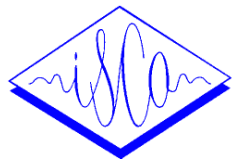

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

THE McGURK EFFECT IS INFLUENCED BY THE STIMULUS SET SIZE.

Jun Amano and Kaoru Sekiyama
Kanazawa University

ABSTRACT

This study examined if the size of the McGurk effect depends on the size of stimulus set presented in a block. The auditory syllables used in the present experiment were eight Japanese monosyllables, /pa/, /ta/, /ma/, /na/, /ba/, /da/, /ga/, and /ka/. Each auditory syllable was dubbed with either a compatible visible syllable or a discrepant visible syllable about place of articulation, resulting in 16 audio-visual stimuli. In the small set condition, two auditory consonants, {/pa/ and /ta/} in one case and {/ma/ and /na/} in another case, appeared in a block. In the medium set condition, four appeared {/pa/, /ta/, /ma/, and /na/}. In the large set condition, eight appeared {/pa/, /ta/, /ma/, /na/, /ka/, /ba/, /da/, and /ga/}. We examined if the size of the McGurk effect for /pa/, /ta/, /ma/, and /na/ varies depending on stimulus set-size. Participants identified consonant in three presentation conditions: audio-visual, audio-only, video-only. Except the video-only condition, auditory white noise was added (S/N=0dB). There was also a clear audio-visual condition in which no auditory noise was added. The results for bimodal discrepant pairs showed that auditory labials differ from auditory nonlabials with respect to the effect of the set size: although the auditory nonlabials (/ta/, /na/) did not show the effect of the set size, the size of the McGurk effect for auditory labials (/pa/, /ma/) depended on the stimulus set size, being larger when a consonant appeared in smaller sets. On the other hand, unimodal identifications were not affected by the set size. The observed effect of the set size on the McGurk effect was argued in terms of the number of dimensions in auditory and visual information.

1. INTRODUCTION

In experiments on the McGurk effect, subjects are typically requested a categorical judgment for speech stimuli in which an auditory and a visual speech segments are synchronized. In such a situation, the subjects are considered to pay attention to both auditory and visual information and to make a probabilistic decision based on some criteria relative to their knowledge of spoken language. During these processes, many factors possibly affect the decision making. Here we investigated attention allocation to auditory and visual cues, by changing the stimulus set size. The set size refers to the number of kinds of consonants which can appear in one block.

Although the set size factor has been neglected in the field of audiovisual speech perception, its influence is suggested by a few pieces of evidence. One of the authors previously showed, by using ten consonants in one block, that native speakers of Japanese are hardly subject to the McGurk effect when auditory speech was perfectly intelligible [1]. Although the Japanese reliance on auditory information was striking in this experiment, it is possible that it was not only due to the nature of the subject group, but also due to the relatively large stimulus set size.

Another line of evidence can be seen in studies by McGurk and MacDonald [2] [3]. The magnitude of the McGurk effect in their 1976 study was much larger than in their 1978 study, as shown in Figure 1. The number of consonants they presented in one block was 4 in 1976, and it was 8 in 1978. Although it is difficult to decide if this sharp contrast is due to the set

size or to the stimuli themselves (differences in talker, recording condition, etc.), the possible effect of the set size draws our attention. In order to test this hypothesis, we manipulated only the set size by using otherwise identical stimuli.

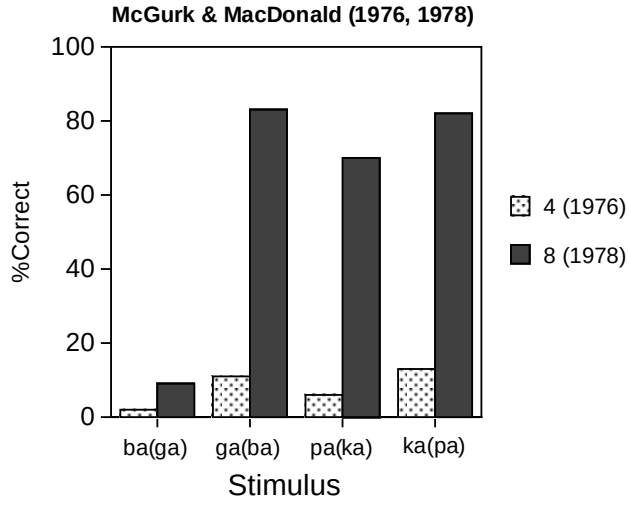

Figure 1: Percent correct (auditorily) for each type of stimulus. The stimulus type "ba(ga)" indicates that audio was /ba/ and video was /ga/.

