Perception & Psychophysics
1997, 59 (1), 73-80

Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects

KAORU SEKIYAMA
Kanazawa University, Kakuma-machi, Kanazawa, Japan

The "McGurk effect" demonstrates that visual (lip-read) information is used during speech perception even when it is discrepant with auditory information. While this has been established as a robust effect in subjects from Western cultures, our own earlier results had suggested that Japanese subjects use visual information much less than American subjects do (Sekiyama & Tohkura, 1993). The present study examined whether Chinese subjects would also show a reduced McGurk effect due to their cultural similarities with the Japanese. The subjects were 14 native speakers of Chinese living in Japan. Stimuli consisted of 10 syllables (/ba/, /pa/, /ma/, /wa/, /da/, /ta/, /na/, /ga/, /ka/, /ra/) pronounced by two speakers, one Japanese and one American. Each auditory syllable was dubbed onto every visual syllable within one speaker, resulting in 100 audiovisual stimuli in each language. The subjects' main task was to report what they thought they had heard while looking at and listening to the speaker while the stimuli were being uttered. Compared with previous results obtained with American subjects, the Chinese subjects showed a weaker McGurk effect. The results also showed that the magnitude of the McGurk effect depends on the length of time the Chinese subjects had lived in Japan. Factors that foster and alter the Chinese subjects' reliance on auditory information are discussed.

Although it is often assumed that speech perception is a purely auditory event, in face-to-face communication it is more of a multimodal process. Humans can read lips fairly well, and this lip-read information plays a role in speech perception. When speech is not clear, looking at a speaker's face normally improves speech perception (e.g., Sumby & Pollack, 1954) because visual information is helpful in knowing the place of articulation, which is difficult to know from auditory information alone (Binie, Montgomery, & Jackson, 1974). Speech sounds are articulated at various places along the vocal tract, such as at the lips, teeth, palate, and glottis. Studies on lipreading (e.g., Sekiyama, Joe, & Umeda, 1988; Walden, Prosek, Montgomery, Scherr, & Jones, 1977) have shown that the human visual system is highly sensitive to the distinction between labials (lip-articulated sounds, such as /b/ and /m/) and nonlabials (sounds articulated behind the lips, such as /d/ and /n/).

This study was supported by a long-term fellowship from the International Human Frontier Science Program (LT-277/93), a grant-in-aid for scientific research from the Japanese Ministry of Education, Science, and Culture (04851019), and a Sasakawa scientific research grant from the Japan Science Society (4-053). The author is grateful to Yoh'ichi Tohkura for his help in creating the video stimuli at ATR Auditory and Visual Perception Research Laboratories. The author also would like to thank Arthur Samuel, Winifred Strange, Joanne Miller, Peter Edwards, Yasuko Hayashi, Edgar Pope, and one anonymous reviewer for their helpful comments on the earlier version of the manuscript. The author's address is as follows: Kanazawa University, Kakuma-machi, Kanazawa 920-11, Japan (e-mail: sekiyama@kenroku.ipc.kanazawa-u.ac.jp).

The McGurk-effect experiment (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978) is a paradigm used to explore how people integrate visual and auditory information during speech perception. The McGurk effect may occur when the auditory and visual stimuli differ in terms of place of articulation. For example, when auditory /pa/ is dubbed onto visual lip movements of /na/, perceivers often report hearing "ta." In this example, auditory /p/ is labial whereas visual /n/ is nonlabial. The perceived sounds ("t" in this example) are often consistent with the visual stimulus in terms of place of articulation, while remaining consistent with the auditory stimulus in terms of manner of articulation.

The McGurk effect demonstrates that lip-read information plays a role even when the speech is clear. While this effect has been shown to be robust in English-speaking cultures (e.g., Dekle, Fowler, & Funnell, 1992; Green, Kuhl, Meltzoff, & Stevens, 1991; Massaro & Cohen, 1983; Rosenblum & Saldaña, 1992), we have found interlanguage differences between native speakers of Japanese and native speakers of American English. In our first study, native speakers of Japanese hardly showed the McGurk effect when listening to very clear Japanese speech, although these Japanese subjects showed a highly increased McGurk effect when auditory noise was added to the stimuli (Sekiyama & Tohkura, 1991). We also conducted a cross-language study in which native speakers of Japanese and native speakers of American English were compared. The Japanese showed a much weaker McGurk effect than did the Americans for clear Japanese stimuli, indicating their tendency to rely on the auditory information (Sekiyama & Tohkura, 1993; also see Sekiyama,

1994a). Another finding of this cross-language study was that the McGurk effect was stronger for foreign speech stimuli than for native speech stimuli. This effect of stimulus language, that is, the native–foreign language effect, seems to be due to some auditory ambiguity induced by the foreign speech. The stronger visual effect for foreign speech stimuli appears to be consistent with our daily experience that listening to foreign speech is more difficult on the telephone than in face-to-face communication although we rarely notice such a difference in the case of listening to native speech.[1]

When the stimuli involved clear Japanese speech and were composed of conflicting auditory and visual syllables, the Japanese subjects often reported incompatibility between what they heard and what they saw, instead of showing the McGurk effect (Sekiyama, 1994a). This implies that the visual information is processed to the extent that the audiovisual discrepancy is detected most of the time. It suggests that, for clear speech, the Japanese use a type of processing in which visual information is not integrated with the auditory information even when they extract some lip-read information from the face of the speaker. The results for the noise-added condition indicate that the auditory ambiguity calls for visual compensation, resulting in a switch to another type of processing where audiovisual integration occurs.

