INFLUENCE OF FACIAL VIEWS ON THE MCGURK EFFECT IN AUDITORY NOISE

Rika Kanzaki[1] *and Takashi Kato*[1, 2]

[1] ATR Human Information Processing Research Laboratories, Kyoto, 619-0288, Japan
[2] Faculty of Informatics, Kansai University, Osaka, 569-1095, Japan

ABSTRACT

To investigate the nature of facial information involved in the integration of audiovisual speech perception, we examined the influence of facial views on the McGurk effect under two auditory conditions. While the speech perception of most of the audiovisual syllables used was little affected by the facial views and the auditory noise, a stronger McGurk effect was obtained for the 3/4-view image uttering labial sounds presented with auditory syllables of alveolars under the auditory noise. However, the facial view did not affect the visual identification of labials in the same way. These results suggest that the information about labial or nonlabial is not the only facial information involved in the McGurk effect. It appears some other information available only in the 3/4-view image might also be involved in the McGurk effect. Implications for the processing of visual, auditory and audiovisual speech are discussed.

1. INTRODUCTION

Although the primary modality for speech perception is hearing, speech perception can be influenced by what listeners see on a speaker's face[1,2,3]. This visual influence on speech perception is demonstrated by a phenomenon known as the McGurk effect[2], in which listeners combine inconsistent auditory and visual information, thereby perceiving different sounds. For example, dubbing an auditory sound /ba/ onto a visual image /ga/ makes an audiovisual stimulus that induces the percept /da/. The McGurk effect occurs when the speech is distinct and unambiguous and even when listeners are completely aware that the presented face image and auditory speech are dubbed and/or asynchronized[4].

Much investigation has been conducted on a variety of situations in which audiovisual speech integration occurs[5,6]. However, relatively little is known about visual speech in audiovisual speech perception. During normal face-to-face interactions, substantial changes in the visual appearance of a speaker's face can occur through changes in facial views as seen by the observer. Although there are some previous studies indicating that presenting the full face is not necessary for integrating visual speech information with auditory speech information[3,7], these studies used only front-view images and so provided little information about the influence of facial views on audiovisual speech perception.

Changes in facial view may affect the perception of audiovisual speech because a shift in the viewing angle of the speaker's face changes the observable parts of the face. For example, consider the appearance of the face for the spoken syllable /ba/, whose articulation begins with a bilabial closure followed by a vertical movement that separates the lips. These features of lip movement can be observed from front and profile views. If the speaker's face is presented as a full face, the whole lip shape, oral cavity, tongue, and teeth can also be seen. However, if the speaker's face is viewed from the side, the listeners can not see the oral cavity or teeth, whereas the jaw rotation can be more saliently recognized than in the full face.

We recently reported that the face information involved in the McGurk effect is dependent on the viewing angle[8]. However, in the following study that employed more intelligible, digitally-recorded voices for the auditory stimuli, we obtained only a slight influence of facial views on the McGurk effect[9]. These results suggest that auditory intelligibility may affect the nature of face information integrated with voice information. Indeed, there is some evidence showing that the magnitude of the McGurk effect is affected by auditory intelligibility[10].

The purpose of this study was to further investigate the influence of facial view on the McGurk effect, by examining whether auditory intelligibility can change the influence of facial views on the McGurk effect.

2. METHOD

The subjects were 18 undergraduate students who were paid for their participation. None of the subjects reported any history of hearing disorders, and all had normal or corrected-to-normal vision. All were native Japanese speakers.

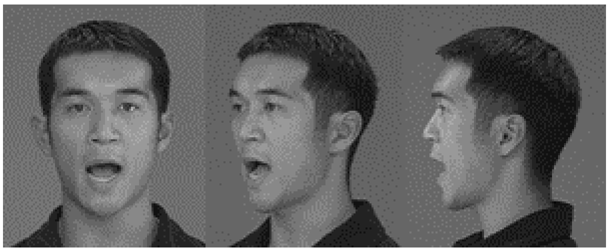

Figure 1: Example of visual stimuli used to study facial view influence on the McGurk effect.

The stimuli used were the utterances of a Japanese male for six Japanese syllables: /ba/, /pa/, /da/, /ta/, /ga/, and /ka/. The visual images were the front, left 3/4 and left profile views of the speaker taken simultaneously by three video cameras (Fig.1). The speaker was seated on a chair with a blue background. Each camera was centered on the speaker's face. The height of the facial images, when shown on a 20-inch video monitor (Sony Trinitron color video monitor PVM-2054Q), was approximately 25 cm. The speech utterances were recorded with a microphone positioned in front of the speaker. These auditory signals were recorded with a DAT and a BETACAM recorder with visual signals.

