Individual differences observed in the McGurk effect

Sumi Shigeno

Abstract

The effects of individual differences on multimodal speech perception were investigated using the McGurk effect paradigm. Two successive experiments were conducted with an interval of three weeks. Two hundred and twenty undergraduate students participated in the experiments, which were carried out under conditions similar to a daily-life audiovisual environment. The experiments examined the amount of visual influence on the perception of audiovisually incompatible syllables and consistency of judgments. The results showed that the distribution of judgments of the participants is wide—ranging from the participants who were strongly influenced to those who were not at all—and that judgment tendencies were consistent. This suggests that factors such as individual differences are very important when conducting McGurk effect experiments and interpreting the results.

Key words: McGurk effect, individual differences, distribution of judgments, speech processing ability, audiovisual information

The perception of phonemes in everyday life plays a very important role in smooth communication behavior. Usually we perceive phonemes through audition but according to the situation we make use of various sensory information other than audition. For example, we use a partner's gestures or lip movements (i.e., articulatory movement) etc. McGurk and MacDonald (1976) found that observers presented with audiovisual stimuli constructed by dubbing a syllable such as /ba/ to the lip movements of the same speaker producing a syllable such as /ga/ were likely to believe that the speaker uttered a fused syllable, such as /da/. On the other hand, observers presented an auditory /ga/ with a visual /ba/ were likely to report the combination syllable /bga/. This “McGurk effect” occurs robustly even when subjects are aware of an incompatibility between the auditory and visual components of consonants (Repp, Manuel, Liberman, & Studdert-Kennedy, 1983).

Since the McGurk effect contains many important suggestions for considering phoneme information processing, various researchers have carried out experiments using the McGurk effect paradigm (Green & Geldeman, 1995; Green, Kuhl, & Meltsoff, 1988; Massaro, Cohen, & Gesi, 1993; Shigeno, 1990, 1993, 1999, 2000, 2002; Smeele, Sittig, & van Heuven 1992; Summerfield, 1987). In McGurk effect experiments for Japanese listeners, it has been reported that some people show strong fusion responses, but others do not (Shigeno, 1996). Most psychological experiments hypothesize participants’ responses as being under the normal distribution and individual differences are regarded as errors. Therefore, if the distribution is not normal but uniform or polarized, experimental results lead to wrong conclusions when derived from a small number of participants. Thus, it is necessary to examine individual differences in the occurrence of the McGurk effect. However, so far, few experiments have been done focusing on such individual differences.

Watson, Qiu, Chamberlain, and Li (1996) examined 90 healthy university students to see if individual differences exist in the relationship between the listening skills of auditory speech stimuli and lip-reading ability for words and sentences. Consequently, it was confirmed that listening performance and lip-reading ability show a correlation coefficient (r) between 0.43 and 0.52, and that those with high lip-reading ability have high listening skills. The results indicate that differences exist in the superiority of linguistic processing ability, independent of the type of sensation. Furthermore, female participants showed significantly better performance in lip reading than male participants did.

Although Watson et al. (1996) showed the existence of individual differences in the ability of audiovisual information processing; the auditory and visual experiments were conducted separately. Therefore, from their results we could not confirm individual differences in visual influences on auditory perception when bimodal linguistic processing was necessary. As seen in the example of visual capture, visual superiority is widely recognized (for example, a ventriloquism effect in sound localization).

The purpose of this research was to investigate whether individual difference can be observed in the occurrence of McGurk effect and how strong this effect is. In addition, since it is generally said that the McGurk effect is robust, the degree of consistency was also examined. In the laboratory, experiments were carried out under strictly-controlled conditions. Therefore, adaptability to the experimental situation will be involved in the performance of trials in addition to perceptual ability.

The experiment was carried out in a large classroom thought to be similar to conditions in our daily life. In addition, a larger sample size comprising hundreds of undergraduate students was used for the experiments.

Method

Participants

Two hundred and twenty Japanese students (48 male and 172 female freshmen) participated in the experiment for a partial course credit.

