A FEW FACTORS WHICH AFFECT THE DEGREE OF INCORPORATING LIP-READ INFORMATION INTO SPEECH PERCEPTION

Kaoru Sekiyama, Yoh'ichi Tohkura**, and Michio Umeda****

*Kanazawa University
**ATR Human Information Processing Laboratories
***Osaka Electro-communication University

ABSTRACT

This paper describes how people incorporate visual lip-read information into speech perception, depending on one's native language/culture and experience of learning a second language. Studies on lipreading show that humans can easily make a distinction between labial consonants and nonlabial ones. Then we investigated how people integrate auditory and visual speech information, by using the "McGurk effect" paradigm in which labial-nonlabial conflict is introduced. Cross-language examinations were done across Japanese, American English, and Chinese. The Japanese and Chinese subjects were less susceptible to the McGurk effect than the American subjects, indicating a cultural/linguistic factor. The results for the Chinese subjects showed a correlation between the magnitude of the McGurk effect and the length of time they lived in a foreign country (Japan), suggesting a change due to second language learning.

1. LIP-READ INFORMATION

1.1. Introduction

In face-to-face communication, speech perception is a multimodal process. Humans can read lips more or less well, and this lip-read information plays a role during speech perception. It is well known that looking at a speaker's face improves speech perception when speech is not clear. This is because face provides us with speech information. In this section, we will describe the nature of lip-read information, showing results of a lipreading experiment.

1.2. Perceptual Features in Lipreading

At the beginning of our series of experiments, we tested how precisely Japanese ordinary adults can lipread [8]. We presented video of a face of speakers speaking Japanese syllables. The video showed 100 mono-syllables which are all that can happen in the Japanese language. Japanesesyllables are CV-syllables in principle, but some syllables have a semi-vowel /y/ between a consonant and a vowel. Thus, in the 100 possible mono-syllables, a consonant or a consonant plus /y/ is followed by one of the five vowels (/a/, /e/, /i/, /u/, /o/). The silent movie of the speakers' articulation of the 100 syllables was shown to 60 untrained adult subjects and they were asked to lipread each visual syllable. Thanks to the fact that there are only five vowels in the Japanese language, the subjects were able to lipread vowels accurately about 90% of the time. However, it was difficult to lipread consonants correctly (accuracy was 20%).

Figure 1 shows the perceptual space of consonants during lipreading. This indicates how well people distinguish various consonants. To get this three-dimensional map, the subjects' performances were analyzed by multidimensional scaling. In this figure, distance between any two consonants represents perceptual dissimilarity between the two. The results show that the Japanese can categorize the visual consonants into five or six groups. The most isolated consonant is /w/, which is the only consonant that has lip protrusion. The second distinct group includes /p/, /b/, and /m/, which have bi-labial lip closure. The third distinct group, /py/, /by/, and /my/, consist of combinations of a bilabial

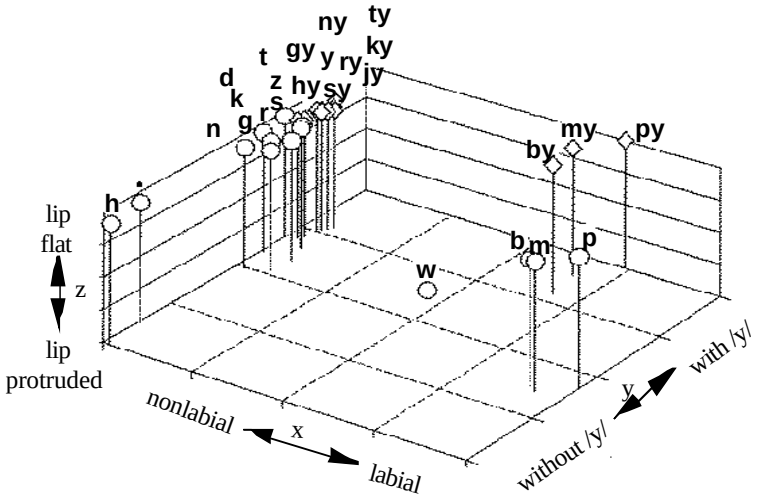

Figure 1: Perceptual space of Japanese consonants in lipreading.

consonant and a semi-vowel /y/. The fourth group includes consonant /h/ and null consonant (in the case of vowel-only presentation). Finally, there is a densely distributed area along the y-axis. They are consonants articulated behind the lips, thus, referred to as nonlabials. This configuration of the stimuli tells that the subjects extracted information about labial-nonlabial distinction (x-axis), existence of the semivowel /y/ (y-axis), and the lip protrusion (z-axis). To summarize, the subjects could not lipread perfectly as in spy movies, but they could easily make a distinction between labials and nonlabials. If we score the Japanese lipreading performances in terms of labial-nonlabial discrimination, the accuracy was 92%.