The six audio signals on the DAT tape and six visual signals on the BETACAM tape were combined on a computer (Power Macintosh 9500/132). The auditory signals were digitized at a 44.1 kHz sampling rate and 16-bit resolution on the computer. To synchronize the audio signals and video images, the auditory signals were dubbed on to the frames of the previous audio signals on a BETACAM tape. The audio and video signals were precisely synchronized by adjusting the dubbing timing with a 33 ms frame unit. For each facial view, six audio and six video stimuli were combined, resulting in 36 audiovisual stimuli. Each audiovisual stimulus was arranged in a 7-sec unit, which included a 3-sec period of no face shown and a 4-sec talking face .

Stimulus sequences for auditory-alone and visual-alone presentations were also created. Six auditory stimuli with blank frames on the video channel were presented for the auditory-alone presentations. Six visual stimuli with no syllables on the audio channel were presented for each facial view for the visual-alone presentation.

Two identical sets of audiovisual, audio-alone, and video-alone stimuli were made and the auditory stimuli of one of them was added with white noise whose S/N ratio was zero. These stimuli were copied onto the BETACAM videotapes for use in the experiments.

Each subject participated in two 2-hour sessions conducted on different days. The subject was given a stimulus set with noise on one day and one without noise on the other day. Each subject was presented with three audiovisual blocks, one audio-alone block, and three visual-alone blocks on one day.

The tests were conducted for each viewing angle in the audiovisual blocks. Thirty-six audiovisual stimuli under one facial view were presented six times in random order. The order of the three blocks was counterbalanced across subjects. The 6 auditory syllables were presented six times each in random order in the audio-alone block. In the same manner, the 6 visual syllables were presented six times each in random order for the visual-alone block,. This design gave a total of six observations per participant per audiovisual, audio, and visual stimulus.

The subjects were individually tested in a dimly lit room, seated on a chair approximately 1.5 m from the video monitor. Initially, each subject participated in the audiovisual blocks. The subjects were instructed to watch and listen to each audiovisual stimulus, and to report what they had heard. In the audio-alone test, the subjects were asked to listen to each syllable and identify it. In the visual-alone blocks, the subjects answered what they thought the speaker had uttered. These tests were run with 5-minute rest periods between each block.

The audio signals were presented via two speakers (JBL CM-40) on both sides of the monitor. The auditory syllables were presented at a comfortable listening level of approximately 65 dB SPL; they were measured for the peak intensity of the syllable at the approximate location of the subject's head using a microphone (B & K, type 4155) with a sound level meter (B & K, type 2231).

3. RESULTS

Audio-alone task: Table 1 shows the results of the audio-alone task under two noise conditions. The numbers indicate the percentages of correct responses in 108 observations (18 subjects * 3 repetitions * 2 days) for each auditory stimulus. The performance was 86 % or better for the auditory syllables with the exception of /pa/ under the noise-added condition where /pa/ was perceived as /ta/ on 23% of trials and /a/ on 22% of trials.

Table 1: Percentage of correct identifications for each of the auditory stimuli employed in the audio-alone task.

Sound	/ba/	/pa/	/da/	/ta/	/ga/	/ka/
Noise-free	93.5	88.0	96.3	96.3	100	100
Noise-added	98.1	55.6	87.0	86.1	100	100

Audiovisual task: As shown in the results of the audio-alone task, some of the auditory syllables were incorrectly identified even without noise. It was expected that /pa/, for example, would be frequently confused with /ta/ and /a/ in the audiovisual test with auditory noise. Therefore, the magnitude of the incorrect speech perception in the audiovisual task was calculated by subtracting the identification errors in the audio-alone task from those in the audiovisual task.

Figure 2 shows the incorrect identifications (%) of auditory stimuli for each visual syllable under two noise conditions. The McGurk effect was mostly obtained for the combinations of the auditory and visual stimuli that differed in the places of articulation. Although the facial view and auditory noise had little effect on the speech perception of most of the audiovisual stimuli, their influence on the speech perception was shown with a face image uttering labial sound and auditory syllables of alveolars. For instance, when the face image uttered /ba/ and the auditory syllable was /ta/, no significant differences were obtained in the magnitude of the McGurk effect across the three facial views under the noise-free condition. However, the McGurk effect under the noise-added condition for all of the viewing angles was stronger than under the noise-free condition. Morever, a stronger McGurk effect was shown for the 3/4-view image than for the full face image.

Visual-alone task: Figure 3 shows the correct identifications (%) of visual stimuli for each facial view under the two noise conditions. When the face image uttered /ba/ or /pa/ (i.e., the labial as the place of articulation), the visual speech was identified as labial with 80% or more accuracy at each viewing angle in both noise conditions. There were no significant differences in the correct identification of visual stimuli between three facial views and the two auditory noise conditions.

