J. Acoust. Soc. Jpn. (*E*) **15**, 3 (1994)

Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility

Kaoru Sekiyama

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

(*Received 30 September 1993*)

In English speaking cultures, it has been reported that when auditory speech is presented in synchrony with discrepant visual (lip-read) speech, the subjects often report hearing sounds that integrate information from the two modalities (the "McGurk effect"). However, Sekiyama and Tohkura recently showed that Japanese subjects are less influenced by discrepant visual cues than Americans. This study examined the inter-language differences in terms of the relationship between the magnitude of the McGurk effect and the frequency of incompatibility. The stimulus materials were ten syllables (/ba, pa, ma, wa, da, ta, na, ga, ka, ra/) pronounced by a Japanese and an American speaker. The ten auditory and ten visual syllables pronounced by a speaker were cross-dubbed resulting in 100 auditory-visual stimuli. Japanese syllables were presented to 14 Japanese and 10 American subjects. English syllables were presented to different groups of subjects, 12 Japanese and 10 American. The stimuli were presented in both quiet and noise-added conditions. The subjects were asked to check incompatibility between what they heard and what they saw as well as to report what they heard. The results showed that the magnitude of the McGurk effect correlated highly negatively with the frequency of incompatibility. It was suggested that the small McGurk effect in the Japanese subjects listening to Japanese speech is due to the higher sensitivity of the Japanese to auditory-visual discrepancy.

Keywords: Speech perception, Auditory-visual integration, McGurk effect, Inter-language difference, Incompatibility

PACS number: 43. 71. Ma, 43. 71. Hw

1. INTRODUCTION

Although speech perception is considered primarily an auditory process, visual information in the form of lipreading is a source of speech perception in face-to-face verbal communication. Lip-read information normally facilitates speech perception especially when auditory speech signals are unintelligible. It is perhaps hearing impaired people that utilize lipreading most to comprehend speech. In fact, recent studies have reported that lip-read information improves speech perception considerably by patients with cochlear implants (inner ear prostheses), that is, people who have to process unintelligible auditory information that the cochlear implants provide (Blamey, Dowell, Brown, Clark, & Seligman[1]; Fukuda, Shiroma, & Funasaka[2]; Funasaka, Shiroma, Yukawa, Iizuka, Yao, Kono, Takahashi, & Kumakawa[3]). Besides the hard of hearing, visual information has been also shown to be helpful for people with normal hearing, especially when the auditory signal is not clear due to added noise (Erber[4], O'Neil[5]; Sumby & Pollack[6]).

The contribution of lip-read information to speech perception is described as compensating for auditory information. In the case of consonant percep-

tion, visual information has the advantage of conveying spatial information, namely, information about the place of articulation that auditory information may fail to convey in some circumstances. Particularly, the visual system is highly sensitive to a distinction between labial (lip-articulated; *e.g.*, /b/, /p/, /m/) and non-labial (articulated inside the mouth; *e.g.*, /d/, /t/, /n/) that the auditory system sometimes misses (*e.g.*, Binie, Montgomery, & Jackson[7]). Using these characteristics, McGurk and MacDonald[8] dramatically demonstrated that visual information influences speech perception even when the auditory signal is highly intelligible. When they presented their normal adult subjects with auditory speech signals of /ba/, they correctly reported hearing "ba" about 99% of the time. However, when the identical auditory speech signals of /ba/ were presented in synchrony with visual lip movements for /ga/, they reported hearing "da" 98% of the time. This "McGurk effect" demonstrates that visual information about place of articulation easily modifies phonetic perception. In this example, visual place information (/ga/: non-labial) biased auditory information (/ba/: labial) with the result that perceptual solution ("da": non-labial) was consistent with the biased place information and the remaining auditory information.

Since McGurk and MacDonald first reported this fusion effect between auditory and visual information, it has been replicated in many studies and established as a robust effect in English speaking cultures (for a review, see Summerfield[9]). These studies have shown that the McGurk effect can be produced under various conditions as follows. The McGurk effect occurs for various combinations of auditory and visual syllables when there is a discrepancy in the place of articulation (MacDonald & McGurk[10]). In most of the cases, auditory labials paired with visual non-labials yield *fused responses* (*e.g.*, auditory /pa/ paired with visual /na/ produces "ta"-response). On the other hand, auditory non-labials paired with visual labials often result in *combined responses* (*e.g.*, auditory /ta/ and visual /pa/ produces "pta"-response). The fusion effect has been also reported for materials of meaningful words (Deckle, Fowler, & Funnell[11]; Dodd[12]) or synthesized speech syllables (Massaro[13]; Massaro & Cohen[14]). It has also been reported that a male speaker's voice paired with a female face can produce the McGurk effect as much as a male voice-male face pair (Green, Kuhl, Meltzoff, & Stevens[15]). Auditory-visual discrepancy yields fused or combined responses and also longer reaction times (Green & Kuhl[16]) and reaction time depends on the extent of ambiguity (Massaro[13]).

However, these studies conducted in English speaking cultures have not yet examined the McGurk effect for linguistic/cultural independence. Although Mills and Thiem (1980, cited in Massaro[13]) asked German listeners to identify syllables consisting of conflicting auditory and visual information, their results also showed that the visual component has strong effects on the identification. In contrast to this, Sekiyama and Tohkura[17] recently reported that Japanese listeners show very little McGurk effect for Japanese syllables when the auditory speech was perfectly intelligible. For example, being presented with auditory /ba/ paired with visual /ga/, their Japanese subjects reported hearing auditory "ba" 100% of the time, in contrast to the fused "da" responses reported for English subjects. They were presented with the speech either in quiet or in noisy condition. In the quiet condition, the visual effects were not substantial except for some auditory syllables that were not perfectly identified in the audio-alone presentation. On the other hand, strong visual effects were observed for each auditory syllable in the noise-added condition where the auditory speech had poor intelligibility. From these results, Sekiyama and Tohkura[17] proposed an intelligibility hypothesis that Japanese listeners hearing Japanese speech do not integrate visual information with auditory information when audition provides sufficient information.