The weaker McGurk effect in the Japanese perception of clear Japanese stimuli shows that there are cultural and/or linguistic factors which affect the manner of audiovisual integration. As for cultural factors, it is often said that the Japanese tend to avoid looking at the face of a person they are listening to because in Japan it is thought to be impolite to stare at a person's face when that person is of a higher status. This cultural habit may cause the Japanese to develop a type of processing which does not integrate visual with auditory information even when they are looking at the speaker's face.

As for linguistic factors, number of phonemes used may be considered. The Japanese phonemes might easily be distinguished without additional visual cues because a smaller number of vowels, consonants, and syllables occur in Japanese than in, for example, English (in Japanese, there are only 100 possible syllables, and no consonant clusters are allowed). Related to this, lip-read information may be less useful in Japanese than in English because the Japanese consonant inventory does not have labiodentals (/f/, /v/, /θ/, /ð/), which are easy to lip-read and form a perceptual group that is distinct from both labials (/b/, /p/, /m/, /w/) and nonlabials (/d/, /n/, /g/, etc.) in English (Walden et al., 1977). These linguistic characteristics also possibly lead to the type of perceptual processing described above.

A direct test of these two hypotheses may be to examine native speakers of such languages as Hawaiian, Spanish, or Italian, because their native languages are phonetically similar to Japanese with respect to the linguistic aspects noted above but their cultures do not emphasize face avoidance. Massaro, Tsuzaki, Cohen, Gesi, and Heredia (1993) tested native speakers of Japanese, Spanish, and American English by using synthesized audible speech and a computer-animation face. In some conditions that were comparable to our natural-speech experiment, the visual (McGurk) effect was weaker in the Japanese subjects than in the native speakers of Spanish or American English. Similarly, Fuster-Duran (1996) found a strong McGurk effect in Spanish subjects. From the viewpoint of phonemic and syllabic structures, the Spanish language seems to resemble Japanese. Spanish has the same number of vowels (five) as Japanese, and it has repeated occurrences of CV (consonant + vowel) syllables in a phrase, as in Japanese. Thus, the strong McGurk effect observed in the Spanish-speaking subjects suggests that the phonological structure of a subject's native language does not determine the strength of the McGurk effect.

If this is true, then the cultural (face-avoidance) hypothesis seems to be plausible. This hypothesis predicts that subjects from a Western culture which does not practice face avoidance will be more susceptible to the McGurk effect than will Japanese subjects. In accordance with this prediction, Massaro, Cohen, and Smeele (1995) reported that the performance of native speakers of Dutch was equivalent to that of native speakers of American English.

To examine the face-avoidance hypothesis further, the present study tested native speakers of Chinese. This language group was chosen on the basis of generally accepted views of cultural similarities between Chinese and Japanese and differences between Eastern and Western cultures. Although it is hard to quantitatively measure the degree of face avoidance in various cultures, cultural anthropologists have observed that Asian cultures generally include face avoidance as an expression of social status (e.g., Nomura, 1984). Also in my informal inquiry, several Chinese people who know both the Chinese and American cultures indicated that they did not stare at the face of the person they are listening to as much as Western people do. Thus, the Chinese are at least more similar to the Japanese than they are to the Americans in terms of face avoidance. The current study tested the face-avoidance hypothesis by testing Chinese subjects on the Japanese and English stimuli used in our earlier study (Sekiyama & Tohkura, 1993). The prediction was that the Chinese, who are culturally close to the Japanese, would also show a weaker McGurk effect than the previously tested American subjects.

In view of the native–foreign language effect mentioned earlier, it might be better to also create Chinese stimuli with which to test the native speakers of Chinese (but also see Note 1). Of course, this should be done further if the Chinese show a strong McGurk effect, because the native–foreign language effect for the Japanese and English stimuli actually works against finding a weak McGurk effect for the Chinese subjects. This fact, however, also implies that using Chinese stimuli is not a necessary condition for testing the face-avoidance hypothesis

if the Chinese show a weak McGurk effect for the Japanese and English stimuli.

Ideally, for the purpose of the current test, the Chinese subjects should have no experience of living in a foreign country. We actually applied this requirement to the American and Japanese subjects in our previous study (Sekiyama & Tohkura, 1993). For practical reasons, however, the subjects in this study were native speakers of Chinese who were living in Japan. Accordingly, close attention was paid to determine if the subjects' various linguistic and cultural experiences affected the results.

