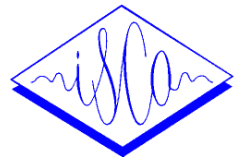


Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

# NATIVE-FOREIGN LANGAGE EFFECT IN THE McGURK EFFECT : A TEST WITH CHINESE AND JAPANESE

*Yasuko Hayashi and Kaoru Sekiyama*
Kanazawa University

## ABSTRACT

This study examined the manners of audiovisual speech perception, using the "McGurk effect", when the speakers were foreigners. The McGurk effect demonstrates that visual (lip movement) information is used during speech perception even when it is discrepant with auditory information. Subjects, 17 Chinese and 23 Japanese reported what they heard while looking at and listening to the speakers' face on the monitor. There were 4 speakers, 2 Chinese and 2 Japanese. All the stimuli were made of one syllable utterance. Half of them were audio-visually compatible stimuli, but half of them were audio-visually incompatible stimuli. The results indicate that the Japanese subjects used more visual information on speech perception when the speakers were foreigners. But the Chinese subjects did not show such language asymmetry.

## 1. INTRODUCTION

The phenomenon known as the McGurk effect demonstrates the influence of vision on audiovisual speech perception [1], [2].

The McGurk effect may occur when an auditory stimulus is presented simultaneously with a visual stimulus whose place of articulation is incompatible with that of the auditory stimulus.

Although in English-speaking cultures, this effect has been shown to be robust [3]-[6], Sekiyama and Tohkura found that native speaker of Japanese showed a much weaker McGurk effect[7].

Comparing native speaker of Japanese with native speakers of American English, a "native-forign language effect" has been suggested. The Japanese subjects show a larger McGurk effect for English stimuli than for Japanese stimuli, and the American subjects show a larger McGurk effect for Japanese stimuli than for English stimuli [8]. It seems that the McGurk effect is stronger for foreign speech stimuli than for native speech stimuli.

To confirm the hypothesis of this native-foreign language effect, we tested native speakers of Chinese and native speakers of Japanese, using Chinese stimuli and Japanese stimuli.

## 2. METHOD

### 2.1. Subjects

Two groups (Chinese group and Japanese group) of subjects under age 30 participated. All subjects had normal hearing and normal (or corrected to normal) vision. Subjects were recruited from Kanazawa University community. Most of the subjects were graduate students of Kanazawa university.

Chinese Group consisted of 17 native speakers of Chinese (11 Male) with a mean age of 27.3, and a range from 23 to 30. Most of them had arrived in Japan after finishing college in China. The mean length of their stay in Japan was 7.6 months, and range from 0 to 23 months. They had never lived in a foreign country except Japan. Although the subjects' native languages included various Chinese dialects, all of them had been educated in Mandarin Chinese since entering elementary school.

There were large differences aomng the subjects' abil-

ity in Japanese. Some subjects could speak Japanese fluently but some of the subjects could understand only few Japanese words.

Five other Chinese participated in this study, but their data were not considered usable because of not looking at the monitor at all (four), or many of his responses in Auditory only condition were unconvincing (one).

Japanese Group consisted of 23 native Speakers of Japanese (11 male) with a mean age of 25.8, and a range from 24 to 30. They had never lived in a foreign country. None of them could speak Chinese. another Japanese participated in this study, but his data were not considered usable because many of his responses in Auditory only condition were unconvincing.

## 2.2. Stimuli

The stimulus materials were eight syllables ([ba], [pa], [ma], [da], [ta], [na], [ga], and [ka]) pronounced by two Japanese speakers (J1, J2) and two Chinese speakers (C1, C2). The speaker C1 was from Beijing and the speaker C2 was from Shanghai. Chinese speakers were asked to pronounce the syllables clearly in high/level tone in Mandarin pronunciation. Japanese speakers, who were professional announcers, were asked to pronounce in Standard Japanese pronunciation.

All the speakers were instructed to close their lips before starting to pronounce. Each speaker's face was recorded on to the videotape through a BETACAM video camera while she pronounced the syllables. Their utterances were rerecorded by a DAT (Digital Audio Tape-corder) in an anechoic room to obtain the auditory stimuli.

Then we normalized the power of auditory stimuli into RMS (Root Mean Square) = 2500. After that, these syllables were dubbed onto the frames where the original speech had been. The onsets of the energy were synchronized.

In dubbing, the auditory syllables were combined only with the visual syllables from the same speaker, but an auditory syllable was dubbed onto the congruent visual syllable and the incongruent visual syllable.

In incongruent version, pairs of audio syllables and visual syllables are as follows; audio[ba]-visual(ga), audio[pa] -visual(ka), audio[ma]-visual(na), audio[da]-visual(ba), audio[ta]-visual(pa), audio[na]-visual(ma), audio[ga]-visual(ba), and audio[ka] -visual(pa).

