Cross cultural studies on audiovisual speech processing: The McGurk effects observed in consonant and vowel perception

著者	Rahmawati Sabrina, Ohgishi Michitaka
journal or publication title	Proceedings of 2011 6th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2011
number	6095406
page range	59-63
year	2011-01-01
URL	http://hdl.handle.net/2297/30109

doi: 10.1109/TSSA.2011.6095406

Cross Cultural Studies on Audiovisual Speech Processing: The McGurk Effects Observed in Consonant and Vowel Perception

Sabrina Rahmawati

School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
sabrina_rahmawati@yahoo.com

Michitaka Ohgishi

College of Science and Engineering
Kanazawa University
Kanazawa, Japan
ohgishi@kenroku.kanazawa-u.ac.jp

Abstract—Twenty-eight University students participated in the experiment on speech perception. The typical McGurk Effects were observed in consonant, vowel and word perception. In addition to consonant condition which had been observed in previous studies, cross-cultural differences can also be found in vowel perception. If the temporal discrepancies between auditory and visual information are large, the McGurk effects decreased.

Keywords—speech perception; the McGurk effect; telecommunication

I. INTRODUCTION

This study is different from the previous ones because in this study we have two speakers from two different country and three nationalities participants.

A. Visual Communication

As the information technology has developed, we can easily communicate with each other through the internet and cellular phones. Recently it has become possible to talk watching other persons' face when we use internet communication media. Human faces convey not only information about the identification and the facial expression of a person but also information about speech.

Before the media of the telecommunication were developed, we could not use visual information such as facial expression and movement of mouths when we talked with other people in other places. But as the media developed, the importance of visual information is increasing.

B. The McGurk Effect

The McGurk effect [1] is a perceptual illusion which demonstrates an interaction between audio information and visual information in speech perception. Because of this illusion, listening to a speech sound while watching lip movements that is corresponding to a different utterance leads to a dramatic alteration of what is actually heard.

The typical McGurk Effect corresponds to consonant condition. In this condition, when an audio information /ba/ is

presented while visual information /ga/ is shown, the subjects are likely to think that they heard /da/ sound.

In cross cultural studies on the McGurk Effect [2, 3], it is found out that speech perception involves the integration of auditory and visual information. The McGurk Effect proved it by showing that mouth movement had some influence on the detection of acoustic speech signal.

In face-to-face conversation, speech perception is affected by various bimodal integration effects. This integration of auditory and visual information affects phonetic identification.

C. Temporal Order (SOA)

When we communicate through the internet, time-lag discrepancies between visual and auditory information sometimes occur. Sometimes the discrepancy causes the information to be not fully delivered so there will be some loss in receiving of essential information. It had become the reason why in the present study the SOA paradigm is employed.

Temporal order judgement (TOJ) task is one of the classic paradigms used by researchers to investigate temporal perception [4].

In a typical TOJ experiment, a pair of stimuli at varying stimuli onset asynchronies (SOAs) is presented, and participants are required to judge which stimulus was presented first (or second).

Multisensory TOJ was employed in this study because there are little number of studies on temporal ordering in speech perception. It is important to get derivation of robust and accurate estimation of people's sensitivity to audiovisual synchrony for some real-world applications such as designing of hearing aids and the derivation of guidelines for satellite telecommunications broadcasting, as well as for the development of new virtual-conferencing technologies.

Another reason why we employed TOJ paradigm is to make participants concentrate on the speaker's mouth movements.

D. Purpose

The main purpose of the current study is to clarify the nature of telecommunication from the viewpoint of the nature of speakers' information. To study the nature of speakers' information, cognitive psychological experiments on the McGurk effects were conducted.

Another purpose of these experiments are to see if the McGurk Effect is also occurs under Vowel and Word condition in addition to Consonant condition as occurred in the typical McGurk effect and to see whether the cross cultural differences occurs in the McGurk effect.

II. METHOD

The experiments were conducted individually.

A. Speakers

The speakers were two male students, a native Japanese master student, 25 years (index ♂) and a native American undergraduate student, 20 years (index ♂).

Picture 1 American Speaker

Picture 2 Japanese Speaker

B. Participants

Total of 28 participants from Kanazawa University community, 6 of American nationality, 11 of Japanese nationality and 11 of Indonesian nationality were recruited.

Most of Indonesian participants have TOEFL score (paper based) around 550 and have been studying Japanese for 6 months. Most of American participants have been studying Japanese for 2 years or more, in their home country and in Japan. Most of Japanese students had never taken such English proficiency test and some who already took the test have TOEIC score around 475.

The Indonesian students are master course students, the American students are undergraduate students and the Japanese students are either master or undergraduate students. These subjects did this experiment voluntarily, without being paid.

