McGurk Effect and Incompatibility : A Cross-Language study on Auditory-Visual Speech Perception

McGurk Effect and Incompatibility:

A Cross-Language study on Auditory-Visual Speech Perception

Kaoru Sekiyama

Abstract

In English speaking cultures, it has been reported that when auditory speech is presented in synchrony with discrepant visual (lip-read) speech, the subjects often report hearing sounds that integrate information from the two modalities (the "McGurk effect"). A cross-language study on the McGurk effect was carried out to examine differences between Japanese and English suggested by Sekiyama and Tohkura (1991). The stimulus materials were ten syllables (/ba, pa, ma, wa, da, ta, na, ga, ka, ra/) pronounced by a Japanese and an American speaker. The ten auditory and ten visual syllables pronounced by a speaker were cross-dubbed resulting in 100 auditory-visual stimuli. Japanese syllables were presented to 14 Japanese and 10 American subjects. English syllables were presented to different groups of subjects, 12 Japanese and 10 American. The stimuli were presented in both quiet and noise-added conditions. The subjects were asked to check incompatibility between what they heard and what they saw as well as to report what they heard. The results showed that Japanese subjects are more sensitive to auditory-visual discrepancy and less prone to the McGurk effect than Americans. The size of the McGurk effect correlated highly negatively with the frequency of incompatibility. These results demonstrate that Japanese listeners tend to separate two modalities of information unless visual support is necessary while Americans easily integrate them.

ACKNOWLEDGMENTS

The author is grateful to Prof. Harry Levitt at the Graduate Center of the City University of New York whose laboratory was used for experiments with American subjects. She also thanks Dr. Yoh'ichi Tohkura for his full support in creating the stimuli at ATR auditory and visual perception research laboratories, Dr. Hwei-Bing Lin at CUNY for her help with arrangements for the experiments. The experiments at CUNY were supported by a grant from Nissan Science Foundation to the author.

McGurk Effect and Incompatibility: A Cross-Language study on Auditory-Visual Speech Perception

Although speech perception is considered primarily an auditory process, visual information in the form of lipreading is a source of speech perception in face-to-face verbal communication. Lip-read information normally facilitates speech perception especially when auditory speech signals are unintelligible. It is perhaps hearing impaired people that utilize lipreading most to comprehend speech. In fact, recent studies have reported that lip-read information improves speech perception considerably by patients with cochlear implants (inner ear prostheses), that is, people who have to process unintelligible auditory information that the cochlear implants provide (Blamey, Dowell, Brown, Clark, & Seligman, 1987; Fukuda, Shiroma, & Funasaka, 1988; Funasaka, Shiroma, Yukawa, Iizuka, Yao, Kono, Takahashi, & Kumakawa, 1989). Besides the hard of hearing, visual information has been also shown to be helpful for people with normal hearing, especially when the auditory signal is not clear due to added noise (Erber, 1975; O'Neil, 1954; Sumby & Pollack, 1954).

The contribution of lip-read information to speech perception is described as compensating for auditory information. In the case of consonant perception, visual information has the advantage of conveying spatial information, namely, information about the place of articulation that auditory information may fail to convey in some circumstances. Particularly, the visual system is highly sensitive to a distinction between labial (lip-articulated; e.g.,/b/, /p/, /m/) and non-labial (articulated inside the mouth; e.g., /d/, /t/, /n/) that the auditory system sometimes misses (e.g., Binie, Montgomery, & Jackson, 1974). Using these characteristics, McGurk and MacDonald (1976) dramatically demonstrated that visual information influences speech perception even when the auditory signal is highly intelligible. When they presented their normal adult subjects with auditory speech signals of/ba/, they correctly reported hearing "ba" about 99% of the time. However, when the identical auditory speech signals of /be/ were presented in synchrony with visual lip movements for /ga/, they reported hearing "da" 98% of the time. This "McGurk effect" demonstrates that visual information about place of articulation easily modifies phonetic perception. In this example, visual place information (/ga/: non-labial) biased auditory information (/ba/: labial) with the

result that perceptual solution ("da" : non-labial) was consistent with the biased place information and the remaining auditory information.

Since McGurk and MacDonald first reported this fusion effect between auditory and visual information, it has been replicated in many studies and established as a robust effect in English speaking cultures (for a review, see Summerfield, 1987). These studies have shown that the McGurk effect can be produced under various conditions as follows. The McGurk effect occurs for various combinations of auditory and visual syllables when there is a discrepancy in the place of articulation (MacDonald & McGurk, 1978). In most of the cases, auditory labials paired with visual non-labials yield *fused responses* (e.g., auditory /pa/ paired with visual /na/ produces "ta"-response) while auditory non-labials paired with visual labials often result in *combined responses* (e.g., auditory /ta/ and visual /pa/ produces "pta"-response). The fusion effect has been also reported for materials of meaningful words (Dekle, Fowler, & Funnell, 1992; Dodd, 1977) or synthesized speech syllables (Massaro, 1987; Massaro & Cohen, 1983). It has also been reported that a male speaker's voice paired with a female face can produce the McGurk effect as much as a male voice-male face pair (Green, Kuhl, Meltzoff, & Stevens, 1991). Auditory-visual discrepancy yields fused or combined responses and also longer reaction times (Green & Kuhl, 1991), and reaction time depends on the extent of ambiguity (Massaro, 1987).

However, these studies conducted in English speaking cultures have not yet examined the McGurk effect for linguistic/cultural independence. Although Mills and Thiem (1980, cited in Massaro, 1987) asked German listeners to identify syllables consisting of conflicting auditory and visual information, their results also showed that the visual component has strong effects on the identification. In contrast to this, Sekiyama and Tohkura (1991) recently reported that Japanese listeners show very little McGurk effect for Japanese syllables when the auditory speech was perfectly intelligible. For example, being presented with auditory /ba/ paired with visual /ga/, their Japanese subjects reported hearing auditory "ba" 100% of the time, in contrast to the fused "da" responses reported for English subjects. They were presented with the speech either in quiet or in noisy condition. In the quiet condition, the visual effects were not substantial except for some auditory syllables that were not perfectly identified in the audio-alone presentation. On the other hand, strong visual effects were observed for each auditory syllable in the noise-added condition where the auditory speech had poor intelligibility. From these results, Sekiyama and

Tohkura (1991) proposed an intelligibility hypothesis that Japanese listeners hearing Japanese speech do not integrate visual information with auditory information when audition provides sufficient information.

As shown above, the Japanese subjects did show strong visual effects in the noise-added condition, which indicates that Japanese can lipread to the extent that should induce the McGurk effect. Therefore, it is unlikely that the cause of the small McGurk effect in Japanese subjects listening to Japanese speech is attributed to the subjects' poor lipreading ability. In our previous study on lipreading (Sekiyama, Joe, & Umeda, 1988), untrained Japanese subjects could visually discriminate labials from non-labials 92% of the time (60 Japanese normal adults were tested with 100 Japanese syllables). This labial vs. non-labial discrimination score is equivalent to that of Americans (91%: calculated based on the results obtained from 31 subjects by Walden, Prosek, Montgomery, Scherr, & Jones, 1977).

Subsequently, Sekiyama extended the above experiment to three new conditions: AJ condition (American subjects, Japanese stimuli) tested native speakers of American English with the identical Japanese speech stimuli. In AE (American subjects, English stimuli) and JE (Japanese subjects, English stimuli) conditions, American subjects and Japanese subjects were tested with English speech stimuli pronounced by a native speaker of American English. Combining the results of these three conditions and those of the earlier study (JJ condition : Japanese subjects, Japanese stimuli), Sekiyama and Tohkura (1993) briefly reported inter-language differences between Japanese and American subjects. The study showed that the McGurk effect in the three new conditions was stronger than that in the JJ condition. This indicates that the small McGurk effect in Japanese subjects for Japanese syllables is due to a perceptual processing of Japanese listeners when they process speech information of their native language.

The purpose of the present study is to give a more penetrating description of the inter-language differences in this phenomenon, by examining the date on sensitivity to the discrepancy between auditory and visual information. In our experiment, the subjects had two concurrent tasks. They were asked to check whether or not they felt the visual lip movements incompatible with what they heard, as well as to report what they heard. This request for reporting incompatibility had two objectives. One was to confirm the results of our pilot experiment where Japanese subjects seemed to experience incompatibility instead of perceptual fusion most of the time when

presented with discrepant auditory-visual stimuli of Japanese speech. This suggests that the small McGurk effect of Japanese subjects for Japanese syllables is related to frequent incompatibility. The second objective of this task was to make the subjects pay attention to both the auditory and visual information.

Although Sekiyama and Tohkura (1993) briefly mentioned the frequency of incompatibility in the four conditions, detailed analyses of the incompatibility data were not done because one condition, the JJ condition, was carried out in a different procedure. Whereas the other three conditions asked the subjects to report only one response (what they heard), the JJ condition encouraged the subjects to report more than one response. Such a procedural difference may have caused different processes for detecting incompatibility between what they heard and what they saw. In this study, the JJ condition was replicated with the identical procedure used for the other three conditions so that a direct comparison was possible.

The main purpose of this study is to compare the frequency of incompatibility in terms of its relationship with the size of the McGurk effect. It was also of interest to see if the JJ condition replicates the small McGurk effect. As Sekiyama and Tohkura (1993) reported only a small set of the data from their experiment, the present study will present the full data set necessary to examine the relationship between the frequency of incompatibility and the size of the McGurk effect.

The results showed that Japanese subjects are less prone to the visual effect and more sensitive to the discrepancy between auditory and visual information. It was also found that the size of the McGurk effect is a negative linear function of the frequency of incompatibility.