METHOD

Subjects

Fourteen native speakers of Chinese, under 30 years of age, were recruited from the Kanazawa University community. They reported that they had normal hearing and normal or corrected-to-normal vision. Most of them were graduate students who had arrived in Japan after finishing college in China. One was from Taiwan. Their ages ranged from 19 to 30 years, and the length of their stay in Japan was between 4 months and 6 years. All of them spoke Japanese, but a few of them had just enough proficiency in Japanese to understand the experimental task. They had never lived in a foreign country other than Japan. Although most of them had taken English classes, from when they were about 12 years old, many of them reported that they did not speak English. One of them complained that even a tape recorder was not available in his school when he started to learn English before the economic liberation by Ten Xiao Ping. Judging from this episode, the English proficiency of these Chinese subjects was probably not much different from that of our previous Japanese subjects (Sekiyama, 1994a; Sekiyama & Tohkura, 1991). Although the subjects' native languages included various Chinese dialects, all of them had been educated in Mandarin Chinese since entering elementary school. They were paid 6,000 yen for the 4-h experiment conducted over 2 days.

Stimuli

The stimuli were the same as those used by Sekiyama and Tohkura (1993). They consisted of 10 syllables (/ba/, /pa/, /ma/, /wa/, /da/, /ta/, /na/, /ra/, /ga/, and /ka/) and were pronounced by a female Japanese speaker for Japanese stimuli and by a female American speaker for English stimuli. Each speaker's face was videotaped while she pronounced the syllables. Her utterances were rerecorded in an anechoic room to obtain the auditory stimuli. The audio signals of the original videotape were replaced by these rerecorded signals. The 10 visual and 10 auditory syllables were combined using a BETACAM video system that can handle frame-by-frame time control (33 msec). To synchronize the rerecorded and original audio signals, the rerecorded signals were dubbed onto the frames where the original audio signals had been. The onsets of the energy were synchronized. For the technical details of the dubbing, see Sekiyama (1994a).

In the dubbing, the auditory syllables were combined only with the visual syllables from the same speaker, but an auditory syllable was dubbed onto all the visual syllables so that all possible combinations of 10 auditory and 10 visual syllables ($10 \times 10 = 100$) were produced. On the final videotapes ("AV tapes"), each stimulus occurred in a 7-sec trial in which the video channel included 3 sec of black frames and 4 sec of the face. The final videotapes were copied onto VHS videotapes for use in the experiments.

Videotapes for auditory-only presentation were also created ("A tapes"), with black frames only on the video channel. For the auditory-only presentation, a videotape included six repetitions of 10 auditory stimuli.

The visual stimuli were presented on a 20-in. color monitor on which an approximately life-sized speaker appeared. Auditory stimuli were presented through two loudspeakers placed at the sides of the monitor. The subjects viewed the monitor from a distance of 1 m.

Experimental Design

Each of the 14 subjects participated in the two sessions of the experiment. One session was for the Japanese stimulus set; the other was for the English set. The order of these two stimulus sets was counterbalanced between subjects. For each stimulus set, there were three conditions: audiovisual (AV), auditory only (A), and visual only (V), which were conducted in that order. The AV tapes and the A tapes were played in the AV and A conditions, respectively. In the V condition, only the video outputs of the AV tapes were played by turning off the main amplifiers of the loudspeakers.

Procedure

The videotapes were played on a VHS videocassette machine located in a control room adjoining the room in which the subjects were tested. The stimuli were presented once every 7 sec in random order. In the AV condition, the subjects were instructed to write down what they thought they had heard while looking at and listening to each syllable. It was an open set response. They were instructed to write in *pin-yin*, which they had been taught in elementary school for spelling Chinese syllables to approximate the Roman alphabet. They were also asked to report any recognized incompatibility between what they heard and what they saw by checking a column on their response sheets. In the A condition, the subjects' task was to report only what they had heard. In the V condition, they were asked to lipread and report what they thought the speaker was pronouncing.

For each stimulus set, the AV condition was conducted in six blocks of 100 trials, and the A and V conditions were conducted in one block of 60 or 100 trials. It took 2 h to conduct the three conditions for one stimulus set. Each subject participated in the experiment for 2 days, each of which was either for the English or the Japanese stimulus set.

RESULTS

Responses in the A Condition

Confusion matrices for the A condition are shown in Figure 1. Each number is the percentage of the responses in a row. Whereas most of the consonants were accurately identified, /r/ in both Japanese and English showed a large number of confusions. English /w/ also showed some confusions with "r" and "l." The data show that English /r/ was perceptually similar to "w" whereas Japanese /r/ was perceptually similar to "l." The "w" responses for the English /r/ had also been observed in Japanese subjects previously (Sekiyama, 1994b). One characteristic was observed for Chinese subjects only: 7%–8% "l" responses for /n/ in both Japanese and English.