2. PURPOSE

The purpose of the present study was to see if the stimulus set size influences the decision making during audio-visual speech perception.

3. METHOD

3.1. Subjects

Twenty undergraduate students at Osaka City University participated as subjects. All had normal hearing and normal (or corrected to normal) vision.

3.2. Stimuli

Eight CV syllables were used : /pa/, /ta/, /ma/, /na/, /ba/, /da/, /ga/, and /ka/. Two female native speakers of Japanese, "Y" and "S," pronounced these syllables. Recorded sound and videotaped faces of each talker were combined within a talker and sixteen pairs in one talker were created. Half of the stimuli were audiovisually incompatible with respect to the place of articulation, and the others were audiovisually compatible.

incompatible pairs

pa(ka), ta(pa), ma(na), na(ma),
ba(ga), ga(ba), da(ba), ka(pa)

compatible pairs

pa(pa), ta(ta), ma(ma), na(na),
ba(ba), ga(ga), da(da), ka(ka)

From these stimulus population, stimuli were selected for presentation according to the set size manipulation.

3.3. Experimental design

We used the three steps of the stimulus set size : In small set size condition, two kinds of auditory syllables appeared in a block. In medium set size, four appeared. In large set size, eight appeared. One block contains stimuli only from one talker.

Small set size

{pa(ka), ta(pa), pa(pa), ta(ta)}
&
{ma(na), na(ma), ma(ma), na(na)}

Medium set size

{pa(ka), ta(pa), ma(na), na(ma), pa(pa), ta(ta), ma(ma), na(na)}

Large set size

{pa(ka), ta(pa), ma(na), na(ma),
ba(ga), ga(ba), da(ba), ka(pa),
ba(ba), ga(ga), da(da), ka(ka),
pa(pa), ta(ta), ma(ma), na(na)}

Among these stimuli, some were dummy. Our main interest was to see if physically identical stimuli in-

duce different magnitude of the McGurk effect due to the set size in which they were presented. We planned to look at the magnitude of the McGurk effect only for four types of the stimuli:

pa(ka), ta(pa), ma(na), na(ma)

The data for these stimuli were compared across set size.

In addition to AV condition in which both audio and video were presented, there were also unimodal conditions (Auditory and Visual). In Auditory condition, stimuli were presented with white noise whose sound level was equal to the peak level of speech sound (S/N ratio=0dB). The noise was added to avoid ceiling effect in audio-only identification. AV condition had two versions of audio clearness: clear (no-noise-added) and noise-added. In the noise added version, the audio was the same as in the Auditory condition.

3.4. Procedure

In AV condition each stimulus was repeated four times in a block, while in unimodal conditions that was repeated basically eight times. Stimuli were presented with about seven seconds cycle by random order in all conditions. The peak intensity of speech sound was 55dB SPL. Whereas the subjects were instructed to report what they heard in AV and Auditory conditions, they were asked to report what they thought in Visual condition. The viewing distance of the subjects was 1m.

4. RESULTS

4.1. Unimodal condition

The results for unimodal conditions are shown in Figure 2 and Figure 3. In both Auditory and Visual conditions, the set size did not affect the accuracy in unimodal identification. These results show that the identification in unimodal condition does not depend on stimulus set size.

