TR－A－0093 0017

MCGURK EFFECT UNDER CONDITIONS WITH OR WITHOUT NOISE :

No Visual Biasing to Hearing Japanese Syllables of High Auditory Intelligibility

Kaoru Sekiyama and Yoh'ichi Tohkura

1990. 11.13

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷 ☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

Inuidani,Sanpeidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

MCGURK EFFECT UNDER CONDITIONS WITH OR WITHOUT NOISE:

No Visual Biasing to Hearing Japanese Syllables of High Auditory Intelligibility

Kaoru Sekiyama and Yoh'ichi Tohkura

In face-to-face communication, the speaker's articulatory movements can be seen. This visual information is called lip-read information. Its role in speech perception has been investigated in several experimental paradigms. It has been reported that lip-read information facilitates speech perception especially when the intelligibility of the speech is not so good in a noisy environment, e.g., when noise was experimentally added to speech (Erber, 1975), and in speech perception by patients with cochlear implants (Fukuda, Shiroma, & Funasaka, 1988). On the other hand, lip-read information interferes speech perception in an artificial situation such as a McGurk effect experiment (McGurk & MacDonald, 1976). The McGurk effect shows a visual influence on auditory perception under audio-visual discrepancy conditions. Dubbed video tapes are used to demonstrate the McGurk effect. For example, dubbing an audio signal "ba" onto a video signal "ga" makes an audio-visual stimulus that often induces a perception of "da", suggesting a perceptual fusion between audio and video information. This fusion phenomenon is interpreted as follows (MacDonald & McGurk, 1978): Audition conveys information about the manner of articulation and voicing. In this case, "ba" is plosive and voiced. Information on the place of articulation, however, is picked up by vision. In this example, visual "ga" is non-labial. Consequently, a perceptual solution is "da", which is plosive, voiced, and non-labial sound, and is characterized in acoustic features between "ba" and "ga".

Thus with audio-visual discrepancy in place of articulation, lip-read information misleads and biases auditory perception while it helps auditory

perception in the natural audio-visual congruent situation. In the McGurk effect, an audio-visual incongruent stimulus is defined as a stimulus where auditory and visual information is incongruent in terms of the place of articulation, and more narrowly, whether the place is labial or non-labial. Studies on lipreading have shown that the visual system can provide information efficiently about the place of articulation so that lip-readers can easily discriminate labials from non-labials (e.g., Summerfield, 1987).

Japanese consonants include four labials (/b/, /p/, /m/, /w/) and ten non-labials (/t/, /d/, /n/, /s/, /z/, /r/, /y/, /k/, /g/, /h/). A confusion analysis study using multi-dimensional scaling by Sekiyama, Joe, and Umeda (1988) showed that untrained Japanese subjects can easily discriminate labials from non-labials in lipreading of Japanese syllables. Therefore the McGurk effect was expected for Japanese syllables as well as for English syllables. One purpose of the present study was to investigate how the McGurk effect occurs for Japanese syllables.

In the original experiments by McGUrk and MacDonald (1976), perceptual fusion was reported with stimuli where an auditory labial was combined with a visual non-labial (more exactly, velar), e.g.., auditory "ba" combined with visual "ga" (perceived as "da") and auditory "pa" combined with visual "ka" (perceived as "ta"). In conversely combined pairs, i.e., an auditory non-labial with a visual labial, "combination errors" were often observed rather than fusion: auditory "ga" and visual "ba" induced perception of "bga", combining both auditory and visual consonants.

After the original report of the McGurk effect, a number of studies were conducted using various kinds of auditory and visual stimuli, establishing this phenomenon (MacDonald & McGurk, 1978; Massaro & Cohen, 1983; Green & Kuhl, 1989; Kuhl, Green, & Meltzoff, 1988; Randolph & Basala, 1982; Rollins & Hendicks, 1980; Summerfield, 1979, 1986). In these previous studies, however, whether the McGurk effect occurred or not and to what extent it

occurred depended on specific experimental conditions and stimuli, leaving one or more necessary conditions for this effect to be discovered.

The main purpose of this study was thus to examine the relation between auditory intelligibility and the McGurk effect. Our hypothesis was that the McGurk effect occurs only when auditory stimuli do not have complete intelligibility. In fact, the literature has suggested this point although it has not been stated explicitly. This is perhaps due to lack of systematic experimentation: in previous studies, the number of the stimuli was small or the auditory intelligibility itself was not measured. To examine this relationship, auditory intelligibility is defined for this paper as the percent of correctly heard responses when auditory-alone stimuli are presented.

Several types of evidence were found for our hypothesis that degraded auditory stimuli are necessary for the McGurk effect. In the experiment by McGurk and MacDonald (1976), average auditory intelligibilities measured in an audio-alone condition were 91%, 97%, and 99% for 3-5 years children, 7-8 years children, and adults respectively. As they used only four stimuli (/baba/, /gaga/, /papa/, /kaka/), these percentages suggest that their auditory stimuli were not highly intelligible. Our preliminary experiment found the McGurk effect for Japanese syllables, using 14 syllables (/ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /ra/, /wa/, /ha/, /ya/, /sa/, /za/), for which average auditory intelligibility was 80%. These auditory stimuli were recorded by paying less attention to a high S/N ratio recording condition.

