Journal of Phonetics (1993) **21**, 427–444

Inter-language differences in the influence of visual cues in speech perception

Kaoru Sekiyama*

Department of Psychology, Faculty of Letters, Kanazawa University, Kakuma-machi Kanazawa 920-11, Japan

Yoh'ichi Tohkura

ATR Human Information Processing Research Laboratories, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Received 26th August 1991, and in revised form 22nd December 1992

This paper reports on inter-language or cultural differences in audio-visual speech perception. When presented an audio signal dubbed onto a video recording of a talking face producing an incongruent syllable, robust visual effects of visual cues ("McGurk effect") have been reported for native English speakers in English-speaking cultures. The present experiment, however, showed that Japanese listeners were less influenced by visual cues than were Americans. We tested 10 Japanese subjects with 10 Japanese syllables, another group of 10 Japanese with 10 corresponding English syllables, 10 American subjects with Japanese syllables, and another group of 10 Americans with English syllables. There was a significant effect of native language (the native English speakers' responses showed much more effect of the incongruent visual cues) and a significant interaction with the stimulus language (each group showed more effect of the visual cues in stimuli from nonnative language). The results also suggested some acoustic and articulatory differences in [r] and [w] between two languages.

1. Introduction

It is known that visual (lip-read) information has a role in face-to-face verbal communication. This was dramatically demonstrated by McGurk & MacDonald (1976): when they showed a film of a speaker in which audio speech signals of [ba] had been dubbed onto visual lip movements for [ga], normal adults reported hearing an acoustically erroneous "da" 98% of the time. This "McGurk effect" phenomenon demonstrates that visual information on place of articulation easily influences phonetic perception. In this example, visual place information ([ga]: non-labial) biased auditory information ([ba]: labial) with a resulting percept ("da": fused response) that was consistent with the biased consonant place information and the remaining auditory information.

* Address until November 1994: Room 36-747, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, U.S.A.

0095-4470/93/040427 + 18 $08.00/0

Although this visual biasing effect on speech perception has been replicated in many studies and established as a robust effect in English speaking cultures (Dodd, 1977; MacDonald & McGurk, 1978; Summerfield, 1979, 1987; Massaro, 1983, 1987; Green & Kuhl, 1989; Dekle, Fowler & Funnell, 1992), it has not yet been examined for language- or culture-independence.

Suggesting a language- or culture-dependent aspect, we recently reported that Japanese listeners show very little visual bias for Japanese syllables when the quality of auditory stimuli is extremely high (Sekiyama & Tohkura, 1991). For example, when an auditory [ba] stimulus was presented in synchrony with a visual [ga] in a "noiseless (no noise added) condition", our Japanese subjects reported hearing acoustically correct "ba" responses 100% of the time, in contrast to the fused "da" responses reported by most English-speaking subjects in the earlier literature on the McGurk effect. On the other hand, when noise was added to the audio signals in a "noise-added condition", the Japanese subjects reported fused responses very frequently. The relationship between the auditory intelligibility score (frequency of correct identification for the audio-alone presentation) and the frequency of visually biased responses indicated that the visual bias for Japanese subjects listening to Japanese syllables depends on the auditory intelligibility score: in the noiseless condition, while three auditory syllables with the auditory intelligibility score less than 100% (95–98%) produced a weak visual bias, the other seven auditory syllables with the auditory intelligibility score of 100% produced very little visual bias. In the noise-added condition, where the Japanese subjects showed a strong McGurk effect for every auditory syllable, the intelligibility of the auditory stimuli was considerably degraded. Based on these results, we raised an "auditory intelligibility hypothesis" that Japanese subjects listening to Japanese speech are hardly influenced by visual cues when audition provides enough information and that the size of the visual bias to their responses depends on the intelligibility score of auditory stimuli. Although the intelligibility score does not give an absolute value of the intelligibility (the intelligibility score varies depending on the number of stimuli and the score of 100% may include a ceiling effect), our results suggested that an intelligibility score of 100% could be a critical point for the absence of the McGurk effect in Japanese subjects listening to Japanese syllables.

In earlier research, a large visual bias has been found for English speaking subjects for highly intelligible auditory syllables (e.g., an average intelligibility score of 99.4% for four stimuli reported by McGurk & MacDonald, 1976; the lowest intelligibility score of 98% for four stimuli reported by Green, Kuhl, Meltzoff & Stevens, 1991). This fact is inconsistent with our results for Japanese subjects. This inconsistency, however, could be attributed to various causes, such as differences in speech production (stimulus factor), in perceptual processing of the listeners (subject factor), or even in specific stimuli and experimental conditions used in the various studies. That is, one possible cause of the inconsistency is that English speech stimuli have acoustic and/or articulatory properties that produce bigger visual effects than Japanese stimuli do. The second possibility is that English-speaking subjects have a perceptual processing through which visual speech cues are easily integrated into the speech percept, whereas the perceptual processing of Japanese subjects does not incorporate visual cues into the percept as long as there is enough auditory information to identify the speech. The third possibility is that

the discrepancy is due to diffences in other experimental variables such as the number of stimuli, the specific response task, and so on.

The purpose of the present study was to determine which of these possible factors accounts for the relatively small McGurk effect in Japanese listeners for Japanese speech, by extending the test to three new conditions. Condition AJ (American subjects, Japanese stimuli) tested native speakers of American English with the identical Japanese speech stimuli under identical viewing and listening conditions. In conditions AE (American subjects, English stimuli) and JE (Japanese subjects, English stimuli), American subjects and Japanese subjects were presented with English speech stimuli pronounced by a native speaker of American English. For convenience in comparison, a part of our previous study (Sekiyama & Tohkura, 1991) is described here as condition JJ (Japanese subjects, Japanese stimuli). We found that American subjects' responses showed a large effect of incongruent visual cues than Japanese subjects' responses. Each group showed a larger visual biasing effect for stimuli of the non-native language than for those of the native language.

