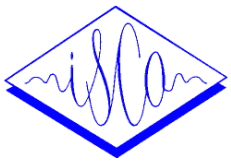

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

THE JAPANESE MCGURK EFFECT: THE ROLE OF LINGUISTIC AND CULTURAL FACTORS IN AUDITORY-VISUAL SPEECH PERCEPTION

Denis Burnham and Sheila Keane

School of Psychology, University of NSW, Sydney, 2052, Australia.

Tel. +61 2 9385 30 25, Fax: +61 2 9385 36 41, email: d.burnham@unsw.edu.au

ABSTRACT

Humans perceive auditory [b] dubbed onto visual [g] as [d] or [ð], as in 'them'. When this is presented with an [a] vowel, "th" responses tend to dominate, while in an [i] vowel context, "d" responses dominate. This "McGurk effect" was used here to investigate humans' integration of auditory and visual speech information. In Experiment 1, Australian English and Japanese subjects viewed McGurk stimuli presented by an English speaker. Despite the phonological irrelevance of [ð] in Japanese, both English and Japanese subjects showed the [a]/[i] x "d" / "th" crossover effect, suggesting a strong language-general (phonetic) influence in auditory-visual integration. Experiment 2 used a Japanese speaker. Here the incidence of "th" responses for Japanese subjects was severely dampened, showing that expectancies based on native phonology may overlay basic phonetic auditory-visual integration.

1. INTRODUCTION

The McGurk effect (McGurk & McDonald, 1976), in which humans perceive auditory [b] dubbed onto visual [g] as [d] or [ð] (as in 'them'), appears to be affected by both cultural and linguistic factors (Burnham, 1997). With regard to cultural factors, Sekiyama suggests that Japanese speakers rely less on visual speech information than do English speakers, partly because Japanese listeners do not typically look at a speaker's face in conversation (Sekiyama, 1994; Sekiyama & Tohkura, 1991). A second cultural factor is that the use of visual information is said to be increased when viewing a foreign speaker, both in Japanese/English (Sekiyama & Tohkura, 1993) and other (deGelder & Vroomen, 1992) inter-lingual situations. Both of these factors are investigated here.

With regard to linguistic factors, two important facts, one phonetic and one phonological, are used in the studies here. First, due to phonetic factors, the relative frequency of "d" and "th" responses in the auditory [b] visual [g] (A[b]V[g]) effect depends on the vowel context: English subjects' fusion responses to A[ba]V[ga] are 85% "tha" and 15% "da", while for A[bi]V[gi] they are 40% "thi" and 60% "di" (Green, 1996). Second, Japanese phonology contains [b], [g], and [d], but not [ð].

Here Australian English speakers and Japanese speakers at three levels of English proficiency, Beginner, Intermediate, Advanced, were tested for their perception of A[b]V[g] in two vowel contexts, [a] and [i]. In each, subjects were presented with auditory-only (AO), visual-only (VO), matching auditory-visual (AV), and mismatching AV trials. Experiment 1 used a native Australian English stimulus person, and Experiment 2, a native Japanese stimulus person. It was expected firstly that if the McGurk effect occurs phonetically before language specific processing is engaged, then the phonetic effect of vowel context on the distribution of "d" and "th" responses should be apparent in both English and Japanese speakers, irrespective of their level of experience with English (and [ð]). Secondly it was expected that there should be less visual influence on speech perception for (a) Japanese than Australian English perceivers, with influence increasing as a function of exposure to English language, and (b) perceivers of a foreign compared with a native speaker.

2. EXPERIMENT 1: ENGLISH SPEAKER

In Experiment 1, stimuli were presented by a native Australian English speaker, chosen from three possible speakers on the basis of intelligibility.

2.1. Method

Subjects: Sixteen native adult Australian English speakers, and 48 native adult Japanese speakers were tested. In the Japanese group there were 16 at each of three levels of English proficiency. The Beginners had been in Australia for a mean duration of 6 weeks at the time of testing, the Intermediate group for 3 months, and the Advanced group for 3 years. The Australian subjects were University of NSW students.