Similar results were obtained for English by Walden and others [11]. When they tested hearing impaired adults before training, there were about six categories (visemes) of consonants, and according to my calculation, their score of labial-nonlabial discrimination was 88%. Thus, in both English and Japanese, untrained people can tell labials from nonlabials most of the time in lipreading.

2. MCGURK EFFECT IN JAPANESE AND AMERICAN PERCEIVERS

2.1. The McGurk Effect

To demonstrate the role of lip-read information in speech perception, the McGurk effect paradigm is useful. The McGurk effect is a biasing effect of incompatible lip-read cues in speech perception when auditory and visual speech conflict concerning labial-nonlabial distinction. When McGurk and MacDonald [4] first discovered this phenomenon, they combined /ga/ lip movements and /ba/ sound on a film, and their subjects reported hearing /da/ 98% of the time. This /da/ response is a reasonable solution, because, as we saw in the lipreading data, visual /ga/ is similar to visual /da/, and at the same time, auditory /da/ is similar to auditory /ba/. The McGurk effect demonstrates that lip-read information is incorporated into speech perception even when lip movements are incompatible with auditory speech.

2.2. A Weak McGurk Effect in the Japanese

While the McGurk effect has been shown to be stable under various conditions in English speaking cultures [1, 2, 5], we have found inter-language differences between native speakers of Japanese and American English. In our first study, native speakers of Japanese hardly showed the McGurk effect when listening to very clear Japanese speech [9]. For example, no /da/ responses were observed for auditory /ba/ combined with visual /ga/. However, these Japanese subjects showed a highly increased McGurk effect when auditory noise was added to the stimuli. These results indicate that the Japanese subjects identified the stimuli based on auditory information unless visual support is necessary due to the noise.

2.3. Recognition of Incompatibility

We also conducted a cross-language study in which native speakers of Japanese and American English were tested [10, 6]. To do so, the above Japanese stimuli (pronounced by a Japanese speaker) and newly created English stimuli (pronounced by an American speaker) were used. The auditory and visual syllables included labial (/ba/, /pa/, /ma/) and nonlabial (/da/, /ta/, /na/, /ga/, /ka/) consonants. The subjects were asked to report what they heard, as well as to report incompatibility between what they heard and what they saw.

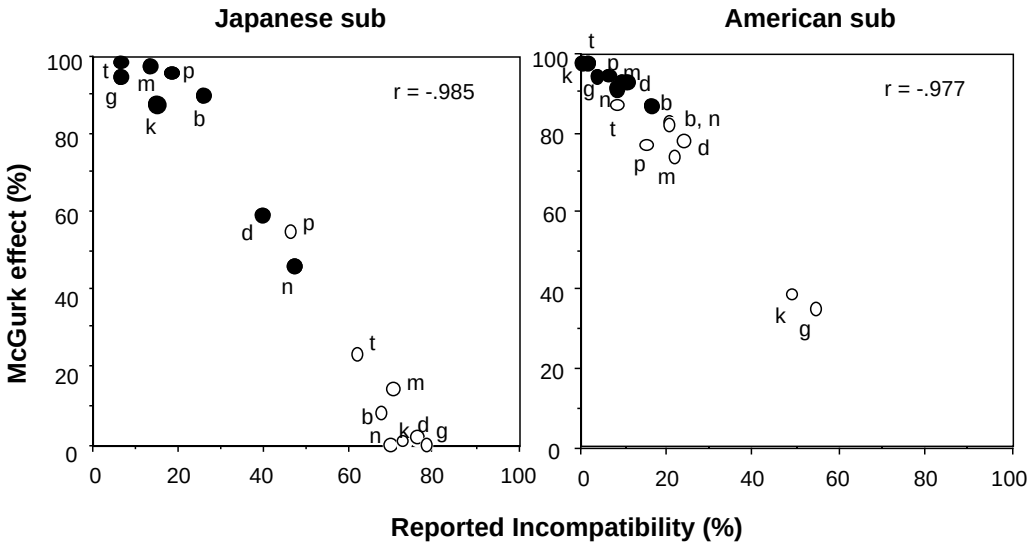

Figure 2: The magnitude of the McGurk effect as a function of recognized incompatibility. (○ quiet condition, ● noise-added condition.)