As shown above, the Japanese subjects did show strong visual effects in the noise-added condition, which indicates that Japanese can lipread to the extent that the McGurk effects is induced. This was more directly shown in our previous study (Sekiyama, Joe, & Umeda[18]) where 60 normal untrained Japanese adults were asked to lipread Japanese nonsense mono-syllables. Although identifying consonants was difficult (this may be partly due to our use of a large set of syllables, that is, the 100 syllables of which the Japanese syllable inventory is composed), the Japanese subjects could visually discriminate labials (/p, b, m, w, py, by, my/) from non-labials 92% of the time. A multidimensional scaling found six perceptual clusters

(/w/, /p, b, m/, /py, by, my/, /h/, /t, s, d, z, n, k, g, r/, as well as these non-labials combined with the semivowel /y/ as in /kya/). Depending on the Japanese phoneme inventory, these clusters are somewhat different from those found in English. For example, six clusters found by Walden, Prosek, Montogomery, Scherr, & Jones[19] were /p, b, m/, /θ, ð/, /f, v/, /s, z, ʃ, ʒ/, /w, r/, and /t, d, n, k, g, j, r, l/. However, as for the labial vs. non-labial discrimination score, which is relevant to the McGurk effect, the score in our study was equivalent to the one obtained in English by Walden *et al.*[19]; my calculation based on their results is 91%. Thus, it is unlikely that the small McGurk effect in Japanese subjects listening to intelligible Japanese speech is due to the subjects' poor lipreading ability.

Subsequently, Sekiyama extended the above audio-visual experiment to three new conditions: AJ condition (American subjects, Japanese stimuli) tested native speakers of American English with the identical Japanese speech stimuli. In AE (American subjects, English stimuli) and JE (Japanese subjects, English stimuli) conditions, American subjects and Japanese subjects were tested with English speech stimuli pronounced by a native speaker of American English. Combining the results of these three conditions and those of the earlier study (JJ condition: Japanese subjects, Japanese stimuli), Sekiyama and Tohkura[20] reported inter-language differences between Japanese and American subjects. The study showed that the magnitude of the McGurk effect was the largest when American subjects were presented with Japanese stimuli. When English stimuli were presented, the magnitude of the McGurk effect in Japanese subjects was comparable to that in American subjects, replicating the substantial McGurk effect as has been reported in literature. These results indicate that the small McGurk effect in Japanese subjects for Japanese syllables is due to a perceptual processing of Japanese listeners when they process speech information of their native language. This difference between the AE and JJ conditions seems to coincide anecdotal evidence. Native speakers of English often dislike dubbed movies (Massaro[13], p. 36); they may experience something like the McGurk effect for the auditory-visual discrepancy that dubbed movies often include. On the other hand, Japanese listeners enjoy them simply although some intellectual people complain of the loss of an atmosphere created in the original language.

The purpose of the present study is to give a more penetrating description of the inter-language differences in this phenomenon, by examining the data on sensitivity to the discrepancy between auditory and visual information. In our experiment, the subjects had two concurrent talks. They were asked to check whether or not they felt the visual lip movements incompatible with what they heard, as well as to report what they heard. This request for reporting incompatibility had two objectives. One was to confirm the results of our pilot experiment where Japanese subjects seemed to experience incompatibility instead of perceptual fusion most of the time when presented with discrepant auditory-visual stimuli of Japanese speech. This suggests that the small McGurk effect of Japanese subjects for Japanese syllables is related to frequent incompatibility. The second objective of this task was to make the subjects pay attention to both the auditory and visual information.

Although Sekiyama and Tohkura[20] briefly mentioned the frequency of incompatibility in the four conditions, detailed analyses of the incompatibility data were not done because one condition, the JJ condition, was carried out in a different procedure. That is, whereas the other three conditions asked the subjects to report only one response (what they heard), the JJ condition encouraged the subjects to report more than one response. Such a procedural difference may have caused different processes for detecting incompatibility between what they heard and what they saw. In this study, the JJ condition was replicated with the identical procedure used for the other three conditions so that a direct comparison was possible.

The main purpose of this study is to compare the frequency of incompatibility in terms of its relationship with the magnitude of the McGurk effect. To do so, the present study will present the full data set necessary to examine the relationship between the two indexes, because our earlier study (Sekiyama & Tohkura[20]) reported only a small set of the data from the experiment. The second purpose of this study was to see if the JJ condition replicates the small McGurk effect.

The results showed that Japanese subjects are more sensitive to the discrepancy between auditory and visual information and that the magnitude of

the McGurk effect is a negative linear function of the frequency of incompatibility.

2. EXPERIMENT

2.1 Subjects

Four groups of subjects participated. Twenty-six native speakers of Japanese were assigned to the JJ (14 subjects) and JE (12 subjects) conditions. Twenty native speakers of American English were assigned to the AJ and AE conditions (ten subjects for each). All subjects had normal hearing and normal or corrected to normal vision. The Japanese subjects were students at Kanazawa University and most of the American subjects were graduate students at the City University of New York. Their ages ranged from 20 to 30 years old. None of the subjects who were given foreign speech stimuli (JE and AJ conditions) had lived in a culture in which the tested language is spoken natively. However, all the Japanese subjedts had taken English classes for at least six years starting at age twelve. None of the American subjects had studied Japanese. The Japanese subjects were volunteers, and the American subjects were paid $55 for the four-hour experiment.

2.2 Stimuli

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

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

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

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

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

On the final videotapes, each syllable occured in a 7 s trial in which the video channel included 3 s of black frames and 4 s of the face. The audio channel included warning tones and speech (see Sekiyama & Tohkura,[17] Fig. 1). There were several tapes of different random orders of 100 stimuli ($7\text{ s} \times 100$: 12 min). For the experimental presentation, the outputs of the BETACAM videotapes were dubbed onto VHS tapes.

Videotapes for auditory-only presentation were also created. On these tapes, the video channel included only black frames. For the auditory-only

presentation, a videotape induced six repetitions of each auditory stimulus (7 s × 10 × 6: 7 min).

