McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility[a)]

Kaoru Sekiyama
Department of Psychology, Kanazawa University, Kakuma-machi, Kanazawa, 920-11, Japan

Yoh'ichi Tohkura
ATR Auditory and Visual Perception Research Laboratories, Inui-dani, Seika-cho, Kyoto, 619-02, Japan

(Received 10 December 1990; accepted for publication 13 June 1991)

The McGurk effect is a phenomenon that demonstrates a perceptual fusion between auditory and visual (lip-read) information in speech perception under the condition of audio-visual discrepancy, created by dubbed video tapes. This paper investigated whether or not the McGurk effect could be extended to Japanese subjects listening to Japanese syllables of different auditory intelligibility. The audio and video signal of a female talker's speech for ten Japanese syllables (/ba/, /pa/, /ma/, /wa/, /da/, /ta/, /na/, /ra/, /ga/, /ka/) was combined on videotapes, giving 100 audio-visual stimuli. These stimuli were presented to ten Japanese subjects who were required to identify the stimuli as heard speech in both noise-added and noise-free conditions. For both conditions, the intelligibility of the auditory stimuli was measured, by presenting the audio-alone stimuli. The results showed that, in the noise-free condition, the McGurk effect was small and almost limited to auditory stimuli of which the intelligibility was less than 100%. In the noise-added condition, the McGurk effect was very strong and widespread. These results indicate that the "Japanese McGurk effect" is less easily induced than the English one, and that it depends on the auditory intelligibility of the speech signal.

PACS numbers: 43.71.Es, 43.71.Hw, 43.71.Ma

INTRODUCTION

In face-to-face verbal communication, the talker's articulatory movements can be seen. This visual information is called lip-read information. Its role in speech perception has been investigated in several experimental paradigms. Lip-read information facilitates speech perception especially when the intelligibility of the speech is reduced, e.g., in a noisy environment, when noise is experimentally added to speech (Erber, 1975), or in speech perception by patients with cochlear implants (Fukuda *et al.*, 1988).

On the other hand, lip-read information interferes with speech perception under audio-visual discrepancy conditions which are created by dubbed video tapes (McGurk and MacDonald, 1976). For example, dubbing an audio signal /ba/ onto a video signal /ga/ makes an audio-visual stimulus that often induces the percept "da," suggesting a perceptual fusion between audio and video information. This fusion phenomenon, the "McGurk effect," has been interpreted as follows (MacDonald and McGurk, 1978): Of the three main phonetic components of speech, i.e., the place and the manner of articulation and voicing, information for place is provided visually; visual information on place (/ga/: nonlabial) replaces the auditory information for place (/ba/: labial) and it is integrated with the remaining auditory information (/ba/: plosive and voiced), resulting in a perceptual solution "da," which is a nonlabial, plosive, and voiced sound, and is acoustically characterized as between /ba/ and /ga/.[1]

Thus with audio-visual discrepancy in place of articulation, lip-read information misleads and biases auditory perception whereas it helps auditory perception in the natural audio-visual congruent situation. In the McGurk effect, an audio-visual incongruent stimulus is defined as a stimulus where auditory and visual information are incongruent in terms of the place of articulation, and more narrowly, whether the place is labial or nonlabial. Studies on lipreading have shown that the visual system can provide information efficiently about the place of articulation so that lipreaders can easily discriminate labials from nonlabials (e.g., Summerfield, 1987).

In the original experiments by McGurk and MacDonald (1976), perceptual fusion was reported with stimuli where an auditory labial was combined with a visual nonlabial (more exactly, velar), e.g., auditory /ba/ combined with visual /ga/ (perceived as "da") and auditory /pa/ combined with visual /ka/ (perceived as "ta"). In conversely combined pairs, i.e., an auditory nonlabial with a visual labial, "combination responses" were often observed rather than fusion: auditory /ga/ and visual /ba/ induced perception of "bga," combining both auditory and visual consonants.

[a)] This study was carried out while the first author was a visiting researcher at ATR Auditory and Visual Perception Research Laboratories.

After the original report of the McGurk effect, a number of studies were conducted using various kinds of auditory and visual stimuli, establishing this phenomenon (Dodd, 1977; Green and Kuhl, 1989; Green *et al.*, 1988; MacDonald and McGurk, 1978; Massaro and Cohen, 1983; Summerfield, 1979, 1987). The previous research, however, studied the McGurk effect only in English-speaking cultures, using English syllables and English-speaking subjects. The main purpose of the present study was to investigate whether the McGurk effect can be extended to non-English cultures, examining how it occurs in Japanese subjects for Japanese syllables.

Japanese consonants include four labials (/b/, /p/, /m/, /w/) and ten nonlabials (/t/, /d/, /n/, /s/, /z/, /r/, /y/, /k/, /g/, /h/). A confusion analysis study using multidimensional scaling by Sekiyama *et al.* (1988) showed that untrained Japanese subjects can easily discriminate labials from nonlabials in lipreading of Japanese syllables. Therefore, the McGurk effect was expected for Japanese syllables as well as for English syllables. In our preliminary experiment, the McGurk effect was found in Japanese subjects listening to 14 Japanese syllables (/ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /ra/, /wa/, /ha/, /ya/, /sa/, /za/). However, the stimuli used there were not highly intelligible: The subjects correctly heard them only 80% of the time when presented as audio-alone stimuli. This was probably due to the video recording process which was done without concern for a high S/N ratio. Thus an experiment with highly intelligible auditory stimuli was necessary to obtain visual effects independent of effects of noise.