For presentation, the audiovisual stimuli were copied onto a laser disk. On the copying process, audio stimuli were copied onto one of the two audio channels and signals to attract attention were dubbed onto the other audio channel.

A tone(800 ms, 800 Hz, RMS = 2000) was presented two seconds before the visual stimulus appeared. One second before the audio stimulus, two tones(33 ms, 800 Hz, RMS = 400) were presented at 150 ms intervals (Figure 1).

The audio was set at comfortable listening level (about 70dB-A visual peak reading with fast scale from Brüel and Kjær Type 2203 sound level meter).

Visual stimuli were presented on a 14 inch color monitor. Auditory stimuli were presented through a loudspeakers placed on the monitor. The subjects viewed the monitor from a distance of 50 cm.

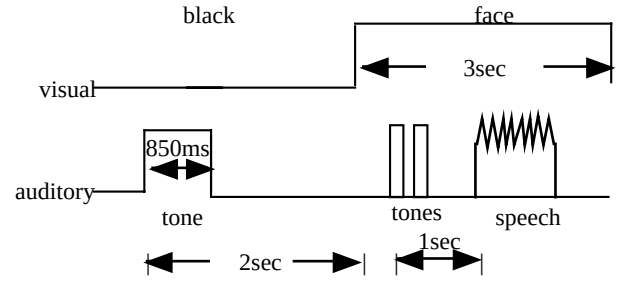



**Figure 1:** Time course of audiovisual stimuli

### 2.3. Experimental Design

All of the subjects participated in the two sessions of the experiment. One session was for the Chinese stimulus set: the other was for the Japanese set. The order of these two stimulus sets was counterbalanced between subjects in each group. For each stimulus set, there were three conditions: audiovisual(AV), auditory only(A), and visual only(V), which were conducted in that order.

### 2.4. Procedure

The laser disk was played on a laser video disk recorder (SONY LVR-300AN) located in a control room next to the soundproof room in which subjects were tested.

The stimuli were presented once every seven seconds in random order. In the AV condition, the subjects were instructed to write down what they thought they had heard while looking at and listening to each syllable. It was an open set response.

Chinese subjects were instructed to write in pin-yin, which they had been taught in elementary school for spelling Chinese syllables to approximate the Roman alphabet. Japanese subjects were instructed to write in Romaji (Roman alphabet). They were also asked to rank their confidence into three levels for each stimulus and write them in a last column on their response sheets.

In the A condition, the subjects' tasks were to report only what they had heard. In the V condition, they were asked to read lips and report what they thought the speaker was pronouncing.

For each stimulus set, the AV condition was conducted in three blocks of 32 trials, and the A and V condition were conducted in three blocks of 16 trials. It took about 30 minutes to conduct the three conditions for one stimulus set.

The experimenter was a native speaker of Japanese and she instructed both language groups in Japanese. Further, Chinese subjects could read written instruction , in which all the procedure was translated into Chinese. So even the subject who didn't understand Japanese at all could also perform the task well.

A video camera was positioned besides the monitor, allowing the experimenter to observe the subjects' eyes and their behavior from the control room.

## 3. RESULTS

### 3.1. Responses in the A condition

Most of the responses were the syllables which were presented. More than 90% of the responses were correct in most of the stimuli. In terms of the percentage of correct answers, there was no difference between the four speakers.

