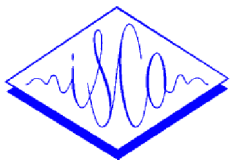
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 – September 3, 2007

Developmental Factor in Auditory-Visual Speech Perception---The McGurk Effect in Mandarin-Chinese and English Speakers

Yuchun Chen[1], *Valerie Hazan* [2]

[1] Department of Human Communication Science, University College London
[2] Department of Phonetics and Linguistics, University College London

yuchun.chen@uclmail.net, v.hazan@ucl.ac.uk

Abstract

This study investigated the developmental factor in the use of visual information in auditory-visual speech perception in Mandarin-Chinese and English native speakers. Adults and 8- to 9-year-old children from two language backgrounds (Mandarin-Chinese and English) were presented stimuli consisting of /ba/, /da/, /ga/ and were asked to make syllable identifications in auditory-only, visual-only and auditory-visual (congruent and incongruent) conditions in clear and in noise. In order to test the second-language learning effect, an additional Chinese child group with a high degree of English exposure was also included. The results showed a developmental increase in the use of visual information and in the magnitude of McGurk effect in both Chinese and English groups. In addition, a positive correlation between the total visual effect and speechreading performance was found, suggesting that the smaller visual influence in the bimodal condition for children might be accounted for by their less sophisticated lip-reading ability. In regard with the foreign language effect, English listeners showed greater visual influence for tokens produced by Chinese speakers, but there was no difference linked to second-language exposure between the two Chinese child groups.

Index Terms: Audiovisual perception, McGurk effect, developmental speech perception

1. Introduction

Seeing the face of a speaker facilitates the comprehension of speech not only in noisy situations [1] but also when auditory signals are clear and intact [2]. Visual speech has also been shown to have a strong influence on perception in situations where it is discrepant with auditory speech as shown in the "McGurk effect" [3]. When phonetic information in the auditory and visual channels is put in conflict (e.g. visual /ga/ and auditory /ba/), listeners report a response (typically /da/) that reflects the influence of both channels of information on phonetic labelling. The McGurk effect has been established as a robust effect in English-speaking cultures and also been tested in speakers from various language backgrounds.

Research on the McGurk effect has also tested young children, and the results suggest that this effect is weaker in children. In their original report, McGurk and MacDonald [3] included 3- to 5- and 7- to 8-year-old children as well as adults and reported that the two groups of English-speaking children displayed a smaller overall visual influence than adults did. Although the three groups all identified the auditory-only stimuli correctly, the visually-influenced responses were 59, 52 and 92% for preschoolers, school children, and adults, respectively. The reduced visual influence in children's auditory-visual speech perception is robust, as it has been confirmed in later studies. For instance, Massaro et al. [4] reported similar weaker visual effects for children than adults using synthesized speech sounds ranging from /ba/ to /da/ combined with a face articulating either /ba/ or /da/. This smaller influence of visual information in children was highly consistent across subjects. In order to further test a possible explanation of this difference, i.e. that children are less sophisticated in lip-reading and therefore use less visual information, children and adults were tested with both auditory and visual sources and were also required to identify speech events on the basis of only the visual source [5]. The results provided a replication in bimodal condition in the previous study. The difference in the proportions of /da/ identifications for a visual /da/ and a visual /ba/ was 0.82 for adults and 0.35 for children. In addition, it was found that children were poorer lip-readers than adults, as the average proportion of correct responses in the visual-only condition was 0.96 for adults and 0.79 for children. What is more, a positive correlation was observed between lip-reading ability and the size of the visual contribution to bimodal speech perception for both the children (r=.641) and adults (r=.670). The results clearly argue in favour of the explanation that children are poorer lip-readers than adults and thus use less visual information in bimodal conditions.

Similarly, Hockley and Polka [6] also reported a developmental increase across ages (5-, 7-, 9-, 11-year-olds, and adults) in speechreading ability and in the degree of visual influence in auditory-visual speech perception by using the McGurk paradigm. The results indicated that the influence of visual information and the integration of auditory and visual information increased across age groups. Furthermore, an adult-like response pattern was observed in only half of the children in the oldest child group (10-12 years old), suggesting that the ability to utilise visual speech information continues to develop beyond the age of 12. Finally, in order to obtain detailed knowledge of the role of linguistic experience in the development of auditory-visual speech processing, Sekiyama et al. [7] combined the developmental and the cross-language McGurk paradigm to investigate the developmental and language factors in audio-visual speech perception. In their experiments, although Japanese adults showed less visual influence than English adults, this difference was not apparent in their 6-year-old counterparts.

