w	100	w	100	w	92	w	100	w	100	w	95	w	96
					v	3								
					l	5					l	5	l	4
r	r	100	r	84	r	99	r	100	r	100	r	100	r	100
			br	16	br	1								
t	t	100	t	85	t	100	t	88	t	100	t	100	t	100
			p	13			p	5						
			k	2			tw	5						
							w	2						

k		k	100	k	98	k	100	k	100	k	100	k	100	k	100
				p	2										

▲p< 05. ■p< .01. •p<.001 (df=1, n=150).
In noise: *p 70■; #p 65■; -f 93■ (Audio-Alone, f=68); +r 77▲; ^t 77▲

Table 7
Relationship Between L1 (Japanese, Korean, Spanish, Malay) and Distribution of Responses Significant Chi-Square Values

	Visual Cue						
Auditory Cue	None	p	f	w	r	t	k
p			21.11•	32.39•	38.07•		
f	23.34•	29.73•	23.13•	38.09•	23.00•	51.62•	27.57•
w		45.05•	16.05■			13.46■	33.44•
r	49.96•	8.48▲	37.98•	22.95•	18.71*	59.38•	74.13•
t		26.36•					
k		19.35*					

▲p<.05. ■p<.01. •p<.001 (df=3, n=270).
*Significance is suspect; >20% of cells have values less than 5.

and identification of /t/ to 77% with visual /w/. The influence of visual /p/ on auditory /f/ was present with this NNS group also, especially with noise, where perception of /f/ declined from 69% to 35% ($p<.001$). Similarly visual /f/ reduced the perception of auditory /p/ in both A-V conditions. Identification of /p/ was also significantly reduced with noise by visual /w/ and /r/ ($p<.01$).

Relationship Between L1 and Distribution of Responses

To investigate whether the distribution of responses for each A-alone stimulus and A-V stimulus pair was related to the L1 of the nonnative groups (Japanese, Korean, Spanish, and Malay), I prepared contingency tables (frequency of response type×L1). Table 7 shows the cells where the chi-square value was significant, demonstrating that the frequencies of responses to the stimuli in those cells varied according to L1. Both the chi-square value and the symbol indicating the level of statistical significance are given. In this analysis, I limited the number of categories for response type to two to maintain cells for all stimuli where the expected frequency would be large enough to compute chi-square with a reasonable assurance of accuracy. Frequencies of "correct" responses, that is, those matching the auditory cue, constituted one category; I pooled all other responses to form the second one. In a number of cells the chi-square value was not significant; that is, responses were quite consistent across the four L1 groups. For example, the values for the relationship between L1 and responses to visual /t/ and /k/ combined with auditory /p/ were not significant.

I observed the following specific L1 patterns. When /t,k/ were the auditory cues, only stimulus pairs involving visual /p/ showed a significantly different distribution of responses according to L1. For the combinations of visual /p/ with auditory /t/, the Koreans showed more responses matching the visual cue compared with the other groups (see Table 4).[9] They also showed a greater influence of visual /p/ on auditory /k/; however, statistical significance is suspect due to the number of cells with values less than

5. Responses to stimulus combinations where the visual cue was /f/ or /w/ and the auditory cue was /p/ as well as visual /p/ and /f/ dubbed onto auditory /w/ provided further evidence of the strong influence of these visual cues for the Korean group. Responses to most stimuli involving matched auditory and visual cues (shown in the diagonal cells) also revealed consistency across L1 groups except for the combination of /f/-/f/ (significance is suspect for visual /r/-auditory /r/). The Korean participants differed from the others in their identification accuracy and use of the visual /f/ cue.

For the A-V pair /r/-/r/, the Japanese showed a greater number of erroneous /l/ responses than the other groups; however, statistical significance is also suspect here due to the number of cells with values less than 5. In the A-alone identification of /r/, the response frequencies for the Japanese differed significantly from the other groups, that is, more incorrect responses (/l/) were given (see Table 3).

The Spanish differed from the other L1 groups most notably where the auditory cue was /f/ and the visual /p/. Here the frequencies of the incorrect responses (/p/ and /pl/) were higher compared to the other L1 groups (see Table 5). This could be attributed to the confusability of AE labial articulations as a result of the lenis production (relatively weak degree of articulatory strength) of bilabial obstruents in Spanish. Both the Spanish and Malaysian responses differed most from the other groups in two areas: auditory /w/ when paired with visual /t/ and /k/, and auditory /r/ when combined with visual /f,w,t,k/. In the case of auditory /w/, there were fewer incorrect /l/ responses from these two groups compared with the others. In addition, /r/, which was 100% intelligible in the A-alone condition, remained a strong cue in the presence of visual nonlabials /t,k/; these did not influence the Spanish speakers' perception compared with the Japanese and Koreans, for whom the more open mouth position for /t,k/ may have contributed to the increase in /l/ responses.

Consider the responses to the stimuli involving only the stops /p,t,k/, present phonemically in all the L1s and identified with 100% accuracy in the A-alone condition. Auditory /p/ was

subject to significant influence from visual /t,k/ in all nonnative groups although, as shown in Table 7, they did not differ significantly in the magnitude of the visual effect. The nonlabials /t,k/ patterned alike both as auditory and as visual cues for them, with the exception of the Korean responses to the combination of visual /p/ and auditory /t/ where findings indicated a significant influence of the visual labial. Other nonnative groups showed some reduction in the identification of the auditory cue with this stimulus pair but not to a significant level. These findings were consistent with the strong influence of the visual cue /p/ in both stimulus pairs for the Koreans as noted above.

Summary

The lack of significant visual effects for the NSs in the noise-free condition in Experiment 1 was not consistent with most previous McGurk effect experiments with English CV syllables. However, previous researchers had confined their studies to consonants identical in manner of articulation (i.e., stops) with the exception of MacDonald and McGurk (1978), whose stimuli also included the nasals /m,n/. Stimulus compellingness and the observers' awareness of the intersensory discrepancy may have played a role, as suggested by Welch and Warren (1980), who noted that the strength of the assumption of a unitary perceptual event is a function of the redundant properties present in the stimulus situation and the relative weighting assigned to them. In the case of speech, the properties are those of the visual and auditory cues. It appears, then, that the observers' task is to assess the degree to which the features of the cues from each modality match the features of a stored representation in memory. The more features of a single representation these cues possess, the more likely they are to be perceived as contributors to the same event, and thus will be integrated. In noise, the auditory signal is less informative and the visual cue contributes more to the percept.

For the NNSs, significant visual effects occurred even in the

absence of noise. The pattern of results indicated that the contribution of a cue to the eventual percept was influenced by the degree to which it was informative relative to the other cue in a given stimulus presentation: /p/ versus /f/ and /t,k/ versus /p,f/. Information value was influenced by the nonnatives' perceptual categories. The Spanish showed a significant decrease, from 62% to 7%, in the perception of auditory /f/ when combined with visual /p/. This was clearly the lowest percentage for this input combination among all the NNS groups. They improved little with the addition of the visual /f/ cue, suggesting their inability to differentiate clearly visual /f/ from other labial articulations, although labiodentals for native speakers constitute a separate visual category (Berger, 1972; Fisher, 1968; Woodward & Barber, 1960). In contrast, the accuracy of the Koreans' identification of /f/ in the A-alone conditions exceeded that of the other NNS groups. In the case of the stops /t,k/, they do not show the same finer language-specific articulatory differences as the labials do. In terms of the stops alone, the influence of visual /p/ on auditory /t/ was evident for all NNS groups but reached significance only for the Koreans. The salience of visual labial cues for the Koreans may be influenced by the Korean consonantal inventory, which includes three phonemes at the bilabial point of articulation (/p, p^h, p'/) and no competing labial fricative sounds. Auditory /k/ was an informative cue for all participants. Nonnative perceptual categories appeared to influence the identification of cues, and thus A-V integration.

