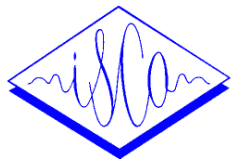

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

FACE OR VOICE? DETERMINANT OF COMPELLINGNESS TO THE MCGURK EFFECT.

Kaoru Sekiyama

Kanazawa University

ABSTRACT

This study examined sources of talker differences in the McGurk effect, by questioning which of auditory speech or visual speech is more determining the size of the McGurk effect. Cross-talker dubbing was done between faces and voices of utterances (/ba/ and /ga/) pronounced by two talkers: one compelling talker (CT) to the McGurk effect, and one less compelling talker (LT), as measured in our previous studies. The two talkers and 18 subjects were native speakers of Japanese. There were three presentation conditions: Audio-only, video-only, and audiovisual. The results of unimodal conditions showed that CT was easier to speechread than LT, but that CT was more difficult to listen to than LT. The results of audiovisual condition showed that, although both the visual and auditory talker affect the size of the McGurk effect, the audio component is more responsible than the video component for the talker differences in the McGurk effect.

1. INTRODUCTION

The McGurk effect is an audiovisual illusion which shows that visual mouth information is integrated with conflicting auditory information during speech perception [1]. For example, auditory /pa/ synchronized with visual mouth movements of /na/ often makes perceptual /ta/. The McGurk effect clearly demonstrates that speech perception is not a solely auditory process in face-to-face communication, but a multimodal process.

Our previous research has shown that native speakers of Japanese are less influenced by visual cues than native speakers of American English [2] [3]. A characteristic result for the Japanese subjects was that they are sensitive to audio quality and the size of visual influence easily depends on auditory intelligibility [4]. From this result we proposed an "auditory intelligibility hypothesis" that native speakers of Japanese take visual cues into perceived speech only when auditory intelligibility is not perfect. The present study examined this hypothesis in relation to talker differences.

It is often observed in laboratory that the absolute size of the McGurk effect differs from one talker to another. Some talkers are very compelling to the McGurk effect while others are not. The question I asked was what is responsible for the differences in compellingness to the McGurk effect. More specifically, is it intelligibility of auditory speech or visual speech? Some talkers are easy to speechread, and this high visual intelligibility generally promotes the McGurk effect (note that the McGurk effect can be defined as a visual biasing effect to the auditory speech perception). On the other hand, highly intelligible auditory speech resists the McGurk effect. It should be noted that easiness in speechreading does not necessarily guarantee easiness in listening in one talker. The auditory intelligibility hypothesis predicts that quality of auditory speech plays a more important role.

2. PURPOSE

The purpose of the present study was to decide which of visual or auditory intelligibility is more determining the size of the McGurk effect. To do so, the present experiment used two talkers: one compelling and one

less compelling. Cross-dubbing was done between faces and voices of the two talkers. The size of the McGurk effect was evaluated for each face and voice to see which component (face or voice) is more responsible for the compellingness to the McGurk effect.

3. METHOD

3.1. Subjects

The subjects were 18 undergraduate students at Osaka City University. They were all native speakers of Japanese and reported having normal hearing and normal (or corrected to normal) vision.

3.2. Stimuli

Stimuli were created by using two talkers. The two talkers were both female native speakers of Japanese in comparable age and profession. Our previous experiments showed that one was fairly compelling to the McGurk effect ([5], talker JF) and the other was hard to induce the McGurk effect [4]. Let the former called compelling talker (CT) and the latter less compelling talker (LT). Stimulus materials were utterances of /ba/ and /ga/ of the two talkers. The videotaped utterances included the talker's whole face and audio signals. The audio signals of each syllable (/ba/ or /ga/) of each talker (CT or LT) were combined with video signals of either syllable of either talker so that all the possible combinations were created. The dubbing was done with frame unit (33-ms) accuracy. The resulting audiovisual stimuli consisted of 2 (auditory syllable) x 2 (visual syllable) x 2 (auditory talker) x 2 (visual talker) = 16, as shown below.