2. Method

2.1. *Stimuli*

Stimulus materials were the 10 Japanese syllables ([ba] [da] [ga] [ka] [ma] [na] [pa] [ra] [ta] [wa]) and the corresponding English syllables. The distinction between Japanese and English syllables was based on the native language of the speaker who pronounced them. Female native speakers, one Japanese and one American, were employed. The speaker's face was videotaped (tape 1) while she pronounced the syllables. Audio signals were recorded separately (tape 2) to avoid degrading factors in the recording process. Using these video and audio signals, the 10 audio and 10 video stimuli were combined, giving 100 audio-visual stimuli for each speaker. To synchronize audio signals with video signals, the dubbing timing was controlled by a 33-ms frame unit using a videocasette recording system of broadcast quality (Sony Betacam). Audio-visual stimuli were created by replacing the original audio signals on tape 1 with the new audio signals on tape 2. For both tapes, the frame numbers on which audio signals existed 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 starting frame was synchronized if the duration of the original and the new audio signals differed, as was often the case for the incongruent audio-visual dubbing.

2.2. *Subjects*

Subjects were 20 native speakers of Japanese and 20 native speakers of American English with normal hearing and normal or corrected to normal vision, ranging in age between 20 and 30. In each language group, subjects were divided into two groups and each group participated in either the Japanese stimuli condition or the English stimuli condition. Thus, each of the four conditions (JJ, AJ, AE and JE) had 10 subjects. Subjects for condition JJ were female secretarial staff at a research laboratory in Osaka, Japan. The other three conditions had approximately equal

number of male and female subjects. Subjects for condition JE were students at Kanazawa University. Most of the subjects for condition AJ and AE were students at the City University of New York. None of the subjects who were given non-native syllables (condition AJ and JE) had lived in a culture of the language to be tested. However, all the Japanese subjects had a knowledge of English through the classes they had taken for at least six years in junior and senior high schools. None of the American subjects had studied Japanese.

2.3. *Procedure*

Subjects were presented with the above audio-visual stimuli every 7 s and asked to write down "what they heard, not what they saw". All the groups were instructed to write in Roman alphabet. In each trial, the subjects were also to report whether or not they perceived a discrepancy between audio and visual stimuli by checking a column on the response sheet. This was done to make the subjects pay attention to both modalities. The experimenter told the subjects that they might hear a syllable with two consonants such as "bga" and encouraged them to report whatever they heard. This information was given because earlier literature on the McGurk effect has reported this type of "combined response" for an auditory non-labial dubbed onto a visual labial.

Each subject participated in six blocks of 100 trials. Visual stimuli were presented on a color monitor in which the speaker's face appeared in approximate life size. The viewing distance was 1 m. Audio signals were presented through two loud speakers attached to sides of the monitor.

After this audio-visual task, the subjects did an audio-alone task in which they reported what they heard when presented with only audio signals. The 10 audio-alone stimuli with the video blacked out were presented six times in a block of 60 trials in random order. Although these audio-visual and audio-alone tasks were carried out in both a noise-added condition (white noise was added to the audio signals) and a noiseless condition (no noise was added), only the latter is reported here.

Although there were a few equipment differences between experiments in the United States and Japan, the equipment was carefully adjusted to make possibly affecting factors consistent.

2.4. *Analysis*

To test whether or not the incongruent visual cues (non-labial visual cues to labial auditory stimuli, and vice versa) significantly biased perception of an auditory syllable, the proportion of labial *vs.* non-labial responses in the audio-visual task and that in the audio-alone task were compared for each auditory syllable. For example, in the audio-alone task in condition JE, a labial sound [pa] was identified as "pa (labial)" 93% of the time and as "ta (non-labial)" 7% of the time. When this auditory [pa] was synchronized with the non-labial lip movements of [ga] in the audio-visual task, the frequency of labial ("pa") responses was 23% and that of non-labial ("ta", "ka" and "ha") responses was 77%. These proportions ("93% *vs.* 7%" and "23% *vs.* 77%") were compared by the chi-square test using the original frequencies in 60 observations ($df = 1$, $n = 120$). Each 'combined response' (such as "mna" and "bla") was counted as half a labial and half a non-labial response.

To test inter-language differences in visual bias, the size of the visual bias was calculated. This was done by subtracting the frequency of "place errors" (e.g., non-labial responses for auditory labials) in the audio-alone task from that in the audio-visual task, assuming that the place errors in the audio-visual task included both auditory errors and visually biased responses. In the above example, the size of visual bias was $77\% - 7\% = 70\%$. On the other hand, in condition AE, auditory [pa] in the audio-alone task produced labial ("pa") responses 88% of the time and non-labial ("ta", "tha") responses 12% of the time. When this auditory [pa] was synchronized with visual [ga] in the audio-visual task, the frequency of labial ("pa") responses was 7% and that of non-labial ("ta", "ka", "tha", "kla") responses was 93%. Thus, the size of the visual bias was $93\% - 12\% = 81\%$. The complement of the visual bias value was regarded as the frequency of auditory responses. That is, the estimated frequency of auditory responses was $100\% - 81\% = 19\%$. The inter-language comparison was carried out based on these values of visual bias: the proportion of visually biased responses *vs.* auditory responses in conditions JE and AE was "70% *vs.* 30%" and "81% *vs.* 19%". These frequencies were compared by the chi-square test.