Another purpose of this study was to examine the relation between auditory intelligibility and the McGurk effect. Our hypothesis was that the McGurk effect is promoted by poorly intelligible auditory stimuli. In fact, the literature has suggested this point although it has not been stated explicitly. Dodd (1977) reported a McGurk effect in audio-visually dissonant words presented in white noise: For example, hearing "tough" and seeing "hole" resulted in the percept "towel." On the other hand, Easton and Basala (1982) found little visual effect for meaningful words. In their experiment, one group of subjects was instructed to attend to auditory information and the other group was instructed to attend to visual information. While the audio-oriented subjects reported auditorily presented words correctly 99% of the time, the video-oriented subjects could not report visually presented words and only 10% of their responses suggested audio-visual fusion. One possible reason for the difference between the results of these two experiments is the presence or absence of noise. This interpretation seemed reasonable because the role of lip-read information, as a facilitator of speech perception, has been shown for speech of low intelligibility, as cited earlier.

This study investigated the "Japanese McGurk effect" with highly intelligible auditory stimuli that are presented in both noise-free and noise-added conditions, and examined the following questions: (1) What is the nature of the McGurk effect for Japanese subjects listening to Japanese syllables? (2) How does the McGurk effect depend on auditory intelligibility? To answer the first question, we used a larger number of Japanese syllables and tested Japanese subjects. For the second question, we used an experimental design in which the auditory intelligibility was measured in an audio-alone condition. To avoid the influence of hearing audio-alone stimuli on the McGurk effect (Massaro and Cohen, 1983), the subjects were given the audio-alone session after the audio-visual session. In this paper, the auditory intelligibility is defined as the percent of correctly heard responses when audio-alone stimuli are presented.

I. EXPERIMENT

A. Stimuli

Ten Japanese syllables were used: /ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /ra/, /wa/. For recording audio and video signals, a female Japanese talker pronounced each syllable once. While she pronounced the syllable, her face was videotaped onto a 1/2-in. tape through a video camera located in front of the talker.

Audio signals were recorded separately to avoid degrading factors in the recording process. The talker was instructed to pronounce clearly and the audio signals were recorded by a digital (DAT) tape recorder (16 bit with a sampling frequency of 20 kHz).

When audio-visual stimuli were made, these audio signals were dubbed onto an audio channel of a 1/2-in. video tape with preceding warning tones for each syllable. Ten audio and ten video stimuli were combined, resulting in 100 audio-visual stimuli. Care was taken to get precise synchronization between audio and video signals, adjusting the dubbing timing by a 33-ms frame unit. Each audio-visual stimulus was arranged in a 7-s unit which included a 3-s video black and a 4-s talking face (Fig. 1).

For presentation, several random order sequences were made by editing the original 100 stimuli. In an experimental session, one of these random order sequences of audio-visual stimuli was reproduced by a video cassette recorder (1/2-in. tape, SONY BVW-40). Visual stimuli were presented on a 20-in. color monitor in which the life-sized talker appeared. The viewing distance was 1 m. Audio signals were presented through two loud speakers attached to sides of the monitor.

B. Method

1. Subjects

Ten female subjects participated in the experiment. They were all native Japanese speakers with normal hearing

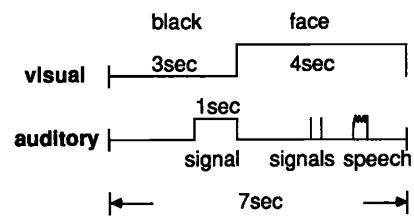

FIG. 1. Time course of audio-visual stimuli.

and normal or corrected vision. Subjects' age ranged from 22 to 27.

2. Procedure

Subjects were given an audio-visual stimulus every 7 s and were asked to look at and listen to each utterance. The task of the subjects was to write "what they heard, not what they saw." In addition, to make the subjects attend to the visual stimuli, instructions asked them to report a perceptual discrepancy between audio and visual stimuli that they noticed. Instructions also suggested that some people might hear syllables that were not included in the Japanese phonological system, like /bda/ or /bga/. The subjects were allowed to report more than two responses with a constraint that a more confident response should precede less confident ones when they were written. If more than two responses were given for a stimulus, their frequencies were weighted by coefficients such as 0.6:0.4 for two and 0.5:0.3:0.2 for three.

3. Experimental design

The subjects were required to do the above task in two conditions: with noise (noise-added audio-visual condition: nAV) and without noise (noise-free audio-visual condition: AV). In the nAV-condition, Gaussian noise with 50-kHz bandwidth was added to the audio signals. Signal to noise ratios, measured by a sound-pressure level meter at the location of the subjects' head, were kept at 0 dB. In the AV condition, no noise was added. To measure auditory intelligibility of the ten syllables for both the nAV and AV conditions, there were two audio-alone conditions where video signals included only black bursts and subjects were required to report what they heard: a noise-added audio-alone (nA) condition, where the auditory stimuli were the same as in the nAV condition, and a noise-free audio-alone (A) condition in which the auditory stimuli were the same as in the AV condition.