< syllable combination>		x <talker combination>	
Audio	Video	Audio	Video
/ba/	/ba/	CT	CT
/ga/	/ga/	CT	LT
/ba/	/ga/	LT	CT
/ga/	/ba/	LT	LT

The talker's face was presented on a 20-in video monitor and the talker's voice was presented through two loud speakers located along the sides of the monitor.

3.3. Procedure

The stimuli were presented in several conditions. The subjects were asked to write, in Roman alphabet, what they thought the speaker was saying. Thus, the response was an open choice. There were three presentation conditions: Audio-only, video-only, and audiovisual. In the audio-only and audiovisual conditions, there were two audio clearness conditions: Noisy audio and clear audio. In the noisy audio condition, white noise was added to audio signals with a S/N ratio of 0 dB. All these conditions were within-subjects factors. Trials for each condition were blocked. All the subjects were tested in a fixed order in which the effect of order was considered to be minimum: Noisy audiovisual, video-only, noisy audio-only, clear audiovisual, and clear audio-only.

4. RESULTS

4.1 Unimodal conditions

The results for the audio-only conditions (Fig. 1) showed that the auditory intelligibility was higher in the less compelling talker (LT) than in the compelling talker (LT). The difference became apparent in the

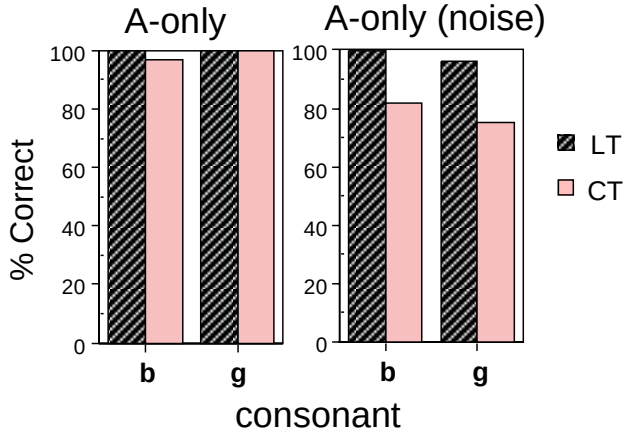

Fig.1 Percent correct in audio-only conditions.

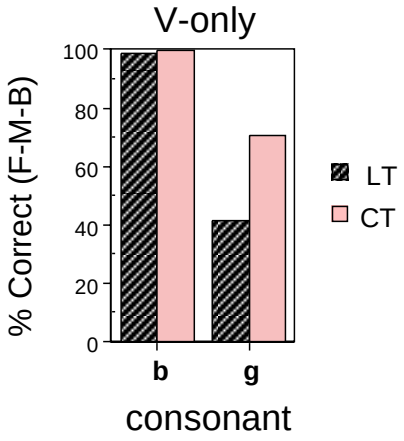

Fig. 2 Percent correct in video-only condition. Scoring was done in terms of the place of articulation, i.e. front, middle, and back consonants.

noisy audio condition (right panel) in which LT's auditory speech was resistant to the noise. In the video-only condition, performances were scored in terms of the accuracy in place of articulation (front, middle, and back consonants). As shown in Fig. 2, CT's mouth movements were better speechread than LT's, especially for visual /ga/. These unimodal results indicate that both visual and auditory components of CT were in favor of the McGurk effect: Her face was clear to speechread and her voice was ambiguous to hear.