3. Results

3.1. *Audio-alone task*

In the audio-alone task, most of the auditory stimuli were heard as what the speaker intended to pronounce. However, auditory [wa] and [ra] in the non-native language (conditions AJ and JE) produced various responses. Therefore, auditory [wa] and [ra] were excluded from the quantitative comparison of interlanguage differences of visual influence. The average intelligibility score for eight auditory syllables excluding [wa] and [ra] was 99.2% for condition JJ, 95.8% for condition AJ, 96.7% for condition AE and 93.5% for condition JE. This variability indicates that the identical stimuli were more correctly identified when the native language of the subject and the speaker was the same. The chi-square test showed that there was a significant difference between conditions JJ *vs.* AJ, ($p < 0.01$, $\mathrm{df} = 1$, $n = 960$), and between conditions AE *vs.* JE, ($p < 0.05$, $\mathrm{df} = 1$, $n = 960$). The intelligibility score of each auditory syllable will be shown with the results for the audio-visual task.

3.2. *Visual bias of* [g*a*] *and* [*ba*]

Based on the response pattern, the results for the audio-visual task in the noiseless condition were categorized into four cases: visual stimuli were (1) labials [ba, pa, ma], (2) non-labials [da, ga, ka, na, ta], (3) [wa], and (4) [ra]. Whereas visual [ba, pa, ma] and [da, ta, na, ga, ka] in two languages produced similar response patterns for both language groups, visual [wa] and [ra] revealed dissimilarities in these two languages. Therefore, a quantitative inter-language comparison of visual bias was possible only for visual [ba, pa, ma] and visual [da, ta, na ga, ka]. Representative results for these two cases are shown in figures below for visual [ba] and [ga]: the response patterns for the other labial *vs.* non-labial visual stimuli were almost identical to those shown in these two cases. The individual data showed results consistent with the group data.

Figures 1–4 give confusion matrices labelled for auditory stimuli and responses. Although these figures include auditory [wa] and [ra], the results for these two

syllables will not be discussed in this section, but in the next section with the results for visual [wa] and [ra]. The number in each cell indicates the percentage of responses in 60 observations (10 subjects × 6 repetitions) for each auditory syllable. The numbers in parentheses in the leftmost column indicate the percentages of correct responses in the audio-alone task (the auditory intelligibility score). If there is no effect of visual cues, the value in each of the diagonal cells should equal the auditory intelligibility score in the leftmost cell. Alternatively, if the McGurk effect occurs, then stimuli with auditory labial initials synchronized with visual [ga] should be perceived as non-labials (fused response), and stimuli with auditory non-labial initials synchronized with visual [ba] should be perceived as labial (fused response) or as having a consonant cluster of labial plus non-labial (combined response). Responses in the shadowed sections in each figure indicate visually biased responses, that is, fused responses and combined responses.

(a)

audition \ response	b	p	m	w	d	t	n	g	k	r'	
b (100)	100										
p (98)		67				33					***
m (100)			96				4				
w (95)				80						20	*
d (100)					100						
t (95)						100					
n (100)							100				
g (100)								100			
k (100)									100		
r' (100)										100	

(b)

audition \ response	b	p	m	w	d	t	n	g	k	r'
b (100)	100									
p (98)		100								
m (100)			100							
w (95)				91						9
d (100)					100					
t (95)		17				83				
n (100)							100			
g (100)								100		
k (100)									100	
r' (100)										100

Figure 1. Stimulus–response confusions (%) in the audio-visual task in condition JJ (Japanese subjects, Japanese syllables) for (a) visual [ga] and (b) [ba]. The leftmost column shows correct responses (%) in the audio-alone task. The shaded sections indicate visually biased responses. In each row, asterisk(s) to the right show that the effect of the visual cue on the auditory syllable was significant (***$p < 0.001$ and *$p < 0.05$), when place of articulation errors in the audio-visual task were compared with the place errors in the audio-alone task. (See Section 2.4 in the text.)

(a)

audition \ response	b	p	m	w	d	t	n	g	k	l	r	others	
b (100)	20				80								***
p (98)		17				83							***
m (100)			32				66					mn2	***
w (38)				0				2		84	2	gr5 dl3 gl2 h2	***
d (90)					100								
t (83)						100							**
n (96)							100						
g (98)								95				bg3 g 2	
k (100)									100				
r' (51)										51	13	gr18 dl10 rr8	

(b)

audition \ response	b	p	m	w	d	t	n	g	k	l	r	others	
b (100)	100												
p (98)		98				2							
m (100)			100										
w (38)	2		13	8						8	5	bl43 ml8 br5 mr2 a3 gr2	***
d (90)	40				35							bd25	***
t (83)		78				17						pt5	***
n (96)			52				25					mn23	***
g (98)	12							76				bg12	***
k (100)		18							74			bk5 pk3	**
r' (51)	2		2	2						10	8	bl41 br10 ml10 bd7 rr7 a2	***

Figure 2. Stimulus–response confusions (%) in the audio-visual task in condition AJ (American subjects, Japanese syllables) for (a) visual [ga] and (b) [ba]. The shaded sections indicate visually biased responses. In each row, asterisk(s) to the right show that the effect of the visual cue on the auditory syllable was significant ($^{***}p < 0.001$ and $^{**}p < 0.01$).

As seen in Fig. 1, Japanese subjects showed only a weak McGurk effect for Japanese syllables, yielding acoustically correct responses 100% of the time for most of the stimuli. The visual influence is noticeable here only with auditory [pa] ("ta" response, 33%), [ma] ("na" responses, 4%), [wa] ("r'a" response, 20%), and [ta] ("pa" response, 17%). In these four, only the visual bias for auditory [pa] and [wa] reached statistical significance based on the chi-square test [$p < 0.001$, and $p < 0.05$, as indicated by rightmost asterisks in Fig. 1(a)]. Although the visual bias to auditory [ta] did not reach significance in this case, it was significant when the auditory [ta] was presented with visual [pa] or [ma].