Stimulus Construction: Stimuli were prepared by videorecording the head and shoulders of a female native Australian English speaker. For the consonant clusters, the schwa vowel, /ə/, was inserted between consonants, eg [bəga:]. In the test situation the visual stimuli were produced from the original videotapes, while the auditory stimuli were digitised versions (Kay CSL 4500 package) of the videotape productions, played from disk. Four exemplars of each visual stimulus, and three exemplars of each auditory stimulus were used to ensure acoustic variability and phonetic invariance. The CAVE (Computerised Auditory-Visual Experiment) package, developed by Burnham et al. (1997), was used for

presenting stimuli and collecting speeded responses. In this, auditory-visual stimuli are created on-line at the time of testing each subject. Based on pre-programmed CAVE software, the sound played from disk on a particular trial is triggered by the original sound from the audio channel of the videotape. In this manner mismatching AV trials are produced. For matching AV trials the same dubbing procedure is used to ensure uniformity. On AO trials the stimulus person's face is motionless, and an auditory stimulus is triggered from disk by a tone pre-recorded on the second audio channel of the videotape. On VO trials the auditory stimulus from the videotape cues the computer to play 'silence' from disk and so just the stimulus person's face and lip movements without sound are presented.

This CAVE package was used to produce stimuli consisting of AO, VO, and matching AV presentations of the syllables [b*V*], [g*V*], [d*V*], [ð*V*], [bg*V*] and [gb*V*] and mismatching presentations of A[b*V*]V[g*V*] and A[g*V*]V[b*V*] (*V* = vowel, [a:] or [i:]). Each trial lasted 4 secs, with 1 sec of black background intervening between trials. For VO and AV trials this consisted of 1 sec of a motionless face, about 1 sec of articulation, and 2 seconds of neutral expression. For the AO trials, the speaker's motionless face was presented for 4 seconds overdubbed with a speech sound.

Stimulus Trials: Each vowel condition was presented separately. In each there were 18 practice trials, one of each of the 6 syllables [b*V*], [g*V*], [d*V*], [ð*V*], [bg*V*] and [gb*V*], in each of the three modes, AO, VO, and AV. Practice trial results were used only to eliminate the data of subjects who responded too slowly or inaccurately. For each condition, there were then two 32-trial test blocks. Each block consisted of exactly the same trial types with trial presentation order varied between blocks, and test block sequence counterbalanced between subjects. In each block there were 2 AO, 2 VO and 2 AV matching presentations of each of [b*V*], [g*V*], [d*V*] and [ð*V*], and 4 each of mismatching A[b*V*]V[g*V*] (McGurk stimulus), and A[g*V*]V[b*V*] (combination stimulus).

Apparatus and Procedure: Subjects were tested individually in a sound-attenuated room, seated in front of a monitor connected to the videorecorder and computer in the control room. A response pad placed in front of the monitor had a central "ready" key with six response buttons (labelled 'b', 'g', 'd', 'th', 'bg' and 'gb') arranged in a semicircle around it. A reward light output from the computer was attached to the left side of the monitor. This flashed when subjects responded correctly (but only during practice trials). An error buzzer output from the computer sounded to inform the subjects and experimenter of any failure to respond appropriately, eg, if response times were too long.

In each condition there was one block of practice trials followed by two test trial blocks. Subjects pressed the ready key to start each trial. On stimulus presentation the subject was required to respond as quickly and accurately as possible, by pressing the response button which "best matched the consonant sound the speaker used". If the subject failed to respond within the maximum limit of 2.2 secs, or took their finger off the ready button prior to the onset of the sound, the error buzzer sounded and a null trial was recorded. Results were recorded on-line and collated by the CAVE package. Testing lasted approximately 30 mins.

2.2 Results

As can be seen in Figure 1 Japanese speakers made more errors than English speakers on [ð] trials in AV, $F(1,60) = 11.16$, AO, $F(1,60) = 24.26$, and VO, $F(1,60) = 7.39$ trials. However, their performance improved linearly in AV, $F(1,60) = 8.03$, and AO trials, $F(1,60) = 10.04$, and quadratically in VO trials, $F(1,60) = 7.43$ as a function of English language experience. In addition, Figure 1 shows the number of cluster responses, "bg" or "gb", to A[g]V[b]. English speakers gave more of these than did Japanese speakers, $F(1,60) = 6.13$, possibly due to the lack of clusters in Japanese.