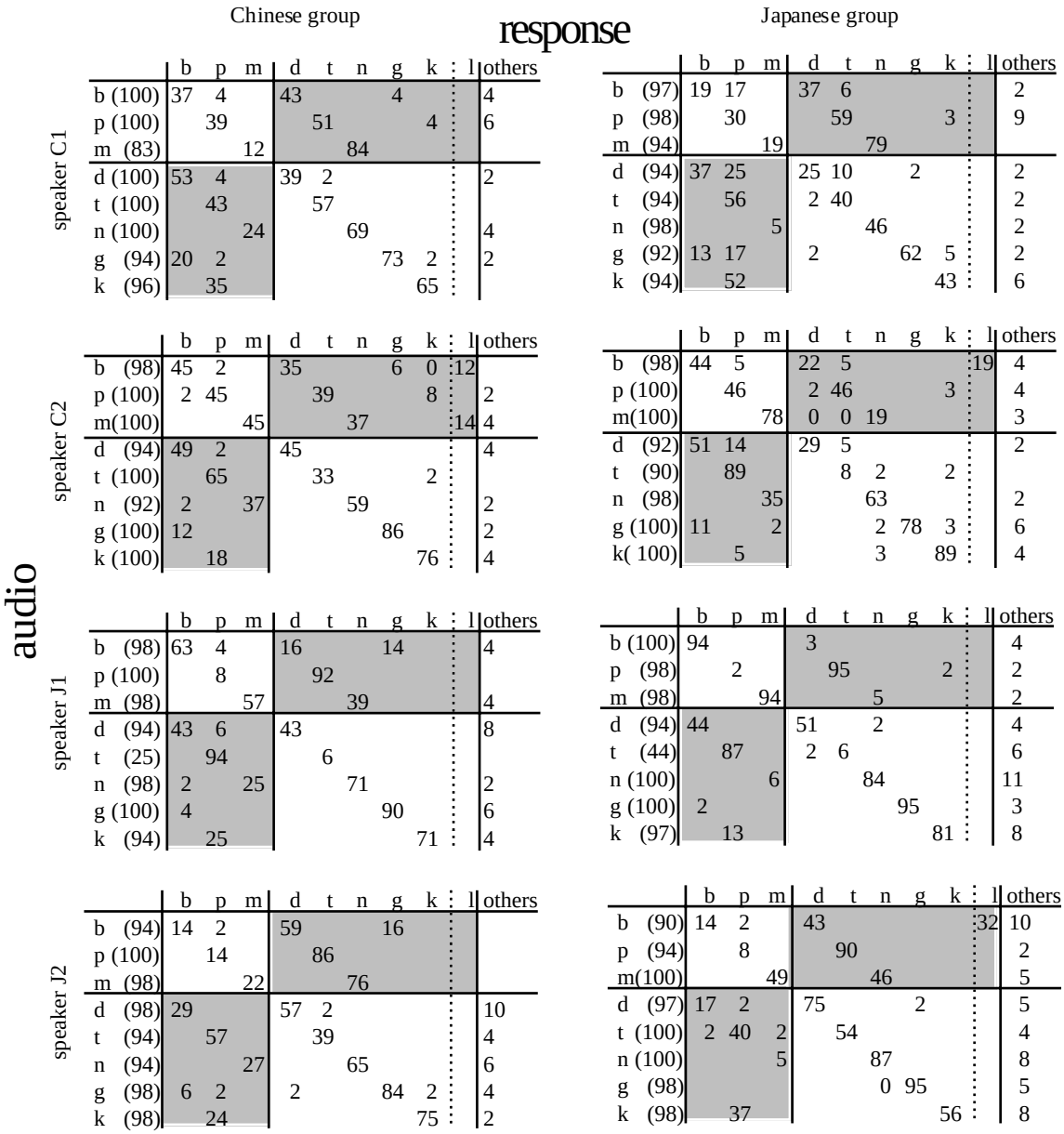
### 3.2. Responses in the V condition

We evaluated the responses in terms of the place of articulation of audio stimuli. So, when the stimulus was [ba], [pa], or[ma], the correct answer would be [ba], [pa], [ma], or other labials. When the stimulus was [da], [ta], [na], [ga], or [ka], the correct answer would be [da], [ta], [na], [ga], [ka], or other nonlabials. More than 90% of the responses were correct in most of the stimuli.

### 3.3. Responses in the AV condition

Table 1 shows the data for the incongruent stimuli in the AV condition. In the confusion matrices in Table 1, the numbers in parentheses in the leftmost column are the percent correct identifications of the auditory syllable in the A condition which provides a baseline for the examination of visual effects in the AV condition. The responses in the shadowed sections indicate the "gross" McGurk effect. These are errors in terms of audition, and their place of articulation (labial of

**Table 1:** Confusion matrices in the audiovisual condition (incongruent stimuli)




nonlabial) is consistent with that of the visual input.

### 3.4. The Magnitude of McGurk effect

The magnitude of the "pure" McGurk effect was calculated by subtracting the auditory place errors from the gross McGurk effect. For example, when combined with visual(na), speaker J2's auditory[ma] produced place errors ("na" responses) 76% of the time in Chinese group. These are counted as the gross McGurk effect. However, this [ma] also produced "na" responses 2% of the time in the A condition. Thus, the magnitude of the pure McGurk effect was 74% (= 76% - 2%).

Figure 2 shows a comparison of the average magnitude of the McGurk effect. To compare the magnitude of the McGurk effect across the language groups, the average magnitudes were calculated. When the stimuli were Chinese, the average magnitudes were 44.87% for the Chinese group and 44.03% for the Japanese group. When the stimuli were Japanese, they were