The McGurk effect has been reported for words as well as isolated CV syllables. Dodd (1977) reported a McGurk effect in audio-visually dissonant words with white noise: for example, hearing "tough" and seeing "hole" resulted in a perceived "towel". On the other hand, Easton and Basala (1982) found little visual effect for meaningful words. In their experiment, a group of subjects were instructed to attend to auditory information and the other group was

instructed to attend to visual information. While the audio-oriented subjects reported auditorily presented words correctly 99% of the time, the video-oriented subjects could not report visually presented words and only 10% of their responses suggested audio-visual fusion. One possible reason for the difference between the results of these two experiments is the presence or absence of noise.

Thus our question arose: Are noisy or degraded auditory stimuli necessary to the McGurk effect? More moderately, do degraded auditory stimuli promote the McGurk effect? To examine this question, the present experiment compared the McGurk effect in two experimental conditions: a noise-free condition and a noise-additive condition. Auditory intelligibility was measured for both the noise-free and the noise-additive conditions, presenting auditory-alone stimuli. We were thus able to examine the following points: (1) How does the McGurk effect depend on the auditory intelligibility? (2) What is the nature of the McGurk effect for Japanese syllables?

Method

subjects. Ten female subjects participated in the experiment. They were all native Japanese speakers with normal hearing and normal or corrected vision. Subjects' age ranged from 22 to 27.

material. Ten Japanese syllables were used: /ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /ra/, /wa/.

stimuli. For recording audio and video signals, a female Japanese narrator uttered each syllable once. While she pronounced the syllable, her face was videotaped onto a 1/2 inch tape through a video camera located in front of the narrator. Audio signals were recorded separately to prevent any degrading factor in a recording process. The narrator was instructed to pronounce clearly and the audio signals were recorded by a digital (DAT) tape recorder (16 bit with

a sampling frequency of 20 kHz). When audio-visual stimuli were made, these audio signals were dubbed onto an audio channel of a 1/2-in. video tape with preceding warning tones for each syllable. Ten audio and 10 video stimuli were combined resulting in 100 audio-visual stimuli. Care was taken to get precise synchronization between audio and video signals, adjusting the dubbing timing by 33 msec frame unit. Each audio-visual stimulus was arranged in a 7 sec unit which included a 3 sec black bursts and a 4 sec speaking face (Fig. 1). For presentation, several random order sequences were made by editing original 100 stimuli. In an experimental session, one of these random order sequences of audio-visual stimuli was reproduced by a video cassette recorder (1/2-in. tape, SONY BVW-40). Visual stimuli were presented on a 20-in. color monitor which subjects looked at from a distance of 1 meter. Audio signals were presented through two loud speakers attached to sides of the monitor.

procedure. Subjects were given an audio-visual stimulus in every seven seconds and asked to look at and listen to each utterance. The task of subjects was to report what they heard, giving written responses. To make sure that the subjects do not ignore visual stimuli, they were instructed to report a perceptual discrepancy between audio and visual stimuli whenever they felt it. Instructions also suggested that some people might hear syllables which were not included in the Japanese phonological system, like /bda/ or /bga/.

design. The subjects were required to do the above task in two conditions: with noise (noise-additive audio-visual condition: nAV) and without noise (noise-free audio-visual condition: AV). In nAV-condition, gaussian noise with 50kHz band width was added to audio signals. Signal to noise ratios, measured by a sound pressure level meter at subjects' hearing position, were kept at 0dB. In AV-condition, there were no additional stimuli. To measure auditory intelligibility of ten syllables for both nAV- and AV-conditions, there were two audio-alone conditions where video signals included only black bursts and subjects were

required to report what they heard: noise-additive audio-alone (nA) condition where auditory stimuli were the same as in nAV-condition, and noise-free audio-alone (A) condition which included the same auditory stimuli as in AV-condition. All the subjects participated in these four conditions in the order: (1) nAV, (2) nA, (3) AV, (4) A. There were six repetitions of trials for each condition. Thus the numbers of trials for each subject were 600 in each of nAV- and AV-conditions, and 60 in each of nA- and A-conditions. The 600 trials for audio-visual conditions were carried out in two days and the 60 trials for audio-alone conditions were done in a day.

Results

The results were analyzed by producing confusion matrices for each visual stimulus.

results for noise-free conditions.

(1) A-condition. Table 1 shows the confusion matrix for the A-condition. The number in each cell indicate the rates of the response to 60 observations (10 subjects X 6 repetitions) for each auditory stimulus. Diagonal cells show the intelligibility of each auditory stimulus. For the noise-free condition, the intelligibility of the auditory stimuli was high: most of the stimuli yielded correct responses 100% of the time. Three syllables, however, show incomplete intelligibility: the rates of correct responses for /pa/, /ta/, and /wa/ were 98%, 95%, and 95%, respectively. /pa/ was perceived as /ta/ 2% of the time, /ta/ was perceived as /pa/ 5% of the time, and /wa/ was perceived as /ra/ 5% of the time.

(2) AV-condition. The ten confusion matrices in Table 2 show the results for the AV-condition where auditory stimuli of high intelligibility (as proved by Table 1) were combined with 10 visual stimuli. In general, these confusion matrices

show high rates in diagonal cells, indicating that subjects perceived most of the auditory stimuli correctly, and that visual biasing effects were fairly weak.