4.2. AV condition

In AV condition, we analyzed only responses to our planned stimuli (as mentioned in section 3.3), that is, audiovisually incompatible stimuli which were commonly used in all set size conditions. The results in AV condition are shown separately for two cases (the auditory consonant was labial or non-labial), because

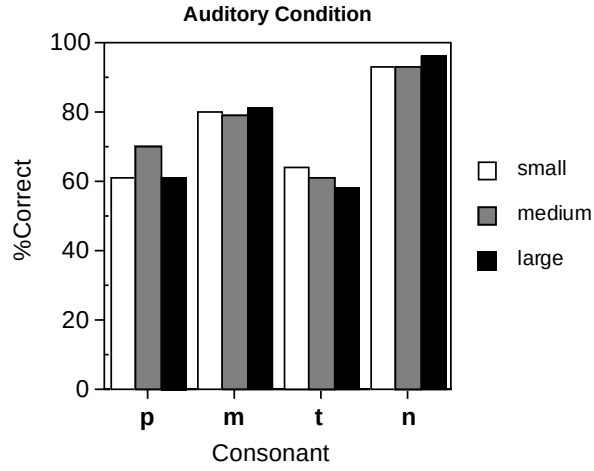

Figure 2: Identification accuracy in the Auditory condition (S/N=0dB) for each stimulus.

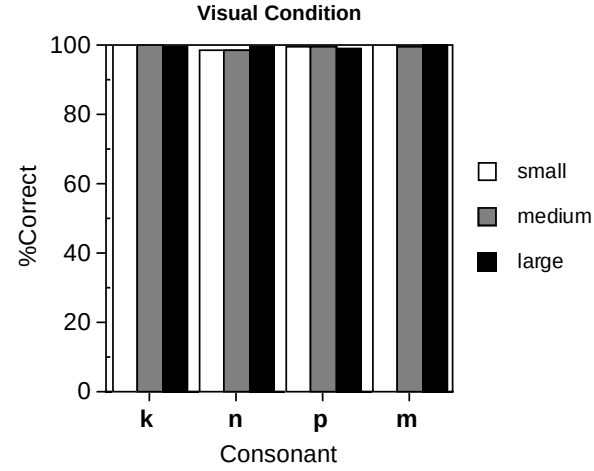

Figure 3: Identification (speechreading) accuracy in the Visual condition. The scoring was done in terms of classification between labials and non-labials.

they seemed different in terms of the effect of the set size.

When the auditory consonant was labial, the main effect of the stimulus set size was significant, indicating that auditory responses increase as the set size gets larger (Figure 4). In the clear AV condition, the proportion of auditory responses in the medium and large set size was significantly larger than that in the small set size ($p<.01$).

On the other hand, this set size effect was not found when the auditory consonant was non-labial (Figure 5). However, as described in the following section, the subjects' confidence ratings showed that judgments for these stimuli were also affected by the set size.

4.3. Confidence rating

In the clear AV condition (but not in the other conditions), the subjects were asked to rate the confidence about their own judgment. In Figure 6, it is clear that the confidence level was higher as the set size gets larger. The main effect of the set size was significant for both auditory labials and auditory nonlabials. As