Compared with the American subjects, the Japanese showed a much weaker McGurk effect and much more frequent detection of auditory-visual incompatibility (Figure 2). In Figure 2, the magnitude of the McGurk effect is plotted as a function of the frequency of recognized incompatibility (only for the Japanese stimuli). Here, open circles are data for the quiet condition, and filled circles are data for the noise-added condition. In the case of the

Japanese subjects, the data for the quiet condition are located in the lower right portion, indicating that the subjects detected auditory-visual incompatibility most of the time and the McGurk effect seldom occurred. When auditory noise was added, the incompatibility was hard to detect, and a strong McGurk effect occurred. In the case of the American subjects, the data for the quiet condition are located in the middle or upper left portion of the graph. That is, the Americans detected incompatibility less frequently, and a strong McGurk effect was induced. From the high frequency of reported incompatibility of the Japanese, it is suggested that the visual information is processed to the extent that the auditory-visual discrepancy is detected most of the time. It suggests that, for clear speech, the Japanese use a type of processing in which visual information is not incorporated into perceived speech even when they extract some lip-read information from the face of the speaker.

The auditory reliance of the Japanese shows that there are cultural and/or linguistic factors which affect the manner of auditory-visual integration. As a cultural factor, it is often said that the Japanese tend to avoid looking at the face of the speaker. This often happens when the speaker is of a higher status. This cultural habit may cause the Japanese to develop a type of processing which does not incorporate visual information into perception even when they are looking at the speaker's face.

Other factors such as linguistic characteristics of Japanese may be responsible for the weak McGurk effect of the Japanese subjects. However, the fact that native speakers of Spanish show a strong McGurk effect [3], the linguistic factor is not plausible, because Spanish and Japanese syllables are phonetically similar.

3. TESTING CHINESE SUBJECTS

3.1. Also Weak in the Chinese

In an attempt to test the face avoidance hypothesis, Chinese subjects were tested [7]. The Chinese were believed to be similar to the Japanese in terms of the face avoidance. Then, the face avoidance hypothesis predicts that the Chinese will also show a reduced McGurk effect.

The subjects were 14 native speakers of Chinese (age 19-30 years old) recruited from the Kanazawa University community. Most of them were graduate students who arrived in Japan after finishing college in China. The length of their stay in Japan was between 4 months and 6 years. Stimuli were the Japanese and English stimuli used in the above experiments.

The results are shown together with the previous results for the Japanese and American subjects (Figure 3). The results appear to support the face avoidance hypothesis. Compared with the results for the American subjects, the Chinese subjects showed a much weaker visual effect, indicating a similarity to the Japanese subjects. Compared with the results for the Japanese subjects, the magnitude of the visual effect for the Chinese subjects was the same (when the stimuli were Japanese), or smaller (when the stimuli were English). The weaker visual effect for the Chinese for the English stimuli suggests that the Chinese rely on auditory information even more strongly than the Japanese.

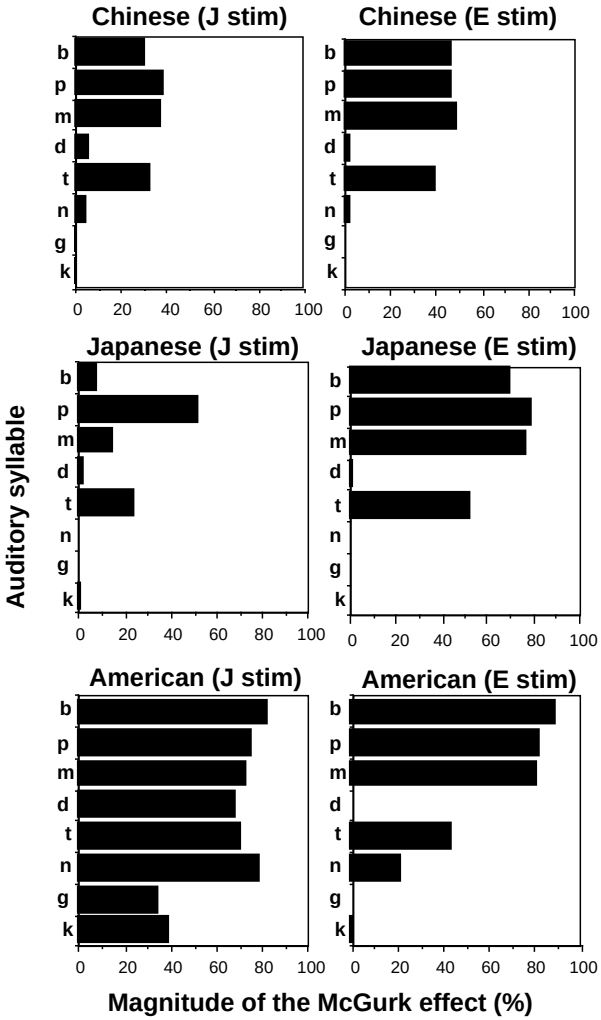

Figure 3: The magnitude of the McGurk effect in the three language groups.

