McGurk Effect in Chinese subjects

McGurk Effect in Chinese subjects

Kaoru Sekiyama

Abstract

The "McGurk effect" is a perceptual blending of auditory and visual (lip-read) information about speech when the two sources of information are discrepant. While this has been established as a robust effect in English speaking cultures, Sekiyama and Tohkura (1993) found that Japanese subjects showed a much weaker McGurk effect than American subjects for Japanese stimuli. The present study examined the magnitude of the McGurk effect in Chinese subjects for Japanese and English stimuli which were the same as those used by Sekiyama and Tohkura (1993). The subjects were 14 native speakers of Chinese. Each subject was presented with both Japanese and English stimuli. The subjects were asked to report what they heard as well as check the incompatibility between what they saw and what they heard. As compared with the previous results, the Chinese subjects as a group showed even a weaker McGurk effect than the Japanese subjects. The Chinese subjects also showed large individual differences, perhaps due to the various lengths of their stay in Japan.

The McGurk effect is a perceptual blending of auditory and visual (lip-read) information of speech when the two sources of information are discrepant (McGurk & MacDonald, 1976). When the place of articulation differs between the auditory and visual stimuli, the McGurk effect may occur. For example, when auditory/pa/is dubbed onto visual lip movements of/na/, perceivers often report hearing "ta." In this example, auditory/pa/is labial (lip-articulated) whereas visual/na/is nonlabial (articulated in the mouth). The perceived sounds ("ta" 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.

ACKNOWLEDGMENTS

This study was supported by the Grant-in Aid for Scientific Research from the Ministry of Education, Science and Culture, and the Sasakawa Scientific Research Grant from the Japan Science Society.

While this effect has been proven to be robust in English speaking cultures (MacDonald & McGurk, 1978; Massaro & Cohen, 1983; Green, Kuhl, Meltzoff, & Stevens, 1991; Dekle & Fowler, 1992; Rosenblum & Saldana, 1992), we found inter-language differences between Japanese and American subjects (Sekiyama & Tohkura, 1993). The Japanese showed a much weaker McGurk effect than the Americans when stimuli were Japanese (that is, the stimuli pronounced by a native speaker of Japanese), whereas there were no differences between the two language groups when the stimuli were English. These results show that the Japanese tend to rely on auditory information as long as they are listening to their native language. The stronger McGurk effect the Japanese showed for English stimuli seem to be due to some acoustical deviations of the stimuli of a foreign language from what they are accustomed to in their native language.

Although the causes of this weak McGurk effect in the Japanese perceiving Japanese stimuli are unknown, there are several possible factors which may cause the Japanese to develop a type of processing which does not integrate visual information into perceived speech. First, 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 the person is of a higher status. Second, due to the smaller number of vowels, consonants, and syllables which can occur (there are only 100 possible syllables and no consonant clusters are allowed) in Japanese, the Japanese phonemes might easily be discriminated without additional visual cues. Related to this second factor, lip-read information may be less useful in Japanese than in English because the Japanese consonant inventory does not have labio-dentals (/f/, /v/, /θ/, /ð/), which are easy to lipread and form a perceptual group distinct from both labials (/b/, /p/, /m/, /w/) and velars (/d/, /n/, /g/, etc.) in English (Walden, Prosek, Montgomery, Scherr, & Jones, 1977). Third, the Japanese use of Chinese characters may affect their manner of processing. Since Chinese characters are pictographic, the Japanese auditory-visual connections might be oriented to visual images of characters, not to those of lip movements.

In an attempt to examine these possibilities, the present study tested Chinese subjects by using the same stimuli as in our previous study, that is, Japanese and English stimuli. Regarding the above factors, the Chinese are considered to be closer to the Japanese than to Americans. First, the phonemic inventory and the syllable structure in Chinese are similar to those in Japanese. According to my informal

inquiry, several Chinese people admitted that they do not stare at the face of the person they are listening to as much as Western people do. Finally, they use Chinese characters.

If any of these factors is responsible for the weak McGurk effect in the Japanese, the Chinese should also show a limited McGurk effect though it is still possible that the stimuli of their foreign language will yield a strong effect as was also observed in the Japanese subjects listening to English stimuli. The results showed that for both Japanese and English stimuli, Chinese subjects as a group showed as weak a McGurk effect as Japanese subjects listening to the Japanese stimuli. However, there was a large individual difference in the magnitude of the effect.