Experiment 2

Experiment 2 explored the hypothesis that a more compelling stimulus set (less discordance between cues) using only the stops /p,t,k/, identical in manner of articulation, would increase the assumption by NSs of a unitary perceptual event, and thus permit A-V integration and the occurrence of the McGurk effect. Although the responses of the NNSs in Experiment 1 revealed significant visual effects, the question arose as to whether the

influence of visual /p/ on /t,k/ would differ if another labial cue close in place of articulation, namely /f/, was removed from the stimulus set. In other words, would the information value of /p/ change for NNSs if potential confusion were eliminated?

Method

Participants

Participants in this experiment included 25 NSs of English and 30 NNSs of the following L1s: Japanese (10), Korean (10), Spanish (10). At the time of this experiment, speakers of Malay were no longer enrolled in this English program. The NNSs were drawn from the same level of proficiency as those in Experiment 1 and were comparable in terms of prior experience with English and minimum TOEFL score; the NSs were drawn from the same population as that in Experiment 1. None had participated in Experiment 1.

Materials

All of the combinations of the syllables /pɑ/, /tɑ/, and /kɑ/ with noise and without (3 repetitions × 9 syllable combinations × 2 conditions), and their A-alone presentations (3 repetitions × 3 syllables × 2 conditions) from the previous videotape were dubbed onto a new tape. The speaker and utterances were identical to those in Experiment 1. Interstimulus interval was also the same.

Procedure

The experimental conditions, equipment, and instructions were identical to those in Experiment 1. I used the same sequence of presentation: nA-V, nA, A-V, A. As before, a tone preceded each trial. I observed all experimental sessions as before. All participants followed the instructions and wrote their responses on answer sheets.

Results and Discussion

As before, I tabulated the types and frequencies of responses in the A-alone and A-V conditions, with and without noise, separately for each L1 group. Participants provided a single response to each stimulus and all responded to every stimulus. Tables 8 through 11 show results for the noise-free condition. Symbols indicate significant differences in the identification accuracy of the auditory stimuli when combined with visual cues compared to the A-alone condition, based again on the chi-square test. Significant differences in responses in the presence of noise occurred only for the NSs and appear at the bottom of Table 8.

Table 8
Response Types and Frequencies (%): Native Speaker, Audio-Alone, and Audiovisual Conditions

			Visual Cue					
Auditory Cue	None		p		t		k	
p	p	100	p	100	p	21•	p	28•*
					t	62	t	20
					k	8	k	36
					h	3	h	5
					pl	3	kp	3
					tp	3	kl	3
							pl	5
t	t	100	t	45•**	t	100	t	91
			p	22			kt	4
			pt	29			kl	4
			pl	4			ht	1
k	k	100	k	37•	k	92	k	100
			p	19	tk	5		
			pk	44	kt	3		

•$p<.001$ ($df=1$, $n=150$).
In noise: *Frequency of /p/ 9%; **Frequency of /t/ 24%. Differences between each of these responses to signals in noise and in the clear were significant by McNemar's test ($p<.001$).

Native Speaker Results

As shown in Table 8, auditory intelligibility of all stimuli remained at 100%. In contrast to the previous findings, these results with a more compelling stimulus set revealed significant visual effects without as well as with noise ($p<.001$, $df=1$, $n=150$). Significant visual effects were evident for the stimulus pairs involving visual /t,k/ and auditory /p/. Visual /p/ also influenced perception of auditory /t,k/. *Fused* responses (e.g., /t/ to the stimulus pair of visual /k/-auditory /p/) and *combination* responses (e.g., /pt/) were evident, as in other McGurk effect experiments. In addition, the frequency of correct responses (those matching the auditory cue) with noise was significantly lower by McNemar's test (repeated measures) for two stimulus pairs: visual /k/-auditory /p/ ($z=3.74$, $p<.001$), and visual /p/-auditory /t/ ($z=4$, $p<.001$). Correct responses to visual /t/-auditory /p/ also showed a reduction in identification accuracy with noise to 13%; however, this was not significantly different from the frequency without noise. Perception of auditory /k/ was not reduced with noise.

Nonnative Speaker Results

For each L1 group, I tabulated frequencies of responses as before and carried out chi-square analysis to determine if the presence of a visual cue significantly influenced identification accuracy of the auditory cue. I applied Yates' correction factor where observed values were less than 5; these cells are indicated on the tables. Unlike the NSs, there were no significant differences between the noise-free and noise-added conditions for the NNSs.

Results (shown in Tables 9–11) revealed some similarities across L1s and some language-specific differences. Across the L1 groups, auditory intelligibility was 100%. In addition, where the auditory and visual cues matched (the diagonal cells), identification accuracy of the auditory cue was 100%. Where these cues did not match, visual /t/ and /k/ significantly affected perception of

Table 9
Response Types and Frequencies (%): Japanese, Audio-Only, and Audiovisual Conditions

Auditory Cue	Visual Cue: None		p		t		k	
p	p	100	p	100	p	7•	p	7•
					t	80	t	43
					h	13	h	20
							k	30
t	t	100	t	100	t	100	t	100
k	k	100	k	93	k	100	k	100
			t	7				

•=$p<.001$ ($df=1$, $n=60$), Yates' correction factor applied to these two cells. No significant differences in noise.

Table 10
Response Types and Frequencies (%): Korean, Audio-Only, and Audiovisual Conditions

Auditory Cue	Visual Cue: None		p		t		k	
p	p	100	p	100	p	10•	p	7•
					t	64		
					h	13	h	26
					k	13	k	67
t	t	100	t	40▪	t	100	t	100
			p	60				
k	k	100	k	67	k	100	k	100
			p	33				

•$p<.001$, Yates' correction factor applied to these two cells; ▪$p<.01$ ($df=1$, $n=60$). No significant differences in noise.

auditory /p/ ($p<.001$, $df=1$, $n=60$) as was noted in Experiment 1. The majority of erroneous responses to the pair of visual /k/-auditory /p/ reflected the visual cue /k/ or /h/; these are visually indistinguishable. There were some *fused* responses (/t/) to this stimulus pair by the Japanese and Spanish. The majority of erroneous responses to the combination of visual /t/-auditory /p/ for all three NNS groups were /t/.

The L1 groups differed in their responses to the combinations of visual /p/ with auditory nonlabials. As shown in Table 9, for the Japanese there was only a slight reduction in identification accuracy of auditory /k/ (93%); the remaining responses were /t/. In general, the visual bilabial was not an informative cue for them. The Koreans continued to show a significant influence of visual /p/ on auditory /t/ ($p<.01$) (Table 10). In addition, the influence of /p/ on auditory /k/ increased although it did not quite reach statistical significance at the .05 level ($\chi^2=3.333$; 3.841 would be needed). For the Spanish, the findings differed from the results of Experiment 1 in their responses to the stimulus pair of visual /p/ -auditory /t/. As shown in Table 11, the visual cue reduced identification of /t/ to 13% ($p<.001$).