In contrast to the Japanese subjects, American subjects were greatly influenced by visual information in perceiving these identical stimuli. As the asterisks in Fig. 2 show, the American subjects showed a significant visual bias for every auditory syllable. For auditory [ba, pa, ma] synchronized with visual [ga], the American subjects showed fused (non-labial) responses much more often ["da" (80%), "ta" (83%), "na" (66%)] than did the Japanese ["da" (0%), "ta" (33%), "na" (4%)]. Because the auditory intelligibility for [ba], [pa], and [ma] was identical for both groups, these differences directly show the differences in the visual bias between language groups. When compared using the chi-square test based on the frequencies of labial *vs.* non-labial responses, the group difference was statistically significant for each of the auditory [ba, pa, ma] with visual [ga] ($p < 0.001$, for each; $df = 1$, $n = 120$). The bigger visual bias to Americans' responses is also found with auditory [da, ta, na, ga, ka] synchronized with visual [ba]. The fused (labial) responses or combined (labial + non-labial) responses that American subjects showed for audi-

tory [da], [ta], [na], [ga], and [ka] were "ba, bda" (40%, 25%), "pa, pta" (78%, 5%), "ma, mna" (52%, 23%), "ba, bga" (12%, 12%) and "pa, bka, pka" (18%, 5%, 3%), respectively. In Japanese subjects, fused responses occurred 17% of the time only for auditory [ta], and the other auditory non-labials did not produce any fused responses or combined responses. In each of these auditory non-labials with visual [ba], the differences in the size of visual bias (calculated by subtracting the frequency of the place errors in the audio-alone task from that in the audio-visual task) between the two groups were again significant ($p < 0.01$ for [ga], and $p < 0.001$ for the others; $df = 1$, $n = 120$).

Although we have considered three possible factors that would account for the small visual bias in the Japanese subjects listening to Japanese speech, the group difference observed between condition JJ and condition AJ rules out two alternatives. The small visual bias in the Japanese subjects is not due to the properties of Japanese speech stimuli nor due to the specific experimental procedure because the American subjects showed a strong visual bias for the identical Japanese speech stimuli when tested with the identical experimental procedure. Thus, the relatively small visual bias in the Japanese subjects must be accounted for by a subject factor: Japanese listeners are less influenced by incongruent lip-read information than American listeners are.

Condition AE tells whether the influence on American listeners of visual cues occurred only for syllables in a non-native language or if it also occurs in their native

(a)

audition \ response	b	p	m	w	d	t	n	g	k	r	l	others	
b (87)	0				13	2					55	th20 s10	***
p (88)		7				66			20			th5 kl2	***
m(100)			22				32				46		***
w (98)				3							97		***
d (98)					98			2					
t (100)						100							
n (100)							100						
g (100)								100					
k (100)									100				
r (100)										37	63		

(b)

audition \ response	b	p	m	w	d	t	n	g	k	r	l	others	
b (87)	100												
p (88)		100											*
m(100)			100										
w (98)			13	38							37	bl12	***
d (98)					100								
t (100)		53				47							***
n(100)			13				74					mn13	***
g(100)								100					
k (100)								2	98				
r (100)										90	2	br8	

Figure 3. Stimulus–response confusions (%) in the audio-visual task in condition AE (American subjects, English syllables) for (a) visual [ga] and (b) [ba]. The shaded sections indicate visually biased responses. In each row, asterisk(s) to the right show that the effect of the visual cue on the auditory syllable was significant (***$p < 0.001$, and *$p < 0.05$).

(a)

audition \ response	b	p	m	w	d	t	n	g	k	r	r'	l	others	
b (58)	12	12			17	22		2		5	7	17	hw5 s2 gw2	***
p (93)		23				72			3				h2	***
m(100)			28				30			5	22	15		***
w (83)				20			2			35	28	13	gw2	***
d (100)					100									
t (97)						97			3					
n(100)							100							
g(100)								100						
k (100)									100					
r (67)				7				10		50	27	3	gw3	

(b)

audition \ response	b	p	m	w	d	t	n	g	k	r	r'	l	others	
b (58)	65	35												
p (93)		100												
m(100)			100											
w (83)	10		5	55						17	5	8		
d (100)	2				98									
t (97)		58				42								***
n(100)							100							
g(100)								100						
k (100)									100					
r (67)				13				10		58	18			

Figure 4. Stimulus–response confusions (%) in the audio-visual task in condition JE (Japanese subjects, English syllables) for (a) visual [ga] and (b) [ba]. The shaded sections indicate visually biased responses. In each row, asterisk(s) to the right show that the effect of the visual cue on the auditory syllable was significant (***$p < 0.001$).

language. As seen in Fig. 3, American subjects showed as strong a McGurk effect for English stimuli as for Japanese stimuli, especially for auditory labials. The result for condition AE replicates existing data reported for native speakers in English-speaking cultures, though the biased responses are somewhat scattered for a few categories (e.g., auditory [ba] was perceived as "da", "la", "tha" and "sa") and there are very few combined responses for auditory non-labials. As compared with the results for condition AJ, the chi-square test showed that, for auditory non-labials, Americans had a significantly larger effect of vision for Japanese syllables than for English syllables ($p < 0.001$ for auditory [da], [na], [ra], [ka] and $p < 0.01$ for [ta] and [ga]; df = 1, $n = 120$ for each).

Condition JE showed that Japanese subjects, who were little influenced by vision for Japanese syllables, were more influenced for English syllables. The results shown in Fig. 4 are very similar to those of American subjects for English syllables in Fig. 3. However, Americans tended to show a stronger visual effect than Japanese in a few English syllables: in auditory [ba] ($p < 0.001$), [pa] ($p < 0.05$), [wa] ($p < 0.01$) and [na] ($p < 0.001$). In condition JE, the auditory intelligibility score of [ba] was only 58%. The other responses for English [ba] by the Japanese subjects in the audio-alone task was "pa" (27%), "ta" (7%), "la" (3%) and so on. Although the poor intelligibility of auditory [ba] seemed to discourage the inter-language comparison of the visual effect, we included auditory [ba] in the comparison because most of the auditory errors were within-place errors ("pa") which probably reflected VOT

differences between the two languages. (At least in our video stimuli, English [b] seemed to have a longer VOT than the corresponding Japanese sound.) In condition AE, the American subjects identified [ba] by audition alone as "ba" (87%), "pa" (5%), "dta" (5%) and "da" (3%).