METHOD

Subjects

Fourteen native speakers of Chinese were recruited from the Kanazawa University community. 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, and the length of their stay in Japan was between 4 months and 6 years. They had never lived in a foreign country other than Japan. Their native languages were various Chinese dialects but all of them were educated in Mandarin Chinese starting in elementary school. All of them had parents whose native dialect was the same as theirs. They reported that they had normal hearing and normal or corrected to normal vision. They were paid 6000 yen for the four-hour experiment conducted in two days.

Stimuli

Stimuli were the same as used by Sekiyama and Tohkura (1993). They consisted of ten syllables (/ba/, /pa/, /ma/, /wa/, /da/, /ta/, /na/, /ra/, /ga/, and /ka/) and were pronounced by a Japanese speaker (for Japanese stimuli) and an American speaker (for English stimuli). Each female speaker's face was videotaped while she pronounced the syllables. Her utterances were re-recorded in an anechoic room. These ten visual and ten auditory syllables were combined using a BETACAM video system that can handle frame by frame time control (33ms). For the technical details of the dubbing, see Sekiyama (1994a).

In the dubbing, the auditory syllables were combined only with the visual syllables of the same speaker, but an auditory syllable was dubbed onto all the visual syllables

so that all possible combinations of ten auditory and ten visual syllables (10 x 10 = 100) were produced. On the final videotapes ("AV-tapes"), each stimulus occurred in a 7 s trial in which the video channel included 3 s of black frames and 4 s of the face.

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 ten 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 loud speakers placed at the sides of the monitor. The subjects viewed the monitor from a distance of 1 m.

Experimental Design

Each of the fourteen subjects participated in the two sessions of experiment. One session was for Japanese stimuli, the other for English stimuli. The order of the stimulus set was counterbalanced between subjects. For each stimulus set, there were three conditions: Auditory-visual (AV), auditory-only (A), and visual-only (V) which were conducted in this 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 loud speakers.

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 s in random order. In the AV condition, the subjects were instructed to write down what they heard while they looked at and listened to each syllable. 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 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 only to report what they heard. In the V condition, they were asked to lipread and report what they thought the speaker was pronouncing.

Each stimulus was repeated six times in the AV and A conditions, and ten times in the V condition. 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 hours to conduct the three conditions for one stimulus set. Each subject participated in the experiment for two 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 Table 1. Whereas most of the consonants were accurately identified, /r/ in both Japanese and English showed large percentages 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." We had also observed the "w" responses for the E nglish/r/in Japanese subjects previously (Sekiyama, 1994b). One characteristic was observed for Chinese subjects only: "l" responses for/n/in both Japanese and English.

Table 1. Confusion matrices for the A condition (% in a row).

Japanese

audition \ response	b	p	m	w	d	t	n	g	k	r	l
b	98	2									
p	1	98				1					
m			100								
w				100							
d	1	2			97						
t		10				90					
n						1	91				8
g								100			
k									100		
r'							8			61	31

English

audition \ response	b	p	m	w	d	t	n	g	k	r	l
b	100										
p	2	97				1					
m			100								
w				85						8	7
d					98			2			
t						100					
n							93				7
g								100			
k									100		
r				17						71	11

Responses in the V condition

Table 2 shows confusion matrices for the V condition. It is clear that in 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. Whereas Japanese/r/was perceived as one of the nonlabials, English/r/was perceived as being in the same category as/w/.

Table 2. Confusion matrices for the V condition (% in a row).

Japanese

vision \ response	b	p	m	w	d	t	n	g	k	r	l
b	46	36	16		1	1					
p	43	39	16			1			1	1	
m	42	42	16								
w			1	97		1				1	
d		2	1	1	19	28	13	4	9	9	14
t	1		1		33	36	10	4	5	5	5
n	1		1	1	14	32	9	16	11	5	10
g	1	1			29	37	7	9	5	6	5
k	2	1			15	14	8	14	21	8	17
r'		1			16	19	20	11	15	11	7

English

vision \ response	b	p	m	w	d	t	n	g	k	r	l	y
b	47	23	30									
p	35	41	22			1				1		
m	36	46	18									
w				93						7		
d		1			27	36	7	11	6	6	6	
t		1			46	32	4	1	9	3	4	
n	1	1		1	10	11	9	11	14	15	27	
g		2			5	9	11	20	30	3	20	
k		1			7	19	9	14	31	5	13	1
r			1	91						8		

Responses in the AV condition

In the analyses of the data for the AV condition, the data for stimuli of 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/). In the confusion matrices in Table 3, the responses in the shadowed portions 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.