All the subjects participated in these four conditions in the order: (1) nAV, (2) nA, (3) AV, and (4) A. There were six repetitions of trials for each condition. Thus the numbers of trials for each subject were 600 in each of the nAV and AV conditions, and 60 in each of the nA and A conditions. The total 1320 trials were carried out in four 50-min sessions on four separate days.

C. Results

The results were analyzed by producing confusion matrices for each visual stimulus. Although the subjects were allowed more than two responses for a stimulus, the majority of the stimuli yielded a single response.

1. Results for the noise-free audio-alone condition

Table I shows the confusion matrix for the A condition. The number in each cell indicates the percentage of the response in 60 observations (10 subjects × 6 repetitions) for each auditory stimulus. Diagonal cells show the percentages of correctly heard responses ("auditory intelligibility score"). Most of the stimuli yielded the score of 100%, indicating the high auditory intelligibility in the noise-free condition. Three syllables, however, showed incomplete intelligibility: the intelligibility score for /pa/, /ta/, and /wa/ were 98%, 95%, and 95%, respectively. The /pa/ was perceived as "ta" 2% of the time, /ta/ was perceived as "pa" 5% of the time, and /wa/ was perceived as "ra" 5% of the time.

TABLE I. Auditory confusions in the *noise-free* audio-alone condition. Indicated in % in 60 observations.

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		98				2					
m			100								
w				95				5			
d					100						
t		5				95					
n							100				
r								100			
g									100		
k										100	

2. Results for the noise-free audio-visual condition

The ten confusion matrices in Table II (a) and (b) show the results for the AV condition where auditory stimuli of high intelligibility (as proved by Table I) were combined with ten visual stimuli. In general, these confusion matrices show high rates in the diagonal cells, indicating that subjects perceived most of the auditory stimuli correctly, and that visual biasing effects were fairly weak.

For visually presented labials [Table II(a)], i.e., visual /b, p, m, w/, it was only with auditory /ta/ that visual effects occurred. Presented with the visual labials, the auditory /ta/ was perceived as "pa" about 20% of the time. This response occurred 17% of the time for visual /ba/, 33% for visual /pa/, 22% for visual /ma/, and 14% for visual /wa/. Compared to the A condition (Table I) where erroneous "pa" responses for /ta/ occurred only 5% of the time, these "pa" responses seemed to be interpreted as involving not only auditory errors but also visual effects. This interpretation was confirmed for visual /pa/ and /ma/ ($p < 0.001$ and $p < 0.05$) by the chi-square test that compared the frequencies of labial ("pa") and nonlabial ("ta") responses in the AV condition with those in the A condition ($df = 1, N = 120$). For visual /ba/ and /wa/, however, the frequencies of the erroneous "pa" response were not significantly different from that of auditory errors in the A condition.

For the visual nonlabials [Table II(b)], i.e., /d, t, n, r, g, k/, it was only with auditory /pa/ and /wa/ that visual effects occurred. When auditory /pa/ were combined with any visual nonlabials, "ta" responses occurred to some extent. Such erroneous "ta" responses were 24%, 24%, 22%, 28%, 33%, and 22% for visual /da/, /ta/, /na/, /ra/, /ga/, and /ka/ stimuli. The chi-square test ($df = 1, N = 120$) revealed that each frequency of these erroneous "ta" responses was significantly greater than auditory errors ($p < 0.001$, $p < 0.001$, $p < 0.01$, $p < 0.001$, $p < 0.001$, and $p < 0.01$, for vi-

TABLE II. Results in the *noise-free* audio-visual condition. Indicated in %. (a) Confusion matrices for the stimuli with *visual labials* (/b,p,m,w/). (b) Confusion matrices for the stimuli with *visual nonlabials* (/d,t,n,r,g,k/).

(a) response

vision = b

audition	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		100									
m			100								
w				91				9			
d					100						
t		17				83					
n							100				
r								100			
g									100		
k										100	

vision = p

audition	b	p	m	w	d	t	n	r	g	k	others
b	98				2						
p	2	97				2					
m			100								
w				91				9			
d	1				99						
t	1	33				66					
n							100				
r								100			
g		1							99		
k		1							2	98	

vision = m

audition	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		100									
m			100								
w				90				10			
d					100						
t		22				78					
n			2				98				
r								100			
g									100		
k			2							98	h1

vision = w

audition	b	p	m	w	d	t	n	r	g	k	others
b	98	2									
p		97				2				2	
m			100								
w				97				3			
d	1			1	97				2		
t		14				86					
n							100				
r								100			
g									100		
k		2		2						97	

(b) response

vision = d

audition	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		76				24					
m			98				2				
w				81				19			
d					100						
t		2				98					
n							100				
r						2		98			
g						2			98		
k										100	

vision = t

audition	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		76				24					
m			93				7				
w				78				22			
d					100						
t						100					
n							100				
r								100			
g									100		
k										100	

vision = n

audition	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		78				22					
m			94				6				
w				82				18			
d					100						
t		2				98					
n							98	2			
r								100			
g								2	98		
									2	98	

vision = r

audition	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		72				28					
m			96				4				
w				77				23			
d					100						
t						100					
n							100				
r								100			
g								2	97	2	
k										100	

vision = g

audition	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		67				33					
m			96				4				
w				80				20			
d					100						
t						100					
n							100				
r								100			
g									100		
k										100	

vision = k

audition	b	p	m	w	d	t	n	r	g	k	others
b	98				2						
p		78				22					
m			88				12				
w				68				32	1		
d					100						
t						100					
n							100				
r								100			
g									100		
k										100	