METHOD

Subjects

Four groups of subjects participated. Twenty-six native speakers of Japanese were assigned to the JJ (14 subjects) and JE (12 subjects) conditions. Twenty native speakers of American English were assigned to the AJ and AE conditions (ten subjects for each). All subjects had normal hearing and normal or corrected to normal vision. The Japanese subjects were students at Kanazawa University and most of the American subjects were graduate students at the City University of New

York. Their ages ranged from 20 to 30 years old. None of the subjects who were given foreign speech stimuli (JE and AJ conditions) had lived in a culture in which the tested language is spoken natively. However, all the Japanese subjects had taken English classes for at least six years starting at age twelve. None of the American subjects had studied Japanese. The Japanese subjects were volunteers, and the American subjects were paid $ 55 for the four-hour experiment.

Stimuli

Stimulus materials were ten syllables (/ba/, /pa/, /ma/, /wa/, /da/, /ta/, /na/, /ra/, /ga/, and /ka/) pronounced by a Japanese speaker (for Japanese stimuli) and an American speaker (for English stimuli). Each female speaker was shown each of the syllables in written form and was asked to pronounce it clearly in her native accent. Because the duration of the American speaker's vowels at the beginning of the recording was apparently much longer than the duration of the Japanese speaker's, the American speaker was instructed to shorten the vowels for inter-language equivalence. The Japanese speaker was a professional announcer and the American speaker had been teaching English in Japan for a year after spending her first 25 years of her life in the U. S.

Visual stimuli were created by videotaping the speaker's face while she pronounced the syllables. The speech was recorded on to a videotape through a broadcast quality (SONY BETACAM) video camera located in front of the speaker.

Though both auditory and visual information was recorded on the videotape, the auditory information was re-recorded to avoid degrading factors in the recording process. To do so, the speaker sat in front of a microphone in an anechoic room. The speaker's utterances recorded on the above videotape were played and presented to the speaker through a headphone. The speaker was asked to mimic the utterances she heard. Her pronunciation was recorded by a DAT (Digital Audio Tape; 16 bit with a sampling frequency of 20 kHz). The ten auditory syllables were processed by a speech analysis package so that the ten syllables had approximately identical peak intensity. Although the American speaker shortened the duration of vowels complying with the instruction, the English syllables still had a longer duration. The duration was approximately 250 ms for the Japanese syllables and 300 ms for the English syllables. The audio signals on the DAT tape were dubbed onto a new BETACAM videotape so that frame by frame time control (33 ms) was possible.

Auditory-visual stimuli were created by replacing the audio signals on the original videotape with the audio signals on the new video tape. For the original and new videotapes, the frame numbers on which audio signals were present had been checked by playing the videotapes frame by frame. Then the new audio signals were dubbed onto the frames where the original audio signals had been.

The dubbing was done only between the Japanese face and the Japanese voice, and between the American voice and the American face. In addition to corresponding auditory-visual dubbing (e.g., /ba/ voice- /ba/ mouth), the auditory syllable was also dubbed onto other visual syllables so that all possible combinations of ten auditory and ten visual syllables ($10 \times 10 = 100$) were produced.

On the final videotapes, each syllable occurred in a 7 s trial in which the video channel included 3 s of black frames and 4 s of the face. The audio channel included warning tones and speech (see Sekiyama & Tohkura, 1991, Fig. 1). There were several tapes of different random orders of 100 stimuli (7 s x 100 : 12 min). For the experimental presentation, the outputs of the BETACAM videotapes were dubbed onto VHS tapes.

Videotapes for auditory-only presentation were also created. On these tapes, the video channel included only black frames. For the auditory-only presentation, a videotape included six repetitions of each auditory stimulus (7 s x 10 x 6 : 7 min).

The JJ and JE conditions were carried out at Kanazawa University, and the AJ and AE conditions at the City University of New York. Visual stimuli were presented on a 20-in. (Kanazawa) and a 21-in. (New York) color monitor in which approximately life-sized speakers appeared. Auditory stimuli were presented through two loud speakers placed at the sides of the monitor. The viewing distance of the subjects was 1 m.

Experimental Design

The subjects were asked to report what they heard in four conditions of two within-subjects factors: (a) quiet (no-noise-added) or noise-added, (b) auditory-only or auditory-visual. In the quiet auditory-only condition (A condition), auditory stimuli were presented with video black. In the quiet auditory-visual condition (AV condition), auditory-visual stimuli were presented. In the noise-added auditory-only condition (nA condition), white noise was continuously added when the videotapes for the A condition were played. In the noise-added auditory-visual condition (nAV

condition), the videotapes for the AV condition were played with the white noise. Signal to noise ratios in the noise-added conditions (nA and nAV), measured by a sound level meter at the location of the subject's head, were kept at 0 dB. The intensity of speech and noise was 55dB SPL (A scale, fast) that was 5dB higher than that in Sekiyama and Tohkura (1991). Background noise was 25-35dB SPL (A scale)[1]. The white noise was presented from a noise generator with its own speaker that was located beneath the monitor (Kanazawa), or it was presented from the loud speakers from which the auditory speech was also presented (New York).

Procedure

The videotapes were played on a VHS videocassette machine located in a control room adjoining the room in which subjects were tested. Subjects were instructed to report what they heard. In the auditory-visual conditions, they were instructed to look at and listen to each syllable. They were asked to write "what they heard, not what they saw" as an open choice. Instructions mentioned that they might occasionally hear a syllable that had a consonant cluster such as "bga". The subjects were asked to make only one response in each trial. In addition, they were asked to report any incompatibility between what they heard and what they saw. Therefore, they had to write down heard speech and to check for incompatibility in the 7 s trial duration. To report incompatibility, they had to mark a column on their response sheets only when they recognized incompatibility. Japanese subjects were instructed to write with *Romaji* (Roman alphabet) not in *Kana* (Japanese orthography) that cannot express any consonant cluster. (*Romaji* spells Japanese syllables approximately in the Roman alphabet. Every Japanese learns it at age eleven). In the auditory-only conditions, the subjects' task was only to report what they heard. The experimenter was a native speaker of Japanese and she instructed both language groups: in Japanese for Japanese subjects and in English for American subjects.

Each stimulus was repeated six times in each condition making 600 trials (100 x 6 blocks) for the AV and the nAV conditions and 60 trials (carried out in 1 block) for the A and the nA conditions. The order of the four conditions was fixed so that perceptually more ambiguous conditions came earlier : nAV, nA (1st day), AV, and A (2nd day). The total sessions for each subject required 4 hours including breaks and were conducted in two days.

RESULTS

Size of the McGurk Effect

Responses in the A Condition

As shown in Table 1, both language groups accurately identified most of the auditory syllables except /wa/ and /ra/, which indicates /wa/ and /ra/ have auditory dissimilarities between the two languages. Thus, /wa/ and /ra/ were excluded from the cross-language comparison of the McGurk effect. As well as the auditory dissimilarities, they had also visual dissimilarities between the two languages, which were revealed in the auditory-visual conditions. For the results for /w/ and /r/ in the AV condition, see Sekiyama and Tohkura (1993). Although /ba/ also had a poor auditory intelligibility score in the JE condition, it was not eliminated because of the visual similarities between the two languages. Another reason for including /ba/ was that, as seen in Table 1, most of the auditory errors ("pa") were within-place (labial) responses that would not prevent us from measuring the size of the McGurk effect.

Table 1. Confusion matrices for the auditory-only presentation in quiet (% in a row).

Japanese sub. (J stim.) N = 84

audition \ response	b	p	m	w	d	t	n	g	k	r'
b	100									
p	98				2					
m			100							
w				98						2
d					100					
t						100				
n							100			
g								100		
k									100	
r'										100

American sub. (J stim.) N = 60

audition \ response	b	p	m	w	d	t	n	g	k	l	r	others
b	100											
p		98				2						
m			100									
w				38						30	18	gr7 bl3 dl2 gw2
d	2				90							bd8
t		15	2			83						
n			2				96					mn2
g								98				dg2
k									100			
r'										51	10	gr17 dl10 rr10 wl2

Japanese sub. (E stim.) N = 72

audition \ response	b	p	m	w	d	t	n	g	k	r	r'	l	others
b	57	31		1	6							3	s1 f1
p		94			6								
m			100										
w	4			86			6					4	
d					100								
t						97			3				
n							100						
g								100					
k									100				
r				13				6		61	11	1	

American sub. (E stim.) N = 60

audition \ response	b	p	m	w	d	t	n	g	k	r	l	others
b	87	5			3							dt5
p		88				10						th2
m			100									
w				98							2	
d					98	2						
t						100						
n							100					
g								100				
k									100			
r										100		

Removing the results for /wa/ and /ra/, the average intelligibility scores across the rest of the eight auditory syllables were 99.7% (JJ condition), 96.0% (AJ condition), 93.6% (JE condition), and 96.7% (AE condition). When these means were compared by a two-factor ANOVA [Listener's Language (2) x Stimulus Language (2)], the main effects of Listener's Language and Stimulus Language were nonsignificant [F (1, 42) = 0.04, p< .90; F (1, 42) = 3.27, p< .10]. However, the interaction was significant [F (1, 42) = 4.94, p< .05]. The source of this interaction was tested by pairwise comparisons using least squares means. The results showed that a significant difference was only between the JJ and JE conditions (t = 3.05, p< .01). No significant differences were found in the other pairs of conditions. Therefore, the accuracy of the identification in the auditory-only condition was about the same across the four conditions with a reservation that the JE condition had a slightly poorer intelligibility score.

Responses in the AV condition

In the auditory-visual condition, the results are presented separately for two cases : the case in which the visual stimuli were labials (/b, p, m/) and in the case in which they were non-labials (/d, t, n, g, k/). The response patterns for the within-group visual tokens were almost identical, agreeing with those shown by Sekiyama and Tohkura (1991).