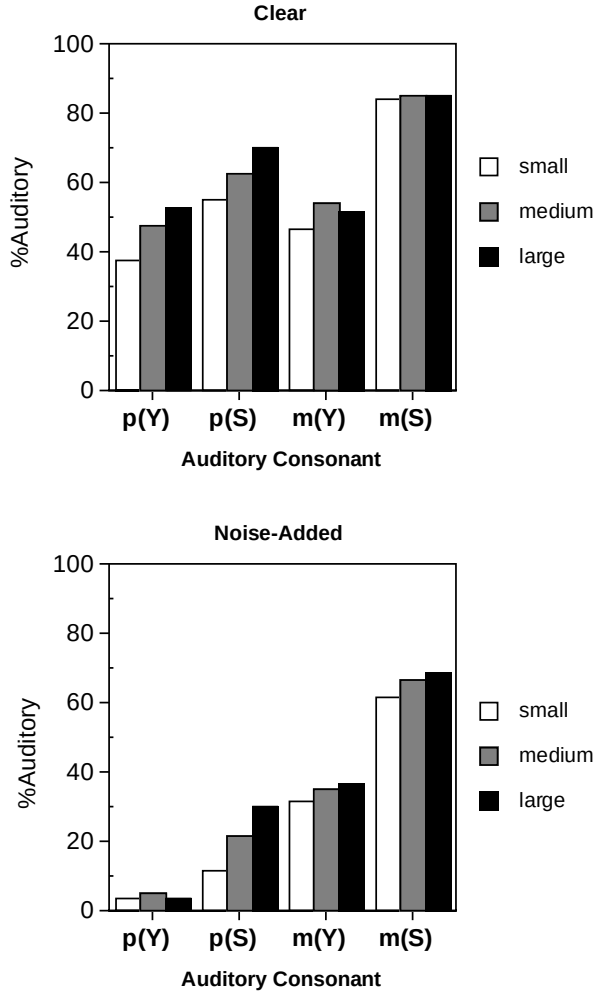

Figure 4: Percent of auditory responses to the stimulus whose auditory consonant was labial. The size of the McGurk effect is calculated by subtracting the percent auditory from 100%. "Y" and "S" in the parenthesis are talkers.

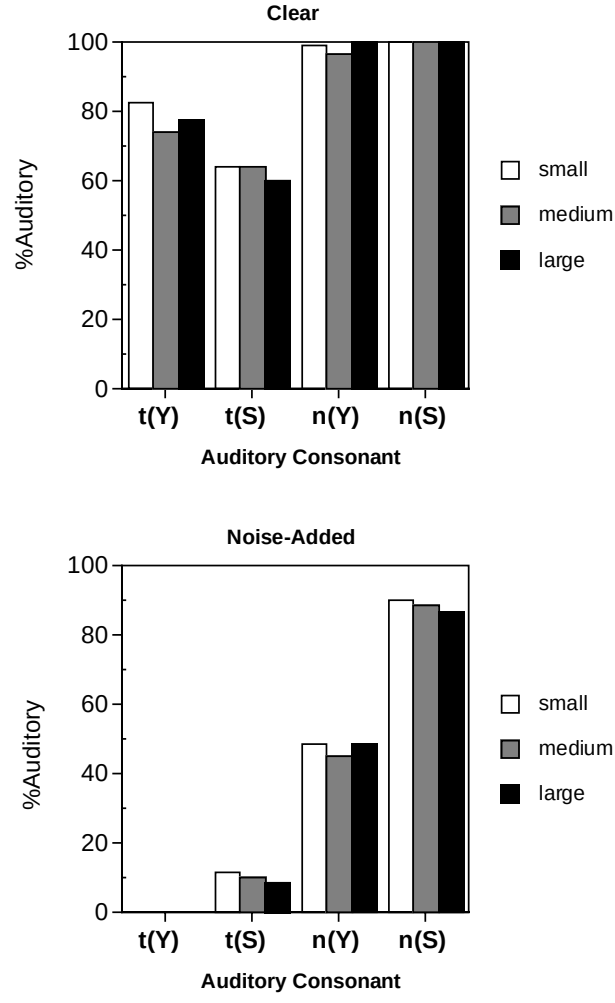

Figure 5: Percent of auditory responses to the stimulus whose auditory consonant was non-labial. The size of the McGurk effect is calculated by subtracting the percent auditory from 100%. "Y" and "S" in the parenthesis are talkers.

compared with the categorical judgments, it is found that the high confidence rating is associated with the occurrence of the auditory responses. These results indicate that the set size affect the process of judgment not only when the auditory consonant was labial, but also when the auditory consonant was non-labial.

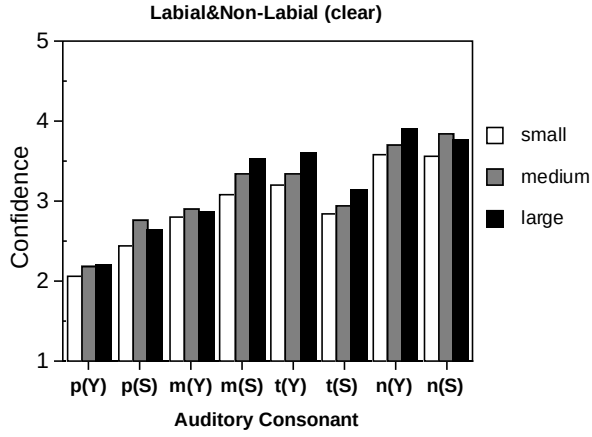

Figure 6: Confidence ratings about the judgment for each stimulus in the clear AV condition.