TABLE III. The frequency of labial responses as an index of place of articulation confusions (noise-free condition), maximum = 60. The leftmost column shows the frequency in the audio-alone condition. The frequencies of labial responses for different visual presentations are given to the right. Each of these frequencies was compared with the frequency for the audio-alone stimulus. Asterisks indicate significant differences ($*p<0.05$, $**p<0.01$, $***p<0.001$) based on the chi-square test ($df=1, N=120$).

		video									
	s	none	b	p	m	w	d	t	n	r	g	k
audio	b	60	60	59	60	59	60	60	60	60	60	60
	p	59	60	59	60	58	45.4***	45.4***	47**	43.4***	40.2***	46.8**
	m	60	60	60	60	60	59	55.6	56.2	57.6	57.6	46.8***
	w	57	54.8	54.8	53.8	58	48.6*	47*	49.4	46.4**	47.8*	40.6***
	d	0	0	0.4	0	0.8	0	0	0	0	0	0
	t	3	10.2	20.4***	13.4*	8.2	1	0	1	0	0	0
	n	0	0	0	1	0	0	0	0	0	0	0
	r	0	0	0	0	0	0	0	0	0	0	0
	g	0	0	0.4	0	0	0	0	0	0	0	0
	k	0	0	0.4	1	1.8	0	0	0	0	0	0

sual /da/, /ta/, /na/, /ra/, /ga/, and /ka/). Similarly, auditory /wa/ combined with any visual nonlabial yielded erroneous "ra" responses; 19%, 22%, 18%, 23%, 20%, and 32% of the time for visual /da/, /ta/, /na/, /ra/, /ga/, and /ka/ stimuli. The chi-square test again revealed that all but one of the frequencies of the erroneous "ra" response involved significant visual effects ($p<0.05$, $p<0.05$, $p<0.01$, $p<0.05$, and $p<0.001$, for visual /da/, /ta/, /ra/, /ga/, and /ka/).

Table III summarizes the results of the chi-square test for each audio-visual stimulus with asterisks indicating significant visual effects. It shows that significant visual effects often occurred for auditory /pa/ and /wa/, and occasionally occurred for auditory /ta/ and /ma/. Subsequently, the ten visual stimuli in Table III were collapsed into visual labials and nonlabials and the resulting frequencies of labial versus nonlabial responses were compared with the frequencies for audio-alone stimuli. The chi-square test ($df=1$) revealed that only auditory /pa/, /wa/, and /ta/ yielded significant visual effects for over all visual nonlabials/labials ($N=420$, $p<0.001$; $N=420$, $p<0.01$; and $N=300$, $p<0.01$).

This result indicates that the McGurk effect in noise-free condition was almost limited to auditory stimuli of incomplete intelligibility: The auditory stimuli in which significant visual effects occurred, i.e., auditory /pa, wa, ta/, are the stimuli of which the intelligibility score was less than 100%, as shown in Table I. For the other auditory stimuli of complete intelligibility, there was little evidence of visual biasing effects, except for auditory /ma/ when combined with visual /ka/ (Table III).

The influence of auditory intelligibility on the McGurk effect was tested for audio-visual incongruent stimuli by the chi-square test, using 2×2 contingency tables (the frequencies of labial versus nonlabial responses $\times$ auditory stimuli of incomplete versus complete intelligibility).[2] When auditory nonlabials were combined with the visual labials, a significant difference in the McGurk effect was found between the incompletely intelligible stimulus (auditory /ta/) and the completely intelligible stimuli (auditory /da/, /na/, /ra/, and /ga/) ($df=1$, $N=1428$, $p<0.001$). Similarly, for the auditory labials combined with the visual nonlabials, a significant difference in the McGurk effect was found between the incompletely intelligible stimuli (auditory /pa/ and /wa/) and the completely intelligible stimuli (auditory /ba/ and /ma/) ($df=1$, $N=1416$, $p<0.001$). These results indicate that the McGurk effect in noise-free condition occurred depending on whether the auditory intelligibility score was 100% or less than 100%.

A striking difference between our results and those of McGurk and MacDonald is that our subjects heard auditory /ba/ combined with visual /ga/ 100% correctly as "ba," indicating no McGurk effect here [Table II(b), vision = *g*] whereas the original McGurk effect was highly significant for this type of incongruent pair so that the perceptual fusion ("da") was found 98% of the time (McGurk and MacDonald, 1976). This inconsistency and the above analyses suggest that the Japanese McGurk effect does not occur for completely intelligible auditory syllables.

3. Results for the noise-added audio-alone condition

Table IV shows the confusion matrix for nA condition. With the addition of the Gaussian noise, the intelligibility of auditory stimuli decreased when compared with that for the A condition. Of the ten syllables, five (/pa/, /ta/, /ra/, /ga/, /ka/) show conspicuous degradation due to noise: Intelligibility is 50%–60% for these syllables (notably it falls to 24% for /ga/) in the nA condition while, in the A condition, it is nearly 100% for each of these cases. The other five syllables (/ba/, /ma/, /wa/, /da/, /na/) have an intelligibility score of more than 90%.