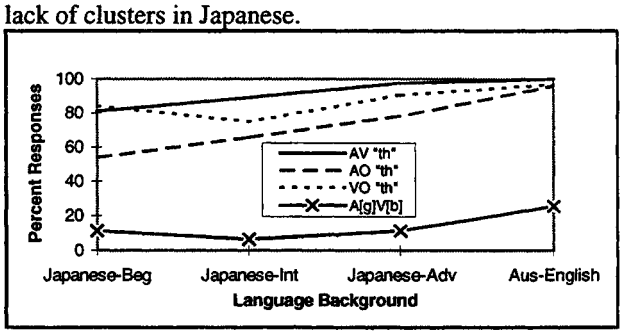

Figure 1. Percent "th" responses on AV, AO, & VO [ð] trials, and of cluster responses on A[g]V[b] trials.

So, in accord with their native phonology, Japanese subjects show a bias toward perceiving [d] for [ð]. What happens when A[b]V[g] is presented? Is there a phonetic effect of the vowel context despite their bias towards perceiving [d] for [ð]? Analysis of the data presented in Figure 2 revealed that Japanese subjects gave more "d" than "th" fusion responses compared to English subjects, $F(1,60) = 31.31$, again indicating their phonological bias. In addition, there was the expected effect of vowel context, [a]/[i], on "d" / "th" responses, and this was statistically equivalent for all four groups. That is there was no significant interaction of this effect with language group, even though a number of post-hoc tests were conducted in order to check this. Thus the vowel context effect occurred for all four language groups, although somewhat differently in some. For example, the Intermediate group increased both "d" and "th" responses from the [a] to [i] vowel context, but nevertheless, in accord with predictions, the increase for "d" was greater than that for "th".

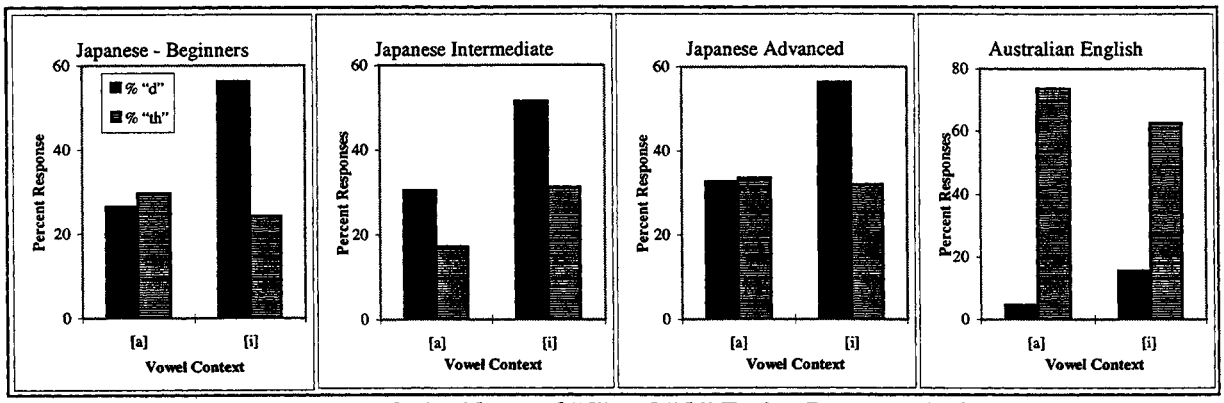

Figure 2. Effect of Vowel Context on the Incidence of "d" and "th" Fusion Responses in Experiment 1

2.3. Discussion

Japanese speakers have a phonological bias to perceive "d" in situations where "th" is correct. Nevertheless, their pattern of "d" and "th" responses in [a] and [i] vowel contexts follows the same pattern as that for English speakers, suggesting that the same phonetic effects are at play in both language groups irrespective of phonological constraints. Nevertheless there were phonological effects: there were less fusions by Japanese subjects inexperienced in English (Beginners and Intermediate) than by Advanced Japanese subjects and native English speakers. This suggests that while there may be a reduced influence of visual speech information for Japanese perceivers, this cultural/linguistic influence can be ameliorated by English language exposure. In addition to these cultural influences, there are phonological/phonetic influences. All three Japanese groups gave more fusion responses (approximately 25%) in the [i] vowel than in the [a] vowel context, of which the majority were "d" responses, whereas for English speakers the incidence of fusions was equivalent in the two vowel conditions. Thus the Japanese McGurk effect appears to be the result of both cultural and phonological factors, for all previous studies with Japanese subjects appear to have involved an [a] vowel.