For visually presented labials (Table 2-a), i.e., /b, p, m, w/, it was only with auditory /ta/'s that visual effects occurred. Presented with the visual labials, about 20% of the auditory /ta/'s were perceived as /pa/. This response occurred 17% of the time for visual /ba/'s, 33% for visual /pa/'s, 22% for visual /ma/'s, and 14% for visual /wa/'s. Compared to the A-condition (Table 1) where erroneous /pa/-responses for /ta/'s occurred only 5% of the time, these /pa/-responses are interpreted as involving visual effects and auditory errors.

For the visual non-labials (Table 2-b), i.e., /d, t, n, r, g, k/, it was only with auditory /pa/'s and /wa/'s that visual effects occurred. When auditory /pa/'s were combined with any visual non-labials, responses of /ta/ occurred to some extent. Such erroneous /ta/ responses were 24%, 24%, 22%, 28%, 33%, and 22% for visual /da/, /ta/, /na/, /ra/, /ga/, and /ka/ stimuli. Similarly, auditory /wa/'s combined with any visual non-labial yielded erroneous /ra/ responses; 19%, 22%, 18%, 23%, 20%, and 32% of the time for visual /da/, /ta/, /na/, /ra/, /ga/, and /ka/ stimuli. These erroneous /ta/- and /ra/-responses in the AV-condition are interpreted as involving visual effects and auditory errors, compared to the A-condition (Table 1) where erroneous /ta/-responses for /pa/'s and /ra/-responses for /wa/'s were 2% and 5% of the time. Comparing Table 2-a and Table 2-b with Table 1, then, it appears that the McGurk effect in noise-free condition was limited to auditory stimuli of incomplete intelligibility. The auditory stimuli in which significant visual effects occurred, i.e., auditory /ta, pa, wa/, are the stimuli that were not completely intelligible as shown in Table 1. For the other auditory stimuli of complete intelligibility, there were no significant visual biasing effects.

A striking difference between our results and those of McGurk and MacDonald is that our subjects heard auditory /ba/'s combined with visual /ga/'s 100%

correctly as /ba/'s, indicating no McGurk effect here (Table 2-b, vision=g) while the original McGurk effect was highly significant for this type of incongruent pairs so that perceptual fusion (perceived "da") was found 90% of the time (McGurk & MacDonald, 1976).

results for noisy conditions.

(1) nA-condition. Table 3 shows the confusion matrix for nA-condition. With the addition of the gaussian noise, the intelligibility of auditory stimuli decreased compared with that for the A-condition. This decrease is seen clearly in Figure 2, where the intelligibility in the A-condition and that in the nA-condition are paired for each auditory stimulus. In these ten syllables, five (/pa/, /ta/, /ra/, /ga/, /ka/) show conspicuous degradation due to noise: the intelligibility is 50-60% for these syllables (notably it falls to 24% for /ga/) in nA-condition while in the A-condition it is nearly 100% for each of these cases.

As for the types of confusion, Table 3 shows that auditory confusions in the nA-condition occurred between consonants of same manner of articulation: e.g., among voiceless stops, /pa/'s were perceived as /ta/ 37% of the time, and /ta/'s were perceived as /pa/ 27% of the time. Note that in this auditory-alone condition, confusion occurred in both directions: from labial to non-labial and from non-labial to labial.

(2) nAV-condition. Firstly, to examine the positive effects of lip-read information on speech perception in noisy situation, we look at results for audio-visual congruent stimuli (where, e.g., auditory /ba/ and visual /ba/ were combined). Preceding research predicts that lip-read information helps speech perception especially in noisy situation. Figure 3 illustrates how lip-read information was helpful in noisy situation, comparing the intelligibility in A-, nA-, and nAV-conditions. For nAV-condition, only stimuli where sound and lip-read information were congruent are depicted here. Figure 3 shows that lip-read information was helpful to a large extent especially for /pa/'s and /ta/'s. Taking

/pa/'s, for example, the intelligibility was 98%, 56%, and 98% for A-, nA-, and nAV-condition, respectively. This indicates that lip-read information increased the correct identification of syllables in noise-additive. Noise decreased the intelligibility of /pa/'s from 98% to 56% in audio-alone conditions. Lip-read information, however, restored it back to 98%. Similarly, the intelligibility of /ta/'s was 95%, 64%, and 98% for A- , nA-, and nAV-conditions, respectively.

Secondly, all the data for nAV-condition are shown in Table 4 to examine how the McGurk effect occurred. In nAV-condition, visual biasing effects were strong and widespread. In each confusion matrix in Table 4, 10 categories of auditory stimuli and responses are divided into labials and non-labials by a vertical and horizontal thick lines.

Table 4-a shows results for visual labials. The perception of auditory non-labial stimuli was quite significantly biased by visual labial stimuli. Every auditory non-labial stimulus type was perceived as corresponding labial type, the predominant response in most cases, almost to the exclusion of the "correct" response. For visual /ba/'s, for example (Table 4-a, vision=b), auditory /da/'s were perceived as /ba/ 60% of the time, while 94% of these auditory /da/'s were correctly perceived by audition alone in nA-condition (Table 3). Similarly, erroneous shifts to labials are as follows: /ta/'s were perceived as /pa/'s 94% of the time, /na/'s as /ma/'s (74%), /ra/'s as /wa/'s (45%) or as /ma/'s (27%), /ga/'s as /ba/'s (88%), and /ka/'s as /pa/'s (94%). This response pattern is also true with visual /pa/, /ma/, and /wa/ stimuli. These results show that auditory non-labials were visually biased to labials with the same manner of articulation.