Responses in the V Condition

Figure 2 shows confusion matrices for the V condition. It is clear that for both the Japanese and English stimuli, visual labials which have a bilabial closure (/b/, /p/, /m/) were well discriminated from nonlabials (/d/, /t/, /n/, /g/, /k/) by lipreading. Visual /w/, which shows lip protrusion, was perceived as a distinct category in both languages. The Japanese /r/ was perceived as one of the nonlabials,

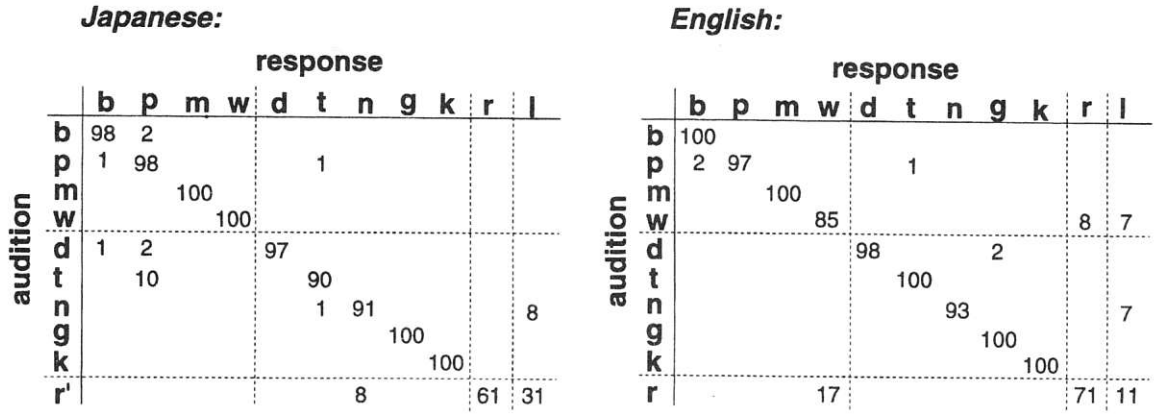

Figure 1. Confusion matrices in the auditory-only (A) condition for Japanese and English stimuli. Each number indicates the percentage of the responses in a row.

whereas the English /r/ was perceived as being in the same category as /w/.

Responses in the AV Condition

In the analyses of the data for the AV condition, the data for stimuli for which the auditory or visual component was /r/ or /w/ were excluded due to the auditory or visual differences between Japanese and English.

The data for the AV condition are presented separately for two cases: the case in which the visual stimuli were labials (/b/, /p/, /m/) and the case in which they were nonlabials (/d/, /t/, /n/, /g/, /k/). The response patterns were almost identical within these two groups (visual labials and visual nonlabials), agreeing with those obtained by Sekiyama and Tohkura (1991). In the confusion matrices in Figure 3, the numbers in parentheses in the leftmost column are the percent correct identifications of the auditory syllable in the A condition which provides a baseline for the examination of visual effects in the AV condition. The responses in the shadowed sections indicate the "gross" McGurk effect. These are errors in terms of audition, and their place of articulation is consistent with that of the visual input. The gross McGurk effect is seen for /b/, /p/, /m/, and /t/ for both Japanese and English about 40% of the time.

The Magnitude of the McGurk Effect

The magnitude of the "pure" McGurk effect was calculated by subtracting the auditory place errors from the gross McGurk effect. For example, when combined with visual labials, Japanese auditory /t/ produced place errors ("p" or "b" responses) 42% of the time, and these are counted as the gross McGurk effect. However, this /t/ also produced "p" responses 10% of the time in the A condition (auditory confusions). Thus, the magnitude of the pure McGurk effect (visual effect) was 42 − 10 = 32%.

Figure 4 (top panel) shows the magnitude of the pure McGurk effect. Also shown in the middle and lower panels are the results for the Japanese and American subjects in Sekiyama (1994b). The McGurk effect was weaker for the Chinese subjects than for the American subjects. When the stimuli are Japanese, the magnitude of the McGurk effect in the Chinese subjects is about the same as it is in the Japanese subjects; when the stimuli are English, it is

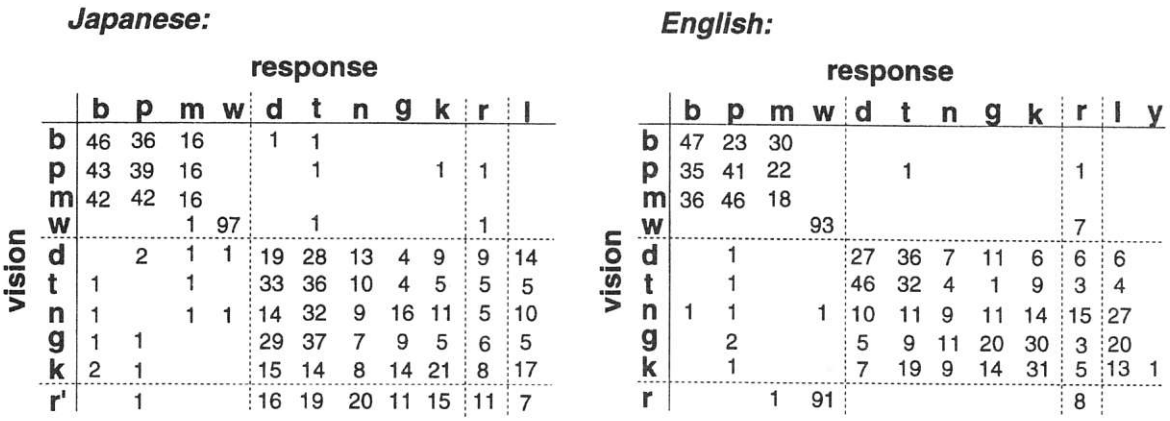

Figure 2. Confusion matrices in the visual-only (V) condition for Japanese and English stimuli. Each number indicates the percentage of the responses in a row.