To summarize, total fequencies of the visual bias in each condition for auditory labials ([ba, pa, ma]) synchronized with visual [ga] and for auditory non-labials ([da, ta, na, ga, ka]) synchronized with visual [ba] were as follows: 11.8% and 2.4% (JJ), 76.4% and 42.5% (AJ), 83.9% and 14.7% (AE), and 81.4% and 12.0% (JE). In each condition, it is obvious that auditory labials were more influenced by incongruent visual cues than were auditory non-labials. For auditory labials, the chi-square test found that the visual bias for condition JJ was significantly smaller than that for the other three conditions ($p < 0.001$, $\mathrm{df} = 1$, $n = 360$, for each pair of conditions JJ *vs.* AJ, JJ *vs.* AE and JJ *vs.* JE). No significant differences were found in the other pairs of conditions. For auditory non-labials, there was a significant difference in each pair of conditions ($p < 0.001$, $\mathrm{df} = 1$, $n = 600$, for each pair of conditions) except in the pair of conditions AE *vs.* JE. When the values of visual bias for auditory labials and that for auditory non-labials were averaged together, there was a significant difference in each pair of conditions ($p < 0.001$, $\mathrm{df} = 1$, $n = 960$, for each pair of conditions) except in the pair of conditions AE *vs.* JE. Thus, the rank of visual bias values for eight auditory syllables was, from the largest to the least, as follows: (1) AJ, (2) AE and JE, (3) JJ. This ranking implies that the American subjects were more often visually influenced than were Japanese subjects, and that non-native stimuli produced a larger visual bias than did native stimuli.

A qualitative difference between condition JJ and the other three conditions is seen in relation to the auditory intelligibility score. As we suggested in our auditory intelligibility hypothesis based on the results in condition JJ, the visual bias for responses of Japanese subjects listening to Japanese stimuli is absent or very weak when the auditory intelligibility score was 100%. The results of the three new conditions in this experiment show that the auditory intelligibility hypothesis cannot be extended to these conditions. In condition JE, in perceiving English auditory [ma] with an auditory intelligibility score of 100%, even Japanese subjects were visually biased (fused responses occurred 72% of the time). In both conditions AJ and AE, the American subjects' responses showed a large visual bias (53–78%) even for stimuli with auditory intelligibility scores of 100% ([ba] and [ma] in condition AJ, [ma] and [ta] in condition AE).

3.3. *Interlanguage differences in* [r] *and* [w]

The results for [ra] and [wa] suggested that there are acoustic and articulatory differences in [r] and [w] between Japanese and English.

It is assumed that Japanese [r], which is an alveopalatal flap, is relatively similar to English [l] or to American English flapped [d]. The results for the audio-alone task showed that Japanese [ra] presented to American subjects was identified as "la" most frequently: It was identified as "la" (51%), "gra" (17%), "dla" (10%), "ra" (10%) and "rra (Spanish "ra")" (10%), and "wla" (2%). When this identical Japanese [ra] was presented to Japanese subjects, the auditory intelligibility score was 100%. When presented to American subjects, English [ra] similarly showed an auditory intelligibility score of 100%. When presented to Japanese subjects, it was

identified as English "ra" (67%), Japanese "ra" (13%), "wa" (13%), and "ga" (7%). Due to their knowledge of English, some of the Japanese subjects voluntarily reported both English "ra" and Japanese "ra", by distinguishing them with "ra & la", "wra & ra", and so on. Because of this spontaneous differentiation of English *vs.* Japanese "r", the experimenter asked the subjects what kind of sound their responses for English [r] implied after the experiment.

An interlanguage acoustic difference was also suggested for [wa]. In the audio-alone task, Japanese [wa] presented to American subjects was identified as "wa" (38%), "la" (30%), "ra" (18%), "gra" (7%), "bla" (3%), "dla" (2%) and "gwa" (2%). When this Japanese [wa] was presented to Japanese subjects, it was identified as "wa" (95%) and "ra" (in Japanese, 5%). English [wa] presented to Japanese subjects was identified as "wa" (83%), "ra" (7%), "la" (5%) and "ba" (5%). When this English [wa] was presented to American subjects, it was identified as "wa" (98%) and "la" (2%).

Since responses by Japanese subjects in condition JE included both Japanese "r" and English "r", Japanese [r] will be symbolized with [r′] in the following.

Figures 5–8 show the results for visual [r/r′] and [w] in the audio-visual task. In the figures, auditory [w] is grouped with labials and auditory [r/r′] is grouped with non-labials for convenience of comparison. However, the response patterns in the confusion matrices suggested that this simple grouping, [w] = labial, [r/r′] = non-labial, was not adequate in terms of visual properties. A clear result was that Japanese visual [r′] can be grouped with visual non-labials for both language groups. In Figs 5(a) (JJ) and 6(a) (AJ), visual [r′] shows the visual biasing effect only on auditory labials. The response pattern of Fig. 5(a) is almost identical to that of Fig. 1(a) (for visual [g]) and the pattern of Fig. 6(a) is almost identical to that of Fig. 2(a) (for visual [g]). Therefore, according to its visual property, Japanese [r′] is categorized with the non-labials.