Conversely, when combined with visual non-labials (Table 4-b), auditory labial stimuli were to a large extent perceived as non-labial. When the visual stimulus was /da/ (Table 4-b, vision=d), auditory /ba/'s were perceived as /da/'s (38%) or /ga/'s (29%), while these same auditory /ba/'s were 91% correctly perceived by audition alone in nA-condition (Table 3). Similarly, the vision-biased erroneous

shift to non-labial is as follows: auditory /pa/'s were perceived as /ta/'s (92%), /ma/'s as /na/'s (77%), and /wa/'s as /ra/'s (37%) or /ga/'s (22%). This response pattern is also found with the other visual non-labial stimuli, showing that auditory labials were visually biased to non-labials with the same manner of articulation.

In the nAV-condition, visual stimuli were the decisive factor so that the same auditory stimuli were perceived quite differently depending on visual stimuli: e.g., auditory /ta/'s were incorrectly perceived as /pa/'s 98% of the time when combined with visual /pa/-stimuli (Table 4-a, vision=p) while they were correctly perceived as /ta/'s 98% of the time when combined with visual /ta/-stimuli (Table 4-b, vision=t). These visual biasing effects are clearly understood as compared with the auditory confusion in nA-condition where both /ta/-response for /pa/'s (37%) and /pa/-response for /ta/'s (27%) were observed. Second, our results in the nAV-condition show stronger visual biasing effects than the original McGurk effect (McGUrk & MacDonald, 1976). The original McGurk effect mainly occurred for pairs of auditory /ba/ and visual /ga/ (resulting in perceptual /da/) or of auditory /pa/ and visual /ka/ (resulting in perceptual /ta/) while conversely paired stimuli, i.e., auditory /ga/ combined with visual /ba/ or auditory /ka/ combined with visual /pa/ yielded few fused responses. In contrast with their results, all audio-visual incongruent stimuli easily gave rise to fused responses in our nAV-condition. For example, most auditory /pa/'s were perceived as /ta/'s with any visual non-labial rather than only with visual /ta/'s, and conversely, most of auditory /ta/'s were perceived as /pa/'s with any visual labials.

A third difference is that there found few "combination errors" in our experiment while McGurk and MacDonald found many for auditory non-labials combined with visual labials.

Figure 4 illustrates the relation between the auditory intelligibility (measured in audio-alone conditions) and the visual biasing effect (observed in audio-visual conditions). For the auditory stimuli with intelligibility of 100% (/ba/, /da/, /na/, /ra/, /ga/, /ka/, /ma/ in noise-free condition), there were no significant confusions between labials and non-labials. On the other hand, when the auditory intelligibility was 99% or less, confusions between labials and non-labials were significant especially in the noise-additive. These results suggest that the incomplete auditory intelligibility and the existence of noise are both related to the McGurk effect.

Discussion

The McGurk effect for Japanese syllables was very different in noisy and noise-free conditions. In the noise-free condition, the visual biasing effect was weak and almost limited to stimuli to which auditory intelligibility measured in the audio-alone condition was less than 100%. For such stimuli, fused responses occurred about 20% of the time. This extent of visual biasing effect is very small as compared with the results by McGurk and MacDonald (1976). In the noise-additive, however, visual biasing effects were observed so strongly for every audio-visual incongruent pair that visual information had the decisive vote for most of the perceptual decisions, resulting in even stronger effects than the original McGurk effect.

Comparing the results of audio-visual conditions and those of auditory-alone conditions, it was shown that the McGurk effect depended on the auditory intelligibility. When the auditory intelligibility was 100%, the McGurk effect was absent or very weak. If intelligibility decreased less than 100%, however, the McGurk effect was easily induced and it was stronger with poorly intelligible auditory stimuli (Table 5). In the noise-free condition, auditory /ba/'s combined

with visual /ga/'s were 100% correctly perceived as /ba/ while this type of stimuli yielded fused /da/-responses 98% of the time for adult subjects in the original report by McGurk and MacDonald (1976). This difference may be due to the auditory intelligibility: Our auditory /ba/'s showed 100% of intelligibility. As mentioned above, we found no McGurk effects for auditory stimuli of 100% intelligibility (i.e., /ba/, /ma/, /da/, /na/, /ra/, /ga/, /ka/ in the noise-free condition). Although McGurk and MacDonald did not describe the auditory intelligibility for each syllable, it would be possible that their /ba/'s did not have 100% intelligibility.

A second difference between our results and those of McGurk and MacDonald (1976) is the type of audio-visual incongruent pair in which visual biasing effects were observed. The original McGurk fusion effect was found only for pairs of auditory labials with visual non-labials. In this experiment, however, the effects were found symmetrically, in both of the auditory labials with visual non-labials and the auditory non-labial with visual labials.

Thirdly, we found very few "combination errors" such as /bda/-responses for an auditory /da/ combined with a visual /ba/. McGurk and MacDonald (1976) reported this type of errors for auditory /ga/'s with visual /ba/'s 54% of the time and for auditory /ka/'s with visual /pa/'s 44% of the time. The reason for these differences may be attributed to either/both of articulatory characteristics of Japanese syllables or Japanese listeners' perceptual organization which represents a fact that Japanese has no consonants cluster.