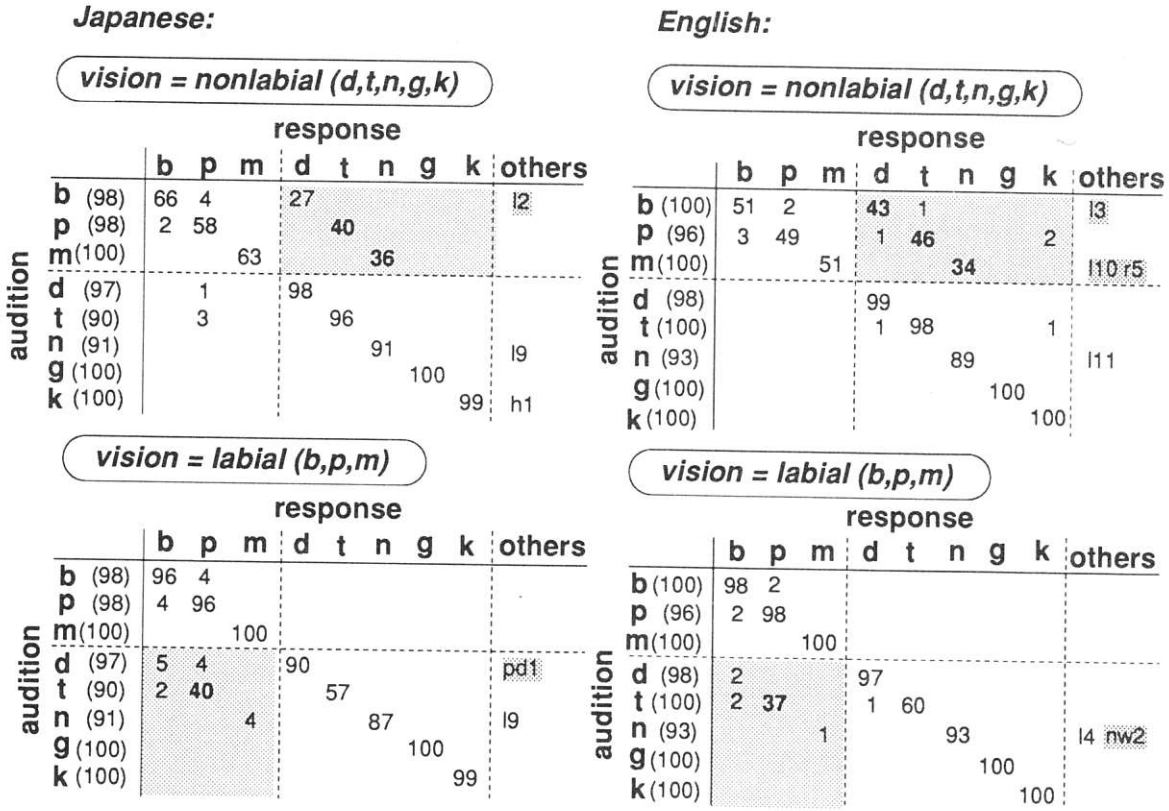

Figure 3. Confusion matrices in the audiovisual (AV) condition for Japanese and English stimuli. The matrices are shown for two cases depending on the visual component of the stimuli. Each number indicates the percentage of the responses in a row. The number in the leftmost columns shows the auditory intelligibility score—that is, the percentage correct in the A condition. The responses in the shadowed sections indicate the "gross" McGurk effect.

even smaller in the Chinese subjects than it is in the Japanese subjects. As found in our previous studies, auditory labials combined with visual nonlabials tend to yield a stronger McGurk effect than conversely combined pairs. To make the comparison across groups easier, the average magnitudes for auditory labials (/b/, /p/, /m/) and nonlabials (/d/, /t/, /n/, /g/, /k/) were calculated for each group. When the stimuli were Japanese, they were 35% and 9% for the Chinese, 25% and 5% for the Japanese, and 77% and 59% for the Americans. When the stimuli were English, they were 48% and 9% for the Chinese, 75% and 11% for the Japanese, and 84% and 13% for the Americans. For the Chinese subjects, who were given both the English and Japanese stimuli in a within-subjects design, there were no significant effects of order on the McGurk effect when comparing the two order groups for a given stimulus set [for Japanese, $F(1,12) = 0.24$, $p > .60$; for English, $F(1,12) = 0.00$—this apparent zero resulted from the almost-zero mean square of the order factor, which represents the almost-equal magnitude of the McGurk effect for the two order groups]. (For the Japanese and American subjects, the two stimulus sets were given in a between-subjects design).

