Interdisciplinary Information Sciences Vol. 18, No. 2 (2012) 83–85

ISSN 1340-9050 print/1347-6157 online
DOI 10.4036/iis.2012.83

Effect of Consonance between Features and Voice Impression on the McGurk Effect

Shuichi SAKAMOTO*, Hiroshi MISHIMA, and Yôiti SUZUKI

Research Institute of Electrical Communication and Graduate School of Information Sciences, Tohoku University, Sendai 980-8577, Japan

Received June 13, 2012; final version accepted August 31, 2012

The McGurk effect is one of the typical phenomena caused by human multi-modal information processing between auditory and visual speech perception. In this paper, we investigated the relation between the degree of the McGurk effect and the perceived impression by speech sounds and moving images of the talker's face. As stimuli, uttered speech sounds were combined with moving images of a different talker's face. These stimuli were presented to observers, who were asked to respond to what the talker was saying. At the same time, they were asked to report their subjective impressions of these stimuli. Matching between the voice and moving image was used as the index of the judgment. Results showed that matching between a voice and a talker's facial movements affected the degree of the McGurk effect, suggesting that audio–visual *kansei* information affects phoneme perception.

KEYWORDS: McGurk effect, audio–visual speech perception, multi-modal processing, lip-reading

1. Introduction

Many researchers have pointed out that visual information strongly affects speech understanding [1, 2]. Using recent broadband networks, not only auditory information (speech signal) but also visual information (moving image of talker's face) can be transferred easily. Therefore, for developing advanced communications systems, it is important to investigate the effect of visual information on speech understanding.

Audio–visual speech signals include not only linguistic information but also rich *kansei* information, such as facial expressions and emotions. Such *kansei* information also affects speech perception. Hakoda *et al.* reported that the impressions of talkers' faces and those of the voices can be expressed using the same factors [3]. Lander *et al.* showed that prosodic variations in rhythmic structure and expressiveness are cues to match identity between the faces and voices of unfamiliar people and languages [4]. However, these studies addressed only the effect of subjective impression from audio–visual speech information related to perceived *kansei* information. It remains unclear whether such subjective impressions affect perceived linguistic information.

Behaviors of human information processing when audio–visual stimuli are given as inconsistent speech sounds and moving images might be useful to illustrate the interactions in the human brain between *kansei* information and speech perception. The most important and clear phenomenon caused by speaker's inconsistent audio and visual signal is the McGurk effect [2]. This is the phenomenon by which listeners hear a different phoneme from an original phoneme because they are affected by visual information. This effect is evoked by a synchronous presentation of a bilabial sound such as /p/ with a speaker's face uttering non-labial sound such as a palatal stop /k/, resulting in a perception of another phoneme, typically a dental stop /t/.

This study examined an aspect of the interaction between subjective impression of audio–visual speech information and speech understanding based on the McGurk effect. First, synthetic stimuli were composed from the respective speech sounds and moving images of five talkers' faces. Then, the occurrence of the McGurk effect as well as speech intelligibility were measured using these stimuli to investigate effects of perceived impression of the stimuli such as consonance between a sound and the moving image of the talker's face and the clearness of sounds on the perception of phonemes.

2. Experiment

Figure 1 depicts a schematic diagram of the experimental setup. Participants sat on a chair in an anechoic room installed in the Research Institute of Electrical Communication, Tohoku University.