In the following analyses, responses[2] will be categorized into *auditory responses* and *visually influenced responses*. If a response to an incongruent auditory-visual stimulus is consistent with its auditory component with respect to the place of articulation, it will be called an auditory response (e.g., "ba"-response for auditory /pa/ with visual /na/). On the other hand, if a response includes a visual component of the stimulus with respect to the place, it will be taken as a visually influenced response. The visually influenced responses include *fused responses* (e.g., "ta"-response for auditory /pa/ with visual /na/) and *combined responses* (e.g., "pta"-response for auditory /ta/ with visual /pa/).

The confusion matrices in Table 2 show the results for the Japanese stimuli by indicating the frequency of each response as a percent of all responses in a row. The numbers in the leftmost column in parentheses indicate the percent of correct responses for the auditory syllable in the auditory-only condition ("auditory intelligibility score"). If the subjects hear the speech without visual influence, each of

the diagonal cells should contain the same response frequency as the auditory intelligibility score in the leftmost cell. Alternatively, if the McGurk effect occurs, it is expected that auditory labials paired with visual non-labials (upper three stimuli in the upper panels) will be perceived as non-labials (fused responses) and that auditory non-labials with visual labials (lower five stimuli in the lower panels) will be perceived as labials (fused responses) or as syllables with consonant clusters (combined responses).

Table 2. Confusion matrices for Japanese stimuli in the quiet auditory-visual condition (% in a row). Shadowed responses indicate the McGurk effect.

(A) Japanese sub. (J stim.)

vision = nonlabial (d,t,n,g,k) N=420

audition \ response	b	p	m	d	t	n	g	k
b (100)	92			8				
p (98)		45			55			
m (100)			86			13	2	
d (100)				100				
t (100)					100			
n (100)						100		
g (100)							100	
k (100)								100

vision = labial (b,p,m) N=252

audition \ response	b	p	m	d	t	n	g	k
b (100)	100							
p (98)		100						
m (100)			100					
d (100)	2			98				
t (100)		23			77			
n (100)						100		
g (100)							100	
k (100)		1						99

(B) American sub. (J stim.)

vision = nonlabial (d,t,n,g,k) N=300

audition \ response	b	p	m	d	t	n	g	k	others
b (100)	18			76			1		th4
p (98)		23			76				
m (100)			26			73			
d (90)				97					th3
t (83)					99				
n (96)						100			
g (98)				9			89		bg1 th1
k (100)					3			90	h7

vision = labial (b,p,m) N=180

audition \ response	b	p	m	d	t	n	g	k	others
b (100)	100								
p (98)		100							
m (100)			100						
d (90)	52			22					bd24 gb1 th1
t (83)		82			13				pt5 bt1
n (96)			59			18			mn23
g (98)	21			1			65		bg13 bl1
k (100)		29						61	pk6 bk4

As shown in Table 2, shadowed portions indicate the responses influenced by discrepant visual input (the McGurk effect). In Table 2A, Japanese subjects showed only a limited McGurk effect, replicating the results by Sekiyama and Tohkura (1991). In Table 2A, it seems that the McGurk effect is almost negligible except for auditory /pa/. No combined responses were observed, perhaps because the Japanese phonological system does not allow any phonological consonant clusters.

In contrast, American subjects were greatly influenced by the visual input. Although the JJ and AJ conditions presented identical stimuli, American subjects

(Table 2B) responded quite differently from Japanese subjects, showing a large McGurk effect for most of the auditoty syllables except /ga/ and /ka/ . In contrast to the JJ condition, combined responses (e.g., "bda", 24% ; "mna", 23% ; "bga", 13%) were observed.

The results for English stimuli are shown in Table 3 . In spite of the lack of distinction between /r/ and /l/ in the Japanese phoneme inventory, the responses of a Japanese subject often included both "ra" and "la" perhaps due to his/her knowledge of English. In the JE condition after the experiment, the experimenter asked the subjects which of Japanese /r/, English /r/, and English /l/ was represented by their "r" or "l"-responses. In Table 3 , the responses as Japanese /r/ are described by /r'/.

Table 3. Confusion matrices for English stimuli in the quiet auditory-visual condition (% in a row). Shadowed responses indicate the McGurk effect.

(A) Japanese sub. (E stim.)

vision = nonlabial (d,t,n,g,k) N=360

audition \ response	b	p	m	d	t	n	g	k	r	r'	l	others
b (57)	11	6		26	33		1		3	3	8	s4 f3 gw1
p (94)		16			82			2				
m(100)			24			34			6	18	19	
d (100)				99					2			
t (97)					97			3				
n (100)						100						
g (100)							100					
k (100)								100				

vision = labial (b,p,m) N=216

audition \ response	b	p	m	d	t	n	g	k
b (57)	61	38						
p (94)		100						
m(100)			100					
d (100)	1			99				
t (97)		53			47			
n (100)						100		
g (100)							100	
k (100)								100

(B) American sub. (E stim.)

vision = nonlabial (d,t,n,g,k) N=300

audition \ response	b	p	m	d	t	n	g	k	l	others
b (87)	2			19	2				38	th20 s17 f1
p (88)		6			80			7		th6 s1
m(100)			18			29			52	
d (98)				99						
t (100)					99			1		
n (100)						100				
g (100)				1			97	1		gl1
k (100)								99		

vision = labial (b,p,m) N=180

audition \ response	b	p	m	d	t	n	g	k	others
b (87)	100								
p (88)	1	99							
m(100)			100						
d (98)				100					
t (100)		44			56				pt1
n (100)			9			79			mn12
g (100)				1			98	1	
k (100)		1					1	99	

In contrast to the JJ condition, Japanese subjects in the JE condition (Table 3A) were greatly influenced by visual input especially for three auditory labials and also for /ta/ to some extent. Combined responses were negligible (only auditory /ba/ produced "gwa"-response 1 % of the time).

The AE condition (Table 3B) showed quite similar results. The McGurk effect was large in the three auditory labials and substantial in /ta/. Table 3B shows that the stimuli and experimental procedure in the present study replicate the large McGurk effect reported in English speaking cultures. To the extent that the visual effects on the auditory non-labials were weak, combined responses were less substantial (only auditory /na/ produced "mna"-response 12% of the time).

Among the four conditions, only the JJ condition showed results consistent with the intelligibility hypothesis. While the substantial McGurk effect was not observed for auditory syllables that had 100% of the intelligibility score in the JJ condition, it occurred even for auditory syllables with the intelligibility score of 100% in the other three conditions (/b/ and /m/ in the AJ condition, /m/ in the JE condition, and /m/ and /t/ in the AE condition).

Size of the Visual Influence in the Quiet Condition

The size of the visual influence was calculated for each auditory syllable by subtracting the place errors (confusions between labials and non-labials) in the A condition from the total place errors (fused and combined responses) in the AV condition. For example, auditory /pa/ in the JJ condition showed place errors ("ta"-response) 55% of the time when presented with visual non-labials in the AV condition (Table 2A) but it also showed place errors ("ta"-response) 2 % of the time in the A condition (Table 1). Thus, 53% (55% − 2 %) was considered as the influence of discrepant visual input. The mean size of the visual influence for each auditory syllable is shown in Fig. 1. It clearly shows that, of the four conditions, the visual influence is the smallest for the JJ condition in which only auditory /p/ shows a substantial visual effect. The strongest visual influence is observed in the AJ condition in which six of eight auditory syllables show 70%−80% visual influence. In the JE and AE conditions, the visual influence is strong if it occurs, but is limited to some auditory tokens (three labials and /t/).

To examine whether the size of visual influence was significantly big for each auditory syllable, a single-factor ANOVA with repeated measures tested the effect of discrepant visual input in each auditory syllable by comparing the frequency of place errors in the A condition with those in the AV condition. In the JJ condition, significant visual effects were found only in auditory /p/ [F (1, 13) = 33. 10, $p<$ .0001] and /t/ [F (1, 13) = 9. 02, $p<$ .05]. In the AJ condition, the visual effect was

significant in all auditory syllables. The significance level was $p< .0001$ for /b/ [F (1, 9) = 122.01], /p/ [F (1, 9) = 102.50], /m/ [F (1, 9) = 67.12], /t/ [F (1, 9) = 85.34], and /n/ [F (1, 9) = 204.57], $p< .001$ for /d/ [F (1, 9) = 38.54], and $p< .05$ for /g/ [F (1, 9) = 5.85] and /k/ [F (1, 9) = 9.42]. In the JE condition, the visual effect was significant in auditory /b/, /p/, /m/ and /t/. The significance level was $p< .0001$ for /b/ [F (1, 11) = 55.08], /p/ [F (1, 11) = 122.77], and /m/ [F (1, 11) = 56.97] and $p< .001$ for /t/ [F (1, 11) = 23.82]. In the AE condition, it was significant also in auditory /b/, /p/, /m/, and /t/. The significance level was $p< .0001$ for /b/ [F (1, 9) = 118.91], /p/ [F (1, 9) = 66.63], and /m/ [F (1, 9) = 53.67] and $p< .01$ for /t/ [F (1, 9) = 16.48]. These results suggest that auditory labials produced stronger visual effects than auditory non-labials.