To compare the magnitude of the McGurk effect across language groups by analyses of variance (ANOVAs), the average magnitudes for auditory labials and nonlabials were calculated for each subject. Two-way ANOVAs—native language × stimulus type (auditory labial vs. nonlabial)—were performed separately for the Japanese and English stimuli. In both cases, the main effect of the subjects' native language was significant [$F(2,35) = 25.04$, $p < .0001$, for the Japanese stimuli; $F(2,33) = 4.46$, $p < .0193$, for the English stimuli], as was the main effect of stimulus type [$F(1,35) = 14.37$, $p < .0006$, for the Japanese; $F(1,33) = 96.05$, $p < .0001$, for the English]. The interaction of the two factors was not significant in either case [$F(2,35) = 0.18$, $p > .80$, for the Japanese; $F(2,33) = 3.01$, $p > .06$, for the English].

Subsequently, t tests, using least squares means, were performed for each pair of language groups. The results were as follows: (1) For the Japanese stimuli, the McGurk effect was significantly weaker in the Chinese and Japanese subjects than in the American subjects ($p < .01$ in both cases), and there were no significant differences between the Chinese and the Japanese ($p > .40$); this was true for both auditory labials and nonlabials. (2) For the English auditory labials, the McGurk effect was significantly weaker in the Chinese subjects than in the Japanese and American subjects ($p < .05$ in both cases), and there were no significant differences between the Japa-

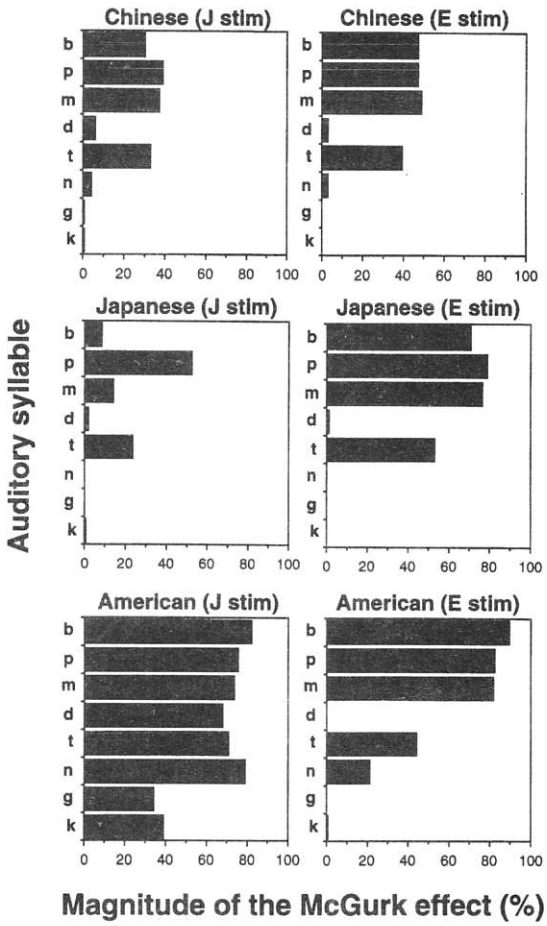

Figure 4. A comparison of the mean magnitude of the McGurk effect across three language groups. The data for the Japanese and American subjects are from Sekiyama (1994b).

nese and the Americans ($p > .50$). For the English auditory nonlabials, the three language groups did not differ significantly ($p > .30$ in each comparison).

These results indicate that the McGurk effect was significantly weaker in the Chinese subjects than in the Americans. It was as weak as that in the Japanese subjects for the Japanese stimuli and even weaker than that in the Japanese for the English stimuli. Perhaps because both stimulus sets are in a foreign language for the Chinese subjects, they did not show a significant difference in the magnitude of the McGurk effect between the Japanese and English stimuli [$F(1,13) = 3.61, p < .08$].

Individual Differences

Compared with the Japanese and the Americans, the Chinese subjects showed much larger individual differences. The left half of Table 1 shows the individual data of the pure McGurk effect in the Chinese subjects for auditory labials (paired with visual nonlabials) and auditory nonlabials (paired with visual labials). The data are averages over the English and Japanese stimuli because the results were similar in the two stimulus sets. The individual differences are obvious for the auditory labials, which yielded a stronger McGurk effect than did the auditory nonlabials.

One might suspect that the subjects who showed the weak McGurk effect were poor lipreaders. However, as shown in the right half of Table 1, the lipreading performance is fairly constant across subjects when scored in terms of the distinction between labials (/b/, /p/, /m/) and nonlabials (/d/, /t/, /n/, /g/, /k/) in the visual-only task. Therefore, the difference in the McGurk effect should be attributed to a difference in intersensory integration rather than to a difference in unimodal performance.