Another issue in auditory-visual speech perception is the second language learning effect. For example, Sekiyama [8] found a positive correlation (r=0.723) between the magnitude of the McGurk effect and the length of the Chinese subjects' stay in Japan: the longer the participants had lived in Japan, the stronger the visual effect had become.

Furthermore, it was found that the French-speakers with only beginning and intermediate skills in English were less likely to be influenced by the unfamiliar visual information (/ða/) and there was a monotonically increasing relation between proficiency in English and visual capture by the interdental fricative [9]. Results of these studies suggest that people might learn to use visual cues to improve their comprehension of a second language and the proficiency of a second language may be another factor that influences the strength of the McGurk effect. Since it is popular to learn English as a foreign language early in Taiwan, the current study also investigated the second language learning effect by including one additional child group in which each child had learned English from kindergarten and was continuously exposed to English after entering elementary school.

In this study, in order to investigate the developmental and second language factors in auditory-visual speech processing, the experimental procedures in the Sekiyama study [7] were partly replicated and modified. The aim of the project was to investigate developmental differences in visual influence in listeners of different language backgrounds tested on the same stimuli. A second aim was to investigate the second-language learning factor in auditory-visual speech processing in children. Our hypothesis was that the Chinese children who had had significant exposure to English as a second language would show greater visual influence than the monolingual Chinese children.

2. Method

2.1 Participants

Two adult groups and three child groups participated in this study. Genders were approximately balanced in each group, all participants had normal (or corrected-to-normal) vision and none of the participants had any reported history of speech or hearing disorders. For the adult groups, 22 Mandarin-Chinese native adults and 18 English controls were tested. They ranged in age from 20 to 54 years old. The Chinese adults were recruited from Taiwan where Mandarin is the formal language and Taiwanese is the most common dialect used at home. Most of the Chinese adult participants had learnt English as a foreign language from the age of 12 in school, with instruction emphasising grammar and reading rather than oral skills, and the average duration of formal English education was six years. None of the Chinese participants had lived in a foreign country for more than six months. As for the child groups, 20 Mandarin-Chinese native children were tested in Taiwan and 18 English children were tested in the UK. In order to test the second-language learning effect, an additional 18 Chinese children with a high degree of English exposure were included. They were all educated in bilingual kindergartens in Taiwan, and had English lessons for at least 5 hours per week after they entered elementary schools. All the Chinese and English children ranged in age from 8 to 9 years old.

2.2 Stimuli

The stimuli consisted of the syllables /ba/, /da/, /ga/ uttered by 4 speakers (2 Chinese and 2 English, one male and one female in each language). Chinese speakers were asked to pronounce the syllables with a falling tone (tone 4) in Mandarin-Chinese. The recorded materials were transferred to a computer and the intensity of the auditory channel was normalised to a fixed level across speakers.

Three kinds of stimuli were prepared: Visual only (V), auditory only (A) and auditory-visual (AV) stimuli. Half of the AV stimuli were audio-visually congruent, and the other half were audio-visually incongruent. In order to construct the incongruent stimuli, tokens of /ba/, /da/ and /ga/ were selected that were most similar to each other in terms of their duration, intonation contours and facial movements. There were three incongruent AV stimuli in this study: (1) auditory-ba/ visual-ga, (2) auditory-da/visual-ba, and (3) auditory-ga/ visual-ba. The auditory and visual channels were aligned to ensure that there was auditory-visual coincidence of the release of the consonant. The V stimuli were created by cutting out the audio track. In the A stimuli, the video of the talking face was replaced by a still face of the talker with the mouth closed. Stimuli were down-sampled once the editing had been completed (250*300 pixels, 25 f/s, audio sampling rate 22.05 kHz). In order to see whether visual influence was greater when the auditory signal is less clear, auditory pink noise was added in AV and A conditions. Experimental conditions were then blocked depending on the modality (A, V, AV) and the Signal to Noise ratio (SNR) of the auditory stimuli (clear, +12dB), and there were 4 repetitions of each stimulus in a block. Thus, the number of trials for each participant were 48 in each of the A, A-noisy and V blocks (3 consonants × 4 talkers × 4 repetitions), and 96 in each of the AV-clear and AV-noisy conditions (3 auditory consonants × 2 congruity types × 4 talkers). The total number of trials for each participant was 336.