The present experiment revealed that there were very few McGurk effect for Japanese syllables of 100%-auditory intelligibility. This fact urges attention to the quality of auditory stimuli for McGurk effect experiments. The strong visual effects observed in the noise-additive condition suggest that degraded auditory stimuli promote the McGurk effect. However, as shown in Figure 4, it seems that both the incomplete intelligibility and the existence of the noise increased the

visual biasing effects. Therefore human beings may depend on eyes in the presence of auditory uncertainty.

Our final question is to what extent these results for Japanese syllables can be generalized: Are there also few visual effects for English syllables of complete intelligibility? Or do Japanese listeners depend on eyes less than English native speakers? A cross-language study will be required to answer this question.

References

Dodd, B. 1977 The role of vision in the perception of speech. *Perception*, **6**, 31-40.

Easton, R. D., & Basala, M. 1982 Perceptual dominance during lipreading. *Perception & Psychophysics*, **32**, 562-570.

Erber, N. 1975 Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, **40**, 481-492.

Fukuda, Y., Shiroma, M., & Funasaka, S. 1988 Speech perception ability of the patients with artificial Inner Ear: Combined use of auditory and visual perception. *Technical Report of IEICE* (The Institute of Electronics, Information and Communication Engineers of Japan), SP88-91. (in Japanese with English abstract)

Green, K. P., & Kuhl, P. K. 1989 The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, **45**, 34-42.

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. *Paper presented to 2nd Joint Meeting of Acoustical Societies of America and Japan*, WW7.

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

Massaro, D., & Cohen, M. 1983 Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **9**, 753-771.

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

Randolph, D. & Basala, M. 1982 *Perception & Psychophysics* **32(6)**, 562-570

Rollins, H. A. & Hendicks R. 1980 *J. Experiment. Psychol.: Human Perception & Performance* **6, 1**, 99-109

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

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

Summerfield, Q. 1986 Some preliminaries to a comprehensive account of audio-visual speech perception. In R. Campbell & B. Dodd (Eds.), *Hearing by eye*. London: Erlbaum. Pp. 3-51.

Table and Figure Captions

Figure 1. Time course of audio-visual stimuli.

Table 1. The auditory intelligibilities in the noise-free (A-) condition. Indicated in % to 60 observations.

Table 2-a. Auditory confusions in the AV-condition for the stimuli with visual labials. Indicated in % to 60 observations.

Table 2-b. Auditory confusions in the AV-condition for the stimuli with visual non-labials. Indicated in % to 60 observations.

Table 3. The auditory intelligibilities in the noisy (nA-) condition. Indicated in % to 60 observations.

Figure 2.Comparison between noise-free and noisy speech intelligibility scores.

Fiuere 3. Speech intelligibility increase by additive visual information.

Table 4-a. Auditory confusions in the nAV-condition for the stimuli with visual labials. Indicated in % to 60 observations.

Table 4-b. Auditory confusions in the nAV-condition for the stimuli with visual non-labials. Indicated in % to 60 observations.

Figure 4. Relationships between syllable intelligibility and syllable confusion caused by misleading visual information.

Table 1. The intelligibility of the auditory stimuli for the noise-free condition(%).

response

audition	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		98				2					
m			100								
w				95				5			
d					100						
t		5				95					
n							100				
r								100			
g									100		
k										100	

Table 2-a Confusion matrices for the sitmuli for which visual components were labials . Noise-free condition. Indicated in %.

vision = b

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		100									
m			100								
w				91				9			
d					100						
t		17				83					
n							100				
r								100			
g									100		
k										100	

vision = p

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	98				2						
p	2	97				2					
m			100								
w				91				9			
d	1				99						
t	1	33				66					
n							100				
r								100			
g		1							99		
k		1							2	98	

vision = m

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		100									
m			100								
w				90				10			
d					100						
t		22				78					
n			2				98				
r								100			
g									100		
k			2							98	h1

vision = w

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	98	2									
p		97				2				2	
m			100								
w				97				3			
d	1			1	97				2		
t		14				86					
n							100				
r								100			
g									100		
k		2		2						97	

Table 2-b Confusion matrices for the stimuli for which visual components were nonlabials . Noise-free condition.

vision = d

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		76				24					
m			98				2				
w				81				19			
d					100						
t		2				98					
n							100				
r						2		98			
g						2			98		
k										100	

vision = t

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		76				24					
m			93				7				
w				78				22			
d					100						
t						100					
n							100				
r								100			
g									100		
k										100	

vision = n

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		78				22					
m			94				6				
w				82				18			
d					100						
t		2				98					
n							98	2			
r								100			
g								2	98		
									2	98	

vision = r

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		72				28					
m			96				4				
w				77				23			
d					100						
t						100					
n							100				
r								100			
g								2	97	2	
k										100	

Table 2-b (2)

vision = g

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	100										
p		67				33					
m			96				4				
w				80				20			
d					100						
t						100					
n							100				
r								100			
g									100		
k										100	

vision = k

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	98				2						
p		78				22					
m			88				12				
w				68				32	1		
d					100						
t						100					
n							100				
r								100			
g									100		
k										100	