For the types of confusion, Table IV shows that auditory confusions in the nA condition occurred between consonants of the same manner of articulation and the same voicing: e.g., among voiceless stops, /pa/ was perceived as "ta" 37% of the time, and /ta/ was perceived as "pa" 27% of the time. Note that in this audio-alone condition, place of articulation confusions occurred in both directions: from labial to nonlabial and from nonlabial to labial.

TABLE IV. Auditory confusions in the *noise-added* audio-alone condition. Indicated in % in 60 observations.

		response										
		b	p	m	w	d	t	n	r	g	k	others
audition	b	91				1				8		
	p		56	1			37				6	
	m			99				1				
	w				94			1	4	1		
	d	3				94				3		
	t		27		1		64				8	
	n			2				98				
	r			1	33			4	51	2		y8
	g	35				41				24		
	k		14				11				61	h7 a7

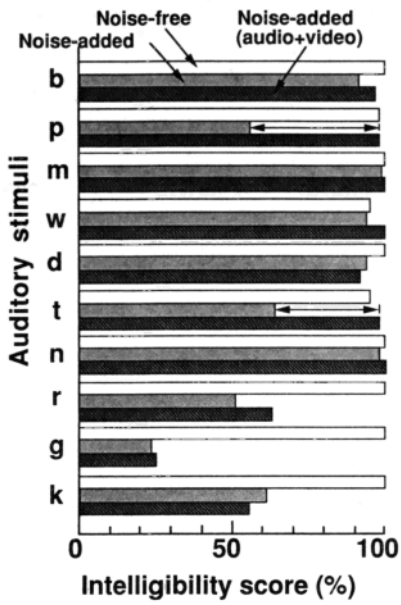

FIG. 2. Speech intelligibility increase by additive visual (lip-read) information in noise-added condition.

4. Results for the noise-added audio-visual condition

First, to examine the positive effects of lip-read information on speech perception in a noisy situation, we can look at results for the audio-visual congruent stimuli. Previous research predicts that lip-read information helps speech perception, especially in noisy situations. Figure 2 illustrates how lip-read information was helpful in our noisy situation, comparing the intelligibility in A, nA, and nAV conditions. For the nAV condition, only stimuli where sound and lip-read information were congruent (e.g., auditory /ba/ with visual /ba/) are depicted. Figure 2 shows that lip-read information was helpful to a large extent especially for /pa/ and /ta/. Taking /pa/, for example, the intelligibility was 98%, 56%, and 98% for A, nA, and nAV conditions, respectively. This indicates that lip-read information increased the correct identification of syllables in a noisy situation. Noise decreased the intelligibility of /pa/ from 98% to 56% in audio-alone conditions. Lip-read information, however, restored it back to 98%. Similarly, the intelligibility of /ta/ was 95%, 64%, and 98% for A, nA, and nAV conditions, respectively.

Second, the confusion matrices for the nAV condition are shown in Table V(a) and (b) to examine how the McGurk effect occurred. In the nAV condition, visual biasing effects were strong and widespread. In each confusion matrix in Table V, ten categories of auditory stimuli and responses are divided into labials and nonlabials by vertical and horizontal thick lines.

Table V(a) shows the results for the visual labials. The perception of auditory nonlabial stimuli was quite significantly biased by visual labial stimuli. Every auditory nonlabial stimulus type was perceived as its corresponding labial type, the predominant response in most cases, almost to the exclusion of the "correct" response. For visual /ba/, for example [Table V(a), vision = *b*], auditory /da/ was perceived as "ba" 60% of the time, whereas the identical auditory /da/ was correctly perceived 94% of the time by audition alone in nA condition (Table IV). Similarly, erroneous shifts to labials are as follows: /ta/ was perceived as "pa" 94% of the time, /na/ as "ma" (74%), /ra/ as "wa" (45%) or as "ma" (27%), /ga/ as "ba" (88%), and /ka/ as "pa" (94%). This response pattern is also true with visual /pa/, /ma/, and /wa/ stimuli. As shown in the lower left of Table VI, each of these labial responses for auditory nonlabials were significantly greater than auditory errors. The response pattens in Table V(a) show that auditory nonlabials were visually biased to labials with the same manner of articulation.

Analogously, when combined with visual nonlabials [Table V(b)], auditory labial stimuli were to a large extent perceived as nonlabial. When the visual stimulus was /da/ [Table V(b), vision = *d*], auditory /ba/ was perceived as "da" (38%) or "ga" (29%), while the identical auditory /ba/ was 91% correctly perceived by audition alone in nA condition (Table IV). Similarly, the vision-biased erroneous shift to nonlabial is as follows: Auditory /pa/ was perceived as "ta" (92%), /ma/ as "na" (77%), and /wa/ as "ra" (37%) or "ga" (22%). This response pattern is also found with the other visual nonlabial stimuli, showing that auditory labials were visually biased to nonlabials with the same manner of articulation. As shown in the upper right of Table VI, all the auditory labials with visual nonlabials yielded highly significant visual effects.