Similarly, Japanese visual [w] presented to Japanese subjects is categorized with the labials. The response pattern of Fig. 5(b) is almost identical to that of Fig. 1(b) (for visual [b]), showing the visual biasing effect on auditory non-labials. Although the response pattern by American subjects for Japanese visual [w] [Fig. 6(b)] was also similar to that for Japanese visual [b] [Fig. 2(b)], a big difference between visual [w] and [b] was found in the influence on auditory [w]. When auditory [wa] was synchronized with visual [wa], 98% of the responses were "wa" (cf. the auditory intelligibility score of [wa] was only 38%). On the other hand, when this auditory [wa] was synchronized with visual [ba], the most frequent response was "bla" (43%) and there were few "wa" responses (8%). This difference between visual [b] and [w] suggests a visual sensitivity of the American subjects to the distinction between a labial closure and a labial protrusion. Although the Japanese subjects did not show such a difference between visual [w] and [b], it is not clear whether they did notice the visual distinction between labial closure and labial protrusion or just did not integrate the visual information into their responses.

For English stimuli, the response patterns for visual [r] and those for visual [w] were very similar, and they were not similar to those for visual [b] or visual [g]. This similarity between visual [r] and [w] was observed in both language groups. In Fig. 7 (AE), the confusion matrix for visual [r] and that for visual [w] are almost identical. Their visual biasing effect is conspicuous on auditory [b]: when auditory [ba] was synchronized with visual [ra], 38% of the responses was "wa" and when auditory

[ba] was synchronized with visual [wa], 53% of the responses was "wa". Although [w] is labial in terms of articulation, it shows few visual biasing effects on auditory non-labials. Conversely, visual labial [w] produced a biasing effect on auditory labials ([b], [p], [m]), showing non-labial responses to some extent. A dissimilarity between visual [w] and [b] is also found in responses for auditory [w]: When synchronized with visual [b] [Fig. 3(b)] auditory [w] produced non-"w" responses such as "l" (37%), "m" (13%) and "bl" (12%), while it was identified as "w" 100% of the time when synchronized with visual [w]. These results indicate that English visual [w] and [b] are so different from each other that their visual effects on auditory non-labials are not the same, so that visual [w] can be an incongruent visual cue for the other auditory labials and visual [b] can be an incongruent visual cue for auditory [w]. These visual properties observed for visual [w] were also found for visual [r] though the primary place of articulation for [r] is not the lips. According to the author's observation of the video stimuli, visual [r] pronounced by the American speaker had a preparatory protrusion of the lips prior to the sound. Thus, the

(a)

audition \ response	b	p	m	w	d	t	n	g	k	r'
b (100)	100									
p (98)		72				28				
m (100)			96				4			
w (95)				77						23
d (100)					100					
t (95)						100				
n (100)							100			
g (100)								97	2	2
k (100)									100	
r' (100)										100

(b)

audition \ response	b	p	m	w	d	t	n	g	k	r'
b (100)	98	2								
p (98)		97				2			2	
m (100)			100							
w (95)				97						3
d (100)	1			1	97			2		
t (95)		14				86				
n (100)							100			
g (100)								100		
k (100)		2		2					97	
r' (100)										100

Figure 5. Stimulus–response confusions (%) in the audio-visual task in condition JJ (Japanese subjects, Japanese syllables) for (a) visual [r'a] and (b) [wa]. The shaded sections indicate visually biased responses.

(a)

audition \ response	b	p	m	w	d	t	n	g	k	l	r	others
b (100)	20				75							th3
p (98)		23				77						
m (100)			32				68					
w (38)				0						87	2	dl5 gr3 gl2 tl2
d (90)	2				98							
t (83)		2				98						
n (96)							100					
g (98)					10			88				dg2
k (100)						2			92			h7
r' (51)					7			2		45	18	gr10 dl10 rr8

(b)

audition \ response	b	p	m	w	d	t	n	g	k	l	r	others
b (100)	95											wb3 bl2
p (98)		98				2						
m (100)			97				3					
w (38)				98							2	
d (90)	23				65			2				bd5 wd5
t (83)		65				32						pt3
n (96)			27				65					mn5 wn3
g (98)	12				2			75				bg5 wg3 bl2 th2
k (100)		18							80			pk2
r' (51)				23						27	17	bl10 wl8 wr3 bd2 gr5 dl3 rr2

Figure 6. Stimulus–response confusions (%) in the audio-visual task in condition AJ (American subjects, Japanese syllables) for (a) visual [r'a] and (b) [wa]. The shaded sections indicate visually biased responses.

(a)

audition \ response	b	p	m	w	d	t	n	g	k	r	l	others
b (87)	32			38	10	2				12		th7
p (88)		83				15			2			
m (100)			72	5			5				18	
w (98)				98							2	
d (98)					100							
t (100)		3				97						
n (100)				2			97					mn2
g (100)								100				
k (100)									100			
r (100)										100		

(b)

audition \ response	b	p	m	w	d	t	n	g	k	r	l	others
b (87)	27	2		53	10					2		th5 s2
p (88)		80				18						th2
m (100)			82	2			5				12	
w (98)				100								
d (98)					98			2				
t (100)		8				92						
n (100)							100					
g (100)					2			97	2			
k (100)									100			
r (100)										100		

Figure 7. Stimulus–response confusions (%) in the audio-visual task in condition AE (American subjects, English syllables) for (a) visual [ra] and (b) [wa]. The shaded sections indicate visually biased responses.

similar visual biasing effects are perhaps attributed to the protruding movements in pronouncing [r] and [w].

In Fig. 8 (JE), the response pattern of Japanese subjects also showed the similarity between English visual [r] and [w]. The confusion matrices for visual [r] and [w] are again almost identical. They show few visual effects on auditory non-labials and some visual effects on auditory labials. As compared with the auditory intelligibility score of auditory [wa] (83%), the acoustically correct "wa" responses increased when it was synchronized with visual [wa] (to 95%) or [ra] (to 92%). Here, "hwa" responses can be seen for auditory [ba], instead of the "wa" responses observed for the American subjects. The increase of "wa" responses and the occurrence of "hwa" responses can be again attributed to the protruding movements for [w] and [r].