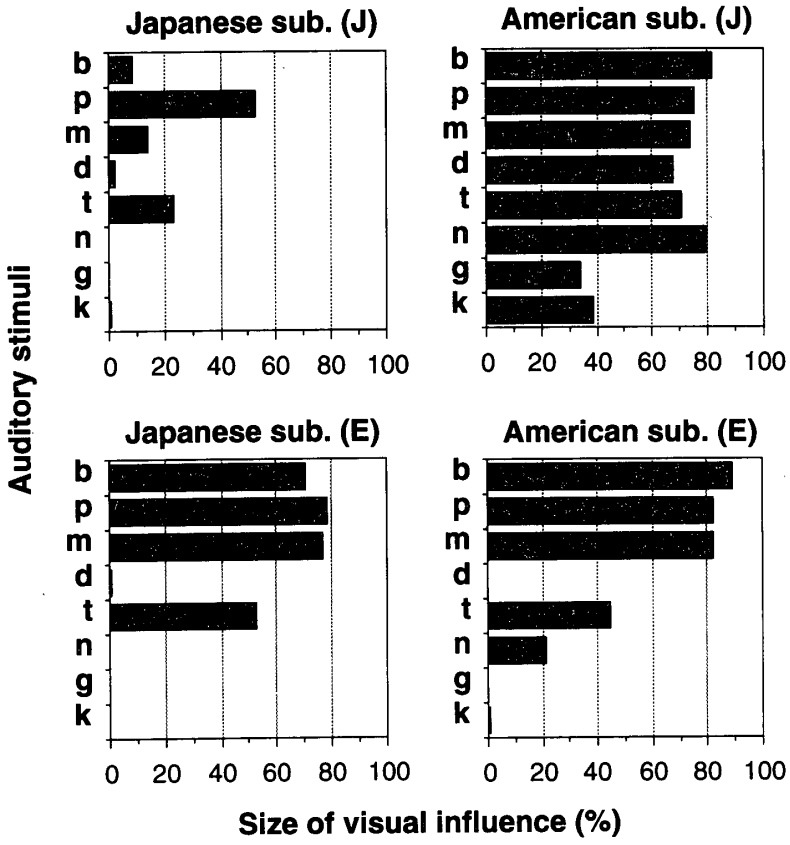

Figure 1. The size of the visual influence (%) in the quiet condition calculated by subtracting the place errors in the auditory-only condition from the total place errors (fused and combined responses) in the auditory-visual condition for incongruent (labial vs. non-labial) auditory-visual stimuli.

To compare the size of visual influence among the four conditions, average sizes were calculated for auditory labials and non-labials in each subject. The mean sizes of visual effects for auditory labials and non-labials were 25% and 5 % in the JJ condition, 77% and 59% in the AJ condition, 75% and 11% in the JE condition, and 84% and 13% in the AE condition. These means were compared by a 2 x 2 x 2 ANOVA [Listener's Language (Japanese vs. American English) x Stimulus Language (Japanese vs. English) x Auditory Stimulus (Labial vs. Non-labial)] with repeated measures. The ANOVA found the following results : (a) Main effects of Listener's Language [F (1, 42) = 38. 20, $p<$ .0001] and Auditory Stimulus [F (1, 42) = 162. 04, $p<$ .0001] were highly significant. That is, the size of visual influence was larger for American subjects than for Japanese subjects, and it was larger for auditory labials than for auditory non-labials. The main effect of Stimulus Language was not significant [F (1, 42) = 0. 84]. (b) The Listener's Language x Stimulus Language interaction was highly significant [F (1, 42) = 24. 08, $p<$ .0001] because visual effects were stronger for non-native speech than for native speech. (c) The Auditory Stimulus x Stimulus Language interaction was significant [F (1, 42) = 50. 98, $p<$ .0001], but the other interactions were nonsignificant.

Subsequently, sources of the above interactions were examined by pairwise comparisons using least squares means. The results were as follows. (1) The size of visual influence was larger for the English stimuli when auditory simuli were labial (t = 3. 95, $p<$ .001), but it was larger for the Japanese stimuli when auditory stimuli were non-labial (t = 5. 01, $p<$ .0001). (2) when auditory stimuli were labial, the size of visual influence in the JJ condition was significantly smaller than that in the other three conditions ($p<$ .0001 for each pair) and there were no significant differences among the other three. When the auditory stimuli were non-labial, the size of visual influence in the AJ condition was significantly larger than those in the other three conditions ($p<$ .0001 for each pair) and there were no significant differences among the other three.

To summarize, these results show that the McGurk effect was (1) stronger for American subjects than for Japanese subject and (2) stronger for non-native speech than for native speech and (3) stronger for auditory labials than for auditory non-labials. In addition, (4) the main effect of Stimulus Language (Japanese or English) was not significant. From the results (2) and (4), we can conclude that the small Japanese McGurk effect in the JJ condition is not because the stimuli were from

Japanese, but due to the fact that the stimuli were from the subjects' native language. The results (1), (2), and (4) indicate that the less vision-dependent processing of Japanese listeners becomes clear when they process speech information of their native language.

Effects of the Noise on the McGurk Effect

Table 4 shows confusion matrices for the nA condition. As they show, the auditory intelligibility score decreased in most of the syllables due to the added noise. However, the degrading effect of the noise was not the same on eight syllables.

Table 4. Confusion matrices for the auditory-only presentation in noise (% in a row).

Japanese sub. (J stim.) — response, N = 84

audition \ response	b	p	m	w	d	t	n	g	k	r'	others
b	98	1	1								
p		79				14			1	1	h5
m			96				2			1	
w			1	93						6	
d	13				87						
t		39				55			4		h2
n			1				98	1			
g	43				39		1	16			z1
k		7				13			21		h46 a12
r'	1		1	26	1		5	4		55	y7

American sub. (J stim.) — response, N = 60

audition \ response	b	p	m	w	d	t	n	g	k	l	r	others
b	98				2							
p		92				8						
m			98				2					
w				10		2				65	5	sl10 tl5 bl3
d	5				93							bd2
t		23				68						h8
n			7		5		93					
g	20				53			25				s2
k		40				3			8			h45 a3
r'				3			8			56		sl8 dl8 bd7 tl2 y2

Japanese sub. (E stim.) — response, N = 72

audition \ response	b	p	m	w	d	t	n	g	k	r	r'	l	others
b	38	25											f38
p		35				17			13				h18 f14 s4
m			100										
w			1	36					1		3		kw35 f19 kr4
d				1	97								y1
t		10				22			36				a14 h14 s3 f1
n							100						
g				1	11			65			8		z8 gr6
k		7				32			60				h1
r	1	3		36				4		6	1	1	kr17 kw10 gw10 br3 z3 --

American sub. (E stim.) — response, N = 60

audition \ response	b	p	m	w	d	t	n	g	k	r	l	others
b	60	7				3						f28 bp2
p		72				13			10			s2 h3
m			98									mb2
w				83							13	kw3
d					97			3				
t		25				45	2		22			h7
n							100					
g					32			67				dg2
k		3				28			67			s2
r										72	28	

Table 5 and Table 6 give the confusions between auditory stimuli and responses in the nAV condition. In the nAV condition, even the JJ condition produced a great number of visually influenced responses (Table 5A) for each of eight auditory syllables. Thus, Japanese subjects who showed only limited visual influence in the quiet condition made great use of visual information in the noise-added condition where visual compensation was necessary for the auditory degradation. This indicates that the small McGurk effect in the quiet JJ condition was not due to the

poorer lipreading ability of the Japanese subjects. For the other three condition, a main contribution of the added noise was to increase the McGurk effect on auditory non-labials paired with visual labials (lower panels in Table 5B, 6A, and 6B) for which the visual effect was comparatively weak in the quiet condition.

For English stimuli, noise induced many fricative responses though there were no fricative stimuli. It is noticeable for English auditory /b/ paired with visual non-labials where many "s" responses are seen (upper panels of Table 6). This fricative "s"-response to auditory /b/ is regarded as a visually influenced version of "f" responses that were frequently observed in the nA condition (Table 4, the JE and AE conditions). However, this type of fricative responses were negligible for Japanese auditory /b/, indicating an acoustical difference between the two languages.

Table 5. Confusion matrices for Japanese stimuli in the noise-added auditory-visual condition (% in a row). Shadowed responses indicate the McGurk effect.

(A) Japanese sub. (J stim.)

vision = nonlabial (d,t,n,g,k) — response — N=420

audition	b	p	m	d	t	n	g	k	others
b (98)	10			59	4		24	1	
p (79)		4			89			1	a4 h2
m (96)			3			95			r1
d (87)				98			2		
t (55)					96			2	h4
n (98)						100			
g (16)				82			17		r1
k (21)					38		2	32	a16 h11

vision = labial (b,p,m) — response — N=252

audition	b	p	m	d	t	n	g	k	others
b (98)	98	1							
p (79)		99			1				
m (96)			100						
d (87)	56			40			2		bd2 nd1
t (55)		98			2				
n (98)			44			54			mn1 my1
g (16)	94			5			1		
k (21)	1	87						1	a11

(B) American sub. (J stim.)

vision = nonlabial (d,t,n,g,k) — response — N=300

audition	b	p	m	d	t	n	g	k	others
b (98)	13			70	1		9		th5 l1
p (92)		5			94				
m (98)			7			88			mn3 l2
d (93)				94			1		th3
t (68)		1			97			1	
n (93)						97			mn3
g (25)				88	1		9		
k (8)					60			11	h28

vision = labial (b,p,m) — response — N=180

audition	b	p	m	d	t	n	g	k	others
b (98)	99	1							
p (92)		100							
m (98)			100						
d (93)	88			7			1		bd4 mb1 th1
t (68)		98			2				
n (93)			82			9			mn7 my1 ml1 nm1
g (25)	93	1		6			0		bd1
k (8)		98			1			0	h2

Table 6. Confusion matrices for English stimuli in the noise-added auditory-visual condition (% in a row). Shadowed responses indicate the McGurk effect.

(A) Japanese sub. (E stim.)

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

audition \ response	b	p	m	d	t	n	g	k	r	r'	l	others	N=360
b (38)	3	4		11	23		4	4	3	3	4	s28 f8 z1 h1 th1	
p (35)		0			43			29				s18 h9 f1	
m(100)			14			50	1		3	16	15		
d (97)				89			1		3	1	3	z2	
t (22)					41			31				a10 f10 s7	
n(100)						92			2	3	3		
g (65)				5		1	54		9	9	3	z12 gr6	
k (60)					42			49				s7 kr1	

vision = labial (b,p,m)

audition \ response	b	p	m	d	t	n	g	k	r	r'	l	others	N=216
b (38)	50	50											
p (35)		100											
m(100)			100										
d (97)	23			61			1					by14	
t (22)		98		1								a1	
n(100)			8			74			2	3		my13	
g (65)	22			2		1	24		12	2		br21 by12 bw2	
k (60)		91			3			4				pr2	

(B) American sub. (E stim.)