3.2. Effect of Second Language Learning

Because the Chinese subjects had lived in Japan for a certain period of time, the relationship between the length of their stay in Japan and the magnitude of the McGurk effect was examined. Figure 4 shows individual data for the Chinese subjects (only for auditory labials for which the McGurk effect ocurred to some extent). There was a positive correlation (r=.723) between the two indexes: The longer the length of the subjects stay in Japan is, the stronger the visual effect becomes. This correlation suggests that the monolingual Chinese (at the far left of the plot) tend to rely on auditory

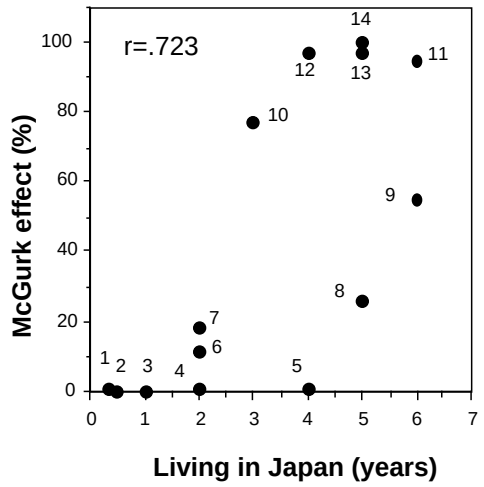

Figure 4: The relationship between the magnitude of the McGurk effect and the length of the subjects stay in Japan.

information and the second language learning causes in the Chinese a shift to a manner of processing where visual information is incorporated into perceived speech. It is plausible that people who are seriously learning a second language in a situation where that language is natively spoken also learn to use any cues including lip-read information to improve their listening comprehension. Presumably the longer stay in Japan corresponds to a higher proficiency of the Japanese language. These results imply that the proficiency of a second language may rely to some extent on the learner's skill of lipreading that language.

4. CONCLUSION

In conclusion, there are a few factors which affect the degree of incorporating visual information into speech perception. A cultural factor, face avoidance was suggested from the weak McGurk effect for the monolingual Japanese and Chinese subjects. Another factor suggested was experience of second language learning. Because speech perception heavily depends on one's perceptual structure which is specialized for one's native language, sounds of non-native language have a certain ambiguity. For this reason, additional visual cues are considered to be helpful in second language learning. The nature of auditory-visual integration should be taken seriously in the field of foreign language education.

5. REFERENCES

1. Dekle, D. J., Fowler, C. A., and Funnell, M. G. "Audiovisual integration in perception of real words," *Percept. & Psychophys.*, Vol. 51, 355-362, 1992.

2. Green, K. P., Kuhl, P. K., Meltzoff, A. N., and Stevens, E. B. "Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect," *Percept. & Psychophys.*, Vol. 50, 524-536, 1991.

3. Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., and Heredia, R. "Bimodal speech perception: An examination across languages," *J. Phonet.*, Vol. 21, 445-478, 1993.

4. McGurk, H. and MacDonald, J. "Hearing lips and seeing voices," *Nature*, Vol. 264, 746-748, 1976.

5. Rosenblum, L. D. and Saladana, H. "An audiovisual test of kinematic primitives for visual speech perception," *J. Exp. Psychol.: Human Percept. Perform.*, Vol. 22, 318-331, 1996.

6. Sekiyama, K. "Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility," *J. Acoust. Soc. Jpn. (E)*, Vol. 15, 143-158, 1994.

7. Sekiyama, K. "Cultural and linguistic factors influencing auditory-visual integration of speech information: The McGurk effect in Chinese subjects," *Percept. & Psychophys.*, 1996 (in press).

8. Sekiyama, K., Joe, K., and Umeda, M. "Perceptual componets of Japanese syllables in lipreading: a multidimensional study," *Trans. Inst. Electr. Inform. Commun. Engineer. Jpn.*, IE87-127, 1988.

9. Sekiyama, K. and Tohkura, Y. "McGurk effect in non-English listenters: Few visual effect for Japanese subjects hearing Japanese syllables of high auditory intelligibility," *J. Acous. Soc. Amer.*, Vol. 90, 1797-1805, 1991.

10. Sekiyama, K. and Tohkura, Y. "Inter-language differences in the influence of visual cues in speech perception," *J. Phonet.*, Vol. 21, 427-444, 1993.

11. Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., and Jones, C. J. "Effects of training on the visual recognition of consonants," *J. Speech Hear. Res.*, Vol. 20, 130-145.

6. ACKNOWLEDGMENTS

The studies presented here were supported by Grant-in Aid for Scientific Research from the Japanese Ministry of Education, Science, and Culture; Research Grant from Nissan Science Foundation; and Sasakawa Scientific Research Grant from the Japan Science Society.