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

2.3 Experimental Design

The subjects were asked to report what they heard in four conditions of two within-subjects factors: (a) quiet (no-noise-added) or noise-added, (b) auditory-only or auditory-visual. In the quiet auditory-only condition (A condition), auditory stimuli were presented with video black. In the quiet auditory-visual condition (AV condition), auditory-visual stimuli were presented. In the noise-added auditory-only condition (nA condition), white noise was continuously added when the videotapes for the A condition were played. In the noise-added auditory-visual condition (nAV condition), the videotapes for the AV condition were played with the white noise. Signal to noise ratios in the noise-added conditions (nA and nAV), measured by a sound level meter at the location of the subject's head, were kept at 0 dB. The intensity of speech and noise was 55 dB SPL (A scale, fast) that was 5 dB higher than that in Sekiyama and Tohkura.[17] Background noise was 25 ~ 35 dB SPL (A scale). The white noise was presented from a noise generator with its own speaker that was located beneath the monitor (Kanazawa), or it was presented from the loud speakers from which the auditory speech was also presented (New York).

2.4 Procedure

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

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

3. RESULTS

3.1 Preparation for Analyses

The results in the auditory-visual condition are presented separately for two cases: the case in which the visual stimuli were labials (/b, p, m/) and the case in which they were non-labials (/d, t, n, g, k/). This is because the response patterns for the within-group visual tokens were almost identical, agreeing with those shown by Sekiyama and Tohkura.[17] Stimuli consisting of auditory or visual /wa/ or /ra/ were excluded from the cross-language comparison because they were dissimilar between the two languages in both auditory and visual aspects, as shown in our previous study (Sekiyama & Tohkura[20]).

Removing the results for /wa/ and /ra/, the results for the A condition showed that both language groups accurately identified most of the syllables in the auditory-only presentation. The average percents of correct responses in the A condition across the rest of the eight auditory syllables were 99.7% (JJ condition), 96.0% (AJ condition), 93.6% (JE condition), and 96.7% (AE condition). When these

means were compared by a two-factor ANOVA [Listener's Language (2) × Stimulus Language (2)], the main effects of Listener's Language and Stimulus Language were nonsignificant [$F(1, 42)=0.04$, $p<.90$; $F(1, 42)=3.27$, $p<.10$]. However, the interaction was significant [$F(1, 42)=4.94$, $p<.05$]. The source of this interaction was tested by pairwise comparisons using least squares means. The results showed that a significant difference was only between the JJ and JE conditions ($t=3.05$, $p<.01$). No significant differences were found in the other pairs of conditions. Therefore, the accuracy of the identification in the auditory-only condition was about the same across the four conditions with a reservation that the JE condition had a slightly poorer intelligibility score.

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

Further details of the results can be found in Sekiyama.[21]

3.2 Weak McGurk Effect in the Quiet JJ Condition

The confusion matrices in Fig. 1 show the results for the quiet JJ condition by indicating the frequency of each response as a row percent of all responses in a row. The numbers in the leftmost column in parentheses indicate the percent of correct responses for the auditory syllable in the auditory-only condition ("auditory intelligibility score"). If the subjects hear the speech without visual influence, each of the diagonal cells should contain the same response frequency as the auditory intelligibility score in the leftmost cell. Alternatively, if the McGurk effect occurs, it is expected that auditory labials paired with visual non-labials (upper three stimuli in the upper panels) will be perceived as non-labials (fused responses) and that auditory non-labials with visual labials (lower five stimuli in the lower panels) will be perceived as labials (fused responses) or as syllables with consonant clusters (combined responses).

As seen in Fig. 1 where shadowed portions indicate the responses influenced by discrepant visual input (the McGurk effect), Japanese subjects showed only a limited McGurk effect, replicating the results by Sekiyama and Tohkura.[17] No combined responses were observed perhaps because Japanese phonological system does not allow any phonological consonant clusters. Figure 1 suggests that the McGurk effect is negligible except for auditory /pa/ and /ta/. This was confirmed by statistical analyses. A single-factor ANOVA with repeated measures tested the effect of discrepant visual input in each auditory syllable, by comparing the frequency of "place errors" in the A condition with that in the AV condition. To do so, responses which included confusions of labials with non-labials were counted as place errors. Therefore, "pa"-response to auditory /ta/ was taken as a place error while "ka"-response to auditory /ta/ was not.

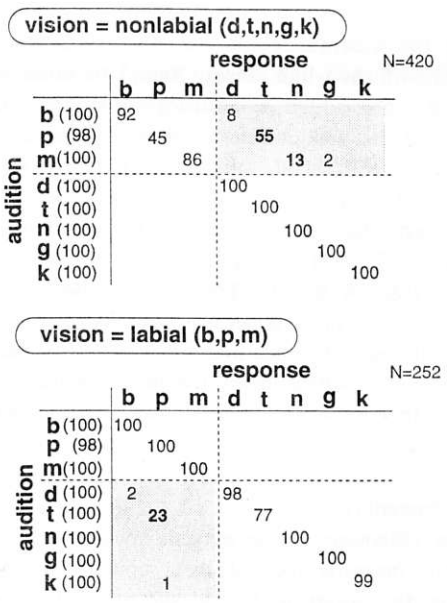

Fig. 1 Confusion matrices for the JJ group in the AV condition (percent in a row). The number of responses in a row is shown at the right upper side of a panel as "N" (subject × repetition × visual stimuli). Shadowed responses indicate the McGurk effect.

In the AV condition, fused and combined responses were counted as place errors. As a result, significant visual effects were found only in auditory /pa/ [$F(1, 13)=33.10$, $p<.0001$] and /ta/ [$F(1, 13)=9.02$, $p<.05$].

These results show that the average magnitude of the McGurk effect was very small for Japanese subjects observing Japanese stimuli. The results also support the intelligibility hypothesis that Japanese listeners hearing Japanese speech do not integrate visual information with auditory information when audition provides sufficient information. As Fig. 1 shows, the substantial (more than 50% of the time) McGurk effect was observed only in aud itory /pa/, and it was only this auditory /pa/ that had an intelligibility score less than 100%.