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

audition \ response	b	p	m	d	t	n	g	k	l	others	N=300
b (60)	1			7	1		1	3	22	s34 th9 st7 kl4 dt4 f2 gl2 tl2	
p (72)		3			54			28	2	th5 s3 h2 tl1 kl1	
m(98)			3			41			54		
d (97)				93			4			st2	
t (45)		1			59			35		h3 th1	
n(100)						99				ny1	
g (67)				36			62	1		dg1	
k (67)					30			67		kl1	

vision = labial (b,p,m)

audition \ response	b	p	m	d	t	n	g	k	others	N=180
b (60)	96	1							bp2 f1	
p (72)		99						1		
m(98)			100							
d (97)	23			62			1		bd11 bl1 bp1 bt1 by1 pd1	
t (45)	4	93			2			1		
n(100)			46			34			mn19 w1 ny1	
g (67)	21			21			35	1	bd7 bl6 br4 y1	
k (67)	3	73			3			19	pl1 bp1	

Although we employed a subtractive operation (subtracting place errors in the A condition from that in the AV condition) to calculate the size of visual influence in the quiet condition, it seemed to be inappropriate in the noise-added condition. For, whereas the noise simply increased place errors in the nA condition, the increase of place errors in the nAV condition seemed to be suppressed by a ceiling effect. The ceiling effect is suggested in many cases where visually influenced responses are observed over 90% of the time. Therefore, as an approximation of the size of the visual influence, an original value of place errors in the AV and nAV conditions was employed in analyses of the effect of the noise. (While the results of statistical analyses for the quiet condition showed no substantial difference between the original value and the subtracted value, those for the noise-added condition showed differences between them.)

For discrepant auditory-visual stimuli, the average percents of place errors are shown in Fig. 2 for both the quiet and noise-added conditions. Fig. 2 shows that the noise increased the size of visual influence (approximated by place errors). It also suggests the ceiling effect in the noise-added condition, showing the frequencies of

over 90% in many cases. In the noise-added condition, the differences among the four groups are smaller than in the quiet condition.

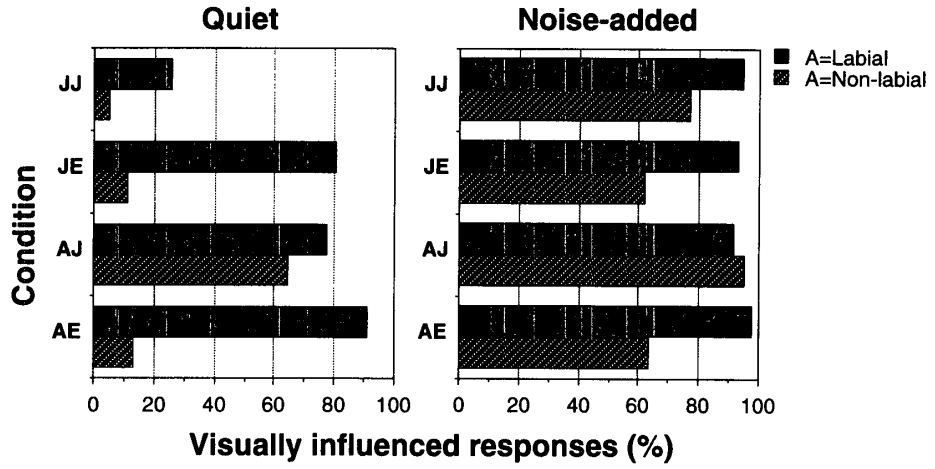

Figure 2. The size of visual influence approximated by the frequency of place errors for discrepant auditory-visual stimuli in the AV and nAV conditions (%).

To examine the effect of the noise on the size of visual influence, a four-factor ANOVE [Noise (2) x Listener's Language (2) x Stimulus Language (2) x Auditory Stimulus (2)] was carried out. The results are summarized as follows. (a) The main effect of Noise [F (1, 42) = 301. 31, p< .0001], Listener's Language [F (1, 42) = 26. 49, p< .0001], and Auditory Stimulus [F (1, 42) = 194. 49, p< .0001] were highly significant. That is, the visual effect was stronger in the noise-added condition than in the quiet condition; the effect was stronger for American subjects than for Japanese subjects, and it was stronger for auditory labials than for auditory non-labials. The main effect of Stimulus Language was not significant [F (1, 42) = 0. 51]. Therefore, the size of the visual influence did not differ for Japanese stimuli and English stimuli. (b) The interaction of Listener's Language x Stimulus Language was significant [F (1, 42) = 14. 26, p< .001]. (c) Reflecting the fact that the increase of place errors was bigger for auditory non-labials, the interaction of Noise x Auditory Stimulus was significant [F (1, 42) = 45. 32, p< .0001]. (d) In the other interactions that included within-subject factors (Noise and Auditory Stimuli), significant interactions were found in Noise x Listener's Language [F (1, 42) = 34. 28, p< .0001], Noise x Stimulus Language [F (1, 42) = 13. 83, p< .001], Noise x Listener's Language x Stimulus Language [F (1, 42) = 25. 48, p< .0001], Auditory Stimulus x Stimulus Language [F (1, 42) = 78. 05, p< .0001], Auditory Stimulus x Listener's Language x Stimulus Language [F (1, 42) = 4. 44, p< .05], and Noise x Auditory Stimulus x

Stimulus Language [F (1, 42) = 17. 03, p< .001].

Subsequently, the sources of the interactions were examined by pairwise comparisons using least squares means. The results are summarized as follows. (1) Whereas the American subjects showed the stronger visual influence than the Japanese subjects in the quiet condition for both auditory labials (t = 4. 50, p< .0001) and non-labials (t = 7. 93, p< .0001), there were no differences between the two language groups in the noise-added condition either for auditory labials (t = 0.28) or non-labials (t = 1. 73). (2) In the quiet condition, the visual influence was stronger for the English stimuli than for the Japanese stimuli when auditory stimuli were labial (t = 4. 89, p< .0001), but it was stronger for the Japanese stimuli when auditory stimuli were non-labial (t = 5. 90, p< .0001). In the noise-added condition, there were no differences between the two stimulus languages when auditory stimuli were labial (t = 0. 65), but the visual influence was stronger for the Japanese stimuli when auditory stimuli were non-labial (t = 4. 23, p< .0001). (3) In the noise-added condition, there were no significant differences in any pairs of the JJ, AJ, JE, and AE conditions when auditory stimuli were labial. When auditory stimuli were non-labial, the visual influence was significantly stronger for the AJ condition than that for the other threes; it was stronger for the JJ condition than that for the JE condition, and the difference between the JJ and AE conditions and the difference between the AE and JE conditions were nonsignificant.

These results showed that, in the noise-added condition, whereas the size of visual influence was the same across the four groups when auditory stimuli were labial (perhaps due to the ceiling effect), it was larger for the Japanese stimuli than for the English stimuli when auditory stimuli were non-labial. This may be related to the extent of the degradation of the auditory intelligibility score due to the noise. As the left most column of each panel in Table 5 and Table 6 shows, the degrading effect was stronger on Japanese non-labials than English non-labials except /t/. At any rate, the results for the noise-added condition also showed the tendency that auditory labials are more prone to visual effect than auditory non-labials.

Frequency of Incompatibility

Fig. 3 shows mean frequency of incompatibility in each condition for discrepant stimuli (i.e., auditory labials paired with visual non-labials, and vice versa) and congruent stimuli (auditory labials paired with visual labials, and auditory non-labials paired with visual non-labials). In both the quiet and noise-added conditions, incompatibility was reported mainly for discrepant stimuli, and its frequencies for congruent stimuli were almost negligible. That is, the subjects reported incompatibility only for the stimuli that had the different places of articulation for their auditory component and visual component. This clear difference between discrepant and congruent stimuli implies that the subjects could lipread the place of articulation.