In summary, English visual [w] was dissimilar to English visual [b] in terms of the visual biasing effect on auditory non-labials, while Japanese visual [w] could be categorized almost into the same group as Japanese visual [b]. Thus, this difference suggests a production difference, perhaps a stronger protrusion for English [w]. While Japanese visual [r′] was categorized with the non-labials in accordance with the fact that auditory [r′] was identified as "l" by the American subjects, English visual [r] was categorized into the same group as English [w] because of its preparatory protrusion. These visual cuing properties were found for both language groups. The only exception was in the responses by the American subjects for Japanese visual [b] synchronized with auditory [w], where the visual similarity between Japanese [w] and [b] broke down.

(a)

audition \ response	b	p	m	w	d	t	n	g	k	r	r′	l	others
b (58)	35	23		7	8	15							hw7 s5
p (93)		83				3			3				hw12 h2
m(100)			90							2	8		
w (83)				92						8			
d(100)					100								
t (97)		7				92			2				
n(100)							100						
g(100)								100					
k(100)									100				
r (67)				8				10		67	13		wr2

(b)

audition \ response	b	p	m	w	d	t	n	g	k	r	r′	l	others
b (58)	38	33		2	2	10							hw15
p (93)		90				7							hw3
m(100)			88				3			2	7		
w (83)				95						3		2	
d(100)					100								
t (97)		10				85			5				
n(100)							100						
g(100)								100					
k(100)									100				
r (67)				12				8		68	10		wr2

Figure 8. Stimulus–response confusions (%) in the audio-visual task in condition JE (Japanese subjects, English syllables) for (a) visual [ra] and (b) [wa]. The shaded sections indicate visually biased responses.

On auditory [w] or [r/r′], the influence of visual cues was also found depending on the group characteristics which we saw in the earlier section. However, the visual influence on auditory [r/r′] presented to non-native subjects was somewhat different from that on the other auditory stimuli. In general, responses for an auditory stimulus were biased by incongruent visual cues and improved by congruent visual cues. When auditory [r/r′] was presented to non-native subjects, some exceptional results were found. When Japanese auditory [r′] was presented to the American subjects, a congruent visual cue did not improve their responses [Fig. 6(a) for visual [r′]], but an incongruent visual cue biased them [Fig. 2(b) for visual [b]]. The absence of the improving effect was perhaps because Japanese visual [r′] was similar to several English non-labials such as [l], [d], [gl] and [dl] to some extent.

When English auditory [r] was presented to the Japanese subjects, neither a biasing effect nor an improving effect of any visual cues was found. In condition JE (Figs 4 and 8), auditory [r] showed an almost constant response pattern for visual [g], [b], [r] and [w]: the most frequent response was "ra" (50% to 68%) and each of "wa", "ga" and "r'a" responses occurred approximately 10% of the time. The absence of a visual influence on auditory [r] even though it had an auditory intelligibility score of only 67% suggests that, though they had a knowledge of English, the Japanese subjects did not have an appropriate representation of English [r] which could correlate its acoustic characteristics with its articulatory characteristics.

4. Discussion

Based on data for auditory [ba, pa, ma] and [da, ta, na, ga, ka] synchronized with either visual [ga] or [ba], interlanguage differences in the influence of visual cues were shown quantitatively. A comparison of condition JJ with condition AJ showed that Japanese subjects were much less prone to visual biasing effects than were Americans for the same Japanese syllables. For English syllables (conditions AE and JE), however, the quantitative difference between Japanese and Americans was not significant. For both language groups, the McGurk effect occurred more for non-native syllables (conditions JJ *vs.* JE, and conditions AE *vs.* AJ). The size of the observed visual biaisng effect can be ranked from the most to the least as, largest in American subjects for Japanese syllables, then Americans for English and Japanese for English, and smallest in Japanese for Japanese.

The results indicate that Americans automatically integrate visual cues with auditory cues in perceiving syllables of native language. This vision-dependent processing was also effective for non-native syllables, where it was even strengthened to some extent. Japanese subjects, in contrast, incorporated visual cues much less than Americans when perceiving native syllables. This vision-independent processing, however, breaks down for non-native syllables, where they were subject to visual biasing effects. Thus, the answer to our initial question is that the relative rarity of the McGurk effect for Japanese subjects must be attributed to the perceptual processing of Japanese listeners, and not to the stimulus characteristics of Japanese syllables.

Another result that indicates a cultural or linguistic difference in perception rather than production is the frequency of combination responses, as for example, a "bda" or "bla" response for auditory [da] with visual [ba]. This type of response, which was found earlier by McGurk & MacDonald (1976) for English subjects, was

reported primarily by the American subjects in our experiment. Although both groups of subjects were informed of the possibility of perceiving consonant clusters, Japanese listeners gave few such combination responses, probably because the Japanese phonological system does not allow any phonological consonant clusters. This is another clear difference in perceptual processing, developed in a specific language environment.

Although the source of these language-specific perceptual differences in the dependence on vision is not clear, a cultural difference suggests an explanation. Whereas Americans commonly look at the face of the person they are listening to, Japanese often avoid this because it may be regarded as impolite in Japan to look at someone's face, even in many circumstances when listening to him/her.

Another possible reason is the simpler structure of the phonological system of Japanese. It contains only five vowels (/i/, /e/, /a/, /o/ and /u/) and does not include several consonants in English such as /v/, /θ/ and /r/. In addition it does not have phonological consonant clusters. This simpler structure may enable Japanese listeners to discriminate one Japanese syllable from all others without additional information provided by vision.