When the face image uttered /ta/ or /da/ (i.e., an articulation at the alveolar), the visual speech was perceived as alveolars with an average accuracy of 70% at all facial views under the two noise conditions. There were significant differences of correct identification of /ta/ as alveolars between the profile images and the other two facial views under the noise-added condition. There was a significant difference between the profile image and 3/4-view images for the

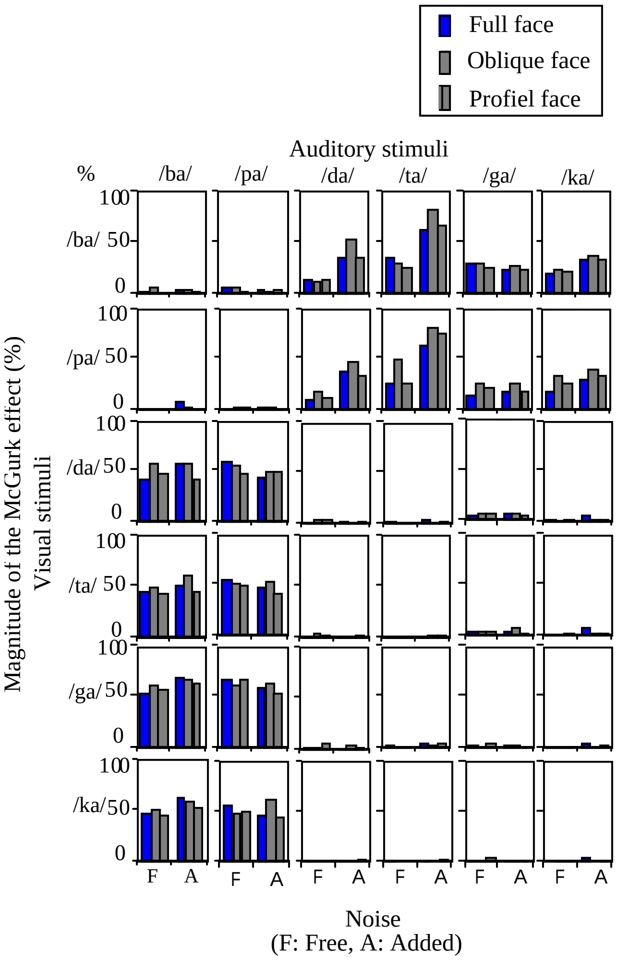

Figure 2: Magnitude of the McGurk effect calculated by subtracting auditory errors in audio-alone task form the incorrect speech perception in audiovisual task.

visual /da/ utterances.

When the face image uttered /ga/ or /ka/ (i.e., an articulation at the velar), there observed poor discrimination of the visual speech as the velars and no significant differences between the three facial views and the two noise conditions.

4. DISCUSSION

While the auditory perception of most of the audiovisual syllables used in the present study was little affected by the facial views and the auditory noise, a relatively strong McGurk effect was obtained for the 3/4-view image uttering labials presented with alveolar sounds under the auditory noise. However, the observed pattern showing the influence of the facial view on the visual perception of labials was not entirely consistent with that on the McGurk effect. Visual labials were well discriminated from non-labials in all of the viewing angles under the both noise conditions.

As shown in Figure 2, the McGurk effect was likely to occur when the auditory and visual stimuli differed in the places of articulation. These results are consistent with those of previous studies of the McGurk effect, which showed that the human visual system can at least distinguish between labials and nonlabials[2,11]. Such a visual system characteristic seems to provide a plausible account of the current finding.

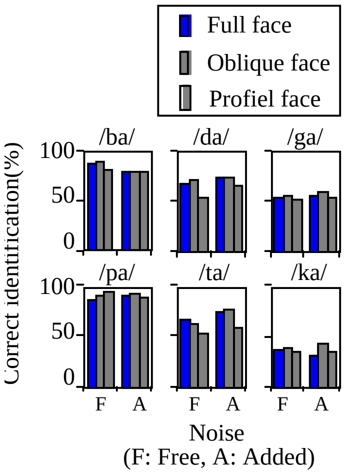

Figure 3: Percentage of correct identifications for each of visual stimuli in visual-alone task. Correct identifications indicate correct responses for the place of articulation.

Visual speech information commonly available in the full-face, oblique-side and profile views includes information concerning how the mouth is opened or closed. While the production of visual labials starts with a closed mouth, that of visual non-labials starts with an opened mouth. It seems that the listeners were able to see the speaker's mouth opened or closed in all of the facial views and obtain information regarding whether the syllables uttered by the face were labial or non-labial.

However, the results of the present study suggest that the information about the place of articulation is not the only facial information involved in the McGurk effect. The results from the face images uttering the labials suggest that some other information available only in the 3/4-view image might also be involved in the McGurk effect under the noise-added condition.