Table 3. The intelligibility of the auditory stimuli for the noisy condition (%).

response

audition	b	p	m	w	d	t	n	r	g	k	others
b	91				1				8		
p		56	1			37				6	
m			99				1				
w				94			1	4	1		
d	3				94				3		
t		27		1		64				8	
n			2				98				
r			1	33			4	51	2		y8
g	35				41				24		
k		14				11				61	h7 a7

Table 4-a Confusion matrices for the stimuli for which visual components were labials. Noisy condition. Indicated in %.

vision = b

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	97	3									h1
p		98				2					
m			100								
w	2		6	86				5			h2
d	60				39						bd2
t	2	94				3				1	
n			74			2	24				
r	8	1	27	45				19			y1
g	88				7				5		
k	1	94				2				2	bp2

vision = p

audition	b	p	m	w	d	t	n	r	g	k	others
b	98								2		
p		98				2					
m			100								
w	2		7	77				11			y3
d	45				51			2	3		
t		98				2					
n		1	67				32				
r	9		29	34				23			y5
g	93				5				2		
k	2	98									

vision = m

audition	b	p	m	w	d	t	n	r	g	k	others
b	96				2				2		h1
p		98				2					
m			100								
w	5	1	7	79				8			
d	64				34		2	1			
t		97				2				2	
n			84				14				h2
r	8		40	29			2	14			y5 wm2
g	93				4				2	2	
k		98				2					h1

vision = w

audition	b	p	m	w	d	t	n	r	g	k	others
b	88			2	1	2			5		h3
p		93				3					h1 f2
m			100								
w				100							
d	33			2	61				3		z1
t		90		1		3				1	f4
n			35				65				
r			2	84				13			a1
g	72			1	17			1	9		
k		92		1		2				2	pk2

Table 4-b Confusion matrices for the stimuli for which visual components were nonlabials. Noisy condition.

vision = d

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	25				38	3			29		h4 z1
p		7				92		1			h1
m			20				77	4			
w				31	2			37	22		y9
d	1				92				8		
t	1					93		2		3	pk2
n			1				96		3		
r				4			8	77	7		y4
g		2			64				34		z1
k		7				47			3	35	a4 h2 pt2

vision = t

audition	b	p	m	w	d	t	n	r	g	k	others
b	11				57	6			25		h1 z1
p		2				97					h2
m			18				80	2			
w				20			6	45	13		
d					97				3		
t						98			1	2	
n			2				98				
r				7			12	63	5		
g	1				82				16		z2
k		4				42			3	39	a5 h3 s2 pt2

vision = n

audition	b	p	m	w	d	t	n	r	g	k	others
b	24		2		33	2			36		h3
p		6				90				1	pt2
m			21				77		2		
w				32	1			40	19		y8
d	2				95				2		
t		4				81				14	h1
n							100				
r				5	1	2	6	63	13		y10
g	2				61			1	34		z2
k		2				36			2	53	a3 h2

vision = r

audition	b	p	m	w	d	t	n	r	g	k	others
b	13				33	1		3	46		h4
p		2				93				3	h1 s1
m			15				82	3			
w				21			2	46	21		y10
d					82				17		h1
t						90			9		h1
n							93		7		
r				3			3	63	12		y20
g	2				52			1	46		
k		2				35			2	47	h6 s2 pt2

Table 4-b (2)

vision = g

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	13				48				35		h3
p		1				98				1	
m			17				78	4			
w				22				47	18		y13
d	1				98				1		
t		1				98				2	
n							98		2		
r				1	2		7	64	13		y13
g	2				71				25	2	h1
k						45			2	47	a3 h2

vision = k

audition \ response	b	p	m	w	d	t	n	r	g	k	others
b	10				30	2		1	46		h6 z2 dg2
p		4				87				5	a2 h1 pt2
m			11				86	3			
w				20			3	43	21		y13
d					91				9		
t		1				90				9	
n							100				
r					5		6	63	12		y14
g					59				40		z1
k		1				38				56	h2 a3