2.3 Procedure

The visual stimuli were presented in a 3 by 3 inch frame on the colour monitor of a laptop and the auditory stimuli were presented to both ears at a comfortable listening level through headphones. All the participants were tested with the same apparatus individually in a quiet room in Taiwan or England. Each participant was presented in the same order of increasing difficulty: AV, AV-noisy (AVn), A, A-noisy (An) and V blocks. Within each block, the stimuli were presented in random order. The test was response-paced, with breaks after every 30 trials in the AV and AVn blocks.

Instructions were given in the participant's native language prior to every experimental block and there were five AV practice trails before the test began. In the AV condition, participants were instructed to click on one of three labels (/ba/, /da/, /ga/) to answer what they had heard while looking at and listening to each syllable. In the A condition, the participants' task was to answer only what they had heard. In the V condition, they were asked to read lips and click on the label they thought the speaker was pronouncing. During the whole test, the experimenter sat next to the participants to make sure they were watching the screen all the time. The entire experiment, including instructions and breaks, lasted about 25 minutes for adult participants and 35 minutes for children.

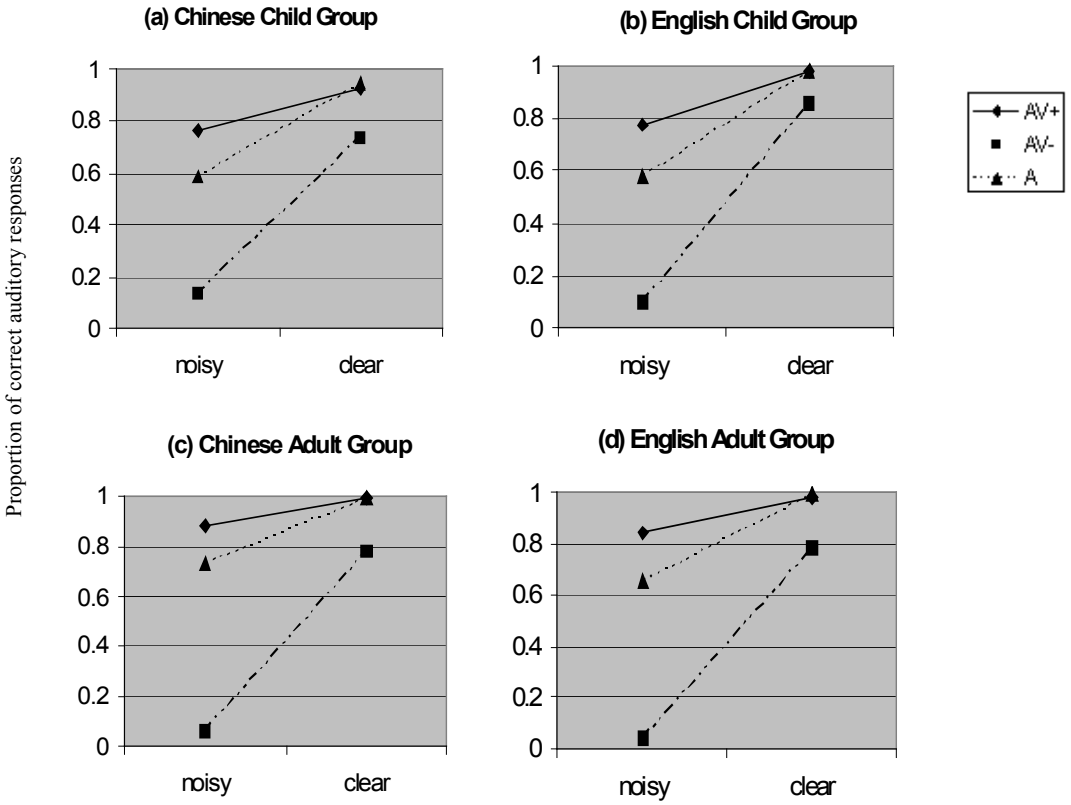


Figure 1: *Proportion of correct auditory response in AV-congruent (AV+), AV-incongruent (AV-) and Auditory (A) conditions in each group*

3. Results

3.1 Degree of visual effect

A preliminary analysis of the results suggested that there were no significant differences in the mean auditory accuracy in AV and A conditions (including clear and with noise) between the two Chinese children groups. Thus, for simplicity of interpretation, these two groups were combined into a single Chinese child group for further analysis.