3. EXPERIMENT 2: JAPANESE SPEAKER

As in Experiment 1 the stimulus person was required to produce various speech sounds, including [ð], which is not native to Japanese. Thus the Japanese person had to be chosen carefully so that - the person was perceived as a native Japanese speaker, fusion responses were obtained, and the person was comprehensible.

3.1 Pilot Studies

Three native Japanese speakers, Sachiko (lived in Australia 20 years), Megumi, and Izumi (both lived in Australia 5 months) were videotaped. In each of two pilot studies perceivers watched videotapes and were asked to make one of 6 written responses, 'b', 'g', 'd', 'th', 'bg' or 'gb'. In the first pilot , 7 native Australian English speakers and 3 native Japanese speakers were tested with 2 A[g], 2 AV[g], 3 A[ð],3 AV[ð], and 6 A[b]V[g] presentations by each speaker both with an [a] vowel, and with an [i] vowel. It was noticed that the overall number of "th" responses to A[b]V[g] (Table 1) was lower than in Experiment 1 (see Figure 2). As we thought this may have been due to contrast with the Japanese person's prominent tongue protrusions for [ð] productions, in the second pilot subjects were presented with 2 A[g], 3 AV[g], 3 A[b], 3 AV[b], and 6 A[b]V[g] (and no [ð] stimuli) by each speaker in each vowel condition. The expected increase of "d" and decrease of "th" responses in shifting from [a] to [i] vowel contexts was obtained for both Australian and Japanese perceivers, and for Australian perceivers both *with* (Pilot 1), and *without* (Pilot 2) productions of [ð] in the blocks (see Table 1).

Table 1: Percent "d" & "th" fusions in Pilot Experiments

Subjects	Vowel	"d"	"th"
Aus	[a]	72.7	27.3
(Pilot 1)	[i]	81.1	18.9
Aus	[a]	58.0	42.0
(Pilot 2)	[i]	76.2	23.8
Jap'ese	[a]	62.5	37.5
(Pilot 1)	[i]	92.1	7.9

Both Australian and the Japanese subjects rated Sachiko to be very anglicised, so it was decided to use either Megumi or Izumi in Experiment 2. Table 2 shows the distribution of fusion responses were more even across [a] and [i] conditions for Izumi than for Megumi. In addition, inspection of correct responses in non-fusion trials (see Table 3) reveals subjects better comprehended Izumi than Megumi. Izumi was selected to be the stimulus person.

Table 2: Percent fusion ("d"/"th") responses in Pilots

Subjects	Vowel	Megumi	Sachiko	Izumi
Aus	[a]	38.0	40.5	57.0
(Pilot 1)	[i]	88.0	62.0	55.0
Aus	[a]	56.0	39.4	56.0
(Pilot 2)	[i]	81.8	80.3	86.4
Jap'ese	[a]	55.5	33.3	44.4
(Pilot 1)	[i]	66.6	83.3	61.0

Table 3: % correct responses on non-fusion trials

Subjects	Megumi	Sachiko	Izumi
Aus (Pilot 1)	73.6	98.6	80.7
Aus (Pilot 2)	75.0	89.0	80.6
Jap'ese (Pilot 1)	51.6	78.3	61.6

3.2 Method

The method mirrored that in Experiment 1 except that a Japanese speaker was used. As in Experiment 1, 16 native Australian English speakers, and 48 Japanese (16 Beginner, 16 Intermediate, and 16 Advanced speakers of English) will be tested. To date the 16 Beginners (mean time in Australia = 4.5 weeks) have been tested.

3.3 Results

There were more non-fusion (78%) than fusion (20%) responses in the [a] vowel condition, and more fusion (89%) than non-fusion (9%) responses in the [i] vowel condition, F (1,15) = 160.6. As shown in Figure 3, there were many more "d" responses in the [a] than the [i] vowel condition, but the opposite difference for "th" responses was minuscule - 0.78% fusions for [a], and 0% for [i], F(1,15)=176.3.