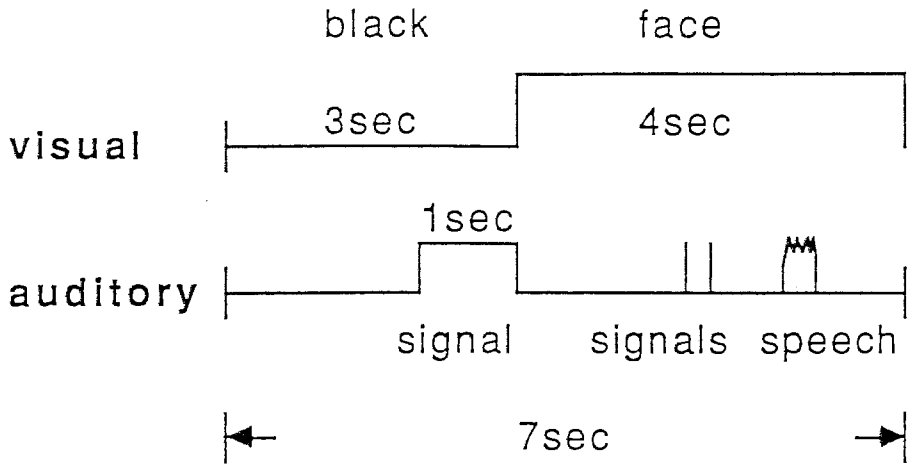

Fig.1 Time course of audio-visual stimuli

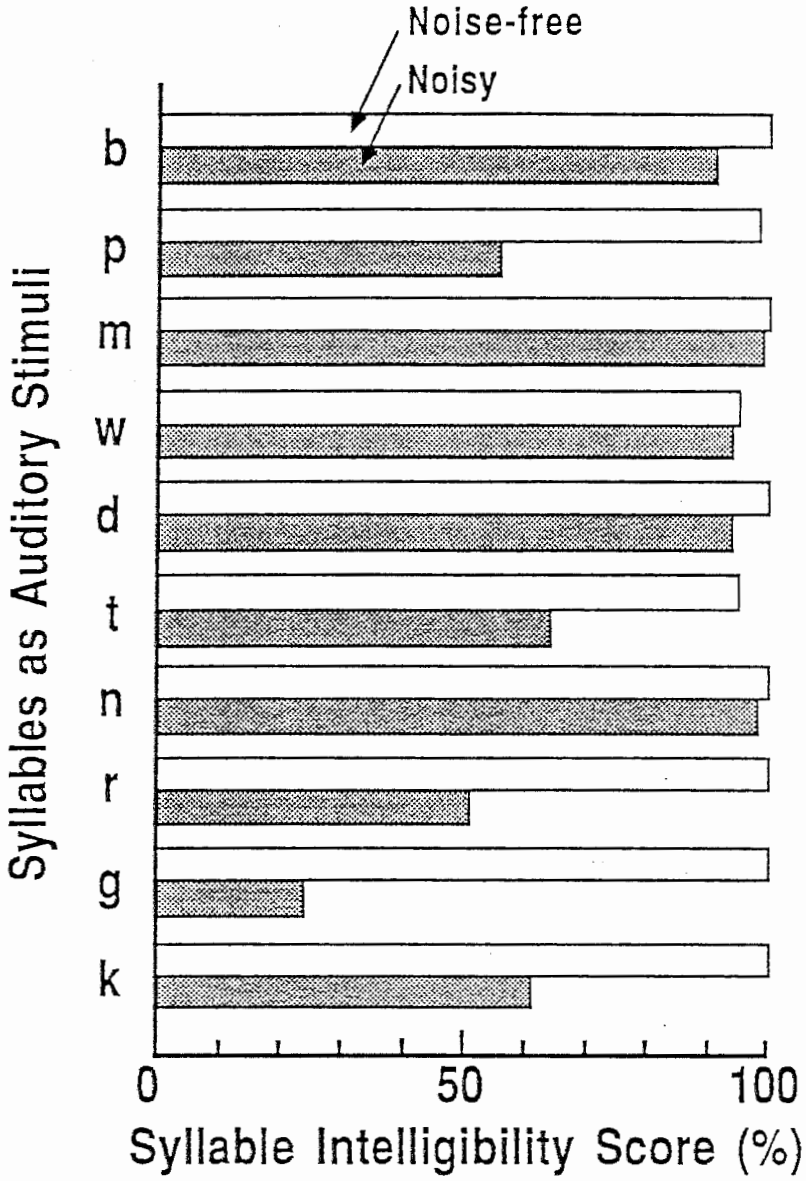

Fig.2 Comparison between noise-free and noisy speech intelligibility scores

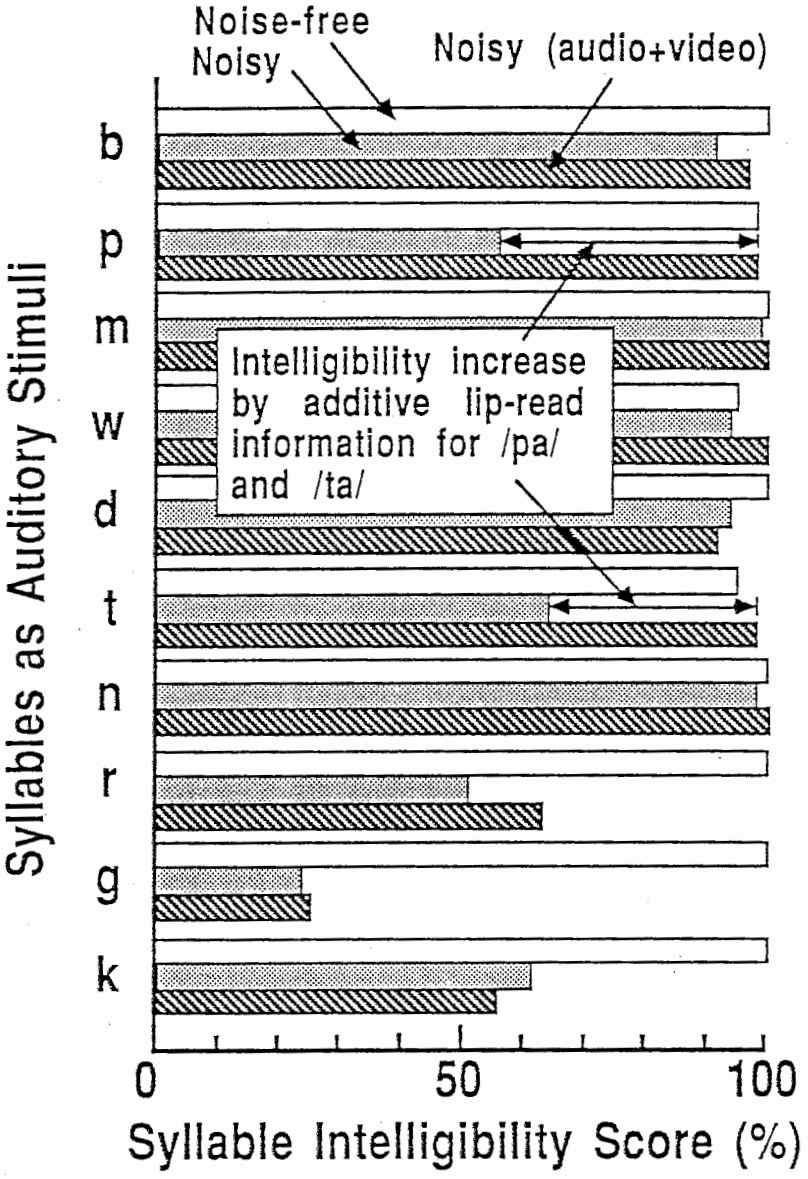