5. DISCUSSION

The above analyses show that the stimulus set size affects the categorical judgment in audiovisual perception. The magnitude of the McGurk effect tended to be larger when a stimulus appeared in smaller sets. Although the set seize effect was not apparent in the percent auditory when the auditory consonant was non-labial (/ta/, /na/), the set seize effect was clearly seen in the confidence rating. The rating of confidence was larger when stimulus set size was larger. Thus, both auditory labials and non-labials were affected by the stimulus set size during categorical judgments.

The primary question is why the McGurk effect is smaller when the set size is larger. It seems to be attention allocation that is affected by the set seize:

The auditory modality is more attended as the set size gets larger. To explain this interpretation, let's look at qualitative differences between auditory and visual speech information.

5.1. Classification of consonants

Consonants can be classified by three dimensions with respect to articulation: place of articulation, manner of articulation and voicing. The consonants used in the present experiment are classified as showed in Table 1.

Table 1:

manner	voicing	place		
		bilabial	dental	palatal
plosive	voiceless	p	t	k
	voiced	b	d	g
nasal	voiceless			
	voiced	m	n	

For example, consonant /p/'s place of articulation is bilabial, its manner is plosive, and its voicing is voiceless. Identification of a consonant needs to judge these three elements. Lack of only one element makes us unable to identify a consonant. About judging these three dimensions, auditory and visual information are very much different. Whereas auditory speech easily provides information about the all three dimensions, only the place information is available from visual speech in many cases [4] [5].

This qualitative difference between visual and auditory information can cause an attentional set of the subjects in response to the stimulus set size. The present results suggest that the auditory modality is more attended when the set size is larger.

5.2. Attention allocation

Basically, we need information about the all three dimensions to identify a consonant. However, judgments of the all three dimensions were not equally requested in this experiment. In the small set size, only

two auditory syllables /pa/, /ta/ (or /ma/, /na/) appeared. The two in a set differed only in the place of articulation. Therefore, the place information was crucial in the small set size. In other words, judgments about manner and voicing dimensions were useless in this set size. This will cause an attentional shift to visual modality because visual modality is good at conveying the place information. In the medium set size, four syllables {/pa/, /ta/, /ma/, /na/} appeared. The four can be classified by two dimensions: place and manner, or place and voicing. In the large set size, eight consonants appeared and they were classified by three dimensions. Thus, in the medium and large set sizes, attention to auditory modality should be more largely allocated, because visual modality does not provide much information about the manner and voicing. Paying attention to auditory modality will cause auditory dominance in integrating bimodal information, This accounts for the results that the McGurk effect in the medium and large set size was weaker than that in the small set size.

6. CONCLUSION

As the stimulus set size increases, the number of information dimensions increases. Presumably due to this dimension increase, the attention allocation is modified during audiovisual speech perception. Our interpretation is that the McGurk effect is weakened in larger sets because the larger sets require attention to more dimensions. An attentional shift to auditory modality takes place when more dimensions must be attended, because auditory modality provides more dimensions of information than visual modality does.

REFERENCES

1. Sekiyama, K. & Tohkura, Y. (1991) McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, **90**, 1797-1805.

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

3. MacDonald, J. & McGurk, H. (1978) Visual influences on speech perception processes. *Perception & Psychophysics*, **24**, 253-257.

4. Sekiyama, K., Joe, K., & Umeda, M. (1988) Perceptual components of Japanese syllables in lipreading: A multidimensional study. *Technical Report of IEICE* (The Institute of Electronics, Information and Communication Engineers of Japan), **IE87**-127. [In Japanese with English abstract]

5. Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977) Effects of training on the visual recognition of consonants. *Journal of Speech & Hearing Research*, **20**, 130-145.