Table 3. Confusion matrices for the AV condition (% in a row).
Shadowed portions indicate the McGurk effect. The number shown in parentheses in the left most column in each row is the percent correct in the A condition.

(A) Japanese stimuli

vision = nonlabial (d,t,n,g,k)

audition \ response	b	p	m	d	t	n	g	k	others
b (98)	66	4		27					2
p (98)	2	58			40				
m(100)			63			36			
d (97)		1		98					
t (90)		3			96				
n (91)						91			9
g (100)							100		
k (100)								99	h1

vision = labial (b,p,m)

audition \ response	b	p	m	d	t	n	g	k	others
b (98)	96	4							
p (98)	4	96							
m(100)			100						
d (97)	5	4		90					pd1
t (90)	2	40			57				
n (91)			4			87			9
g (100)							100		
k (100)								99	

(B) English stimuli

vision = nonlabial (d,t,n,g,k)

audition \ response	b	p	m	d	t	n	g	k	others
b (100)	51	2		43	1				3
p (96)	3	49		1	46			2	
m(100)			51			34			10 5
d (98)				99					
t (100)				1	98			1	
n (93)						89			11
g (100)							100		
k (100)								100	

vision = labial (b,p,m)

audition \ response	b	p	m	d	t	n	g	k	others
b (100)	98	2							
p (96)	2	98							
m(100)			100						
d (98)	2			97					
t (100)	2	37		1	60				
n (93)			1			93			4 nw2
g (100)							100		
k (100)								100	

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.

Thus, the magnitude of the pure McGurk effect (visual effect) was 42−10=32%.

Fig. 1 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 my previous study. As compared with the American subjects, the Chinese subjects showed a weaker McGurk effect. The magnitude of the McGurk effect in the Chinese subjects is about the same as that in the Japanese subjects when the stimuli are