Fig. 3 Speech intelligibility increase by additive visual (lip-read) information

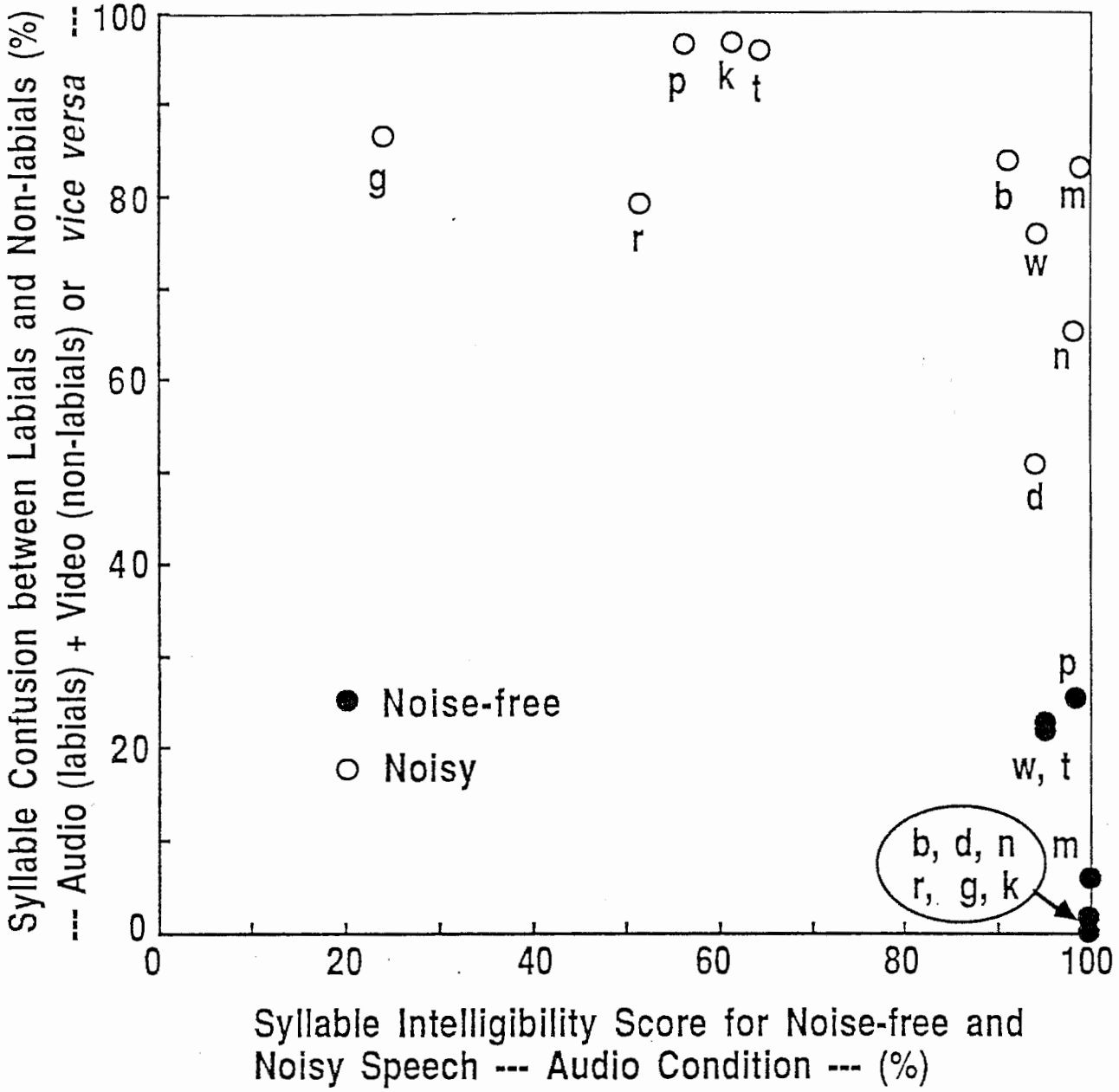

Fig.4 Relationships between syllable intelligibility in audio-only conditions and syllable confusion caused by misleading visual information