Stimuli

Nine kinds of audiovisual stimuli were used (Table 1). The face of a Japanese female speaker was presented on an 8 mm videotape recorder (SONY, ACP-80). She clearly pronounced the Japanese CV-syllables /ba/, /da/, and /ga/. The best auditory stimuli were dubbed on to the best visual stimuli. The timing of the dubbing was accommodated to give the most natural impression possible. The duration of one syllable was about 370-400 ms.

Table 1. Audiovisual stimuli used for the experiment.

Auditory	Visual	Notation
	ba	AbVb
ba	da	AbVd
	ga	AbVg
	ba	AdVb
da	da	AdVd
	ga	AdVg
	ba	AgVb
ga	da	AgVd
	ga	AgVg

The speaker’s voice was recorded in the PCM format on a track for post-recording. Then, the speaker’s shot was recorded from the neck to the head top with the same 8 mm video recorder. The three syllables were dubbed on to the 1/2-inch S-VHS tape, respectively. Another nine participants received the voice listening test to select the clearest sounds and the three syllables were chosen. Then, the sound stimulus selected as the most distinct was synchronized with the video stimulus. The acoustic stimuli were dubbed on to the visual stimuli at an accuracy of 33.3 ms. The dubbing was done again and again until two listeners felt that the synchronized stimuli were natural. Nine kinds of combinations for the experiment were determined.

Nine audiovisual stimuli were presented in random order with two dummy stimuli adding to the beginning and the end of the stimulus series. Then, a total of 94 trials were played to the participants. A face appeared every 6 s and after the onset of the face 0.9 s the sound was presented. A face was presented for 1.8 s. The interval from the end of a face to the onset of the next face was 4.2 s and the participants were required to write down the syllable that they heard as it was.

Procedure

The experiment was conducted in a large classroom capable of accommodating about 300. The stimulus was reproduced with a video deck (SONY, VPH 2042 QJ), and the speaker's face was presented to the participants on a screen (3 m long × width 3 m). Voices were presented from loudspeakers (Victor, SMBM-330) on both sides of the screen at a comfortable listening level (65-75 dB A). The observation distance of the participant was determined by seating position, which ranged from about 3 m to about 15 m. The sound pressure level was 75 dB (A) (the seat in the center of the front row) from 66 dB (A) (the seat at the right end of the last row). The speaker's face was presented to the participants on a screen (3 m × 3 m) and the voice was presented through loudspeakers. The interval between the two experiments was three weeks.

The audiovisual stimuli were presented to the participants, who were strictly instructed to keep their eyes on the screen except for when writing down their responses. The participants were asked to write down what they heard, not what they saw, looking at the lip movements of the speaker. The response was written in Roman letters such as ba, da, ga. Even if they did not exist in Japanese, they were required to fill in what they heard correctly, for example, bga.

The first experiment was conducted on November 21 in 1995 and the second experiment on December 12 in the same year, after three weeks. A few trials were performed as practice before the first experiment. In each of the two experiments, ten judgments were obtained per participant for nine types of audiovisual stimuli, respectively. Although the seating position was free, most of the participants were seated in the same seat in the two experiments.

Results

The results of AbVg and AbVd were summed because their response tendencies were very similar. The judgments in the first and second experiments were compared to see how the occurrence of the McGurk effect changed through the time interval of 3 weeks. For comparison, we calculated the correlation coefficient of the rate of identifying the auditory component for each stimulus between the first and second experiments. In addition, to see the individual difference caused by the McGurk effect, the distribution of the number of participants was examined regarding whether the distribution was uniform or not, in the fusion response and in the combination response, respectively.

1. Comparison between the first and the second experiments

Table 2. The response categories and the percentage for each audiovisual stimulus, and the correlation coefficients between the first and second experiment.