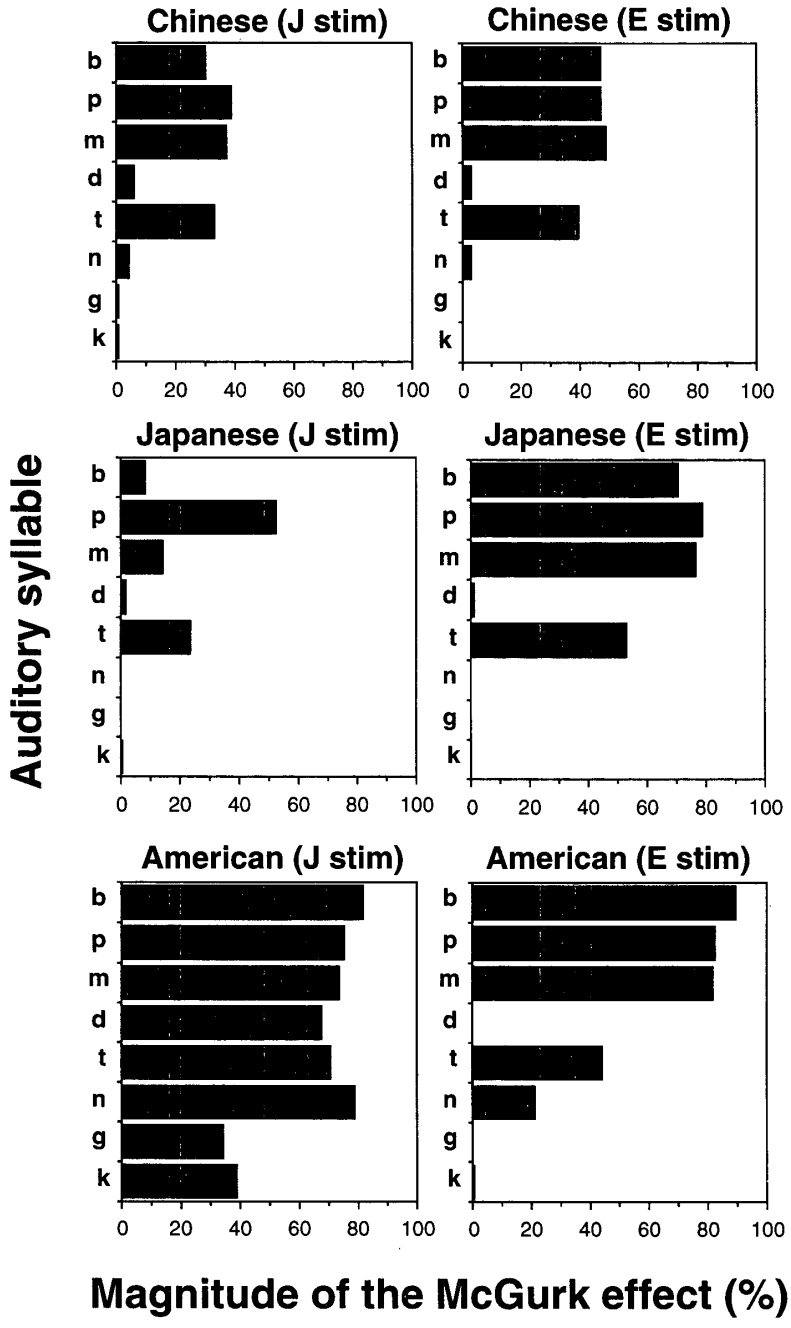

Figure 1. The magnitude of the McGurk effect for each auditory syllable. The data for the Japanese and American subjects are from Sekiyama (1994b).

Japanese, and it is even smaller in the Chinese subjects than in the Japanese subjects when the stimuli are English. 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 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. In the Chinese subjects, there were no significant effects of order on the McGurk effect when comparing two order groups in a stimulus set [for Japanese, $F(1, 12)=0.24$; for English, $F(1, 12)=0.0$].

To compare the magnitude of the McGurk effect across language groups by ANOVAs, the average magnitudes for auditory labials and nonlabials were calculated for each subject. Two-way ANOVAs [native language x stimulus type (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 well as the main effect of the 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 nonsignificant in both cases [$F(2, 35)=0.18$ for the Japanese; $F(2, 33)=3.01$ 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 in the Chinese and Japanese subjects was significantly weaker than in the American subjects and there were no significant differences between the Chinese and the Japanese. This was true for both auditory labials and nonlabials. (2) For the English auditory nonlabials, the Chinese subjects showed a significantly weaker McGurk effect than the Japanese and American subjects and there were no significant differences between the Japanese and the Americans. For the English auditory nonlabials, the three language groups did not differ significantly.

These results indicate that the McGurk effect in the Chinese subjects was as weak as that in the Japanese subjects for the Japanese stimuli and it was even weaker than that in the Japanese for the English stimuli, even though both of them are of their foreign language. Perhaps because both the two stimulus sets are in a foreign

language, the magnitude of the McGurk effect for the Japanese stimuli and that for the English stimuli did not differ significantly in the Chinese [$F(1, 13) = 3.61, p < .08$].

Individual differences

Compared with the Japanese and the Americans, the Chinese subjects showed much larger individual differences in the magnitude of the McGurk effect. These differences are not only in the magnitude of the McGurk effect itself, but also in the relationship between the McGurk effect and incompatibility. In my previous study (Sekiyama, 1994a), negatively high correlations stronger than $r = -.90$ were observed between these two indexes in both the Japanese and American subjects. This implies that the McGurk effect was absent when the subjects detected the auditory-visual discrepancy. In the present study using Chinese subjects, seven who produced a strong McGurk effect (left panel of Fig. 2) also showed a high correlation ($r = -.97$). However, the other seven subjects who produced only a weak McGurk effect (right panel) did not show such a meaningful correlation ($r = -.55$). The WEAK group often failed to detect the auditory-visual discrepancy even when the McGurk effect was absent. For stimuli which produced a negligibly weak McGurk effect, the frequencies of reported incompatibility are densely distributed at about 40% in the WEAK group while they are 60-70% in the STRONG group.