Effect of Living in a Foreign Country

With respect to the cause of the large individual differences, Figure 5 depicts the average magnitude of the McGurk effect for auditory labials as a function of the length of the subject's stay in Japan. (Note that the data are shown only for auditory labials, because the McGurk effect was almost absent for auditory nonlabials in all of the subjects, as shown in Table 1.) There was a positive correlation ($r = .723$) between the magnitude of the McGurk effect and the length of the subject's stay in Japan. The subjects who had lived in Japan for less than 2 years all show a weak McGurk effect (the magnitude is smaller than 20%). Those who had stayed for more than 3 years, with the exception of Subjects 5 and 8, show a strong McGurk effect (the magnitude is larger than 50%). This relationship suggests that the strong reliance on auditory information might be true for the monolingual Chinese who have never lived in a foreign country. However, given the small number of the subjects who had lived in Japan for less than 1 year, this speculation about the monolingual Chinese should be tested

Table 1
The Magnitude of the McGurk Effect (%) and the Rate of Lipreading Errors in Terms of Labial–Nonlabial Discrimination (%) in Chinese Subjects

	McGurk Effect Audition		Lipreading Errors Vision	
Subject	Labial	Nonlabial	Labial	Nonlabial
1	0.55	7.25	0.00	0.00
2	0.00	13.90	0.00	13.00
3	0.00	11.15	0.00	2.00
4	0.55	10.00	1.67	1.00
5	0.55	32.80	0.00	10.00
6	11.65	3.35	0.00	3.00
7	18.35	8.35	0.00	0.00
8	26.10	1.65	0.00	0.00
9	55.00	6.65	0.00	7.00
10	77.20	7.20	0.00	1.00
11	95.00	11.65	0.00	0.00
12	97.25	1.10	0.00	4.00
13	96.65	10.00	0.00	0.00
14	100.00	4.45	10.00	0.00

Note—The percentages are averages over English and Japanese stimuli. The magnitude of the McGurk effect is shown for auditory labials (paired with visual nonlabials) and auditory nonlabials (paired with visual labials).

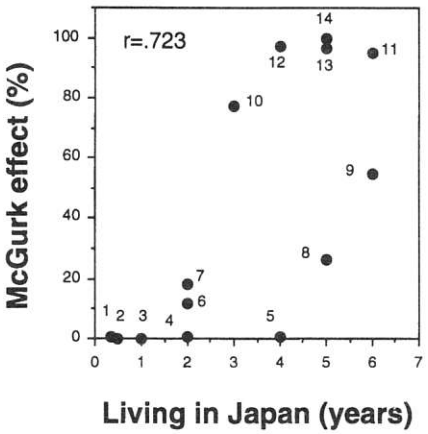

Figure 5. The relationship between the magnitude of the McGurk effect and the length of time the subject had lived in Japan. The number beside each dot indicates subject number. For a period longer than a year, the length of the stay was reported only on a yearly basis. The magnitude of the McGurk effect is the average over Japanese and English auditory labials.

using more Chinese subjects—ideally, those who will have had no experience of living in a foreign country.

DISCUSSION

The Chinese subjects as a group showed a much weaker McGurk effect than did the American subjects, displaying a similarity to the Japanese subjects. Given that both Japanese and English stimuli were foreign speech for the Chinese subjects, the weak McGurk effect clearly indicates that the Chinese are less susceptible to the visual influence than are the Americans. As noted earlier, the native–foreign language effect for these stimuli works against finding a weak McGurk effect for the Chinese subjects. Thus, the results show a strong reliance of the Chinese on auditory information. These results are consistent with the face-avoidance hypothesis.

The Chinese subjects showed large individual differences in the magnitude of the McGurk effect. One possible reason for the large individual differences is the diversity of their native dialects. However, the relationship between the magnitude of the McGurk effect and the subject's native dialect was not obvious. For example, even a couple who came from the same area showed a large difference in the McGurk effect. Rather, the data suggested that the cause of the individual differences was the variation in the extent of the subject's experience of living in Japan, as shown in Figure 5. It is plausible that people who are seriously learning a second language also learn to use any cues, including lip-read information, to improve their listening comprehension. The data suggest that the monolingual Chinese tend to rely on auditory information and that they learn the utilization of visual cues within 3 years if exposed to a foreign language.

As mentioned in the introduction to this paper, Massaro et al. (1993) tested native speakers of Japanese, Spanish, and American English by using synthesized audible speech and a computer-animation face. Although they concluded that "the underlying mechanisms for speech perception are similar for the three languages" (p. 477), their interpretation can be questioned. In some conditions, which were comparable to our natural-speech experiment, the visual (McGurk) effect was weaker in the Japanese subjects than in the native speakers of Spanish or American English. More specifically, when treating magnitude of the visual effect as the outcome variable, Massaro et al. found a significant interaction ($p < .001$) between auditory ambiguity (measured on a five-level scale) and language group. This was because the visual effect was weaker in the native speakers of Japanese than in the other two groups when the auditory stimuli were less ambiguous. Although the authors insist that this quantitative difference does not mean a qualitative difference in manner of processing, they showed some evidence of a qualitative difference in Experiment 2 when they tested different models against the results for the native speakers of Japanese and American English. The results for the native speakers of American English were best explained by their fuzzy logical model of perception (FLMP), which hypothesizes that the visual and auditory information is integrated in a form of multiplication under various conditions. On the other hand, the results for the Japanese speakers could be explained by both the FLMP and the auditory dominance model, which assumes that the visual information is integrated into the perceptual decision only when the auditory speech itself is not identified. Therefore, their results do not deny the tendency of the Japanese to rely more than the Americans on auditory information.