stimulus		Response				*r*
Auditory	Visual	ba	da	ga	others	
ba	ba	88.8	0.5	0.1	10.6(7.1)	0.67
		87.4	0.7	0.3	11.5(8.1)	
	da	26	53.0	2.6	18.4(11.0)	0.77
		28.2	52.0	2.0	17.9(11.9)	
	ga	21.4	44.2	5.6	28.2(13.5)	0.73
		23.7	48.3	4.0	24.0(11.5)	
da	ba	19.2	41.1	1.8	37.9(34.0)	0.78
		13.8	42.5	2.0	41.6(42.6)	
	da	0.7	88.9	5.1	5.4(3.5)	0.67
		0.5	92.3	3.0	4.1(3.1)	
	ga	0.4	79.5	9.6	10.4(3.7)	0.65
		0.5	86.0	4.7	8.9(4.1)	
ga	ba	5.2	0.3	48.8	45.7(42.6)	0.71
		3.2	0.4	43.6	52.9(52.2)	
	da	0.1	1.9	88.9	9.1(7.0)	0.63
		0.3	1.0	86.5	12.3(10.7)	
	ga	0.0	0.8	93.9	5.3(2.6)	0.53
		0.0	1.1	94.0	4.9(2.6)	

The upper row shows the first experiment and the lower row shows the second experiment.
The figures in the parentheses show the percentages of combination response.

Table 2 shows the identification results in the nine kinds of stimuli. The upper level indicates the response category and the rate of judgments for the auditory component in the first experiment and the lower level indicate those in the second experiment. The figures were all averaged among the participants. In the first and second experiments, there was almost no difference in judgment rate for each category. The rightmost column in Table 2 showed the correlation coefficients between the first and second judgment rates for the auditory component shown for each stimulus. The correlation coefficient was in the range of 0.53 to 0.78 and a high positive correlation was recognized. Figure 1 shows the correlation diagram of the number of “da” responses for AbVd and AbVg (noted in the text as AbVd,g) between the first and second experiment. In other words, although there are a few exceptions, most of the participants’ judgments did not change, not even after three weeks. The seating position effect was not significant.

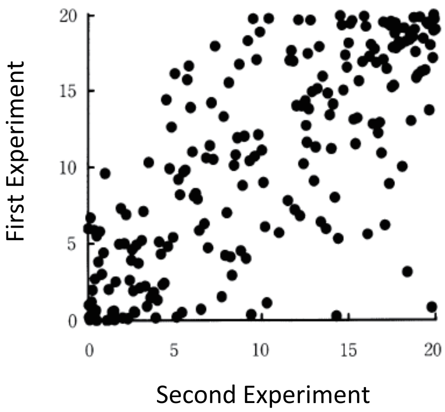

Fig. 1. The correlational diagram of the number of "da" responses for AbVd and AbVg (noted in the text as AbVdg) between the first and second experiment.

To investigate how consistently the fusion response and complex response were observed, the number of fusion response /da/ in the case of AbVd and AbVg and the number of combination response in the case of AdVb and AgVb were compared between the first and the second experiment. Since the reaction tendency is very similar between visual /da/ and visual /ga/, the number of “da” responses in the case of AbVd and AbVg was summed. When comparing the first and the second experiment, the correlation coefficient was 0.76, which shows a high positive correlation. Similarly, when the auditory condition was /da/ and /ga/, the total number of combination responses in the case of AdVb and AgVb was summed up and compared between the first and second experiment. The correlation coefficient was 0.78, also showing a high positive correlation.

2. Response category and identification rate

From Table 2, it can be seen that the percentage of identification as the same category as the auditory information (the shaded parts in the table) tended to be higher in the congruent condition than in the incongruent condition, especially when the auditory information was /ba/ and visual information was /da/ or / ga / (AbVd, AbVg), with identification as “ba” decreasing greatly and increasing as a “da” response. The x^2-test was performed on the frequency of identification to “da” and other than “da” between AbVb and AbVd. Consequently, the “da” response increased significantly, $x^2 = 3103.53$, $df = 1$, $N = 13200$, $p < .005$. On the other hand, in the case of AdVb and AgVb, the complex responses increased in which two or more consonants such as /bga/ were combined ($x^2 = 3814.38$, $df = 1$, $N = 17600$, $p < .005$).