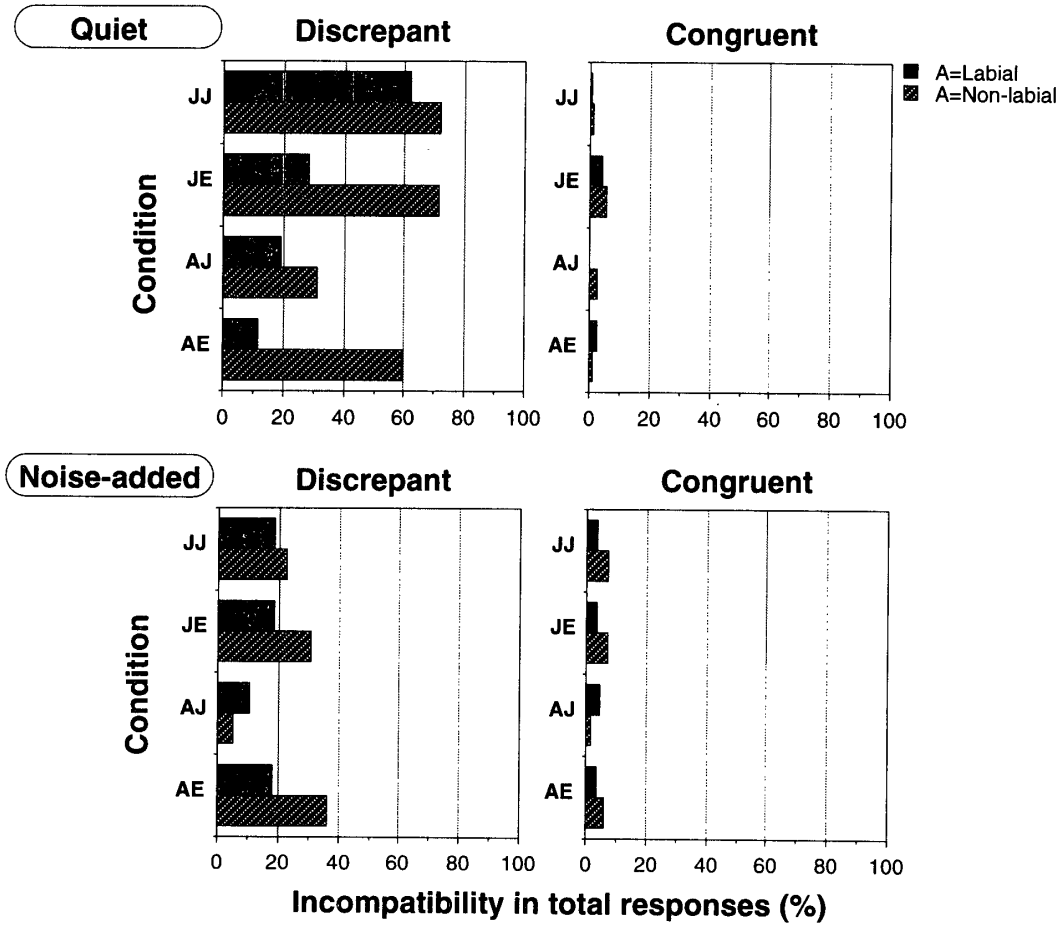

Figure 3. The frequency of reported incompatibility (%) in total responses.

For the discrepant stimuli, however, incompatibility was not always reported. In the quiet condition, whereas the JJ condition showed incompatibility for the majority of the responses, the other three conditions showed less incompatibility especially for auditory labials. Incompatibility was reported more frequently by the Japanese subjects, and it was reported more frequently for stimuli whose auditory component was non-labial. In the noise-added condition, each group reported less incompatibility.

Statistical analyses were done only for the case of the discrepant stimuli. The results of a four-factor ANOVA (Noise x Listener's Language x Stimulus Language x Auditory Stimulus) are summarized as follows. (a) The main effects of Noise [F (1, 42) = 73.21, p< .0001], Listener's Language [F (1, 42) = 13.91, p< .001], and Auditory Stimulus [F (1, 42) = 41.39, p< .0001] were significant. That is, the frequency of incompatibility was higher in the quiet condition than in the noise-added condition; it was higher in the Japanese subjects than in the American subjects, and it was higher for auditory non-labials than for auditory labials. As compared with the results in the earlier section, these relations among the conditions are opposite to those in the place error data. The main effect of Stimulus Language was not significant [F (1, 42) = 0.82], agreeing with the place error data. (b) The interaction of Listener's Language x Stimulus Language [F (1, 42) = 5.90, p< .05] was significant. This interaction reflects the fact that the incompatibility was reported more frequently for the stimuli from the subjects' native language. (c) Among the interactions that included within-subject factors, significant interactions were found in Noise x Listener's Language [F (1, 42) = 15.68, p< .001], Noise x Stimulus Language [F (1, 42) = 6.77, p< .05], Auditory stimulus x Stimulus Language [F (1, 42) = 20.90, p< .0001], Noise x Auditory Stimulus [F (1, 42) = 39.58, p< .0001], and Noise x Auditory Stimulus x Stimulus Language [F (1, 42) = 7.53, p< .01].

Subsequently, the sources of these interactions were examined by pairwise comparisons using least squares means. The results are summarized as follows. (1) In the quiet condition, the Japanese subjects reported significantly more frequent incompatibility than the American subjects, both for auditory labials (t = 4.45, p< .0001) and non-labials (t = 3.54, p< .001). In the noise-added condition, there were no differences between the two language groups either for auditory labials (t = 0.87) or non-labials (t = 1.16). (2) In the quiet condition, more frequent incompatibility

was reported for the Japanese stimuli when auditory stimuli were labial ($t = 3.11$, $p < .01$), but there were no differences between the two stimulus languages when auditory stimuli were non-labial ($t = 1.86$). In the noise-added condition, there were no differences between the two stimulus languages when auditory stimuli were labial ($t = 0.69$), but more frequent incompatibility was reported for the English stimuli when auditory stimuli were non-labial ($t = 3.62$, $p < .001$). (3) In the quiet condition, when auditory stimuli were labial, the JJ condition showed significantly more frequent incompatibility than the other three conditions, and there were no differences among the other three. when auditory stimuli were non-labial, the AJ condition showed the significantly less frequent incompatibility than the other three conditions, and there were no differences among the other three. (4) In the noise-added condition, when auditory stimuli were labial, there were no differences among the four conditions. When auditory stimuli were non-labial, the AJ condition showed significantly less frequent incompatibility than the other three conditions, and there were no differences among the other three.

As compared with the results in the earlier section, these results suggest that more frequent incompatibility was reported in conditions where the size of visual influence was smaller.

Correlation Between Incompatibility and Place Errors

Comparing the panels for discrepant stimuli in Fig. 3 with corresponding panels in Fig. 2 , it was suggested that the frequency of the incompatibility is negatively correlated with the frequency of place errors which are regarded as the size of visual influence. Fig. 4 plots the relationship between the two indexes for each of the eight auditory syllables in the quiet and noise-added conditions. As shown in Fig. 4 , the 16 data points in each condition fitted to a negative linear function with a high correlation coefficient ranging from $r = -.906$ to $r = -.985$. When the correlation was calculated based on individual data, it was $r = -.809$ (JJ), $r = -.816$ (AJ), $r = -.820$ (JE), and $r = -.751$ (AE). The results clearly show that the more frequent incompatibility there is, the smaller the size of visual influence becomes.

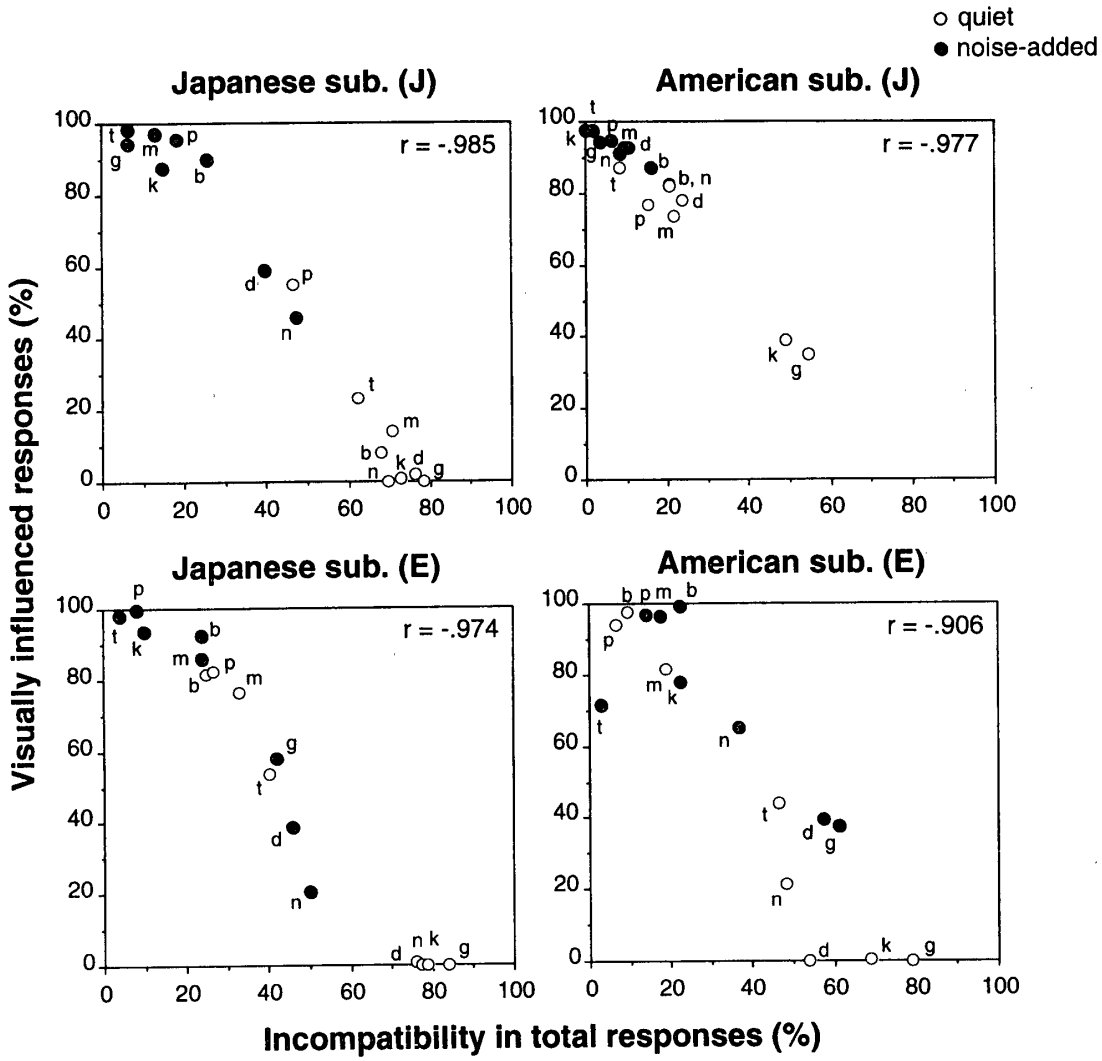

Figure 4. The relationship between the frequency of incompatibility (%) in total responses and the size of the visual influence approximated by the frequency of place errors (%).