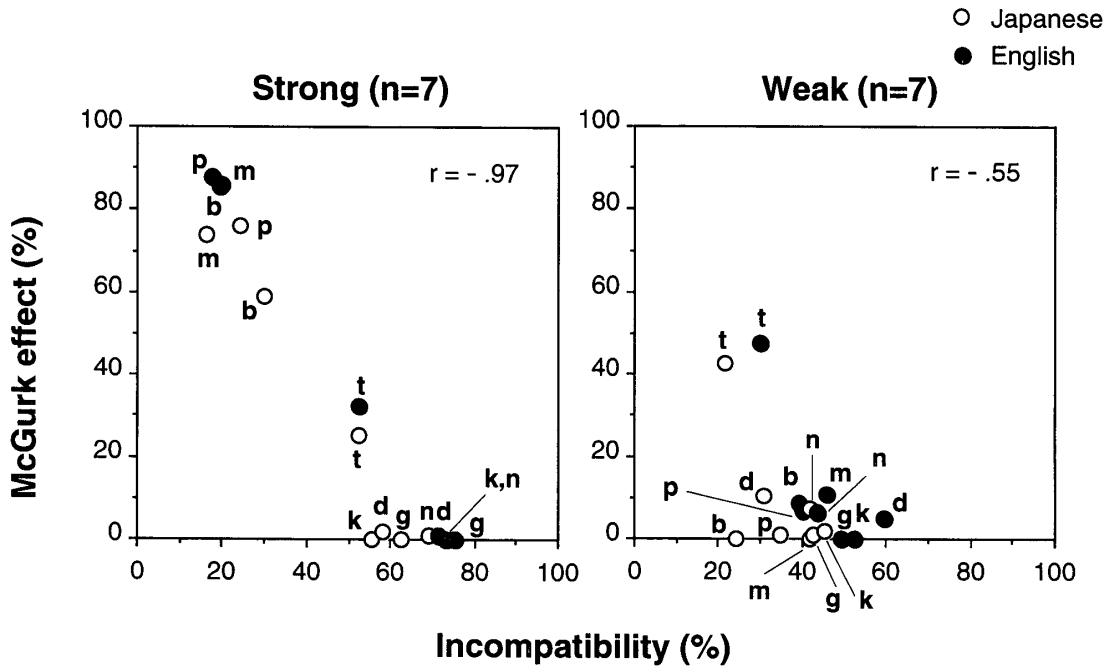

Figure 2. The relationship between the magnitude of the McGurk effect and incompatibility. The left panel is for the subjects who showed a strong McGurk effect and the right is for those who showed a weak McGurk effect.

One might suspect that the WEAK group consisted of poor lipreaders. However, the two groups did not differ in lipreading performance in terms of discrimination between labials (/b, p, m/) and nonlabials (/d, t, n, g, k/). The percents of correct categorization for visual labials and nonlabials for the Japanese stimuli were 98.1% and 97.4% in the STRONG group and 99.5% and 95.4% in the WEAK group. The percents for the English stimuli were 99.0% and 99.1% in the STRONG group and 100% and 96.3% in the WEAK group. An ANOVA did not find a significant difference between the two groups [$F(1, 12) = 0.62$)]. Therefore, the difference in the McGurk effect cannot be attributed to a difference in unimodal performances, but to a difference in inter-sensory integration. The WEAK group seems to have difficulty with the parallel processing of visual and auditory information.

DISCUSSION

Although there were individual differences, it should be noted that all the stimuli presented to the Chinese subjects were in foreign languages. Had we used Chinese stimuli, the McGurk effect might have been weaker even in the STRONG group. For, as seen in the Japanese and American subjects in Fig. 1, stimuli in the subjects' foreign language yield a stronger McGurk effect than those in their native language. As compared with the cases where the Japanese subjects were presented with English stimuli and the American subjects were presented with Japanese stimuli, the Chinese subjects as a group showed a much weaker McGurk effect. For both stimulus sets, the magnitude of the McGurk effect in the Chinese was the same as that in the Japanese listening to Japanese speech. This means that the Chinese reliance on auditory information is even stronger than the Japanese. It may be related to the tonal characteristic of the Chinese language. Auditory information must certainly be more effective to identify tones than visual information.

It may be possible that the large individual differences in the Chinese subjects are due to 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 because even a couple who came from the same area showed a large difference in the McGurk effect.

Another possible cause of the individual differences is the length of their stay in Japan and the extent to which they have been exposed to a foreign language and culture. It is plausible that people who are seriously learning a second language learn

to use visual cues to understand the language. Supporting this hypothesis, the length of the subjects' stay in Japan showed a relationship with the magnitude of the McGurk effect: All the seven subjects in the STRONG group had been in Japan for more than 3 years, whereas all but one subjects in the WEAK group for less than 2 years. The one who came to Japan latest arrived just 4 months before the experiment began. This subject's performance was perfect in the A and V conditions and hardly showed the McGurk effect. This reliance on auditory imformation might be true for the "pure" Chinese who have never lived in a foreign country. It would be of great interest to test more number of Chinese subjects with various lengths of stay in Japan.

REFERENCES

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

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.

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

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 and Philosophy (Kanazawa University)*, **14**, 29-62.

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

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.