The participants' age varies between 18 to 35 years old, with average of 23 years old.

C. Material

There are three kinds of condition used in this experiment: Vowel, Consonant and Word condition. On each condition three stimuli were used: /ba/, /bo/ and /bu/ for Vowel condition, /ba/, /da/ and /ga/ for Consonant condition, and /but/, /duck/ and /gut/ for Word condition.

Each of the stimuli was pronounced by both American and Japanese speaker. Each speaker's face and voice were digitally recorded while pronouncing the syllables.

The three stimuli in each speaker and condition was dubbed with each other, and combined with SOA variation -600 ms, -300 ms, +000 ms, +300 ms and +600 ms, which means in total there are 270 video combinations (2 speakers x 3 conditions x 3 visual stimuli for each condition x 3 audio stimuli x 5 SOA). The minus SOA means that the auditory information precedes the visual information while the plus SOA means vice versa. For example, +600 ms SOA means the visual information precedes the auditory information for 600 ms.

On the final video clips, each stimulus occurred in a 3-sec trial in which the video channel included 1 sec of black frames presented before the tone burst, and 60 frames of the speaker's face. The stimuli itself contains still image and the video clip of the speaker's pronouncing a syllable.

For stimuli with +600 SOA, there were 12 frames of still image, a video clip of the speakers pronouncing syllable, followed with another still image until total stimuli of 60 frames. For stimuli with +300 SOA, there were 21 frames of still image, video of speakers pronouncing syllable, and another still image. For stimuli with +000 SOA, the still image was presented for 30 frames, followed by the syllable, and another still image. In -300 ms and -600 ms SOA, there were different numbers of frames presented before the video clip.

D. Apparatus

A digital video camera (Sony TRV70K) was used to record the speakers' face, and a desktop PC (Dell Dimension 9100)

and a 17 inch CRT monitor (Sony Trinitron) were used to present stimuli.

E. Procedure

There were 270 trials in total, which were divided into three sections of condition. At first they were provided with instruction sheet to read. Before the trial, 10 practice trials of vowel stimuli were presented. While doing the first trial on practice trial, oral instruction was given.

After answering the practice trial, the participant was asked whether they had gotten accustomed to the trials or not. If they had not been accustomed yet, the practice trials were supposed to be repeated. After the participant confirmed that they had gotten accustomed to answering the practice trial, the first condition was presented.

The order of the conditions presentation was 90 trials of Vowel condition, followed by 90 trials of Consonant condition and 90 trials of Word condition. In each section, a break was inserted every 20 trials. If the participant wanted to continue answering the trials when they already had enough break, they had to press the space key.

The participant was seated in front of the monitor. After a stimulus had been presented in the monitor, the participant was supposed to decide which of the audio information or visual information appeared first, before deciding what sound was heard. Questions and options were appeared on the monitor. The questions for Japanese subjects were presented in Japanese, and for American and Indonesian subject in English. To choose if the auditory information appears first, the key "1" on the right side of the keyboard should be pressed. To choose if the visual information appears first, the key "2" should be pressed. After either of the buttons was pressed, there appeared a question of which sound was heard, along with the options of the answer.

For Vowel condition, the answer options were /ba/, /bo/ and /bu/, which means that key "1" was needed to be pressed if the participant think /ba/ syllable had been heard, key "2" for /bo/ syllable and key "3" for /bu/. For Consonant condition, key "1" would be for /ba/, key "2" for /da/ and key "3" for /ga/. For word condition, key "1" for /but/, key "2" for /duck/ and key "3" for /gut/.

III. RESULT AND DISCUSSION

Vowel Condition

ANOVA for Vowel condition

First, a 4-factor analysis of variance, speaker x visual information x auditory information x SOA, was conducted.

There was a mistake occurred in this experiment that a wrong video clip was presented three times. A stimulus of

audio information /ba/, visual information /ba/ and SOA +000 was presented when the stimulus with audio information /da/, Visual information /ba/ and SOA -600 was supposed to be presented. This fault had caused three participant picked answers of /ba/ instead of /da/. These mistakes were ignored because it was not significant (three mistakes out of 28 trials in total).

The main effects of speaker, visual information and auditory information are significant ($F[1, 25] = 25.175$, $p < .001$; $F[2, 50] = 3.479$, $p < .05$; $F[2, 50] = 31.980$, $p < .001$) while the main effect of SOA is not significant.

Visual information x auditory information is significant ($F[4, 100] = 5.236$, $p < .001$). Further analysis showed that there was a significant difference between three visual stimuli under the presentation of auditory information /bo/ as shown in Figure 1.