In summary, in an attempt to test the face-avoidance hypothesis, the current study examined the extent to which Chinese subjects integrate visual and auditory information during speech perception. The Chinese, who are believed to be culturally similar to the Japanese in their avoidance of gazing at the face of a speaker, were also less susceptible to visual influences than are Americans. This indicates that the face-avoidance hypothesis is not rejected. However, further research will be needed to test this hypothesis against another possible factor that may influence the strength of the McGurk effect. This is because there was some evidence in the current study that the auditory reliance of the Chinese might be even stronger than that of the Japanese. The average magnitude of the McGurk effect was smaller in the Chinese subjects than in the Japanese subjects when the stimuli were English. Moreover, the Chinese subjects who had lived in Japan for less than 3 years showed little McGurk effect for both English and Japanese stimuli, even though both were nonnative languages for them. If the native–foreign language effect holds, the results imply a very strong auditory reliance. This strong auditory reliance may be related to the tonal characteristics of the Chinese language.

In the Chinese language, the meaning of a spoken word is determined not only by its syllabic structure but also by its tone. For example, /ma/ means "mother," "flax" (this tone also means "to numb"), "horse," or "to revile,"

depending on the tone. Auditory cues must certainly be more effective for identifying tones than visual information. This language characteristic may foster in the Chinese a strong reliance on auditory information.

Regarding tonal properties, the Japanese language is also described as tonal to some extent, though the extent is much lower than that for the Chinese language. In Japanese, some syllables do have more than one meaning which can be distinguished by the tone. For example, /hashi/ means "bridge," "chopstick," or "edge," depending on the tone. If one's auditory reliance in audiovisual speech perception depends on the extent to which the subject's native language is tonal, it is possible to explain the reduced McGurk effect in the Japanese and the Chinese only by the tonal characteristics of one's native language. Thus, the critical test of the face-avoidance hypothesis awaits further research. Of course, both of the factors—the tonal properties and face avoidance—may account for the reduced visual influence in the Chinese and Japanese subjects.

Finally, the current study found a positive correlation between the magnitude of the McGurk effect and the amount of time that the Chinese subjects have spent in Japan. This suggests that even people who normally do not use visual information do learn to use visual cues in the process of learning a second language. Thus, the proficiency of a second language may be another factor that influences the strength of the McGurk effect.

REFERENCES

Binie, 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.

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

Fuster-Duran, A. (1996). Perception of conflicting audio-visual speech: An examination across Spanish and German. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 135-143). Berlin: Springer-Verlag.

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

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

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., & Smeele, P. M. (1995). Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition*, **23**, 113-131.

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

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

Nomura, M. (1984). *Body language wo yomu* [Understanding body language]. Tokyo: Heibonsha.

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

Sekiyama, K. (1994a). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E)*, **15**, 143-158.

Sekiyama, K. (1994b). McGurk effect and incompatibility: A cross-language study on auditory-visual speech perception. *Studies and Essays in Behavioral Sciences & Philosophy (Kanazawa University)*, **14**, 29-62.

Sekiyama, K., Braida, L. D., Nishino, K., Hayashi, M., & Tuyo, M. (1995). The McGurk effect in Japanese and American perceivers. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences* (Vol. 3, pp. 214-217). Stockholm: Congress organizers at Kungliga Tekniska Högskolan and Stockholm University.

Sekiyama, K., Joe, K., & Umeda, M. (1988). Perceptual components of Japanese syllables in lipreading: A multidimensional study. *Technical Report of IEICE* (The Institute of Electronics, Information, and Communication Engineers of Japan), **IE87**-127. [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.

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

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 & Hearing Research*, **20**, 130-145.

NOTE

1. Although the native–foreign language effect is often found in cross-language studies on the McGurk effect (e.g., Fuster-Duran, 1996), this effect is not always obtained. In our recent study, we tested Japanese and American subjects with syllables pronounced by two Japanese and two American speakers. Although the weaker McGurk effect in the Japanese subjects was replicated, the native–foreign language effect was not obvious in the Japanese subjects but was found in the American subjects (Sekiyama, Braida, Nishino, Hayashi, & Tuyo, 1995). This was because the speech of the American speakers was clearer than that of our previous American speaker, whereas the speech of the Japanese speakers was less clear than our previous Japanese speaker. These results indicate that the magnitude of the visual effect depends not only on the native–foreign language factor, but also on the quality of each speaker's speech. Both factors may be responsible for auditory ambiguity.

(Manuscript received March 17, 1995;
revision accepted for publication February 27, 1996.)