The influence of the auditory intelligibility on the McGurk effect was again tested for audio-visual incongruent stimuli, using 2×2 contingency tables (the frequencies of labial versus nonlabial responses $\times$ auditory stimuli with intelligibility score more versus less than 90%).[3] For the auditory nonlabials with the visual labials, the auditory stimuli with intelligibility scores less than 90% (auditory /ta/, /ra/, /ga/, and /ka/) yielded significantly more McGurk effects than the more intelligible auditory stimuli (auditory /da/ and /na/) ($df = 1$, $N = 1161.2$, $p < 0.001$). Similarly, for the auditory labials with the visual nonlabials, the auditory stimulus with an intelligibility score less than 90% (auditory /pa/) was influenced by visual cues significantly more than the more intelligible ones (auditory /ba/, /ma/, and /wa/) ($df = 1$, $N = 1330$, $p < 0.001$). These results indicate that the amplitude of the McGurk effect was dependent on the auditory intelligibility in the nAV condition as well as the AV condition.

Summarizing the results for the nAV condition, the main findings were as follows: First, the same auditory stimuli were perceived quite differently depending on the visual stimuli: e.g., auditory /ta/ was incorrectly perceived as "pa" 98% of the time when combined with visual /pa/ stimuli [Table V(a), vision = *p*] while it was correctly perceived as "ta" 98% of the time when combined with visual /ta/ stimuli [Table V(b), vision = *t*]. These visual biasing effects are clearly understood as compared with the auditory confusion in nA condition where both "ta" responses for /pa/ (37%) and "pa" responses for /ta/ (27%) were observed.

Second, we found few "combination responses" in our experiment whereas the original McGurk effect included many for auditory nonlabials combined with visual labials. For this type of stimulus, our results in nAV condition show

TABLE V. Results in the *noise-added* audio-visual condition. Indicated in %. (a) Confusion matrices for the stimuli with *visual labials* (/b,p,m,w/). (b) Confusion matrices for the stimuli with *visual nonlabials* (/d,t,n,r,g,k/).

(a) response

vision = b

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	97	3									h1
p		98				2					
m			100								
w	2		6	86				5			h2
d	60				39						bd2
t	2	94				3				1	
n			74			2	24				
r	8	1	27	45				19			y1
g	88				7				5		
k	1	94				2				2	bp2

vision = p

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	98								2		
p		98				2					
m			100								
w	2		7	77				11			y3
d	45				51			2	3		
t		98				2					
n		1	67				32				
r	9		29	34				23			y5
g	93				5				2		
k	2	98									

vision = m

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	96				2				2		h1
p		98				2					
m			100								
w	5	1	7	79				8			
d	64				34		2	1			
t		97				2				2	
n			84				14				h2
r	8		40	29			2	14			y5 wm2
g	93				4				2	2	
k		98				2					h1

vision = w

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	88			2	1	2			5		h3
p		93				3					h1 f2
m			100								
w				100							
d	33			2	61				3		z1
t		90		1		3				1	f4
n			35				65				
r			2	84				13			a1
g	72			1	17			1	9		
k		92		1		2				2	pk2

(b) response

vision = d

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	25				38	3			29		h4 z1
p		7				92		1			h1
m			20				77	4			
w				31	2			37	22		y9
d	1				92				8		
t	1					93		2		3	pk2
n			1				96		3		
r				4			8	77	7		y4
g		2			64				34		z1
k		7				47			3	35	a4 h2 pt2

vision = t

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	11				57	6			25		h1 z1
p		2				97					h2
m			18				80	2			
w				20			6	45	13		
d					97				3		
t						98			1	2	
n			2				98				
r				7			12	63	5		
g	1				82				16		z2
k		4				42			3	39	a5 h3 s2 pt2

vision = n

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	24		2		33	2			36		h3
p		6				90				1	pt2
m			21				77		2		
w				32	1			40	19		y8
d	2				95				2		
t		4				81				14	h1
n							100				
r				5	1	2	6	63	13		y10
g	2				61			1	34		z2
k		2				36			2	53	a3 h2

vision = r

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	13				33	1		3	46		h4
p		2				93				3	h1 s1
m			15				82	3			
w				21			2	46	21		y10
d					82				17		h1
t						90			9		h1
n							93		7		
r				3			3	63	12		y20
g	2				52			1	46		
k		2				35			2	47	h6 s2 pt2

vision = g

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	13				48				35		h3
p		1				98				1	
m			17				78	4			
w				22				47	18		y13
d	1				98				1		
t		1				98				2	
n							98		2		
r				1	2		7	64	13		y13
g	2				71				25	2	h1
k						45			2	47	a3 h2

vision = k

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	10				30	2		1	46		h6 z2 dg2
p		4				87				5	a2 h1 pt2
m			11				86	3			
w				20			3	43	21		y13
d					91				9		
t		1				90				9	
n							100				
r					5		6	63	12		y14
g					59				40		z1
k		1				38				56	h2 a3

TABLE VI. The frequency of labial responses as an index of place of articulation confusions (noise-added condition), maximum = 60. Asterisks indicate significant differences in the frequency ($*p<0.05$, $**p<0.01$, $***p<0.001$) between the audio-visual stimulus and the audio-alone stimulus, based on the chi-square test ($df=1$, $N=120$).