What might be the visual speech information available only in the oblique face? Why might such information be more available for the visual labials and perhaps more salient under noise? Although the full face shows the whole lip shape, oral cavity, tongue, and teeth, the profile shows only some part of the lips, one side cheek, and jaw rotation. The oblique face might show both types of information to some extent, resulting in a stronger visual influence on the McGurk effect. Or, it could be that the 3/4-view presents important information in visual speech perception. The oblique face uttering labials might present more salient movement for visual speech, while noise might encourage the processing of such information.

However, these interpretations are not consistent with the results of the visual-alone task, where visual labials were well discriminated from non-labials in all the viewing angles under both the noise conditions. Further, although there might have been a ceiling effect, adding auditory noise did not lead to a better identification of the visual labials. Neither the performance in the visual alone task provide a satisfactory account of either the effect of 3/4-view images or auditory intelligibility on the McGurk effect.

We also found inconsistent results of the audiovisual tasks and the visual-alone task for the visual alveolars. When the face image articulated /ta/ and the auditory syllable was /ba/ or /pa/, no significant differences were obtained in the magnitude of the McGurk effect across the viewing angles. However, in the visual alone task, the visual alveolars were better identified in the front and oblique side views than in the profile view under the noise-added condition. These front and 3/4-view advantages for speech reading are in accord with previous findings in speech reading studies[12,13]. In those studies, full face and oblique side face provided more information for speech reading.

A previous study reported that the McGurk effect depended on the auditory intelligibility of the speech signal and that auditory uncertainty induced a stronger McGurk effect[10]. However, in the present study, whereas the auditory syllable /pa/ showed extremely low intelligibility under the auditory noise (Table 1), there were no differences in the magnitude of the McGurk effect between the two auditory conditions. Furthermore, although the auditory syllables /da/ and /ta/ showed a slightly reduced intelligibility under the auditory noise, the auditory syllable under the noise resulted in a stronger McGurk effect. As such, the McGurk effect results could not be accounted for by the results of the auditory-alone task.

The results of this study suggest that facial view can influence the McGurk effect and that this influence is dependent on the combined syllables and the phonemic features of the auditory syllable. The results also suggest that visual or auditory information involved in the McGurk effect might be different from the visual or auditory information available for unimodal speech perception. These interpretations are consistent with a recent study that examined the influence of facial orientation on the McGurk effect[14]. However, they are not consistent with the assumption of a bimodal speech model proposed by Massaro[15,16], where the magnitude of the McGurk effect can be predicted from measurements made under unimodal conditions. It appears that a careful examination is warranted for the relation of unimodal and bimodal speech perception.

5. REFERENCES

1. Sumby, W.H. and Pollack, I. "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am., 26, 212-215, 1954.

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

3. Summerfield, Q. "Use of visual information for phonetic perception," Phonetica, 36, 314-331, 1979.

4. Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. "Temporal constraints on the McGurk effect," Percept. Psychophy., 58, 351-362, 1996.

5. 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., 50, 524-536, 1991.

6. Driver, J. "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," Nature, 381, 66-68, 1996.

7. Smeele, P.M.T., Hahnlen, L.D., Stevens, E.B., and Kuhl, P.K.. "Investigating the role of specific facial information in audio-visual speech perception," J. Acoust. Soc. Am., 98, 2983, 1995.

8. Kanzaki, R., Kato, T., and Tohkura, Y. "Influence of facial views on the McGurk effect, " J. Acoust. Soc. Jpn. (E), 19, 1998.

9. Kanzaki, R. "Integration process of face and voice information," Vision, 9, 233-239, 1997. (in Japanese).

10. Sekiyama, K and Tohkura, Y. "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.

11. MacDonald, J. and McGurk, H. "Visual influences on speech perception processes," Percept. Psychophys., 24, 253-257, 1978.

12. IJsseldijk, F.J.. "Speechreading performance under different conditions of video image, repetition, and speech rate," J. Speech Hearing Res., 35, 466-471, 1992.

13. Eber, N. P. "Effects of angle, distance, and illumination on visual reception of speech by profoundly deaf children," J. Speech Hearing Res., 17, 99-112, 1974.

14. Jordan, T. and Bevan, K. "Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition," J. Exp. Psycho.: Human Percept. Perform., 25, 388-403, 1997.

15. Massaro, D. W. "Speech perception by ear and eye." In B. Dodd & R. Campbell (Eds.), *Hearing By Eye: The Psychology of Lip-Reading*. (pp.53-83). Hillsdale, NJ: Erlbaum, 1987.

16. Massaro, D. W., Cohen, M. M., and Smeele, P. M. "Cross-linguistic comparisons in the integration of visual and auditory speech," Memory Cognit., 23, 113-131, 1995.