3.3 Comparison of the Magnitude of the McGurk Effect across Four Groups

Figure 2 gives a comparison of the magnitude of the McGurk effect across four groups (for the confusion matrices for the AJ, AE, JE conditions, see Sekiyama[21]). Because the results in the AV condition showed an apparent difference between auditory labials paired with visual non-labials and conversely paired stimuli, the results are shown for the two cases. Bars indicate the frequency of visually influenced responses (place errors), that is, the sum of fused and combined responses. In the quiet condition, it is clear that the visual effect was the weakest in the JJ condition especially for auditory labials. However, for auditory non-labials, the differences among the JJ, JE, AE conditions are not obvious. It was only in the AJ condition that a substantial visual effect on the auditory non-labials was observed.

In the noise-added condition, place errors greatly increased especially in the JJ condition. In the other three groups, the increase was large for auditory non-labials for which the visual effect was relatively weak in the quiet condition. As a result, the differences among the four groups and between auditory labials and non-labials are smaller in the noise-added condition than in the quiet condition. However, the results for the noise-added condition suggest a ceiling effect, showing the frequencies of over 90% in many cases.

Concerning this ceiling effect, there was a difficulty in handling the data for the noise-added condition. In Fig. 2, the visually influenced responses are indicated by "place errors" in the auditory-visual condition. As shown in Fig. 2, noise increased place errors in the auditory-visual condition. However, these place errors seemed to include both auditory errors and visual effects because the noise also increased place errors in the auditory-only condition. Percents of place errors in the nA condition averaged over auditory labials and auditory non-labials were 5% and 21% for the JJ condition, 34% and 4% for the JE condition, 4% and 19% for the AJ condition, and 20% and 6% for the AE condition. One possible way to estimate the magnitude of visual effects could be to subtract the frequency of auditory errors from the frequency of the "visually influenced responses." However, it seemed to be inappropriate at least in the noise-added condition. For, whereas the noise simply increased place errors in the nA condition, the increase of place errors in the nAV condition seemed to be suppressed by the ceiling effect described above. Therefore, as an ap-

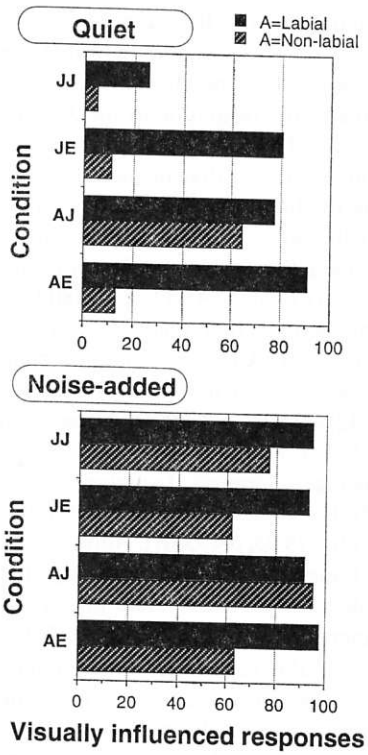

Fig. 2 The magnitude of visual influence approximated by the frequency of place errors (%), *i.e.*, the sum of fused and combined responses.

proximation of the magnitude of the visual influence, the frequency of "visually influenced responses" shown in Fig. 2 was adopted.

To compare these mean frequencies, a four-factor ANOVA [Noise (2) × Listener's Language (2) × Stimulus Language (2) × Auditory Stimulus (2)] was carried out. The results are summarized as follows. (a) The main effect of Noise [$F(1, 42)=301.31$, $p<.0001$], Listener's Language [$F(1, 42)=26.49$, $p<.0001$], and Auditory Stimulus [$F(1, 42)=194.49$, $p<.0001$] were highly significant. That is, the visual effect was stronger in the noise-added condition than in the quiet condition; it was stronger for American subjects than for Japanese subjects, and it was stronger for auditory labials than for auditory non-labials. The main effect of Stimulus Language was not significant [$F(1, 42)=.51$]. Therefore, the magnitude of the visual influence did not differ for Japanese stimuli and English stimuli. (b) The interaction of Listener's Language × Stimulus Language was significant [$F(1, 42)=14.26$, $p<.001$]. This seems to reflect mainly the differences among the four groups in the quiet condition where the magnitude of visual influence was larger for the subjects' non-native speech than for their native speech (compare the magnitude in the JE condition with that in the JJ condition and the magnitude in the AJ condition with that in the AE condition). (c) Reflecting the fact that the increase of the visual effect due to the noise was larger for auditory non-labials, the interaction of Noise × Auditory Stimulus was significant [$F(1, 42)=45.32$, $p<.0001$]. (d) In the other interactions that included within-subject factors (Noise and Auditory Stimuli), significant interactions were found in Noise × Listener's Language [$F(1, 42)=34.28$, $p<.0001$], Noise × Stimulus Language [$F(1, 42)=13.83$, $p<.001$], Noise × Listener's Language × Stimulus Language [$F(1, 42)=25.48$, $p<.0001$], Auditory Stimulus × Stimulus Language [$F(1. 42)=78.05$, $p<.0001$], Auditory Stimulus × Listener's Language × Stimulus Language [$F(1, 42)=4.44$, $p<.05$], and Noise × Auditory Stimulus × Stimulus Language [$F(1, 42)=17.03$, $p<.001$].

Subsequently, the sources of these interactions were examined by pairwise comparisons using least squares means. The results were as follows. (1) Whereas the American subjects showed the stronger visual influence than the Japanese subjects in the quiet condition, for both auditory labials ($t=4.50$, $p<.0001$) and non-labials ($t=7.93$, $p<.0001$), there were no differences between the two language groups in the noise-added condition either for auditory labials ($t=0.28$) or non-labials ($t=1.73$). (2) In the quiet condition, the visual influence was stronger for the English stimuli than for the Japanese stimuli when auditory stimuli were labial ($t=4.89$, $p<.0001$), but it was stronger for the Japanese stimuli when auditory stimuli were non-labial ($t=5.90$, $p<.0001$). In the noise-added condition, there were no differences between the two stimulus languages when auditory stimuli were labials ($t=0.65$), but the visual influence was larger for the Japanese stimuli when auditory stimuli were non-labial ($t=4.23$, $p<.0001$). (3) In the quiet condition, when auditory stimuli were labial, the visual influence in the JJ condition was significantly weaker than that in the other three conditions ($p<.0001$ for each pair), and there were no significant differences among the other three. When the auditory stimuli were non-labial, the visual influence in the AJ condition was significantly larger than that in the other three conditions ($p<.0001$ for each pair), and there were no significant differences among the other three. (4) In the noise-added condition, there were no significant differences in any pairs of the JJ, AJ, JE, and AE conditions when auditory stimuli were labial. When auditory stimuli were non-labial, the visual influence was significantly stronger for the AJ condition than that for the other three; it was stronger for the JJ condition than that for the JE condition, and the difference between the JJ and AE conditions and the difference between the AE and JE conditions were nonsignificant.