To determine if the distribution of responses for each of the A-V stimulus pairs was related to the L1 of the nonnatives, I prepared contingency tables (frequency of response type × L1). Responses to stimulus pairs where the A-V cues matched and where visual /t/ or /k/ was paired with auditory /p/ were quite consistent across L1 groups. However, the chi-square value was significant for two cells where /p/ was the visual cue. Responses to visual /p/ paired with auditory /t/ differed by L1 ($\chi^2=33.34$, $df=2$, $n=90$, $p<.001$). This can be attributed to both the relatively smaller visual effects for the Japanese compared to the other groups, and the strong influence of the visual cue for the Spanish. In addition, responses to visual /p/-auditory /k/ differed by L1 ($\chi^2=6.43$, $df=2$, $n=90$, $p<.05$). In this case, the frequency of erroneous responses (/t/) by the Japanese was lower than that for the other groups.

Table 11
Response Types and Frequencies (%): Spanish, Audio-Only, and Audiovisual Conditions

Auditory Cue	Visual Cue: None		p		t		k	
p	p	100	p	100	p	10•	p	7•
					t	80	t	7
					k	10	k	86
t	t	100	t	13•	t	100	t	87
			p	84				
			k	3			k	13
k	k	100	k	80	k	100	k	100
			p	10				
			pk	10				

•$p<.001$ ($df=1$, $n=60$), Yates' correction factor applied to these cells. No significant differences in noise.

Summary

A comparison of the results of Experiments 1 and 2 strongly suggests that stimulus compellingness and the associated awareness of intersensory discrepancy condition the occurrence of the McGurk effect, and depend on the observers' perceptual categories which, in turn, influence the observers' ability to identify cues and assess their similarity. In contrast to Experiment 1, when the stimulus set included only the stops, the NSs' responses in Experiment 2 exhibited visual effects in noise and in the clear for both the labial and nonlabial cues. For the NNSs, Experiment 2 again demonstrated the influence of visual /t,k/, as in Experiment 1. However, in Experiment 2, L1 groups differed in the effect of visual /p/. For the Spanish, the increased contribution of visual /p/ may be attributed to the absence of another competing labial cue such as /f/ in this stimulus set. For the Koreans, influence of visual /p/ increased in strength in combination with both auditory /t/ and /k/. For the Japanese, visual /p/ remained a weaker cue compared with /t/ or /k/.

General Discussion

In the past, most accounts of cross-language differences in speech perception have focused on the use of auditory information. For example, Best and colleagues (Best et al., 1988; Best & Strange, 1992) proposed a perceptual assimilation model to address variations in the ability to discriminate nonnative contrasts. L2 sounds similar in phonetic-articulatory features to L1 categories will be assimilated to the native phonemic categories, resulting in perceptual difficulties. Those not sharing features with a native category are not subject to assimilation and may be more easily perceived.

Flege (1988, 1991) described a model of second language phonetic perception based on the comparison of L1 phonetic categories and L2 phonetic input. There are three major directions of development. If the new phone is dissimilar to L1 categories, it will be difficult for the learner to categorize, and phonetic representation will require increased L2 experience. If an L2 phone is identical or very similar to an L1 category, it will be easily perceived. But an L2 phone that is similar to one in the L1, but not identical, will be the most problematic.

These models lack a framework for the integration of information from more than one source to represent the perception of speech in face-to-face communication; they share the element of similarity. In fact, theories of category learning generally depend on some similarity between stimulus and stored representation (Estes, 1986). In the present study, which unites both nonnative and visual factors in speech perception, similarity plays a role in the process of integrating auditory and visual cues. Auditory and visual inputs are evaluated independently for their similarity, in terms of features (phonetic and visual), to stored representations in memory. In addition, similarity between the auditory and visual cues contributes to the observers' assumption of a unitary perceptual event, a condition necessary for audiovisual integration.

To provide a framework for understanding bimodal speech

perception and the variable influence of visual cues on speech sounds, we turn to a model that describes bimodal speech perception within the context of a general theory of perceptual recognition—FLMP or fuzzy logical model of perceptual recognition (Massaro, 1987; Massaro & Cohen, 1983; Oden & Massaro, 1978). Within this model, similarity also plays a role. FLMP provides a model for evaluating and integrating multiple sources of information in pattern classification that can be applied to both L1 and L2 speech perception. The McGurk effect in NSs and NNSs has implications for both the process of A-V integration in this model and the nature of the representation to which input is matched for identification. Before discussing the application of this model to the present findings, I give a brief overview of it.

FLMP consists of three stages: feature evaluation, integration, and pattern classification. The fundamental assumption is that information from auditory and visual sources is independently evaluated in terms of the degree to which each feature in the stimulus matches the corresponding feature in each stored representation in memory to determine its overall goodness-of-fit to one representation relative to all others. Each representation has auditory and visual attributes that are graded rather than binary in nature.[10] The contribution of a given cue depends on its relative information value. A more ambiguous source contributes less to the eventual categorization decision.

In the present study, critical features that serve as retrieval cues for specific patterns from memory involve acoustic characteristics, such as: the transitions of the second and third formants (F_2 and F_3) as cues to differences in place of articulation for the stops; the frequency range of the random noise pattern for the fricative (e.g., Ladefoged, 1982); the formant structure characteristic of /w/ (low F_1 and F_2 with rise in F_2 to the following vowel) (Lehiste, 1964); and a lowered third formant for AE /r/ (Lindau, 1985). Characteristics that contribute to the evaluation of these visual stimuli include lip rounding for /w/, some lip protrusion for some speakers in the articulation of /r/ (Daniloff & Moll, 1968; Lindau, 1985), degree of aperture, position of articulators relative to one

another (as in the production of /f/), point of closure visually distinguishing the labial from the nonlabial stops, and so on.

The second stage of FLMP is integration of the cues from both modalities. Here the more informative cue contributes the most to the process. The more features the stimuli share, the more similar they are. This, in turn, creates the assumption of perceptual unity; that is, the assumption that the two sources of information specify a single perceptual event. Information from two or more sources can be processed efficiently and integrated if it can be perceived as contributing to a single event (Massaro, 1985, 1987; Welch & Warren, 1980). As Neisser (1976) pointed out, observers attend to objects and events, not to sensory inputs themselves. The result of integration consists of the degree to which the input (e.g., a syllable) matches the stored representation, with the least ambiguous features contributing the most. The contribution of one cue increases with a decrease in the information value of the other.

Pattern classification, the third stage, is the perceptual outcome. According to the original design of FLMP, input is matched against stored prototypes. Categorization depends upon how well the stimulus pattern matches one prototype relative to all others. Some previous cross-language studies have also referred to prototypes (Jamieson & Morosan, 1986, 1989; Strange & Dittman, 1984) or templates (Henly & Sheldon, 1986). In the categorization literature, the prototype is an abstract summary representation of a category as a whole, based on the characteristic attributes (central tendency) of a category's exemplars or the ideal attribute (Medin & Barsalou, 1987). Category membership depends on an evaluation of similarity between stimulus and prototype above some threshold. In turn, the degree of similarity to the prototype influences ease of classification. In Rosch's (1975, 1978) terms, prototypes are *focal instances*, that is, the clearest cases of category membership that serve to constrain but not specify membership.