Fig. 4 also describes the differences among the four conditions. In the JJ condition, the contrast between the quiet and noise-added conditions is very clear : The data points for the quiet condition are characterized by the weak visual influence and the frequent incompatibility, while the data for the noise-added condition shows the opposite tendency. In the AJ condition, the contrast is also clear, but the data points gather at the upper left of the panel, indicating the strong visual influence and the infrequent incompatibility even in the quiet condition.

In the JE and AE conditions, the contrast is not clear. As we saw in the earlier section (Fig. 1), these two groups showed the large McGurk effect only on auditory /ba/, /pa/, /ma/, and /ta/ in the quiet condition. The data points for these four syllables intrude into the upper left region where the data points for the noise-added condition exist. In these two groups, the distance between auditory labials (the

strong visual influence and the infrequent incompatibility) and non-labials (the weaker visual influence and the more frequent incompatibility) was larger than the distance between the noise-added and quiet conditions. In this sense, the responses in the JE and AE conditions were similar, suggesting a stimulus characteristic of the English stimuli. It is suggested that English auditory labials paired with visual non-labials have some characteristics that are prone to visual effects.

In summary, Fig. 4 clearly shows that the small McGurk effect in the quiet JJ condition was highly correlated with the frequent incompatibility: The subjects experienced incompatibility instead of the perceptual fusion effect. In each condition, the McGurk effect did not occur on a stimulus for which the subjects perceived incompatibility more than 70% of the time. The fact that the Japanese subjects reported more frequent incompatibility suggests that Japanese listeners are more sensitive to the auditory-visual discrepancy.

Sensitivity to the Auditory-Visual Discrepancy

In the quiet condition, finer analyses of the incompatibility data were possible because the subjects reported substantial incompatibility. In Fig. 5, the incompatibility data for discrepant stimuli in the quiet condition are shown for two cases: The left panel is for the case when the subjects made auditory responses and the right panel is for the case when the subjects made visually influenced (fused and combined) responses. Comparing the right panel with the left one, it is clear that the auditory responses were accompanied with much more incompatibility than the visually influenced responses. In the auditory responses, American subjects show less incompatibility than Japanese subjects especially for auditory labials.

The difference between the auditory and visually influenced responses can be explained as follows. Suppose that auditory /pa/ is presented with visual /na/. If the subject's response is "ba", that is, an auditory response, "ba" and /na/ are different with respect to the place of articulation. Thus, it is reasonable to report incompatibility between what the subject heard and what he/she saw if he/she lip-reads information about the place. In this sense, the frequency of incompatibility in the auditory responses is regarded as a direct index of the sensitivity to the auditory-visual discrepancy. On the other hand, when a response is visually influenced (e.g., "ta"-response for auditory /pa/ and visual /na/), incompatibility is not necessarily expected because what the subject heard is not incongruent with what he/she saw

with respect to the place of articulation ("ta" and /na/ are both non-labial).

Based on this explanation, the auditory responses were examined further. The frequency of incompatibility in the auditory response was considered an index of the sensitivity to the auditory-visual discrepancy. As shown in the left panel of Fig. 5, the frequency of in compatibility in the auditory responses was apparently higher in the Japanese subjects than in the Americans. This indicates that the Japanese subjects were more sensitive to the discrepancy than the Americans. It is striking that the results were almost the same irrespective of the stimulus language, when the subjects' native language was the same. In the JJ and JE conditions, the Japanese subjects detected the discrepancy about 70% of the time for both auditory labials and non-labials. In the AJ and AE conditions, the American subjects reported less incompatibility especially for auditory non-labials. In these cases, the frequency of incompatibility did not reach 50% (39% in the AJ condition, and 31% in the AE condition) even though the responses were auditory ones. These low frequencies of incompatibility in the American subjects are notable, because studies in English speaking cultures have been reported that the McGurk effect is stable for this type of stimuli (i.e., auditory labials paired with visual non-labials). This coincidence suggest that this type of stimuli has a nature which encourages native speakers of English to integrate auditory and visual information. On the whole, judging from the striking resemblance of the data between the JJ and JE conditions and between the AJ and AE conditions, it seems that this index can assess each language group's sensitivity to the auditory-visual discrepancy.

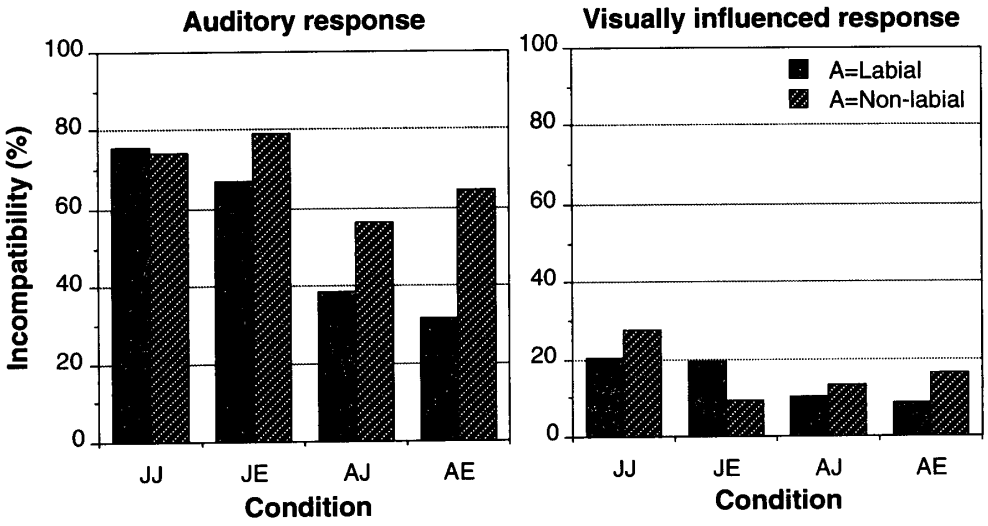

Figure 5. The frequency of incompatibility (%) in the auditory responses and the visually influenced responses (fused and combined responses) for discrepant auditory-visual stimuli in the AV condition.

A 2 x 2 x 2 ANOVA (Listener's Language x Stimulus Language x Auditory Stimulus) was carried out on the data for the auditory responses. (It should be noted that for auditory labials, there were missing observations for several subjects who showed no auditory responses. The number of such subjects were two for the AJ condition, four for the JE condition, and four for the AE condition.) The ANOVA found main effects of Listener's Language [F (1, 32) = 8. 39, $p<$ .01] and Auditory Stimulus [F (1, 32) = 5. 44, $p<$ .05] significant. These main effects show that (a) Japanese subjects were more sensitive to the auditory-visual discrepancy than American subjects and that (b) more frequent incompatibility was reported for auditory non-labials than for auditory labials. The main effect of Stimulus Language and the interactions were not significant.

Although the above ANOVA failed to find significance for the interaction of Listener's Language x Auditory Stimulus, this may be due to the data exclusion in the ANOVA due to missing observations (missing observations occurred when all of a subject's responses were visually influenced). For, the above ANOVA excluded all the data of subjects who had a missing observation, even though the missing observations were found only for auditory labials. Therefore, separate analyses were done for auditory labials and non-labials by two-factor ANOVAs (Listener's Language x Stimulus Language). When auditory stimuli were labial, the main effect of Listener's Language was significant [F (1, 32) = 9. 02, $p<$ .01], but the main effect of Stimulus Language and the interaction was nonsignificant [F (1, 32) = 0. 43; F (1, 32) = 0. 00, respectively]. When auditory stimuli were non-labial, none of the effects were significant [F (1, 42) = 3. 62, $p<$ .10, for Listener's Language; F (1, 42) = 0. 64, for Stimulus Language; F (1, 42) = 0.04, for the interaction]. These results imply that the difference in sensitivity between the two language groups was reliable only when auditory stimuli were labial. That is the American subjects were less sensitive to the auditory-visual discrepancy than the Japanese subjects only when auditory stimuli were labial.

DISCUSSION

The present study replicated the small McGurk effect in Japanese subjects listening to Japanese stimuli in the quiet condition. Specifically, the size of the

McGurk effect in the quiet JJ condition was smaller than any of the other three conditions when auditory stimuli were labial.

For our primary concern, the results clarified that the size of the McGurk effect is a negative linear function of the frequency of incompatibility in total responses. When responses were plotted on this function, the small McGurk effect in the quiet JJ condition can be expressed as points in the right lower part of the function. That is, in most of the cases in the quiet JJ condition, the subjects perceived incompatibility, and the McGurk effect was absent. This clearly shows that the small McGurk effect in the quiet JJ condition cannot be explained by an account that the Japanese subjects ignore visual information. They paid attention to the visual information in an early stage of processing so that they feel it incompatible with the auditory information. Their smaller McGurk effect and higher frequency of incompatibility indicate that Japanese listeners tend to separate the discrepant visual information from the auditory information. On the other hand, the American subjects' larger McGurk effect and lower frequency of incompatibility indicate that Americans tend to integrate them.

As described in the earlier report (Sekiyama & Tohkura, 1993), we raise two hypothetical explanations of this perceptual processing to separate two modalities of information. One is less frequent eye contact of Japanese listeners and the other is a simpler structure of the Japanese phonological system. These two factors may encourage Japanese listeners to form a perceptual processing in which visual speech information is not integrated with auditory information unless visual compensation is necessary. Such a processing would easily detect the auditory-visual discrepancy. The present analyses of the incompatibility data clearly described that the Japanese subjects in the quiet JJ condition paid attention to the visual information in an early stage of the processing so that they feel it incompatible with the auditory information. Therefore, their small McGurk effect implies that they discard the discrepant visual information in a later stage of the processing.