* Corresponding author. E-mail: saka@ais.riec.tohoku.ac.jp

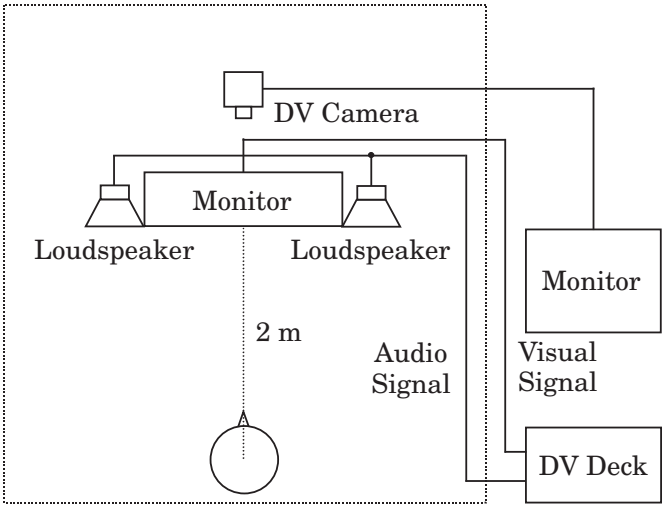

Fig. 1. Schema showing the experiment setup.

Experimental stimuli of three kinds were prepared. The first group of stimuli comprised visual–auditory pairs of /ke/-/pe/, which were designed to elicit fusion responses. The other two were visual–auditory pairs of /ke/-/ke/ and /pe/-/pe/, which were for reference purposes. We expected to observe the McGurk effect when the visual–auditory pair of /ke/-/pe/ was presented, resulting in a typical perception of /te/. These stimuli were generated as follows. First, speech sounds uttered by five female speakers were recorded via a microphone (Type 4165; Brüel & Kjær). At the same time, the moving images of their talking faces were also recorded using a DV camera (DCR-TRV7; Sony Corp.). Next, the sounds and moving images of all talkers were extracted independently. Then each signal was combined with moving image of each talker's face using audio–visual editing software (Adobe Premiere 6.0; Adobe Systems Inc.). By this processing, 25 stimuli were generated for each kind of visual–auditory pair.

Each of these stimuli was presented 10 times. The sounds were presented at 60 dB (A-weighted sound pressure level) at the point of the center of observer's head via two loudspeakers (Diatone DS-7; Mitsubishi Electric Corp.). The moving images were presented via a 42-inch monitor (TH-42PWD4: 920 × 518 mm; Panasonic Inc.) that was set 2 m in front of an observer. The size of a talker's face in the monitor was set to be as large as the actual size of the talker standing at the point of the monitor.

Nine naïve observers (with no information about the talker or the McGurk effect) with normal hearing acuity participated in the experiment. They had normal or corrected-to-normal vision. Participants were asked to perform the following three tasks: (1) identify what a talker was saying, (2) rate the consonance between the impression received from the sound and that received from the moving image, and (3) rate the clearness of sounds. The latter two tasks were rated with a numerical scale from 1.0 (low) to 5.0 (high).

3. Results and Discussion

Figure 2 shows the percentage of occurrence of the McGurk effect as a function of the consonance between the impression of speech sound and that of moving image of talker's face. The occurrence of the McGurk effect denotes how often participants answer /te/ when the visual–auditory pair of /ke/-/pe/ is presented. This percentage and the consonance were calculated for each of the 25 stimuli by averaging the scores of the nine participants. In this figure, the correlation coefficient between the percentage of the occurrence of the McGurk effect and the consonance was 0.66 and statistically significant ($p < .01$). This result shows that the percentage of the occurrence of the McGurk effect increases as the consonance of the visual–auditory pair grows. This fact indicates that, as the impression of speech sound and that of moving image of talker's face become better matched subjectively, visual and auditory information more strongly affect each other and thereby cause the McGurk effect more frequently.

Figure 3 presents the percentage of occurrence of the McGurk effect as a function of sound stimuli clearness. The correlation coefficient between the percentage of the occurrence of McGurk effect and the clearness of sound stimuli was 0.86 and statistically significant ($p < .01$). These results show that the percentage of the occurrence of the McGurk effect increases as the clearness of speech sound increases. As we mentioned previously, talkers utter /pe/ in the stimuli. Therefore, if participants hear the voice of the talker correctly, they would answer /pe/. However, they perceived the voice clearly and answered the voice as /te/ clearly, when the occurrence of the McGurk effect is high. As we mentioned in the introduction, the McGurk effect is that phenomenon by which participants hear a different phoneme from an original phoneme because they are affected by visual information. The results mean that visual–auditory information's tightly coupling would induce the high occurrence of the McGurk effect and the high clearness