Figure 1 The mean number of accurate responses (Maximum is 1) in the vowel condition for visual information x auditory information

This result suggests that the McGurk effects occurred under the vowel presentation as expected. Japanese pronunciation of /bo/ is in intermediate between their pronunciation of /ba/ and /bu/, so we can predict that the McGurk Effect might appear while audio information /ba/ is combined with visual information /bu/. With this combination, the subject might guess that they heard /bo/ sound. But it turned out that the experiment result for this condition was not significant.

Figure 2 The mean number of accurate responses (Maximum is 5) in the Vowel condition for speaker's nationality x participant's nationality

Since the main effect of SOA is not significant, a four factor ANOVA, participants' nationality x speakers' nationality x visual information x auditory information was conducted by eliminating the factor of SOA.

An interaction of participants' nationality x speakers' nationality was significant ($F[2, 23] = 6.294, p < .01$), showing that Japanese participants are better at discriminating vowels than the other two groups of the participants. This interaction can be seen in Figure 2.

An interaction of participants nationality x auditory information was significant ($F[4, 46] = 3.526, p < .05$), showing that there is a significant difference between three group of participants under the presentation of the auditory stimuli /bo/.

Figure 3 The mean number of accurate responses (Maximum is 5) in the vowel condition for Participant's nationality x Auditory information

Figure 3 shows the relationship between participant's nationality and auditory information. The x-axis of the figure

shows the auditory information, while the color differences distinguish the participants' nationalities.

From Figure 3, we can see that Indonesian participants made more mistakes in deciding which sound they had heard than Japanese and American participants. The reason why such thing occurred was probably because most of Indonesian participants had just been learning Japanese for 3-4 months, so they are not used to Japanese pronunciation yet, while Japanese speakers are certainly used to their native pronunciation and most of the American participants had been learning Japanese for 2 years or more.

The other thing that can be concluded from Figure 3 is that Indonesian participants made more mistakes in deciding which sound they had heard while listening to auditory information /bo/ than while listening to /bu/ or /ba/. The reason was probably because most of Indonesian participants are not used to /bo/ pronunciation that was pronounced by American and Japanese speaker, as there is only one way to pronounce /bo/ in Indonesian language which are different from Japanese pronunciation of /bo/. For Indonesian participants, the Japanese speaker's pronunciation of /bo/ is somehow similar to /bu/ pronunciation in Indonesian language. In addition, as explained before, Japanese pronunciation of /bo/ is in intermediate between their pronunciation of /ba/ and /bu/. The American participants are better in judging these syllables because there are several kinds of /bo/ and /bu/ pronunciation in English, so that American subjects are used to the pronunciation used in this experiment.

Consonant Condition

Figure 4 The mean number of correct responses (Maximum is 1) in Consonant condition for Visual information x Auditory information x SOA

A 4-factor analysis of variance, speaker's nationality x visual information x auditory information x SOA, was conducted for Consonant condition. The main effect was only in the auditory factor ($F[2, 54] = 4.500, p < .05$). An

interaction was significant ($F[16, 432] = 1.696, p < .05$). This interaction shows that the typical McGurk Effect occurred in this condition (visual information /ga/, audio information /ba/, SOA +000) (see Figure 4).

The results of Consonant condition differ from Vowel condition, that is, the factor of the participants' nationality did not affect the results of Consonant condition very much while it did have significant effect in Vowel condition. It is because vowel pronunciation is different in every language but consonant pronunciation does not really much different with each other.

Word Condition

The ANOVA shows that there is a significant interaction between visual and auditory information ($F[4, 108] = 3.293, p < .05$). This results indicate that there is a tendency that the word "but" was heard as "duck" when visual information was "gut", suggesting that we can observe the McGurk effect in the word condition.

Figure 5 The mean number of accurate responses (Maximum is 1) in the Word condition for visual information x auditory information

IV. CONCLUSION

The present study confirmed that the McGurk effects were observed in the perception of vowels and words as well as consonants. Cross-cultural differences were also observed only in vowel perception. If the temporal discrepancies between auditory and visual information are large, the McGurk effects decreased. These findings contribute to designing telecommunication systems.

V. REFERENCES

- [1] H. McGurk., and J. MacDonald, "Hearing lips and seeing voices," *Nature*, 1976, pp. 264, 746-748.
- [2] K. Sekiyama, "Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subject," *Perception & Psychophysics*, 1997, pp. 59, 73-80.
- [3] H. Traunmüller, "Factors affecting visual influence on heard vowel roundedness : Web experiments with Sweds and Turks," *Proceedings, FONETIC 2009*, 2009.
- [4] M. Zmpini, D. I. Shore, and C. Spence, "Audiovisual temporal order judgments," *Experimental Brain Research*, 2003, pp. 152, 198-210.