audio \ video	none	b	p	m	w	d	t	n	r	g	k
b	54.4	59.6	59	57.6	53.6	15***	6.6***	15.2***	7.6***	7.6***	6***
p	33.8	59***	59***	59***	55.8***	4***	1***	3.8***	1***	0.4***	2.6***
m	59.6	60	60	60	60	11.6***	11***	12.6***	9***	10.4***	6.4***
w	56.4	56	51.6	55	60	18.4***	12***	19.2***	12.4***	13.2***	11.8***
d	2	35.8***	26.8***	38.2***	20.8***	0.4	0	1.4	0	0.4	0
t	16.6	57.6***	59***	58***	54.2***	0.4***	0***	2.4***	0***	0.4***	0.4***
n	1	44.4***	40.6***	50.4***	21***	0.4	1	0	0	0	1
r	20.8	48.2***	43***	46.6***	51.8***	2.4***	4.4***	3***	1.6***	0.8***	3.4***
g	21.2	53***	55.6***	55.6***	43.6***	1***	0.6***	1.4***	1***	1***	0***
k	8.4	58***	60***	59***	55.6***	4	3.4	1*	2	0**	0.4*

fused responses instead of combination responses, as well as for the conversely combined stimulus type. For example, 72%–93% of auditory /ga/ were perceived as "ba" with visual labials [Table V(a)]. On the other hand, for this stimulus type, McGurk and MacDonald (1976) reported no fused ("ba") responses and "bga" responses 54% of the time.

5. Relationship between intelligibility and visual biasing effect

Figure 3 illustrates the relation between the auditory intelligibility (measured in audio-alone conditions) and the McGurk effect (place of articulation confusions for audio-visual incongruent stimuli, reduced by auditory errors). For the auditory stimuli with intelligibility scores of 100% (/ba/, /da/, /na/, /ra/, /ga/, /ka/, /ma/ in noise-free condition), there were few visual biasing effects. On the other hand, when the auditory intelligibility was 99% or less, the McGurk effect was significant, especially in the noise-added condition. Figure 3 suggests a nonlinear function between the auditory intelligibility and the McGurk effect with the exception of /ra/ in the noise-added condition. As indicated earlier, the chi-square test revealed that the McGurk effect occurred more for less intelligible auditory stimuli when we compared (a) completely (100%) intelligible auditory stimuli with incompletely intelligible ones in the noise-free condition, and (b) auditory stimuli of more than 90% intelligibility with those of less than 90% intelligibility in the noise-added condition. These results indicate that the Japanese McGurk effect is dependent on the auditory intelligibility.

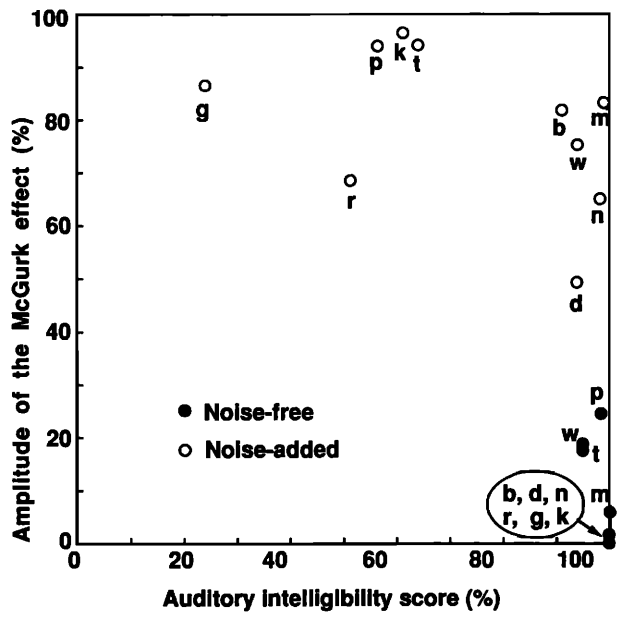

FIG. 3. Relationship between the auditory intelligibility and the McGurk effect.

However, as seen in Fig. 3, the influence of the intelligibility is not equivalent for the noise-free and noise-added conditions: Even when the intelligibility score was the same, stimuli in the nAV condition showed much stronger McGurk effects than those in the AV condition (compare /na/ in the nAV condition with /pa/ in the AV condition: The intelligibility score was 98% for both). This fact indicates that both the auditory intelligibility and the existence of noise influence the McGurk effect.

II. DISCUSSION

There were several inconsistencies between the results of the present study and those of McGurk and MacDonald (1976).

First, the results of our noise-free condition indicated that it is much less easy to induce the Japanese McGurk effect than the English one. In our noise-free condition, the visual biasing effect was weak, and if any, fused responses occurred only about 20% of the time. The difference between the Japanese and the English McGurk effect is clearly shown by the fact that our auditory /ba/ combined with visual /ga/ was correctly perceived as "ba" 100% of the time whereas this type of stimulus yielded fused "da" response 98% of the time for the adult subjects of McGurk and MacDonald. Possible sources for this inconsistency are differences in cultural factors (in terms of production or perception of speech) and in the auditory intelligibility. Although we cannot know the intelligibility of each syllable used by McGurk and MacDonald, their description on the averaged intelligibility for four syllables (99% for /ba/, /ga/, /pa/, and /ka/) leaves a possibility that the intelligibility score of their /ba/ was less than 100%.[4]

Comparisons of the results of audio-visual conditions with those of audio-alone conditions indicated that the Japanese McGurk effect depended on auditory intelligibility. When the auditory intelligibility was 100%, the McGurk effect was absent or very weak. If intelligibility was less than 100%, however, the McGurk effect could be induced.