As in Sekiyama [7], the positive effect of visual information was described as the difference between the auditory accuracy in the congruent AV stimuli (AV+) and A, and negative (interference) effect was represented by the difference between the auditory accuracy in the incongruent AV stimuli (AV-) and A (see Figure 1). The total visual effect was measured by combining the positive and negative visual effects, and this measure was used in the analyses. Group difference in the size of the visual effect was analyzed via a repeated-measure ANOVA, with age and participant language as across-subject factors and noise condition as within-subject factor. The main effect of age ($F[1,92]=21.02$, $p<.0001$) and noise ($F[1,92]=836.74$, $p<.0001$) were highly significant. There was also an interaction effect between age and noise factors ($F[1,92]=8.66$, $p=.004$), showing evidence of greater visual influence for adults in the noisy condition. However, there was no significant main effect of participant language backgrounds ($F[1,92]=0.254$, $p>.05$) so no evidence of a lower degree of visual influence for Chinese participants than English participants. To identify the source of age factor, two separate ANOVAs were further conducted for clear and noisy conditions. It was found that the source of difference for age was from the noisy condition ($F[1,92]=41.403$, $p<.001$), while there were no group differences in the clear condition ($F[1,92]=1.995$, $p=.161$). Thus, these results suggested that the size of visual influence was the same between Chinese and their English counterparts at age 8 as well as in adulthood irrespective of clear or noisy conditions. However, there was a developmental increase of visual effect within both the Chinese and English groups but only under noisy condition.

3.2 McGurk Effect

Next, the extent of the McGurk effect was examined in more detail. The proportion of auditory correct responses for the 'auditory-ba/ visual-ga' tokens were compared across the four participant groups. This stimulus was chosen because it has been found most likely to give the McGurk effect.

Figure 2 shows the mean of the auditory correct response for the V[ga]A[ba] stimuli across four groups under clear and noisy conditions. In Figure 2, a strong McGurk effect (i.e. high degree of visual influence in the phonetic decision) would be shown by a low rate of auditorily-correct responses (i.e. /ba/ response). It must be noted that, compared to related studies, the magnitude of the McGurk effect in clear condition in this study is generally lower than expected,

especially for English participants. This might be related to some methodological issues. For example, wearing headphones might induce higher auditory dependence and thus have caused more auditorily-correct responses. Using a mouse to click the buttons on the screen rather than oral report may also cause visual distraction. In addition, the instruction given for AV condition in this study was "to answer what sound you hear", and this type of instruction could have increased the influence of the auditory modality.

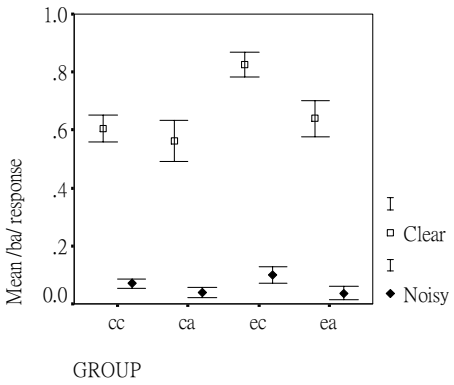


Figure 2：*Mean proportion of /ba/ response given for V[ga]A[ba] stimuli under clear and noisy conditions (Error bars show 1 s.e.; cc=Chinese children, ca=Chinese adults, ec=English children, ea=English adults)*

Despite of these methodological constraints and the weak McGurk effect in clear condition, an obvious McGurk effect emerged when noise was added, and the rate of auditorily-correct response reduced to around 10%. Results of a 2 (participant language) x 2 (age) x 2 (noise: clear and +12dB) ANOVA showed that the age factor was marginally significant in clear condition but significant in noisy condition ($F[1,92]=4.766$, $p=0.032$). This confirms the fact that a greater degree of visual influence was seen for adults than for children.