There is evidence for phonetic prototypes. In selective adaptation experiments, Samuel (1982) found that adaptation

effects were greater with a prototypical adapter than with nonprototypical ones. Kuhl (1991) reported that some tokens of the vowel /i/ were given higher ratings of category goodness and functioned as perceptual magnets (i.e., perceptually assimilating close members in the category). However, Lively (1993) demonstrated that a perceptual magnet effect was not replicated with stimuli based on Kuhl's tokens of /i/ nor with prototypes selected by the listeners. Lively's findings were not compatible with the definition of a prototype as an abstract idealized representation of a category. Rather, the perceptual categories, phonetic in this case, demonstrated context-dependent variability.

In contrast to the use of prototypes as a basis for classification, the exemplar view approaches classification on the basis of similarity to stored, context-dependent representations of previously experienced category members (Estes, 1986; Jusczyk, 1993; Medin & Barsalou, 1987; Medin & Schaffer, 1978; Nosofsky, 1986, 1987; Smith & Medin, 1981). Matching input to stored exemplars versus abstract prototypical representations is consistent with the findings that Japanese listeners at a category-learning stage develop context-dependent, talker-specific representations during training to identify English /r/ and /l/ (Lively et al., 1993; Logan et al., 1991). Massaro (1987) noted that FLMP and the exemplar model of Medin and Schaffer make mathematically similar predictions in category learning experiments, and that FLMP can be extended to represent both summary-based and exemplar-based information.

Applying the principles of FLMP to the findings of the current research, I begin with the NS results. A comparison of NS responses in Experiments 1 and 2 supported the hypothesis that a more compelling stimulus set, such as the one used in Experiment 2 (sounds identical in manner of articulation, the stops /p,t,k/) was necessary in the absence of noise to increase the NSs' assessment of similarity between the auditory and visual cues, thereby permitting the cues' integration and the McGurk effect. This follows from the fact that NSs' perceptual categories facilitated accurate identification of the cues. The varied stimuli of

Experiment 1 were not perceived as contributing to the same event (Massaro, 1985; Welch & Warren, 1980). Under those circumstances, participants also would have become aware of the intersensory discrepancy (Welch & Warren, 1980), and there were few visual effects. Experiment 2 found significant visual effects for labial and nonlabial cues with and without noise. Effects were greater with noise, where the auditory cue could not contribute as much information to the integration process. The results supported the claim that the assumption of perceptual unity underlies A-V integration.

In contrast, Green et al. (1991) argued that the unity assumption may not be required for audiovisual integration. Their claim was based on findings that the McGurk effect occurred despite gender differences between the speaker whose voice provided the auditory cue and the speaker whose face provided the visual cue. The assumption was that such a difference constituted a reduction in the perception of a single event. Their findings, however, do not necessarily disprove the need for a compelling stimulus set to establish an observer's assumption that the auditory and visual cues are contributing to a unitary perceptual event, thereby permitting integration and the McGurk effect. It is not clear that gender differences, irrelevant to the task of reporting the content of what was heard, contribute significantly to the establishment of perceptual unity. It has been demonstrated that indexical properties of speech (the physical characteristics of the speaker) are not discarded by the listener, but rather become a part of the internal representation of utterances (e.g., Pisoni, 1993). However, the linguistic properties of the speech signal that are critical to the identification of a sound may rank much higher in a hierarchy of establishing a compelling situation for the observer. Thus, gender discrepancies between auditory and visual cues would not reduce the assumption of perceptual unity in such a task.

The question also arises as to the possible influence of selective attention in experiments involving more than one source of input. In other words, is it possible for an observer to attend to

only one source, completely ignoring the other despite instructions? Experiments on the McGurk effect in NSs by Massaro (1987), involving a synthesized /bɑ/-/gɑ/ continuum and traditional tasks of selective and divided attention, revealed that participants could not selectively attend to one source without influence from the other. Results, including reaction times, provided evidence for the integration of information from auditory and visual sources in speech perception, without an increase in processing time for integration compared with the processing of input from a single source. This is consistent both with Welch and Warren's (1980) observation that the influence of one modality of speech on the other (i.e., modality bias) is an actual perceptual effect, not the result of postperceptual decision processes or learning strategies, and with the comment by Neisser (1976) that observers attend not to sensory inputs themselves but to the events they specify.

In the NNS findings of the present Experiment 1, responses to /r/ revealed less than 100% intelligibility for the Koreans and Japanese, especially the latter. The Spanish and Malaysian speakers were able to identify /r/; the Spanish phone in initial position, a trill, is sufficiently different to avoid confusion; and Malay has two distinct liquids, /l/ and /r/. For the Japanese, the development of two new L2 categories is necessary to accommodate AE /r/ and /l/. Miyawaki et al. (1975) noted that Japanese speakers perceive these sounds as the same consonant, their /r/; however, the latter is realized most often as a dental or alveolar flap. In the present study, the presence of both visual /w/ and /r/ significantly increased the Japanese participants' identification of auditory /r/. These participants apparently assessed the visual articulation of /w/ as similar to /r/. This contrasts with the findings of Sekiyama and Tohkura (1993), which revealed no increase in identification accuracy for AE auditory /r/ with a matched visual cue, or with visual /w/. For the Koreans, A-alone identification of /r/ was better than for the Japanese, and increased in the presence of the corresponding visual cue. In syllable-initial position, AE /r/ is also a new phone for the Korean speakers, although they had

much less difficulty forming a new L2 category for this context, because there is only one similar L1 category with no phonetic realization in initial position.

In general, the NNSs in Experiment 1 showed more widespread significant visual effects with clear audio signals. In both experiments, for the stimulus pairs involving visual /t,k/ and auditory /p/, the nonlabial stops were clearly the most informative cues in the evaluation process; therefore, they contributed substantially to the categorization decisions across L1 groups. In some cases, responses matched the visual cue; in others, *fused* responses occurred. As noted above, the nonlabials do not show the finer language-specific differences that sounds with labial involvement do. In Experiment 1, visual /t,k/ also influenced perception when the auditory cue was /f/. Across the groups, the lower the intelligibility of /f/ (A-alone condition), the greater the contribution of the informative visual nonlabial cues to the percept. The majority of the responses reflected the three stops. In fact, all incorrect responses to /f/ in the A-alone condition involved /p/. Its proximity to /p/ in terms of place of articulation was problematic for many participants at the feature evaluation stage, especially for the Spanish and Japanese. AE /f/ has been described as more fricative in nature than the closest L1 category for the Spanish (Stockwell & Bowen, 1965); it may represent too great a similarity auditorily and visually to the generally lenis articulation of L1 bilabial phones, influencing its categorization as /p/, especially when presented with /p/ as the visual cue. Confusion for the Japanese may stem from the occurrence in Japanese of weak bilabial and labiodental fricatives (Vance, 1987). These findings are consistent with the premise of FLMP that the contribution of a given cue to the percept depends not only on its information value, but also on its value relative to another cue.