It is apparent, however, that the vision-independent processing of Japanese is not because they are poor lip-readers. First, it is evident from the results for condition JE that Japanese subjects did use visual cues for English syllables. Another piece of evidence comes from our previous study (Sekiyama, Joe & Umeda, 1988) where the lip-reading ability of Japanese subjects for 100 Japanese syllables was tested, using two female narrators and 60 normal hearing subjects without any training experience of lip-reading. The results showed that they discriminated labials from non-labials correctly 92% of the time. This score is comparable to the data for American subjects (e.g., 91%, calculated from the confusion matrix reported by Walden, Prosek, Montgomery, Scherr & Jones, 1977, for 31 hearing impaired adults). Thus, the lipreading ability of Japanese subjects is enough to pick up the visual information which gives rise to the McGurk effect.

Moreover, it is unlikely that the Japanese subjects ignored visual information for Japanese syllables, because in complying with the instruction to report any audio-visual discrepancy, they gave reasonable responses for many cases: When their responses were auditory (i.e., not biased by visual cues) for incongruent audio-visual stimuli, the majority (60% for condition JJ, 64% for condition AJ, 63% for condition AE and 79% for condition JE) of the responses were accompanied by a report of perceived discrepancy. According to a chi-square test, condition JE showed a significantly higher frequency of reported discrepancies than did the other three conditions ($p < 0.001$, $\mathrm{df} = 1$, $n = 2684$, for conditions JJ *vs.* JE; $p < 0.001$, $\mathrm{df} = 1$, $n = 1540$, for conditions AJ *vs.* JE; $p < 0.001$, $\mathrm{df} = 1$, $n = 1874$, for conditions AE *vs.* JE). The frequencies of reported discrepancies in the other three conditions did not differ significantly. Thus, the Japanese subjects presented with Japanese stimuli noticed the auditory–visual incongruence as accurately as the American subjects did when their responses were not biased by visual cues. In spite of the fact that the Japanese subjects in condition JJ noticed the incongruent visual information in many cases, their responses were not biased by the visual cue. It suggests that Japanese listeners tend to separate visual information from auditory information as long as audition provides enough information. In contrast, the large

visual bias found in condition AE suggests that American listeners tend to integrate the two sources of information.

For both Japanese and American subjects, the McGurk effect occurred more for the subjects' non-native syllables than for their native ones. It may be attributed to the acoustic deviation of these stimuli from what the subjects are accustomed to hearing for corresponding phonetic categories in their native languages. Perhaps the acoustic deviations obliged Japanese listeners to give up the vision-independent processing that they applied to their native syllables.

The present study also showed some acoustic and articulatory differences in [w] and [r] between Japanese and American English. Japanese [r′a] presented to the American subjects by audition alone was identified as "la" (51%) and other non-labial sounds. Japanese auditory [wa] presented to the American subjects was heard as "wa" (38%), "la" (30%), and so on. English auditory [ra] presented to the Japanese subjects was heard as "ra" (67%), "r′a" (13%), and so on. English auditory [wa] presented to the Japanese subjects was heard as "wa" (83%), "ra" (7%), and so on. In addition to the well known acoustic difference between Japanese [r′] and English [r], the results showed a difference between Japanese [w] and English [w]. Japanese [w] presented to American subjects was identified as "w" only 38% of the time while English [w] presented to Japanese subjects showed an auditory intelligibility score of 83%. This suggests that the acoustic property of Japanese [w] is less distinct than that of English [w].

In accordance with this idea, an articulatory difference between Japanese [w] and English [w] was found. When visual stimuli [b, g, w, r/r′] were categorized based on their biasing effect on auditory stimuli, English stimuli were categorized into three groups [b], [w, r], and [g] while Japanese visual stimuli were categorized into [b, w] and [g, r′]. This categorization was suggested for both language groups. The [w, r] category in English was characterized by its protruding movement of lips. Although Japanese [w] is also somewhat protruded, the subjects did not seem to pay particular attention to the protrusion: the visual biasing effect of [w] on auditory non-labials was similar to that of [b]. This suggests that the protrusion of Japanese [w] is weaker than the protrusion of English [w]. The implication of this finding is that teachers should teach this difference in English classes in Japan so that the students can pronounce an intelligible English [w].

When English auditory [r] with the auditory intelligibility score of only 67% was presented to Japanese subjects, no influence of any visual cues was found. This suggests that, though they had learned English for at least six years, the Japanese subjects did not have an appropriate representation of English [r] where its auditory information was correlated with its visual information. This shows a constraint from the native language in that the Japanese phoneme repertoire does not have a consonant that resembles English [r]. In our video stimuli, English visual [r] had a preparatory protrusion. It is necessary to examine whether the protrusion for [r] is common in American English or was specific to the speaker we employed.

The auditory intelligibility hypothesis that Japanese listeners are hardly biased by visual cues when auditory information is complete was applicable only to the results in condition JJ. It did not fit responses of the Japanese subjects listening to English syllables, probably due to the acoustic deviation of the stimuli from what the subjects are accustomed to hearing for corresponding phonetic categories in native language. It did not fit responses of the American subjects in either condition AJ or

condition AE, indicating that American listeners have a perceptual processing through which visual information is integrated with auditory information even when the auditory information is complete. It is desirable to accumulate more data for Japanese subjects to establish the interlanguage difference in perceptual processing shown in the present study.

We would like to thank Prof. Harry Levitt and Dr Hwei-Bing Lin at the Graduate Center of the City University of New York whose laboratory was used to collect the data of American subjects. This part of the experiment in CUNY was supported by a grant from Nissan Science Foundation to the first author. We also thank Dr Quentin Summerfield, Prof. Mary Beckman, and two anonymous reviewers for helpful comments on a previous version of this manuscript.

References

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

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

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

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

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

Massaro, D. W. (1987) *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: 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 and Performance*, **9**, 753–771.

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

Sekiyama, K., Joe, K. & Umeda, M. (1988) Perceptual components of Japanese syllables in lipreading: a multidimensional study, *Technical Report of IECE* (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.

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

Summerfield, Q. (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In *Hearing by eye* (R. Campbell & B. Dodd, editors) 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.