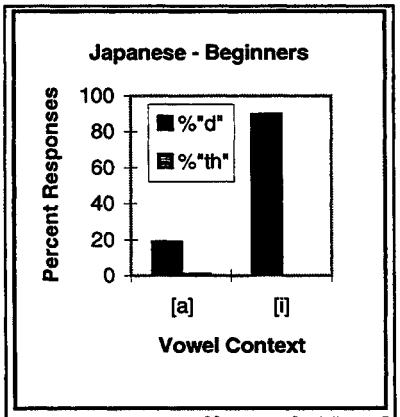

Figure 3. Effect of Vowel Context on "d" and "th" Fusions for Japanese (Beginner English) Speakers.

3.4 Discussion

The expected effect of vowel context was found, but only for "d" responses, due to the paucity of "th" responses. This is surprising given that Japanese Beginners in Experiment 1 showed a large proportion of "th" responses and a change in these with vowel context (see Figure 2). The paucity of "th" responses here was presumably due to the only difference between the experiments - the use of a Japanese speaker. Could this be an instance of the foreign speaker effect? As Japanese subjects here were presented with a native speaker, there may have been less visual influence (and therefore less fusions) than in Experiment 1. However, this explanation alone cannot be correct because Japanese subjects here *did* perceive "d" fusions. The specific paucity of "th" responses could have had two causes. First the Japanese perceivers may not have expected a Japanese speaker to utter the phonologically irrelevant [ð]. Second the Japanese speakers' exaggerated [ð] in the AV and the VO trials may have, by contrast, precluded "th" responses to A[b]V[g]. Certainly in the pilot studies the number of "th" responses to A[b]V[g] is greater when there are *no* [ð] trials (Pilot 2) than when there *are* [ð] trials (Pilot 1) (see Table 1). Further resolution awaits data from the English speakers and the other two Japanese groups.

4. CONCLUSIONS

Experiment 1 shows that auditory-visual integration occurs phonetically *before* the influence of language-specific phonological effects, and supports similar cross-language findings in an English/Thai study (Burnham & Dodd, 1996). To date neither Experiment 1 or 2 provide evidence for the foreign vs native speaker effect, but further cross-experiment analyses are required. The Japanese McGurk effect appears to be based on phonetic, phonological, and cultural expectancy variables. Japanese subjects perceive less fusions than English subjects in the usual [a] vowel context because (i) phonetic conditions are more conducive to the perception of [ð], *and* (ii) Japanese subjects do not respond "th" due to its phonological irrelevance, especially when spoken by a Japanese speaker. This difference between *perceiving* and *responding* and the early integration of auditory and visual speech information before phonological influences is in accord with Burnham's (1997) Phonetic Plus Post-Categorical (3PC) model of auditory-visual speech perception.

4. REFERENCES

Burnham, D. (1997) Language specificity in the development of auditory-visual speech perception. In R.Campbell, B. Dodd, D.Burnham (Eds) *Hearing by Eye II :* Psych. Press.

Burnham, D. & Dodd, B. (1996) Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In D.Stork, M.Hennecke (Eds.) *Speechreading by humans & machines.*Berlin: Springer-Verlag.

Burnham, D., Fowler, J. & Nicol, M. (1997) CAVE: An on-line procedure for creating and running auditory-visual speech perception experiments - Hardware, software and advantages. In *Proceedings of the 5th ECSCT,* Rhodes.

de Gelder, B. & Vroomen, J (1992) Auditory and visual speech perception in alphabetic and non-alphabetic Chinese-Dutch bilinguals. In R.J. Harris (Ed.) *Cognitive processing in bilinguals.* North Holland: Elsevier Science.

Green, K. (1996) The use of auditory and visual information in phonetic perception. In D. Stork & M.E.Hennecke (Eds.) *Speechreading by humans & machines.* Berlin: Springer-Verlag.

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

Sekiyama, K. (1994) Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *J. Ac.Soc.Jap, 15,* 143-158.

Sekiyama, K. & Tohkura, Y. (1991) McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Amer., 90,* 1797-1805.

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