The bilabial stop also influenced perception of auditory /t/—however, to a significant degree only for the Koreans in Experiment 1, and for the Koreans and Spanish in Experiment 2. In the case of the Koreans, the presence of three bilabial stop phonemes

(/p,p^h,p'/) and the absence of sounds close in place and manner of articulation in the L1 contributed to the information value of visual /p/ and thus, to the perceptual outcome. Visual effects for this stimulus pair were absent for the Japanese in both experiments and for the Spanish in Experiment 1, where /f/ was present in the stimulus set. Only when /f/ was removed in Experiment 2 was visual /p/ more informative and a significant contributor to the perceptual outcome. When the auditory cue was /k/, the Japanese showed a similar lack of visual contribution. Although the responses from the Spanish revealed some visual influence, the reduction in identification of the auditory cue /k/ did not reach statistical significance as it did for the NSs. (The chi-square value for the Koreans, 3.33, was close to, but did not reach significance at the .05 level.)

To summarize the data on the stops thus far, Experiment 1 showed that the nonlabials /t,k/ patterned similarly as auditory and visual cues and contrasted with the labial /p/, with the exception of the Korean responses to visual /p/ combined with auditory /t/. In general, analysis of the relationship between L1 and the distribution of response frequencies revealed that the Koreans showed more visual effects than the other groups. Visual /p/ was an informative cue for them. In Experiment 2 (using /p,t,k/), results revealed an effect of visual nonlabials /t,k/ on auditory /p/ for NSs and the three NNS groups (Japanese, Korean, Spanish). For the Japanese, visual /p/ was not sufficiently informative in either experiment regardless of whether the auditory cue was /t/ or /k/. Only when the auditory cue was relatively ambiguous, as with /f/ and /r/ in Experiment 1, did visual /p/ have a substantial influence. In contrast, /p/ influenced auditory /t/ to a significant degree for the NSs, the Koreans as before, and now the Spanish in the absence of /f/. A similar influence was found on auditory /k/ for the NSs. Visual effects for that stimulus pair did not quite reach statistical significance for the Koreans and were less evident for the Spanish. A greater visual effect for /p/ on /t/ versus /k/ also occurred for the NSs in Experiment 2 in the noise-added condition, just as in the study by MacDonald and McGurk

(1978). Results from that study showed identification accuracy of auditory /t/ was reduced to 58% versus 82% for /k/ when combined with visual /p/. Repp et al. (1983) also found less influence of conflicting visual information on the perception of auditory /d/ compared to /b/; however, there was no velar stop in their stimuli for comparison. Massaro (1987) suggested that alveolar and velar stops may have more robust acoustic properties than /b/ and thus may be less influenced by discrepant visual articulations. The major concentration of energy during the noise burst is higher in frequency and greater in intensity for /t/ and /k/ versus /p/. Velars are further distinguished by the fact that the formant transitions, which are powerful cues to place of articulation, have a longer duration than corresponding alveolar or bilabial sounds (Ladefoged, 1982). These features may be contributing to the assessment of the most informative cue.

The influence of the L1 is evident in the NNS findings from both Experiments 1 and 2. Comparisons of the L1s and the L2 have pointed to areas of difficulty. In terms of the variability in the responses to combinations of the stops, the potential role of typological markedness relations remains unclear. Bimodal speech perception is a challenge, inasmuch as it involves not only phonological markedness but also its acoustic and visual correlates. Applying the notion of markedness to the FLMP framework suggests that the more marked sound would be more ambiguous as a source of information, in turn having less information value and contributing less to the percept. It is commonly held that coronals are the least marked consonants. Kean (1980) proposed that /t/ be considered the universally unmarked consonant and coronal the unmarked articulation.[11] Evidence for the special status of coronals has been provided by frequency and acquisition facts as well as by the behavior of coronals in phonological processes across various languages (for a review of internal and external evidence, see Paradis & Prunet, 1991; for the status of coronals in Korean, see Cho, 1991; Kim-Renaud, 1974).

Of the three stops, /p/ can be considered the most marked in phonological terms. Acoustically, /p/ is lower in frequency and

weaker in intensity than the other stops. From an articulatory standpoint, sounds with labial involvement show finer language-specific differences than do nonlabials, and could be considered more marked visually. As an auditory cue in the present experiments, /p/ certainly contributed less to the percept for NNS groups in noise and in the clear. As shown by the NS findings in Experiment 1, the contribution of /p/ (as with other cues) was conditioned by the assumption that the cues from both modalities contributed to the same perceptual event (i.e., a compelling condition). However, variations were evident in responses to /p/ as a visual cue. For NSs, Experiment 2 presented a compelling condition, and /p/ exerted a significant influence on auditory /t/ and /k/ in noise-free and noise-added conditions. Among the NNS groups, there was also variability. Responses from the speakers of Malay, who were present only for Experiment 1, revealed some influence of visual /p/ on auditory /t/, but not to a significant degree. For the Japanese, the cue was relatively uninformative throughout. These participants', as well as the Spanish and Koreans', responses are attributable to L1 influence. As noted earlier, weak bilabial and labiodental fricatives occur in Japanese and Spanish, contributing to the ambiguity of labial articulations. In A-V integration, the contribution of a specific cue is a function of its ambiguity relative to others. Only when /f/ was removed from the stimulus set in Experiment 2 did the Spanish show any visual effects for /p/. For the Koreans, whose L1 has three bilabial phonemes and no potentially confusable labial sounds, /p/ was an informative cue in both experiments.

In contrast to /p/, the nonlabials /t,k/ were informative cues visually. These articulations are less marked than the labial; for some observers, they are not visually distinguishable from one another. They clearly contributed the most to perception for NNSs when presented with auditory /p/ and, in some cases, /f/ and /w/. Acoustically, given their higher frequency and stronger intensity, they appear to possess more robust acoustic properties. As auditory cues, they underwent variable visual influence of /p/ for those groups for whom /p/ was an informative cue—the Kore-

ans, and the Spanish in Experiment 2.

In these two experiments, the NNS groups showed consistency in their evaluation of auditory /p/ and visual /t,k/ but varied in their assessments of the information value of visual /p/. The fact that /p/ in general was a less informative cue than /t,k/ across L1 groups may be related to its relative markedness; however, the variable performance across groups in response to visual /p/ appears related to L1 factors.

In summary, for the NNSs, the influence of a visual cue was dependent upon its information value, the intelligibility of the auditory cue, and the assessment of similarity between the cues. On this point, the McGurk effect in adult L2 learners is comparable to that in young children whose L1 is English. Massaro (1984, 1987) found differences in the information value of visual cues in speech perception for children, when compared with adults, presented with synthetic speech syllables along a /bɑ/-/dɑ/ continuum. Auditory information was as effective for the children (averaging 6 years of age) as for the adults; however, smaller visual effects obtained for the children. On an additional task, Massaro (1984, 1987) also required the children to indicate whether or not they saw the speaker's lips move. Drawing attention to the visual modality did not increase the relative influence of the visual information. This result was compatible with the poorer performance of children in a separate lipreading task using the same syllables (Massaro, Thompson, Barron, & Laren, 1986). Visual cues were less informative alone; consequently, they contributed less to bimodal speech perception. At the initial feature evaluation stage of FLMP, the visual source of input is less informative for young children, but the information value increases with age and experience.