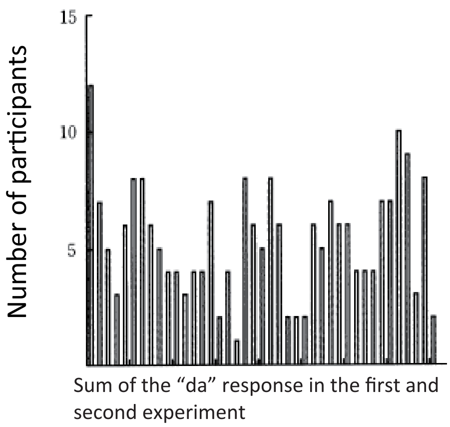

Fig. 2. Distribution of the number of responses judged as "da" for AbVd and AbVg.

3. Individual difference in the McGurk effect

Figure 2 shows the distribution of the number of responses of “da” to AbVd and AbVg. The correlation between the first time and the second experiment was high; therefore, the data of the two times were summed. The Kolmogorov-Smirnov test was conducted to confirm the fitness to the uniformity of distribution, using the number of “da” responses. Consequently, the uniformity of distribution to the “da” responses was confirmed, $D = 0.05$, $N = 220$, $p > 0.2$.

Figure 3 shows the distribution of the number of combination responses to AdVb and AgVb. About 1/4 of the participants showed no combination responses at all. The Kolmogorov-Smirnov test confirmed that bias was generated to a significant degree in the distribution of combination responses, $D = 0.2$, $N = 220$, $\rho < 0.01$. It was further confirmed that when the Kolmogorov-Smirnov test was conducted, excepting participants who did not show combination responses, $D = 0.088$, $N = 170$, $p > 0.10$. The result indicates that the distribution of those who show few combination responses is uniform.

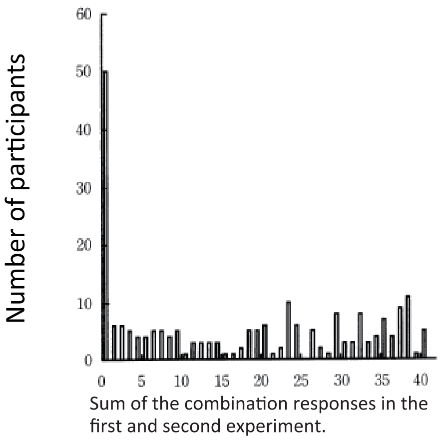

Fig. 3. Distribution of the number of "combination responses" for AbVd and AbVg.

Discussion

There were few differences in the identification percentages for each category between the first and second experiment. Both fused responses and combination responses showed high positive correlations between the two experiments: correlation coefficients were in the range of 0.53 to 0.78. On the other hand, even when auditory information was /da/ or /ga/ with visual information /ba/ (AdVb, AgVb), the rate of identification as auditory /ba/ decreased.

When the experiments were conducted in the laboratory, the rate of identifying /da/ for AbVg, for example, was 56.2% (average of 24 participants) (Shigeno, 1993). On the other hand, in the present experiments the fusion responses were 44.2% (the first experiment) and 48.3% (the second experiment), which were lower than that obtained in the laboratory experiment. Although the McGurk effect is not so strong, it can be said that when we watch movies or TV, we experience the McGurk effect.

As for individual differences in the McGurk effect, there seem to be three groups: those who almost always perceive the illusion, those who almost never perceive the illusion, and those who partially (i.e., sometimes) perceive the illusion. This seems to be similar to cases of absolute pitch possessors. The correlated coefficients of the McGurk effect between two temporally separated experiments differ according to the situation. Although we obtained the correlated coefficient ($r = .78$) from a three-week separation, Mallick, Magnotti, and Beauchamp (2015) obtained the correlated coefficient ($r = .91$) from a one-year separation. In general, Japanese people showed the McGurk effect more weakly than English native speakers did, often below 50 %.

The reason Japanese people showed a less significant McGurk effect might be because Japanese has only five vowels and a smaller number of consonants than English has. Therefore, Japanese participants do not need much visual information to understand what a conversation partner says. They focused on articulatory movements because they were instructed by the experimenter to do so; however, that is not what they typically do in their daily lives. Japanese people are sensitive to auditory information more than North American people are, not only in the audiovisual perception of speech but also in the audiovisual perception of emotion (Shigeno, 1998). The degree of learning and the ability to perform a task in a laboratory might be one basic factor influencing the use of visual information in speech perception.