3.4 Frequency of Incompatibility

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

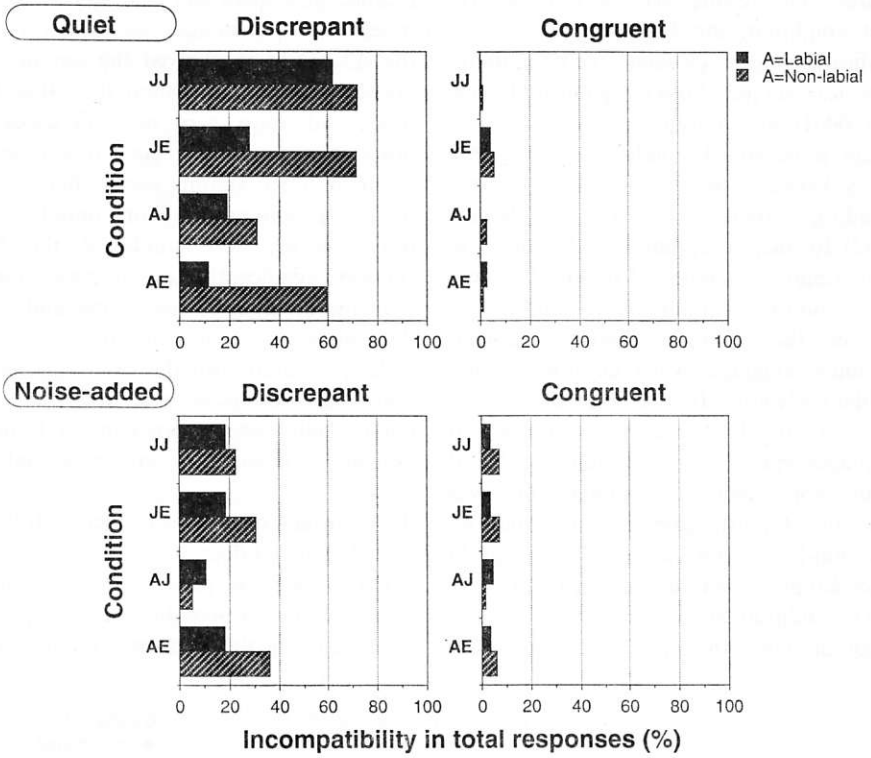

Fig. 3 The frequency of reported incompatibility (%) in total responses in each condition.

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

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

Subsequently, the sources of these interactions were examined by pairwise comparisons using least

squares means. The results were as follows. (1) In she quiet condition, the Japanese subjects reported significantly more frequent incompatibility than the American subjects both for auditory labials ($t=4.45$, $p<.0001$) and non-labials ($t=3.54$, $p<.001$). In the noise-added condition, there were no differences between the two language groups either for auditory labials ($t=0.87$) or non-labials ($t=1.16$). (2) In the quiet condition, more frequent incompatibility was reported for the Japanese stimuli when auditory stimuli were labials ($t=3.11$, $p<.01$), but there were no differences between the two stimulus languages when auditory stimuli were non-labial ($t=1.86$). In the noise-added condition, there were no differences between the two simulus languages when auditory stimuli were labial ($t=0.69$), but more frequent incompatibility was reported for the English stimuli when auditory stimuli were non-labial ($t=3.62$, $p<.001$). (3) In the quiet condition, when auditory stimuli were labial, the JJ condition showed significantly more frequent incompatibility than the other three conditions, and there were no differences among the other three. When auditory stimuli were non-labial, the AJ condition showed the significantly less frequent incompatibility than the other three conditions, and there were no differences among the other three. (4) In the noise-added condition, when auditory stimuli were labial, there were no differences among the four conditions. When auditory stimuli were non-labial, the AJ condition showed significantly less frequent incompatibility than the other three conditions, and there were no differences among the other three.

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

3.5 Correlation between Incompatibility and the McGurk Effect

Comparing the panels for discrepant stimuli in Fig. 3 with corresponding panels in Fig. 2, it was suggested that the frequency of the incompatibility

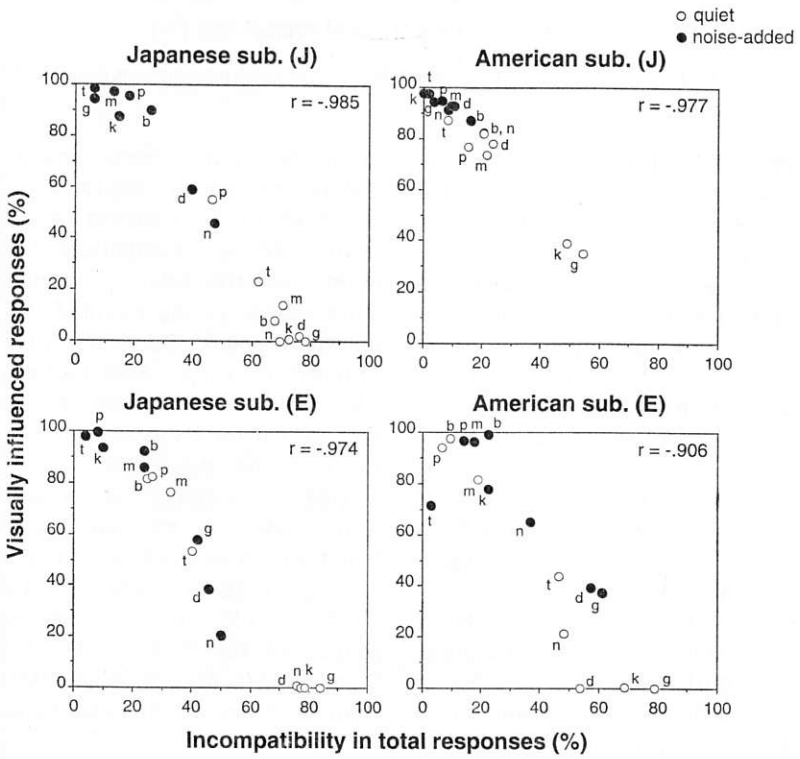

Fig. 4 The relationship between the frequency of incompatibility (%) and the magnitude of the visual influence approximated by the frequency of place errors (%).