The information values of cues and linguistic experience are factors to consider in interpreting Sekiyama and Tohkura's (1993) findings, especially in comparing them with the present study's. Sekiyama and Tohkura claimed that Japanese perceptual processing is different; that is, Japanese speakers do not incorporate visual cues into the percept as long as there is enough auditory

information to identify the speech. They attributed minimal visual effects for the Japanese listening to Japanese syllables to the listeners' "vision-independent" perceptual processing, not to the stimulus characteristics of the syllables.

However, the present study and those by Massaro (1987) and Massaro et al. (1993) bring several issues to bear on the above claim. First, examining the NS findings of my Experiment 1, while disregarding previous McGurk effect experiments with English speakers, could lead to the same conclusion, that is, when auditory intelligibility is 100%, few visual effects are noted. However, comparing the results of Experiments 1 and 2 provides evidence of the role of compellingness in establishing perceptual unity for the occurrence of the McGurk effect (Easton & Basala, 1982; Massaro, 1987). Perhaps a stimulus set limited to stops in Japanese would also show a different pattern of results for Japanese speakers.

Second, assuming that eye gaze is directed toward the speaker's face and lip movements are observed, *and* a compelling condition is present, a lack of visual effects would suggest that individuals can selectively attend to one modality, completely ignoring the other. Experiments using selective and divided attention techniques with English-speaking adults have demonstrated that this is not the case (Massaro, 1987). In addition, Massaro et al. (1993) reported a similar McGurk effect for Japanese, English, and Spanish speakers, although these experiments involved synthesized syllables and visual articulations from an animated face. Perhaps attentional factors need to be explored for Japanese speakers.

Third, it has been stated that Japanese speakers do not generally use visual cues in face-to-face communication (Sekiyama & Tohkura, 1993). Although Japanese speakers are apparently able to distinguish between labials and nonlabials in speechreading (Sekiyama et al., 1988), we must consider the information value of the individual visual articulations of specific sounds. This factor is crucial to the integration process. Perhaps as a result of cultural differences in speech communication style, phonetic

features rather than visual ones contribute more to the classification of a given sound for the Japanese; thus, what appear to be processing differences can be attributed to differences in the information value of visual cues as a function of experience. The variable information value of cues for Japanese speakers observing English syllables is evident in a comparison of the responses to AE visual /r/ matched with auditory /r/ given by Japanese speakers in Japan and by those in the U.S. who participated in my Experiment 1. For those in Experiment 1, visual /r/ had greater value and thus increased identification accuracy of auditory /r/, even in the absence of explicit instruction. This difference may well be attributable to their daily contact with spoken English, analogous to the greater magnitude of the McGurk effect in English-speaking adults with more linguistic experience than in children (Massaro, 1987). The survey results from Hattori (1987) further support the role of linguistic experience in an English-speaking environment as a factor in determining the information value of speech cues for Japanese learners of English.

Given the above, is it possible to enhance the information value of visual cues to speech perception for nonnatives? Auditory training studies have shown significant improvement over only a three-week period in the identification of /r/ and /l/ in various word positions by Japanese speakers in the U.S. (Lively et al., 1993; Logan et al., 1991), and by Japanese in Japan, whose accuracy decreased only 2% when performance on the posttest at the end of training was compared with the posttest given three months later (Lively et al., 1994). Moreover, auditory training in the perceptual identification of /r/ and /l/ significantly improved /r/ and /l/ production for Japanese speakers in Japan in the absence of explicit production training (Bradlow, Pisoni, Yamada, & Tohkura, 1995).

From these studies and from results of the present study, it becomes apparent that enhancing the information value of visual articulations through training is an area of L2 instruction to be explored further. In Experiment 1, the addition of the visual /f/ and /r/ cues with auditory /f/ and /r/, respectively, increased the accuracy of identification for the Japanese and Korean learners

for whom auditory intelligibility had been less than 100%. For the Japanese, the presence of both visual /w/ and /r/ contributed significantly to the identification of /r/ ($p<.05$). These findings point to the learners' attending to visual cues to identify speech sounds, even though they had not received any explicit training in the visual recognition of English sounds. Although the number of visual distinctions that can be made among the consonants is smaller than the total number of consonants in the language (Walden et al., 1977), training studies in the speechreading literature (e.g., Walden et al., 1977), preliminary findings of present research, and my own ESL teaching experience, all suggest that the critical attributes of visual articulations lend themselves well to training.

Nevertheless, several textbooks for ESL instruction at the beginning, intermediate, and advanced levels variably use visual articulatory information for production purposes, and make few references to the visual modality of speech for perception purposes. Some authors have provided for production assistance a few diagrams of tongue positions typical of certain vowels (e.g., Baker & Goldstein, 1992), or diagrams of lips and teeth as guidelines to sound production (e.g., Gilbert, 1993). Generally, books targeted toward more advanced ESL learners contain more information. Lane (1993) included two exercises on "Mouth Shapes", with the suggestion that the instructor produce the words without sound for students to lipread. These exercises involve labial consonant contrasts in initial position in words (e.g., /b/-/v/-/w/) and the vowels /ɑ/ and /ə/. Prator and Robinett (1985) provided descriptions of movements and diagrams of lip positions for pronunciation purposes. With descriptions of vowel production, they suggested that students watch their lips in a mirror to compare their articulations with the descriptions given in the book. They gave two exercises on lipreading. One focuses on distinctions between /p/ and /f/ in CV syllables with different back vowels and /ə/; the other concerns consonants /b,v,w,ð,d/ in CV syllables with the above vowels. According to Prator and Robinett, these exercises were designed "to help fix in your mind the position

in which the consonants . . . and the back vowels are formed" (p. 170). Many such books are accompanied by audio cassettes; thus, use of visual information, live or on videotape, is up to the instructor (see also Samuda, 1993, for a collection of reviews on pronunciation textbooks).

Summary

The two experiments on the McGurk effect reported here clearly demonstrate the power of visible speech for NSs and NNSs of English. However, the effect's occurrence is subject to conditions. Integration of auditory and visual cues in speech perception occurs when observers can assume these cues contribute to a unitary perceptual event; that is, that they create a compelling condition. Further, the NS-NNS contrast shows that the determination of what constitutes a compelling condition is influenced by the observers' perceptual categories, native or nonnative. These representations, with visual and auditory information, permit evaluation of cues, assessment of similarity, and integration. Results are compatible with the framework of FLMP (e.g., Massaro, 1987) in that the most informative (least ambiguous) cue contributes the most to the perceptual outcome. For the NNSs, the L1 influenced not only the auditory intelligibility of the consonants, but also the information value of the visual cues which, in turn, affected A-V integration.

The literature in the field of L2 speech perception has been dominated by an interest in exploring only one of the two modalities of speech. Although the visual effects on speech perception found in the present study involved ESL learners, the contribution of visual input in audiovisual contexts does not appear to be limited to English. Given that AE speakers with no knowledge of Japanese showed visual effects on the perception of Japanese syllables (Sekiyama & Tohkura, 1993), visual cues to speech sounds represent a potential second channel of input for L2 learners in general. However, their use crucially rests upon their information value to the learners. Additional studies are needed

to evaluate the use of visual cues for learners of other languages. A better understanding of how L2 learners use not only acoustic-phonetic input but also visual input will provide a foundation on which to extend the enhancement of audiovisual input from speech perception to spoken word recognition in connected speech.