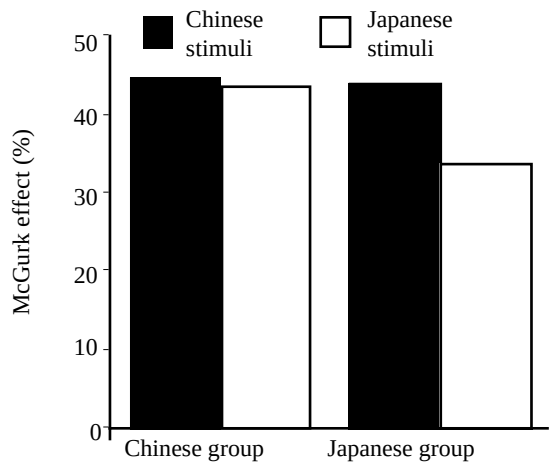



**Figure 2:** A comparison of the mean magnitude of the McGurk effect between Chinese group and Japanese group.

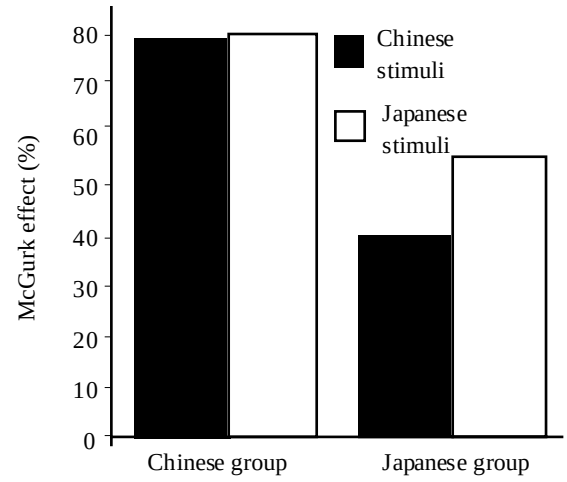



**Figure 3:** A comparison of the confidence for responses between two language groups.

43.67% for the Chinese group and 33.63% for the Japanese group.

There were no difference in the magnitudes of the McGurk effect between the Chinese group and the Japanese group. The magnitude of the McGurk effect was weaker in Japanese group when the stimuli were Japanese ($F[1,36] = 13.93$, $p<.01$). There were no significant effects of native-foreign language in the Chinese group.

### 3.5. Confidence for responses

To compare the degree of confidence for the responses across language groups, the average percent of confidence in AV condition were calculated for each subjects. When the stimuli were Chinese, the average percent of confidence were 78.82% for the Chinese group and 39.69% for the Japanese group. When the stimuli were Japanese, they were 79.79% for the Chinese group and 55.25% for the Japanese group.

The Chinese subjects had stronger confidence for the responses than the Japanese subjects ($F[1,36] = 30.44$, $p<.01$). In the Chinese group, there were no difference between Chinese stimuli and Japanese stimuli. In the Japanese group, subjects had stronger confidence when the stimuli were Japanese ($F[1,20] = 28.94$, $p<.01$). There were no correlation between the detree of confidence and the magnitude of McGurk effect.

## 4.DISCUSSION

There were no difference in the magnitude of McGurk effect between the Chinese group and the Japanese group. The result is agreeing with the result by Sekiyama that Chinese subjects , similar to Japanese, showed weaker McGurk effect than Americans[9].

The Japanese subjects showed stronger McGurk effect when the stimuli were Chinese, consistent with the native-foreign language effect hypothesis. On the other hand, the Chinese subjects didn't show such a difference between Chinese stimuli and Japanese stimuli, against the native-foreign language effect hypothesis.

As Chinese is a tone language, the meaning of a spoken word is determined not only by its syllabic structure but also by its tone. Sekiyama indicated that this language characteristic might foster in the Chinese a strong reliance on auditory information and showed less McGurk effect[9]. The attitude of some Chinese

subjects, during the experiment, supports the interpretation, further. There were four subjects, out of 21, who didn't watch the monitor at all (their data were not included in this study). But there wasn't such a subject in the Japanese group.

If visual information had little influence to Chinese subjects, that may explain the lack of native-foreign language effect in Chinese. But, the result of the magnitude of the McGurk effect, which was same degree as (or somewhat much degree than) Japanese, is not suitable for this interpritation.

Other possible interpritation that Chinese didn't show the difference between native-foreign language are (1) The Chinese subjects had been exposed to nonnative language. Moreover, Mandarin Chinese is not native language for some of the subjects. (2) The stimuli used in this study were real words in Chinese (meaningless in Japanese). The stronger confidence in the Chinese subjects, without correlation between the magunitude of the McGurk effect, might be the reflection of the effect of meaningfulness.

## REFERENCES


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

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

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

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

5. Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **9**, 753-771.

6. Rosenblum, L. D., & Saldaña, H. M. (1992). Discrepant tests of visually influenced syllables. *Perception & Psychophysics*, **52**, 461-473.

7. Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, **90**, 1979-1805.

8. Sekiyama, K. (1994) McGurk effect and incompatibility: A Cross-Language study on Auditory-Visual Speech Perception. *Studies and Essays in Behavioral Sciences & Philosophy (Kanazawa University)*, **14**, 29-62.

9. Sekiyama, K. (1997) Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, **59**, 73-80.