is negatively correlated with the magnitude of visual influence. Figure 4 plots the relationship between the two indexes for each of the eight auditory syllables in the quiet and noise-added conditions. As shown in Fig. 4, the 16 data points in each condition fitted to a negative linear function with a high correlation coefficient ranging from $r=-.906$ to $r=-.985$. When the correlation was calculated based on individual data, it was $r=-.809$ (JJ), $r=-.816$ (AJ), $r=-.820$ (JE), and $r=-.751$ (AE). The results clearly show thas the more frequent incompatibility there is, the smaller the magnitude of visual influence becomes.

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

In the JE and AE conditions, the contrast is not clear. These two groups showed the large McGurk effect only on auditory /ba/, /pa/, /ma/, and /ta/ in the quiet condition. The data points for these four syllables intrude into the upper left region where the data points for the noise-added condition exist. In these two groups, the distance between auditory labials (the strong visual influence and the infrequent incompatibility) and non-labials (the weaker visual influence and the more frequent incompatibility) was larger than the distance between the noise-added and quiet conditions. In this sense, the responses in the JE and AE conditions were similar, suggesting a stimulus characteristic of the English stimuli. It is suggeted that English auditory labials paired with visual non-labials have some characteristics that are prone to visual effects.

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

3.6 Sensitivity to the Auditory-Visual Discrepancy

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

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

Based on this explanation, the auditory responses were examined further as an index of the sensitivity to the auditory-visual discrepancy. As shown in the left panel of Fig. 5, the frequency of incompatibility in the auditory responses was apparently higher in the Japanese subjects than in the Americans. This indicates that the Japanese subjects were more sensitive to the discrepancy than the Americans. It is striking that the results were almost the same irrespective of the stimulus language, when the subjects' native language was the same. In the JJ and

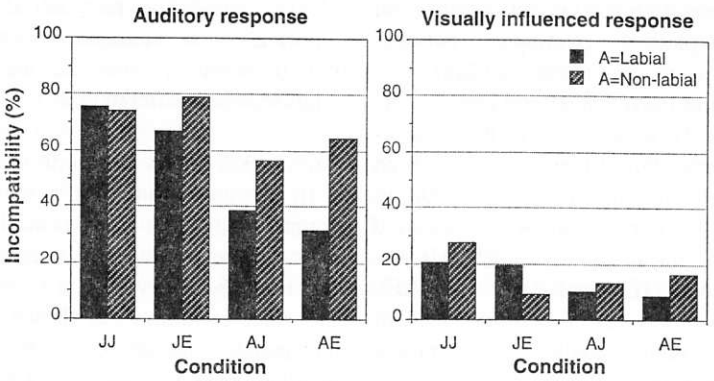

Fig. 5 The frequency of incompatibility (%) in the auditory responses and the visually influenced responses (fused and combined responses) for discrepant auditory-visual stimuli in the AV condition.

JE conditions, the Japanese subjects detected the discrepancy about 70% of the time for both auditory labials and non-labials. In the AJ and AE conditions, the American subjects reported less incompatibility especially for auditory labials. In these cases, the frequency of incompatibility did not reach 50% (39% in the AJ condition, and 31% in the AE condition) even though the responses were auditory ones. These low frequencies of incompatibility in the American subjects are notable, because studies in English speaking cultures have been reported that the McGurk effect is stable for this type of stimuli (*i.e.*, auditory labials paired with visual non-labials). This coincidence suggest that this type of stimuli has a nature which encourages native speakers of English to integrate auditory and visual information. On the whole, judging from the striking resemblance of the data between the JJ and JE conditions and between the AJ and AE conditions, it seems that this index can assess each language group's sensitivity to the auditory-visual discrepancy.

A $2\times2\times2$ ANOVA (Listener's Language × Stimulus Language × Auditory Stimulus) was carried out on the data for the auditory responses. (It should be noted that for auditory labials, there were missing observations for several subjects who showed no auditory responses. The number of such subjects were two for the AJ condition, four for the JE condition, and four for the AE condition.) The ANOVA found main effects of Listener's Language [$F(1, 32)=8.39$, $p<.01$] and Auditory Stimulus [$F(1, 32)=5.44$, $p<.05$] significant. These main effects show that (a) Japanese subjects were more sensitive to the auditory-visual discrepancy than American subjects and that (b) more frequent incompatibility was reported for auditory non-labials than for auditory labials. The main effect of Stimulus Language and the interactions were not significant.

Although the above ANOVA failed to find significance for the interaction of Listener's Language × Auditory Stimulus, this may be due to the data exclusion in the ANOVA due to missing observations (missing observations occurred when all of a subject's responses were visually influenced.) For, the above ANOVA excluded all the data of subjects who had a missing observation, even though the missing observations were found only for auditory labials. Therefore, separate analyses were done for auditory labials and non-labials by two-factor ANOVAs (Listener's Language × Stimulus Language). When auditory stimuli were labial, the main effect of Listener's Language was significant [$F(1, 32)=9.02$, $p<.01$], but the main effect of Stimulus Language and the interaction was non-significant [$F(1, 32)=0.43$; $F(1, 32)=0.00$, respectively]. When auditory stimuli were non-labial, none of the effects were significant [$F(1, 42)=3.62$, $p<.10$, for Listener's Language; $F(1, 42)=0.64$, for Stimulus Language; $F(1, 42)=0.04$, for the interaction]. These results imply that the difference in sensitivity between the two language groups was reliable only when auditory stimuli were labial. That is, the American subjects were less sensitive to the auditory-visual discrepancy than the Japanese

subjects only when auditory stimuli were labial.