Revised version accepted 12 September 1995

Notes

[1]Omar (1982) refers to Malay as the national and official language of Malaysia. It is the language of education, with English being accorded the status of the second most important language. Omar noted that Malay has been given other political designations, specifically, *bahasa Malaysia* and *bahasa Indonesia*. In this paper, I have adopted the term used on the response sheets in this study by the majority of the speakers who participated (i.e., Malay).

[2]This paper uses a dichotomous designation of labial-nonlabial to classify the sounds involved in the study. A reviewer has suggested that /r/ be considered labial. The term *labial* refers to the active use of one or both lips in the production of a sound. According to Catford (1977), there must be a marked movement from resting position for a sound to qualify as labial (e.g., /w/). In traditional feature terms, /r/ is [–labial]. In addition, many consonants are labialized in the context of a rounded vowel (Ladefoged, 1982). Labialization, as a secondary articulatory phenomenon, is context-dependent. In the CV syllable used in this experiment, the context created by the open vowel /a/ is minimally coarticulatory. Measurements made by Daniloff and Moll (1968) revealed small protrusion effects and speaker variation for /r/ before unrounded vowels. Although the articulation of /r/ involves some degree of lip protrusion in the syllable /ra/ in this experiment, it is not considered here to be labial.

[3]Silva (1992) reported that the intersonorant lax stop in Korean has two percepts—voiced or voiceless—depending upon its prosodic position. This was based on native speaker judgments and three acoustic measures (duration of voicing during closure, degree to which the stop closure is voiced, post-release VOT). At the edge of a phonological phrase (or "minor phrase"), the stops are lightly aspirated, have the longest VOT values, and are judged to be voiceless. Word-internal lax stops are unaspirated, have shorter VOT values, and are judged to be voiced (see Silva for details).

[4]The term *homophenous* comes from *homophene* referring to a word that looks like another in its production (Jeffers & Barley, 1971). It should not be confused with *homophone*, a word that sounds like another.

[5]*Auditory intelligibility* refers to the identification of a sound. It is con-

trasted with the term *audibility*, which refers simply to a determination of presence or absence (see Miller, Heise, & Lichten, 1951). In the visual domain of speech, *discernability* is frequently used to refer to the visual distinctiveness of an articulatory gesture.

[6]At the advanced level in this program, students may take academic classes part time along with English classes, depending upon several factors that are evaluated on an individual basis. Some of the participants in this experiment had reached the required TOEFL score for college admission; others were preparing to return to their countries after completing the final session in our English program.

[7]In the chi-square test for statistical significance here, expected frequencies (f_e) are the correct responses obtained in the audio-alone condition. This test was selected to determine if, in the presence of a given visual cue, the frequency of correct responses (accuracy of identification) was significantly different than that in the audio-alone condition. Analysis of erroneous responses rather than correct responses would have provided empty cells in many cases in the native and nonnative data.

[8]The majority of the Spanish speakers who participated in this study were from Venezuela; five were from Colombia. A feature of the Spanish spoken along the coastal areas of these countries and others in Latin America is the loss of a distinction between /r/ and /l/ in the speech of some individuals (Resnick, 1975). The data presented by Resnick were compiled from previously published sources. With regard to these sounds, the data excluded word-initial position following a pause that is the environment of /r/ in the present study. The loss of the /r/-/l/ distinction is apparent primarily in medial position and is often associated with persons having less formal education (B. Ruiz, personal communication, July, 1995). No dialectal variants were given by Resnick for the other sounds in this study.

[9]In the chi-square analysis of the relationship between L1 and distribution of responses to each stimulus, the largest $^{(O-E)^2}/_E$ value for each cell indicated where the difference between the observed and expected frequencies was the greatest (where O=observed frequency, and E=expected frequency). This also indicated the frequency that made the largest contribution to the chi-square value (see Hatch & Lazaraton, 1991).

[10]In the categorization literature, a distinction sometimes is made between the term *feature* to designate binary-valued contrasts and *dimension* to describe concepts where quantitative variations need to be described (see e.g., Smith & Medin, 1981). In this paper, following Estes (1986) and Massaro (1987), no distinction is made between these terms.

[11]There is a less common view that velars are the least marked. Readers are referred to Paradis and Prunet (1991) for a thorough discussion of the evidence in support of coronals as least marked.

References

Baker, A., & Goldstein, S. (1992). *Pronunciation pairs*. Cambridge: Cambridge University Press.

Berger, K. W. (1972). *Speechreading: Principles and methods*. Baltimore: National Education Press.

Berger, K. W (1973). Consonant confusions in speechreading. *Ohio Journal of Speech and Hearing*, 8, 10–18.

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

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics, 20*, 305–330.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech & Hearing Research, 17*, 619–630.

Bloch, B. (1950). Studies in colloquial Japanese IV: Phonemics. *Language, 26*, 86–125.

Borden, G., Gerber, A., & Milsark, G. (1983). Production and perception of the /r/-/l/ contrast in Korean adults learning English. *Language Learning, 33*, 499–526.

Bradlow, A. R., Pisoni, D. B., Yamada, R., & Tohkura, Y. (1995). Acquisition of the English /r/-/l/ contrast by Japanese speakers: Effects of training in perception on production. In *Proceedings of the XIII International Congress of Phonetic Sciences* (Vol. 4, pp. 562–565). Stockholm: Stockholm University.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press.

Cho, Y.-M. Y. (1987). Phrasal phonology of Korean. In S. Kuno, I.-H. Lee, J. Whitman, S.-Y. Bak, & Y.-S. Kang (Eds.), *Harvard studies in Korean linguistics II* (pp. 328–340). Cambridge, MA: Harvard University, Department of Linguistics.

Cho, Y.-M. Y. (1991). On the universality of the coronal articulator. In C. Paradis & J.-F. Prunet (Eds.), *The special status of coronals: Internal and external evidence* (pp. 159–179). San Diego, CA: Academic Press.

Daniloff, R., & Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech & Hearing Research, 11*, 707–721.

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

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

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

Estes, W. K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General, 115*, 155–174.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech & Hearing Research, 11*, 796–804.

Flege, J. E. (1988). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), Human communication and its disorders, A review (pp. 224–401). Norwood, NJ: Ablex.

Flege, J. E. (1989). Chinese subjects' perception of the word-final English /t/-/d/ contrast: Performance before and after training. *Journal of the Acoustical Society of America, 86*, 1684–1697.

Flege, J. E. (1991). Perception and production: The relevance of phonetic input to L2 phonological learning. In T. Huebner & C. A. Ferguson (Eds.), *Crosscurrents in second language acquisition and linguistic theories* (pp. 249–289). Amsterdam: John Benjamins.

Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics, 15*, 67–83.

Gilbert, J. (1993). *Clear speech: Pronunciation and listening comprehension in North American English* (2nd ed.). New York: Cambridge University Press.

Gillette, S. (1980). Contextual variation in the perception of L and R by Japanese and Korean speakers. *Minnesota Papers in Linguistics and the Philosophy of Language, 6*, 59–72.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "l" and "r". *Neuropsychologia, 9*, 317–323.