Although a correlation does not necessarily determine causality, additional evidence suggests that incompatibility is a cause of the small McGurk effect. For, as we saw in the frequency of incompatibility in *auditory responses*, the data suggested that the Japanese subjects were more sensitive to the auditory-visual discrepancy. To discriminate two indexes, let the incompatibility in the auditory responses be IA, and let the incompatibility in the total responses be IT. Note that the auditory responses

imply the absence of the McGurk effect, whereas the total responses include both the absence and the presence of the McGurk effect. Hence, the frequency of IT can be higher for the Japanese subjects even if incompatibility is a consequence of the absence of the McGurk effect, rather than a cause of it, because the absence of the McGurk effect was more frequent for the Japanese subjects. However, the frequencies in % of IA should be equal for the two language groups if incompatibility is a mere consequence of the absence of the McGurk effect, because the percents of IA is independent of the frequency of the auditory responses. On the contrary, the results showed a clear difference between the two language groups in IA. Accordingly, the frequency of IA should be considered to indicate a factor that influences the size of the McGurk effect. In the last section, we called it sensitivity to the auditory-visual discrepancy. It is plausible that the higher sensitivity to the discrepancy interrupts integration of the two sources of information. It should be also noted that the difference between the two language groups in sensitivity was reliable only for auditory labials. This corresponds with the fact that the difference in the size of the McGurk effect between the JJ and AE conditions was significant only for auditory labials.

The results of the JJ condition also replicated the results by Sekiyama and Tohkura (1991) in terms of the intelligibility hypothesis that the McGurk effect hardly occurs for completely intelligible auditory stimuli. In the JJ condition, only auditory /p/ showed the intelligibility score less than 100% (98%) in the A condition, and it was only one that produced a considerable McGurk effect (53%) in the AV condition. One might argue that the intelligibility difference between 100% and 98% is too trivial. Perhaps 100% and 98% do not make a statistically reliable difference. However, they are theoretically different because an intelligibility score of 100% includes a ceiling effect. In fact, whenever presented with this auditory syllable with intelligibility score of 98%, the author notices a kind of poor quality to the speaker's pronunciation though the syllable can be accurately identified. The poor quality would cause perceptual ambiguity that can easily induce visual effects. This might be clarified by another index than an intelligibility score that is more appropriate to describe the quality of the speech. The fact that the Japanese subjects employed the visual information depending on the ambiguity of the auditory information also suggests that their discard of the visual information occurs at a relatively later stage of the processing.

The other three conditions did not support the intelligibility hypothesis. In the AJ, AE, and JE conditions, strong McGurk effects occurred even for the auditory tokens with an intelligibility score of 100%. Thus, the auditory intelligibility hypothesis was applicable only to Japanese subjects listening to their native speech. However, the problem of the quality of speech should also be addressed to the overall results. The fact that the foreign speech stimuli produced a larger visual effect than the native stimuli suggests that the quality of speech in the foreign language was different from that in the native speech though there were no substantial differences in the intelligibility score. It should also be noted that in the noise-added condition, a small decrease of the intelligibility score sometimes increased the size of visual influence a great deal (e.g., in the JJ condition, the auditory intelligibility score of /ba/ decreased only 2 % due to the noise, but the size of visual influence increased more than 80%). This suggests that even though the decrease of the intelligibility score is small, the noise can cause a substantial degradation in the quality of speech, forcing the subjects to depend on the visual information. It remains to find an index that can describe the quality of speech in foreign speech and noise-added stimuli.

In contrast to the present results, Massaro, Tsuzaki, Cohen, Gesi, and Heredia (1993) tested native speakers of Japanese, American English, and Spanish and drew the conclusion that there was no difference among language groups in the manner of auditory-visual integration. However, it should be noted that there are many differences between their study and the present study. The biggest difference would be the nature of the stimuli: Their auditory stimuli were a continuum of sounds covering the range from /ba/ to /da/ that were synthesized from natural English speech. Although they used artificial speech to equate the stimuli across language groups, this will cause a problem. First, artificial speech would resemble foreign speech because of its acoustical deviation from natural speech in the subject's native language. Secondly, these stimuli might have selectively strengthened the visual influence on Japanese subjects because the stimuli were generated from English speech. If so, their results do not disagree with the present results that Japanese subjects used visual information as much as American subjects for English stimuli.

Various analyses in the present study showed differences between labials and non-labials. Fist, the McGurk effect was stronger for auditory labials than for auditory non-labials in both the quiet and noise-added conditions. Secondly, incompatibility was reported less frequently for auditory labials than for auditory

non-labials in both the quiet and noise-added conditions. Thirdly, the subjects were more sensitive to the auditory-visual discrepancy for auditory non-labials paired with visual labials than conversely paired stimuli. One possible explanation for this labial vs. non-labial asymmetry is the salience of visual cues. Visual labials (/b, p, m/) that include lip closure are considered as more salient than visual non-labials. Thus, the salient visual labials will more often let the subjects detect an auditory-visual discrepancy than visual non-labials. If so, auditory non-labials paired with visual labials will more often yield incompatibility than conversely paired stimuli. Because of the more frequent incompatibility, visual labials will more often interrupt integration of incongruent auditory-visual information than will visual non-labials. In that case, auditory non-labials paired with visual labials will produce weaker McGurk effects than conversely paired stimuli. The high correlation between the size of visual influence and the frequency of incompatibility supports this explanation.

Another possible explanation for the asymmetry is the number of consonants in a phoneme inventory. Both the English and the Japanese language have many more non-labials than labials. Therefore, when the information contains ambiguity, responses may be biased to non-labials because the non-labials have a higher probability of occurrence than labials.

As we saw in the data for auditory responses, the American subjects showed particularly low sensitivity to the discrepancy between auditory labials and visual non-labials (in both the AJ and AE conditions). In accordance with this, they produced the large McGurk effect for auditory labials. Because it was observed in both the AJ and AE conditions, we should interpret this low sensitivity to the discrepancy between auditory labials and visual non-labials as a constraint from the subject's native language. Although the cause is not clear, it may also be related to the number of consonants in a phoneme inventory. The English phoneme inventory has more non-labials than the Japanese.

As shown in Fig. 4, there were common observations in the AE and JE conditions that the distance on the graph between auditory labials (the large visual influence and the infrequent incompatibility) and non-labials (the smaller visual influence and the more frequent incompatibility) was larger than the distance between the noise-added and quiet conditions. This similarity suggests that English auditory labials paired with visual non-labials are especially prone to visual effects compared with corresponding Japanese stimuli. This remains for further analyses.

In the AE condition, auditory non-labials produced a much weaker McGurk effect than that shown in the English literature whereas auditory labials replicated the strong McGurk effect. However, considering that the present study used a large number of syllables (ten), the results are not inconsistent with the earlier research. In fact, MacDonald and McGurk (1978) reported only a weak visual effect for auditory non-labials by using eight (/ba, pa, ma, da, ta, na, ga, ka/) auditory syllables dubbed onto eight visual syllables, while studies using small numbers of syllables have reported strong visual effects for auditory non-labials as well as for auditory labials (Green, Kuhl, Meltzoff, & Stevens, 1991; Massaro & Cohen, 1983; McGurk & MacDonald, 1976). It may be that the small number of stimuli (e.g., only two such as /ba/ and /ga/) make subjects sensitive to differences between congruent auditory-visual stimuli and incongruent ones, so that the responses to incongruent stimuli are easily biased toward visual information.

The fact that the intelligibility hypothesis is applicable to the JJ condition but not to the AE condition also suggests that the manner of processing discrepant auditory-visual information is different between the two language groups. This means that American listeners integrate discrepant visual information with auditory information even when the auditory information is completely intelligible. On the other hand, Japanese listeners do not do so unless visual support is necessary. This contrast coincides anecdotal evidence. That is, native speakers of English often dislike dubbed movies (Massaro, 1987, P.36); they may experience something like the McGurk effect for the auditory-visual discrepancy that dubbed movies often include. On the other hand, Japanese listeners enjoy them simply although some intellectual people complain of the loss of an atmosphere created in the original language. The results of the present study show that a model of auditory-visual speech perception should take this inter-language difference into account.

REFERENCES

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

Blamey, P. J., Dowell, R. C., Brown, A. M., Clark, G. M., & Seligman, P. M. (1987). Vowel and consonant recognition of cochlear implant patients using formant-estimating speech processors. *Journal of the Acoustical Society of America*, **82**, 48–57.

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.

Erber, N. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, **40**, 481−492.

Fukuda, Y., Shiroma, M., & Funasaka, S. (1988). Speech perception ability of the patients with artificial inner ear. *Technical Report of IEICE* (The Institute of Electronics, Information and Communication Engineers of Japan), SP88−91. (in Japanese with English abstract)

Funasaka, S., Shiroma, M., Yukawa, K., Iizuka, N., Yao, M., Kono, J., Takahashi, H., & Kumakawa, K. (1989). Consonant information transmitted through 22-channel cochlear implant. *Audiology Japan*, **32**, 146−151. (in Japanese)

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

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across 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. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. New Jersey: Lawrence 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., 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.

O'Neil, J. J. (1954). Contribution of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders*, **19**, 429−439.

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.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In R. Campbell and B. Dodd (Eds.), *Hearing by eye: The psychology of lip-reading* (Pp. 3−51). London: Lawrence Erlbaum.

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.

NOTES

1) The back ground noise was about 5dB higher in New York [30−35dB SPL (A)] than in Kanazawa. To check the effect of the 5dB difference in background noise, four subjects in New York were tested in an anechoic room in the Lexington Center in which the background noise was about 20 dB SPL (A): two subjects in the AE experiment and two in the AJ experiment. For them, three of the six sessions in the AV condition were conducted at CUNY and the remaining three were conducted at the Lexington Center. There were no differences in the size of the McGurk effect between two places. Therefore, the 5dB difference in the background noise could be ignored at least for American subjects.

2) A few subjects in the AJ and AE conditions occasionally made a response with a final consonant such as "bap". They perhaps took the speaker's post-pronunciation lip closure as a bilabial closure. As the final consonant is not important in the present study, it was ignored and a "bap"-response was counted as "ba".