4. DISCUSSION

In the present study, the results of the JJ condition replicated those of the earlier experiment by Sekiyama & Tohkura.[17] That is, when no noise was added, the Japanese subjects listening to Japanese speech were less influenced by discrepant visual information than the subjects in the other three conditions. The contrast with the other three conditions was sharp when the auditory stimuli were labial (/b, p, m/).

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

Although a correlation does not necessarily determine causality, additional evidence suggests that incompatibility is a cause of the small McGurk effect. For, as we saw in the frequency of incompatibility in *auditory responses*, the data suggested that the Japanese subjects were more sensitive to the auditory-visual discrepancy. To discriminate two indexes, let the incompatibility in the auditory responses be IA, and let the incompatibility in the total responses be IT. Note that the auditory responses imply the absence of the McGurk effect, whereas the total responses include both the absence and the presence of the McGurk effect. Hence, the frequency of IT can be higher for the Japanese subjects even if incompatibility is a consequence of the absence of the McGurk effect, rather than a cause of it, since the absence of the McGurk effect was more frequent for the Japanese subjects.

As for IA, however, incompatibility should be equally frequent for the two language groups if incompatibility is a mere consequence of the absence of the McGusk effect, since this index deals only with responses where the McGurk effect was absent, thus there is no reason to expect higher frequencies in the Japanese subjects. On the contrary, in IA the results showed a clear difference between the two language groups. Accordingly, the frequency of IA should be considered an indication of a factor that influences the magnitude of the McGurk effect. In the last section, we called it "sensitivity to the auditory-visual discrepancy." It is plausible that the higher sensitivity to the discrepancy interrupts the integration of the two sources of information. It should also be noted that the difference found between the two language groups in sensitivity was reliable only for auditory labials. This corresponds with the fact that the difference in the magnitude of the McGurk effect between the JJ and AE conditions was significant only for auditory labials.

In a study on the McGurk effect in English, Manuel, Repp, Liberman, & Studdert-Kennedy[22] also reported a similar relationship between the magnitude of the McGurk effect and the subjects' incompatibility ratings. It is misleading that this study is sometimes cited as showing that the McGurk effect can occur even when subjects are aware of acoustic-optic discrepancy (Summerfield[9]). Actually, their data show that the subjects were keenly aware of the discrepancy only when the McGurk effect was absent (*e.g.*, auditory /da/ paired with visual /ba/). For stimuli which produced a strong McGurk effect (*e.g.*, auditory /ba/ paired with visual /da/), their subjects showed only moderate incompatibility ratings.

As described in the earlier report (Sekiyama & Tohkura[20]), we raise two hypothetical explanations of the Japanese subjects' perceptual processing to separate two modalities of information. One is less frequent face-to-face contact of Japanese listeners and the other is a simpler structure of the Japanese phonological system. Growing up in such a cultural/linguistic environment, Japanese listeners may be encouraged to form perceptural processing in which visual speech information is not integrated with auditory information unless visual compensation is necessary. Such processing would easily detect the

auditory-visual discrepancy.

In contrast to the present results, Massaro, Tsuzaki, Cohen, Gesi, and Heredia[23] tested native speakers of Japanese, American English, and Spanish and drew the conclusion that there was no difference among language groups in the manner of auditory-visual integration. However, it shoud be noted that there are many differences between their study and the present study. The biggest difference would be the nature of the stimuli: Their auditory stimuli were a continuum of sounds covering the range from /ba/ to /da/ that were synthesized from natural English speech by altering the parametric information. Although they used artificial speech to equate the stimuli across language groups, this will cause a problem. First, artificial speech would resemble foreign speech because of its acoustical deviation from natural speech in the subjects' native language. Secondly, these stimuli might have selectively strengthened the visual influence in Japanese subjects because the stimuli were generated from English speech. If so, their results do not disagree with the present results that Japanese subjects used visual information as much as American subjects for English stimuli.

Nonetheless, it should be pointed out that the results by Massaro *et al.*[23] also showed that Japanese subjects tended to be less subject to visual effects than the other two groups, though this tendency was statistically marginal. It seems that their use of ambiguous speech favored finding the group difference statistically nonsignificant. On the continuum of sounds, the sound at one end was most /ba/-like, at the other end most /da/-like, and at the mid point it was ambiguous (halfway between /ba/ and /da/). I believe that their results suggest a group difference in visual effect at the two end points but no group difference at the mid point. Supporting this interpretation, they found a significant interaction between group and auditory level.

Various analyses in the present study showed differences between labials and non-labials. First, the McGurk effect was stronger for auditory labials than for auditory non-labials in both the quiet and noise-added conditions. Secondly, incompatibility was reported less frequently for auditory labials than for auditory non-labials in both the quiet and noise-added conditions. Thirdly, the subjects were more sensitive to the auditory-visual discrepancy for auditory non-labials paired with visual labials than conversely paired stimuli. One possible explanation for this labial vs. non-labial asymmetry is the salience of visual cues. Visual labials (/b, p, m/) that include lip closure are considered as more salient than visual non-labials. Thus, the salient visual labials will more often let the subjects detect an auditory-visual discrepancy than visual non-labials. If so, auditory non-labials paired with visual labials will more often yield incompatibility than conversely paired stimuli. Because of the more frequent incompatibility, visual labials will more often interrupt integration of incongruent auditory-visual information than will visual non-labials. In that case, auditory non-labials paired with visual labials will produce a weaker McGurk effect than conversely paired stimuli.

Another possible explanation is the number of consonants in a phoneme inventory. Both the English and the Japanese language have many more non-labials than labials. It is plausible that the subjects make a probabilistic decision based on their knowledge that non-labial sounds have a higher probability of occurrence than labials in their native language. Then, when the information contains ambiguity, responses may be biased to non-labials which have a higher probability of occurrence.

The results of the JJ condition also replicated the results by Sekiyama and Tohkura[17] in terms of the intelligibility hypothesis that the McGurk effect hardly occurs for completely intelligible auditory stimuli. On the other hand, as we showed in the earlier study (Sekiyama & Tohkura[20]), the other three conditions did not support the intelligibility hypothesis. In the AJ, AE, and JE conditions, strong McGurk effects occurred even for the auditory tokens with an intelligibility score of 100%. Thus, the auditory intelligibility hypothesis is applicable only to Japanese subjects observing their native speech.