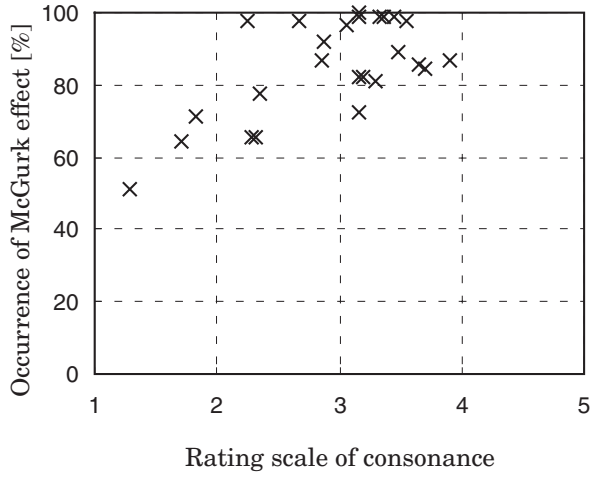

Fig. 2. Occurrence ratio of McGurk effect as a function of consonance between audio and visual impressions of talkers.

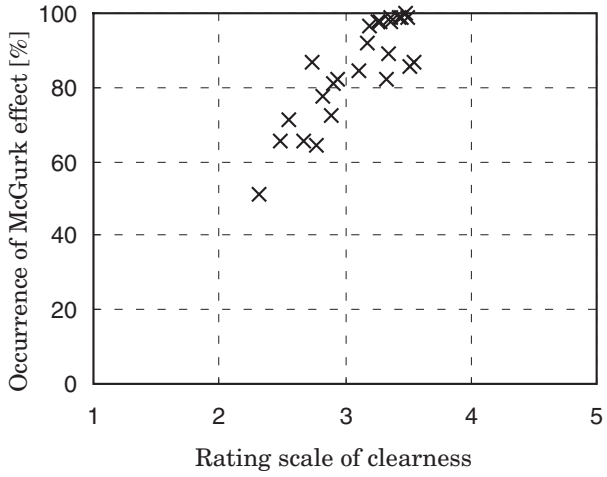

Fig. 3. Occurrence ratio of McGurk effect as a function of the clearness of sound stimuli.

of sound stimuli. These results suggest that speech understanding can be changed not only by auditory information, but also by subjective impression, i.e., the consonance between a speech sound and the motion of the talker's face.

4. Conclusions

We investigated the relation between the occurrence rate of the McGurk effect and the perceived impression by speech sound and moving image of talker's face, in terms of audio–visual consonance and clearness. Results showed that consonance, i.e., matching, between a speech sound and the moving image of talker's face as well as the clearness of speech sound influenced the occurrence of the McGurk effect. These results indicate that the subjective impression obtained by visual and auditory information affect phoneme perception.

Acknowledgment

This research was partially supported by JSPS, Grant-in-Aid for Scientific Research (C), No. 23500252, 2011.

REFERENCES

[1] McGurk, H., and MacDonald, J., "Hearing lips and seeing voices," *Nature*, **264**: 429–439 (1976).
[2] Erber, N. P., "Interaction of audition and vision in the recognition of oral speech stimuli." *Journal of Speech Hearing Research*, **12**: 423–425 (1969).
[3] Hakoda, Y., Oda, M., Haraguchi, M., Yoshizaki, S., and Akamatsu, S., "Physical features of faces and personality impression," *Technical Report of IEICE*, HIP99–68 (2000).
[4] Lander, K., Hill, H., Kamachi, M., and Vatikiotis-Bateson, E., "It's not what you say but the way you say it: Matching faces and voices," *Journal of Experimental Psychology, Human Perception and Performance*, **33(4)**: 905–914 (2007).