4.2. Bimodal conditions

In the audiovisual condition, only the results for the McGurk-type incompatible syllable combinations were analyzed. Fig. 3 shows percent auditory for each stimulus. There were four types of talker combinations of voice(face): LT(LT), LT(CT), CT(LT), CT(CT). With respect to the size of the McGurk effect (opposite to the percent auditory), the talker combination differences were significant only when auditory syllable was /ba/. In such cases, the size of the McGurk effect varied, from smaller to larger, LT(LT), LT(CT), CT(LT), CT(CT). When no noise was added (left panel), auditorily more intelligible LT's voice

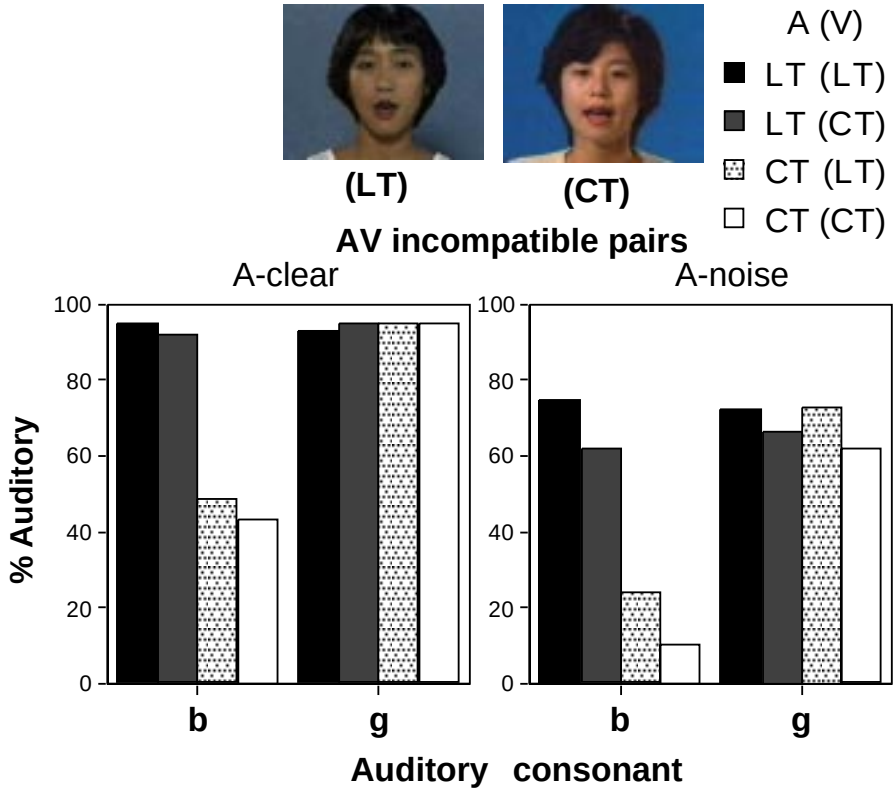

Fig. 3 Percent auditory in audiovisual incompatible pairs.

was hardly influenced by visual input, whereas auditorily more ambiguous CT's voice easily induced the McGurk illusion. It seems that the talker difference in the audio-only condition tends to be magnified in the audiovisual condition.

About the face differences, visually more intelligible CT's face tended to produce a larger McGurk effect when the voice was the same. However, as a quantitative comparison, the size of the McGurk effect varied more largely due to the voice than to the face.

5. DISCUSSION

These results suggest that the audio component is more responsible than the video component for the talker differences in the McGurk effect. Although the generalization of the present results may need to wait for experiments using other talkers, the auditory intelligibility hypothesis [4] is in accordance with the current conclusion that the more determining component is auditory speech. It is also of interest to see if the present results are replicated across languages.

6. REFERENCES

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

2. Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E)*, 15, 143-158.

3. Sekiyama, K. & Tohkura, Y. (1993). Interlanguage differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21, 427-444.

4. Sekiyama, K. & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, 90, 1797-1805.

5. Sekiyama, K., Braida, L. D., et al. (1995). The McGurk effect in Japanese and American perceivers. *Proceedings of 13th International Congress of Phonetic Sciences (Stockholm)*, Vol.3, 214-217.

7. ACKNOWLEDGEMENT

This study was supported by a grant-in-aid for scientific research from the Japanese Ministry of Education, Science, and Culture, and a research grant from International Communications Foundation. The author is grateful to Prof. Yoichi Sugita at Toyohashi University of Technology for his help in preparing the experiment, and to Jun Amano and Naoko Yanagida for their help in running the experiment.