However, the overall results suggest that the quality of speech influences the magnitude of visual effects. It should be stressed that there is a conceptual difference between the *intelligibility* and the *quality of speech*. The quality of speech may be defined as the proximity of given speech information to the ideal value of the speech. The fact that the foreign speech stimuli produced a larger visual effect than the native stimuli suggests that the quality of speech in the foreign language was different from that in the native speech though

there were no substantial differences in the intelligibility score. It should also be noted that in the noise-added condition, a small decrease of the intelligibility score sometimes increased the magnitude of visual influence a great deal (*e.g.*, in the JJ condition, the auditory intelligibility score of /ba/ decreased only 2% due to the noise, but the magnitude of visual influence increased more than 80%). This suggests that even though the decrease of the intelligibility score is small, the noise can cause a substantial degradation in the quality of speech, forcing the subjects to depend on the visual information. It would be desirable to find an index that can describe the quality of speech.

The present analyses of the incompatibility data clearly describes that Japanese subjects in a quiet JJ condition must pay attention to the visual information in an early stage of the processing because they report it as being incompatible with the auditory information. This implies that they must discard the discrepant visual information in a later stage of the processing. Also suggested was that the small McGurk effect in the JJ condition is a consequence of their higher sensitivity to auditory-visual discrepancy, which interrupts the integration of the two sources of information. This means that each modality of information is processed at the time of integration to the extent that mismatching can be detected. Therefore, integration is considered a relatively later stage of processing, at least for Japanese subjects perceiving Japanese speech. Further research needs to be carried out to incorporate this inter-language difference in the magnitude of the McGurk effect into a model of auditory-visual speech perception.

ACKNOWLEDGMENTS

The author is grateful to Prof. Harry Levitt at the Graduate Center of the City University of New York whose laboratory was used for experiments with American subjects. She also thanks Dr. Yoh'ichi Tohkura for his full support in creating the stimuli at ATR Auditory and Visual Perception Research Laboratories, and Dr. Hwei-Bing Lin at CUNY for her help with arrangements for the experiments. The experiments at CUNY were supported by a grant from Nissan Science Foundation to the author. The writing of this paper was supported by Human Frontier Science Program fellowship LT-277/93.

REFERENCES

1) P. J. Blamey, R. C. Dowell, A. M. Brown, G. M. Clark, and P. M. Seligman, "Vowel and consonant recognition of cochlear implant patients using formant-estimating speech processors," J. Acoust. Soc. Am. **82**, 48–57 (1987).
2) Y. Fukuda, M. Shiroma, and S. Funasaka, "Speech perception ability of the patients with artificial inner ear," IEICE Tech. Rep. SP88-91 (1988) (in Japanese with English abstract).
3) S. Funasaka, M. Shiroma, K. Yukawa, N. Iizuka, M. Yao, J. Kono, H. Takahashi, and K. Kumakawa, "Consonant information transmitted through 22-channel cochlear implant," Audiol. Jpn. **32**, 146–151 (1989) (in Japanese).
4) N. Erber, "Auditory-visual perception of speech," J. Speech Hear. Dis. **40**, 481–492 (1975).
5) J. J. O'Neil, "Contribution of the visual components of oral symbols to speech comprehension," J. Speech Hear. Dis. **19**, 429–439 (1954).
6) W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215 (1954).
7) C. A. Binie, A. A. Montgomery, and P. L. Jackson, "Auditory and visual contributions to the perception of consonants," J. Speech Hear. Res. **17**, 619–630 (1974).
8) H. McGurk and J. MacDonald, "Hearing lips and seeing voices," Nature **264**, 746–748 (1976).
9) Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, R. Campbell and B. Dodd, Eds. (Lawrence Erlbaum, London, 1987), pp. 3–51.
10) J. MacDonald and H. McGurk, "Visual influences on speech perception processes," Percept. Psychophys. **24**, 253–257 (1978).
11) D. J. Dekle, C. A. Fowler, and M. G. Funnell, "Audiovisual integration in perception of real words," Percept. Psychophys. **51**, 355–362 (1992).
12) B. Dodd, "The role of vision in the perception of speech," Perception **6**, 31–40 (1977).
13) D. W. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Lawrence Erlbaum, New Jersey, 1987).
14) D. W. Massaro and M. M. Cohen, "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol., Hum. Percept. Perform. **9**, 753–771 (1983).
15) K. P. Green, P. K. Kuhl, A. N. Meltzoff, and E. B. Stevens, "Integrating speech information across speakers, gender, and sensory modality: Female faces and male voices in the McGurk effect," Percept. Psychophys. **50**, 524-536 (1991).
16) K. P. Green and P. K. Kuhl, "Integral processing of visual place and auditory voicing information during phonetic perception," J. Exp. Psychol.,

Hum. Percept. Perform. **17**, 278–288 (1991).
17) K. Sekiyama and Y. Tohkura, "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," J. Acoust. Soc. Am. **90**, 1797–1805 (1991).
18) K. Sekiyama, K. Joe, and M. Umeda, "Perceptual components of Japanese syllables in lipreading: a multidimensional study," IEICE Tech. Rep. IE87-127 (1988) (in Japanese with English abstract).
19) B. E. Walden, R. A. Prosek, A. A. Montgomery, C. K. Scherr, and C. J. Jones, "Effects of training on the visual recognition of consonants," J. Speech Hear. Res. **20**, 130–145 (1977).
20) K. Sekiyama and Y. Tohkura, "Inter-language differences in the influence of visual cues in speech perception," J. Phonet. **21**, 427–444 (1993).
21) K. Sekiyama, "McGurk effect and incompatibility: A cross-language study on auditory-visual speech perception," Stud. Essays Behav. Sci. Philos., Kanazawa Univ. **14**, 29–62 (1994).
22) S. Y. Manuel, B. H. Repp, A. M. Liberman, and M. Studdert-Kennedy, "Exploring the 'McGurk effect'," Paper presented at the 24th Annu. Meet. Psychonomic Society, San Diego, California (Nov. 1983).
23) D. W. Massaro, M. Tsuzaki, M. M. Cohen, A. Gesi, and R. Heredia, "Bimodal speech perception: An examination across languages," J. Phonet. **21**, 445–478 (1993).