Green, K. P., Kuhl, P. K., & Meltzoff, A, N. (1988). Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. *Journal of the Acoustical Society of America, 84*(S1), S155.

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 & Performance, 17*, 278–288.

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

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Hattori, T. (1987). A study of nonverbal intercultural communication between Japanese and Americans—Focusing on the use of the eyes. *Japan Association of Language Teachers, 8*, 109–118.

Henly, E., & Sheldon, A. (1986). Duration and context effects on the perception of English /r/ and /l/: A comparison of Cantonese and Japanese speakers. *Language Learning, 36*, 505–521.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /θ/-/ð/ contrast by francophones. *Perception & Psychophysics, 40*, 205–215.

Jamieson, D. G., & Morosan, D. E. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology, 43*, 88–96.

Jeffers, J., & Barley, M. (1971). *Speechreading (lipreading)*. Springfield, IL: Charles C. Thomas.

Jung, M.-W. (1962). *A contrastive study of English and Korean segmental phonemes with some suggestions toward pedagogical application*. Unpublished master's thesis, Georgetown University, Washington, DC.

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics, 21*, 3–28.

Kean, M.-L. (1980). The theory of markedness in generative grammar. Bloomington, IN: Indiana University Linguistics Club Publications. (Doctoral dissertation, MIT, 1975).

Kim-Renaud, Y.-K. (1974). Korean consonantal phonology (Doctoral dissertation, The University of Hawaii, 1974). Dissertation Abstracts International, 36/02, p. 869.

Koo, J. H. (1970). Language conflict and its specification: Experimental study (Doctoral dissertation, Indiana University, Bloomington, 1970). Dissertation Abstracts International, 32/02, p. 948.

Koo, J. H. (1972). Language universals and the acquisition of an unfamiliar sound. *International Review of Applied Linguistics in Language Teaching, 10*, 145–152.

Kuhl, P. K. (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototype of speech categories, monkeys do not. *Perception & Psychophysics, 50*, 93–107.

Kuhl, P. K., Green, K. P., & Meltzoff, A. N. (1988). Factors affecting the integration of auditory and visual information in speech: The level effect. *Journal of the Acoustical Society of America, 83*(S1), S86.

Ladefoged, P. (1982). *A course in phonetics* (2nd ed.). New York: Harcourt Brace Jovanovich.

Lane, L. (1993). *Focus on pronunciation: Principles and practice for effective communication*. New York: Longman.

Lapoliwa, H. (1981). *A generative approach to the phonology of bahasa Indonesia*. Canberra: Australian National University, Research School of Pacific Studies, Department of Linguistics.

Lehiste, I. (1964). Acoustic characteristics of selected English consonants. *International Journal of American Linguistics, 30*, 10–115.

Lindau, M. (1985). The story of /r/. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 157–168). Orlando, FL: Academic.

Lively, S. E. (1993). *On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception* (Res. Rep. No. 96). Bloomington: Indiana University, Cognitive Science Program.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*, 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III: Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America, 96*, 2076–2087.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89*, 874–886.

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

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics, 2*, 369–390.

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English 'l' and 'r'. *Cognition, 24*, 169–196.

Martin, S. E. (1951). Korean phonemics. *Language, 27*, 519–533.

Massaro, D. W. (1984). Children's perceptions of visual and auditory speech. *Child Development, 55*, 1777–1788.

Massaro, D. W. (1985). Attention and perception: An information-integration perspective. *Acta Psychologica, 60*, 211–243.

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

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

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

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology, 41*, 93–113.

McGurk, H. (1981). Listening with eye and ear. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 336–337). Amsterdam: North-Holland.

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

Medin, D. L., & Barsalou, L. W. (1987). Categorization processes and categorical perception. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 455–490). Cambridge: Cambridge University Press.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review, 85*, 207–238.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of test materials. *Journal of Experimental Psychology, 41*, 329–335.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338–352.

Miyawaki, K., Strange, W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of /r/ and /l/ by native speakers of Japanese and English. *Perception & Psychophysics, 18*, 331–340.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics, 9*, 283–303.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco: W. H. Freeman.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13*, 87–108.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85*, 172–191.

Omar, A. H. (1982). *Language and society in Malaysia*. Kuala Lumpur: Dewan Bahasa dan Pustaka, Kementerian Pelajaran.

Paradis, C., & Prunet, J. F. (Eds.). (1991). *The special status of coronals: Internal and external evidence*. San Diego, CA: Academic Press.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication, 13*, 109–125.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception & Performance, 8*, 297–314.

Prator, C. H., Jr., & Robinett, B. W. (1985). *Manual of American English pronunciation* (4th ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Price, P. J. (1981). A cross-linguistic study of flaps in Japanese and in American English (Doctoral dissertation, University of Pennsylvania, 1981). *Dissertation Abstracts International, 42*, 1128A.

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

Resnick, M. C. (1975). *Phonological variants and dialect identification in Latin American Spanish*. The Hague, The Netherlands: Mouton.

Robson, B. (1982). Teaching English pronunciation to speakers of Korean. *Focus, 11*, 1–8.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192–233.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.) *Cognition and categorization: Papers from a conference sponsored by Social Science Research Council, Lake Arrowhead, CA, Spring 1976* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Samuda, V. (Ed.). (1993). [Reviews of pronunciation textbooks]. *TESOL Quarterly, 27*, 757–776.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics, 31*, 307–314.

Sekiyama, K., Joe, K., & Umeda, M. (1988). *Perceptual components of Japanese syllables in lipreading: A multidimensional study* (Tech. Rep. IE87–127). The Institute of Electronics, Information and Communication Engineers of Japan. (In Japanese with English abstract)

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America, 90*, 1797–1805.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21*, 427–444.

Sheldon, A. (1985). The relationship between production and perception of the /r/-/l/ contrast in Korean adults learning English: A reply to Borden, Gerber, and Milsark. *Language Learning, 35*, 107–113.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics, 3*, 243–261.

Silva, D. J. (1989). Korean obstruent voicing: Investigations into prosodic structuring. In S. Kuno, I.-H. Lee, J. Whitman, S.-Y. Bak, & Y.-S. Kang (Eds.), *Harvard studies in Korean linguistics III* (pp. 177–188). Cambridge, MA: Harvard University, Department of Linguistics.

Silva, D. J. (1992). A phonetically based analysis of [voice] and [fortis] in Korean. In P. M. Clancy (Ed.), *Japanese/Korean linguistics* (Vol. 2, pp. 164–174). Stanford, CA: Center for Study of Language and Information, Stanford Linguistic Association/Stanford University.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Stockwell, R. P., & Bowen, J. D. (1965). *The sounds of English and Spanish*. Chicago: University of Chicago Press.

Strange, W., & Dittman, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics, 36*, 131–145.

Strange, W., & Jenkins, J., Jr. (1978). Role of linguistic experience in perception of speech. In R. D. Walk & H. L. Pick (Eds.), *Perception and experience* (pp. 125–169). New York: Plenum.

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

Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany, NY: State University of New York Press.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20*, 130–145.

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics, 30*, 557–564.

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

Woodward, M. F., & Barber, C. G. (1960). Phoneme perception in lip reading. *Journal of Speech & Hearing Research, 3*, 212–222.

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics, 52*, 376–392.