In the noise-added condition, visual biasing effects were observed so strongly for every audio-visual incongruent pair that visual information had the decisive vote for most of the perceptual decisions. The results indicated that both the poor intelligibility and the existence of noise promoted the visual biasing effects. Human beings may depend on eyes in the presence of auditory uncertainty.

Another difference between our results and those of McGurk and MacDonald (1976) is the type of audio-visual incongruent pair in which visual biasing effects were observed. The original McGurk fusion effect was found only for pairs of auditory labials with visual nonlabials and the fusion effect was not found for the conversely combined audio-visual pairs. In the present experiment, however, the fusion effects were found symmetrically, in both the auditory labials with visual nonlabials and the auditory nonlabials with visual labials.

Instead, we found very few "combination responses" such as "bda" responses for auditory /da/ combined with visual /ba/. On the other hand, McGurk and MacDonald reported this type of responses for auditory /ga/ with visual /ba/ 54% of the time and for auditory /ka/ with visual /pa/ 44% of the time. The reason for these differences may be attributed to either/both the articulatory characteristics of Japanese syllables or Japanese listeners' perceptual organization for speech which takes into account the fact that the Japanese phonological system does not allow any consonant clusters in a "systematic phonemic level" (Chomsky and Halle, 1967).

III. CONCLUSION

This study showed that there were very few McGurk illusions in Japanese subjects listening to Japanese syllables of 100%-auditory intelligibility and that the illusions depended on the auditory intelligibility.

If any, the Japanese McGurk effect was less easily induced than the English one. Although this difference seemed to be attributed to differences in production of speech (stimulus factor) of two languages or/and perception of speech (subject factor), the present experiment was not designed to examine these factors. Do Japanese-speaking subjects depend on eyes less than English-speaking subjects? Or do English-speaking subjects also show few McGurk effects for Japanese syllables? Is there also little McGurk effect for English syllables of complete intelligibility? Does the English McGurk effect also depend on the auditory intelligibility? Does noise promote the English McGurk effect? To answer these questions, a cross-language study in which audio-alone conditions are incorporated is necessary.

ACKNOWLEDGMENTS

The authors would like to thank Professor T. R. Hofmann, Professor D. W. Massaro, and an anonymous reviewer, and the editor for their critical comments on the earlier version of this manuscript.

[1] Later work has demonstrated that the idea that place information is provided visually and manner information is provided auditorially cannot fully explain the McGurk effect. More recent work shows that visual speech information affects the identification of both the place and manner features in auditory-visual speech perception (Green and Kuhl, 1989, 1991).

[2] In this test, the frequencies of the place of articulation confusions (such as "pa" response for auditory /ta/) in the AV condition were analyzed after being reduced by those in the A condition (auditory errors).

[3] In this test, the frequencies of the place of articulation confusions in the nAV condition were analyzed after being reduced by those in the nA condition (auditory errors).

[4] Although 100% and 99% are not statistically different, there is a theoretical difference between these because the intelligibility score of 100% often includes a ceiling effect that makes it impossible to measure the "true" intelligibility. Another inconvenience with the intelligibility score is that it is not an "absolute" measure of the intelligibility. As the studies based on information theory have shown (Miller *et al.*, 1951), the intelligibility score decreases as the number of test items increase. Thus, a score of 99% obtained in our experiment from ten materials does not represent the same intelligibility as a 99% score reported by McGurk and MacDonald which was obtained for four items.

Chomsky, A. N., and Halle, M. (**1967**). *The Sound Pattern of English* (Wiley, New York).

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

Easton, R. D., and Basala, M. (**1982**). "Perceptual dominance during lipreading," Percept. Psychophys. **32**, 562–570.

Erber, N. (**1975**). "Auditory-visual perception of speech," J. Speech Hear. Dis. **40**, 481–492.

Fukuda, Y., Shiroma, M., and Funasaka, S. (**1988**). "Speech perception ability of the patients with artificial Inner Ear: Combined use of auditory and visual perception," Technic. Rep. IEICE (The Institute of Electronics, Information and Communication Engineers of Japan), SP88-91 (in Japanese with English abstract).

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

Green, K. P., Kuhl, P. K. (**1991**). "Integral processing of visual place and auditory voicing information during phonetic perception," J. Exp. Psychol: Hum. Percept. Perform. **17**, 278–288.

Green, K. P., Kuhl, P. K., and Meltzoff, A. N. (**1988**). "Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment," J. Acoust. Soc. Am. Suppl. 1 **84**, S155.

MacDonald, J., and McGurk, H. (**1978**). "Visual influences on speech perception processes," Percept. Psychophys. **24**, 253–257.

Massaro, D. W., and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol: Hum. Percept. Perform. **9**, 753–771.

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

Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). "The Intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol. **41**, 329–335.

Sekiyama, K., Joe, K., and Umeda, M. (**1988**). "Perceptual components of Japanese syllables in lipreading: a multidimensional study," Technic. Rep. IECE (The Institute of Electronics, Information and Communication Engineers of Japan) IE87-127 (in Japanese with English abstract).

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*, edited by R. Campbell and B. Dodd (Erlbaum, London), pp. 3–51.