Both fusion and combination responses should be tested to explore the mechanism of the McGurk effect and the individual differences of perceivers. Therefore, Table 2 shows the correlation coefficients between two experiments for all stimulus pairs (both congruent and incongruent between audition and vision). Consequently, there was little difference between the correlation coefficients of fusion responses and those of combination responses. However, having looked into the combination responses in more detail, it was found that a quarter of all the participants showed no combination responses and the rest were scattered broadly. This tendency was quite different from that of the fusion responses (Figures 2-3).

Conclusion

The present results clearly show that the distribution of the participants regarding fusion responses ranged from those who were significantly susceptible to the influence of visual information to those who were not at all. On the other hand, the distribution of the participants regarding combination responses showed that the distribution can be divided into two classes: those who occurred and those who did not occur at all. In the case of those who occurred, the distribution was found to be uniform. In addition, we confirmed that such reaction tendencies were almost constant even after three weeks, and that the McGurk effect has great reproductivity. Furthermore, we found the McGurk effect occurred in the perceptual situation in a manner similar everyday life even if the situation was not strictly controlled such as in the laboratory. Thus, the McGurk effect is very robust. In conclusion, the present results indicate that notable individual differences exist regarding the extent of the McGurk effect.

References

Green, K. P., & Gerdeman, A. (1995). Crossmodal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1409-1426.

Green, K. P., Kuhl, P. K., & Meltzoff, A. N. (1988). Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. *Journal of the Acoustical Society of America*, *84,* S155.

Mallick, D. B., Magnotti, J.F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review, 22*, 1299-1307.

Massaro, D. W., Cohen, M. M., & Gesi, A. T. (1993). Long-term training, transfer, and retention in learning to lipread. *Perception & Psychophysics, 53*, 549-562.

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

Repp, B. H., Manuel, S. Y., Liberman, A. M., & Studdert-Kennedy, M. (1983). Exploring the "McGurk effect." *Paper presented to the 24th annual meeting of the Psychonomic Society*, *358* (*abstract*), San Diego, California.

Shigeno, S. (1990). On the effect of visual information upon the perception of voiced stop consonants. *Proceedings of the Spring Meeting of the Acoustical Society of Japan*, Tokyo, The Acoustical Society of Japan, 325-326. (In Japanese.)

Shigeno, S. (1993). Effects of stimulus context on the integration of auditory and visual information. *Proceedings of the Autumn Meeting of the Acoustical Society of Japan*, Kofu, The Acoustical Society of Japan, 385-386. (In Japanese.)

Shigeno, S. (1996). Effects of stimulus context on the integration of auditory and visual information. *Proceedings of the Autumn Meeting of the Acoustical Society of Japan*, Okayama, The Acoustical Society of Japan, 347-348. (In Japanese.)

Shigeno, S. (1998). Cultural similarities and differences in the recognition of audio-visual speech stimuli. Fifth International Conference on Spoken Language Processing., *Proceedings of International Congress on Spoken Language Processing, Vol.2,* pp. 281-284, Sydney.

Shigeno, S. (1999). The use of short-term memory in the discrimination of visually-influenced CV-syllables. *The Japanese Journal of Psychonomic Science, 18(1)*, 62-72.

Shigeno, S. (2000). Influence of vowel context on the audio-visual speech perception of voiced stop consonants. *Japanese Psychological Research, 42(3)*, 155-167.

Shigeno, S. (2002). Anchoring effects in audiovisual speech perception. *The Journal of the Acoustical Society of America, 111(6)*, 2853-2861.

Smeele, P. M. T., Sittig, A C., van Heuven, V. J. (1992). Intelligibility of audio-visually desynchronized speech: Asymmetrical effect of phoneme position. *Proceedings of 2nd International Congress on Spoken Language Processing*, pp. 65-68.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. *Hearing by eye: The psychology of lip-reading* (edited by Dodd, B. and Campbell, R.), Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 3-51.

Watson, C. S., Qui, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: confirmation of a modality-independent source of individual differences in speech recognition. *Journal of the Acoustical Society of America, 100*, 1153-1162.