3.3 'Foreign-language' effect

The next analysis examined whether there was a foreign language effect across language and age groups. In regard to the total visual effect, the stimuli were further divided according to the two speaker groups, i.e. Chinese and English. Combining the visual effect under clear and noisy conditions, it was found that all participants showed a greater visual effect when the stimuli were Chinese, which might be due to the individual voice characteristics of the Chinese speakers. However, the difference between the visual effect for the two language speaker groups was significantly greater in the English listener groups ($F[1,92]= 21.086$, $p<.001$) which suggests that both the child and adult English participants were more influenced by visual information when the speakers were Chinese (see Figure 3). There was no evidence of a greater magnitude of this speaker-language effect in children than adults.

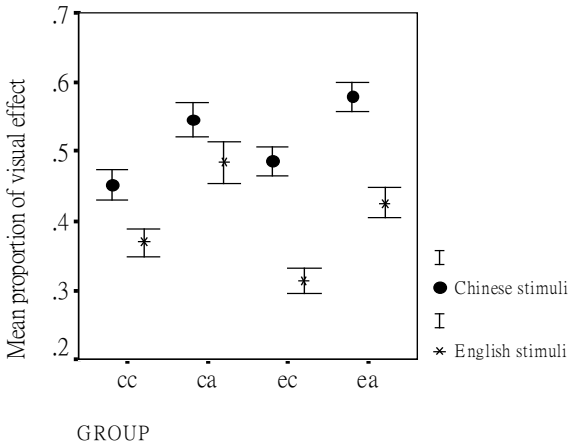


Figure 3：*Comparison of the mean visual effect between the two speaker groups for the four listener groups (Error bars show 1 s.e.)*

3.4 A-only and V-only conditions

With respect to unimodal conditions, the percentages of correct responses were calculated in the auditory and visual alone conditions and shown in Table 1. Again, a 2 (participant language) x 2 (age) x 2(noise) ANOVA was carried out for the A-only conditions. As can be expected, all the participants performed worse in the noisy condition. As for the age factor, the age-related difference was significant ($F[1,92]=31.049$, $p<.001$). The results revealed that adults showed better performance than their children counterparts in identifying auditory signals without visual information, especially in the noisy condition.

Group	A clear	A noise	V
Chinese Children	94.2	58.7	61.1
(n=38)	(7.2)	(10.5)	(11.6)
Chinese Adults	99.3	73.2	75.3
(n=22)	(1.6)	(6.7)	(10.2)
English Children	98.3	57.9	61.1
(n=18)	(2.2)	(6.7)	(9.2)
English Adults	99.5	65.3	72.6
(n=18)	(1.1)	(8.1)	(9.1)
ALL (n=96)	97.1	63.1	66.6
	(5.3)	(10.5)	(12.1)

Table 1. *Percent correct scores in the auditory alone (in clear and in noise) and visual alone conditions for the four listener groups. Standard deviation measures are given in parenthesis*

Similarly, in the V-only (speechreading) condition, the results revealed a age-related difference only ($F[1,92]=32.66$, $p<.001$), suggesting that there is an developmental increase in lip-reading ability. To explicitly test the hypothesis that children are poor lip-readers and thus use less visual information in the AV condition, a correlation between the proportion correct responses in V-only condition and the total visual effect was calculated. A positive correlation was observed between lip-reading ability and the size of the

visual effect (r=.543, p=.01). The pattern of result is similar to that of a previous study [5] and thus confirmed that the less sophisticated lip-reading ability in children might be the reason for the poorer performance in the bimodal conditions.

4. Conclusions

The current study aimed to provide a better understanding of the role of language experience in the development of auditory-visual speech processing. This study replicated and extended an experimental procedure conducted by Sekiyama et al. [7] and tested the McGurk effect in Mandarin-Chinese native speakers in Taiwan and English controls, including adults and children. Compared with the study of Sekiyama et al., the developmental increase of visual influence was not significant in clear condition in the current study. However, in the noisy condition, the increase of visual effect across the two age groups was significant and showed a similar pattern to the previous study. The differences in the performance of English speakers across the two studies might be accounted for by the generally smaller size of visual influence (difference between AV-congruent and AV-incongruent stimuli) in English adults as well as some methodological issues (e.g. speakers used and differences in experimental procedures). In sum, an evident developmental increase of the use of visual information and in the magnitude of McGurk effect was found in both Chinese and English groups in noisy conditions, which is in accordance with previous studies.

One of the hypotheses of the current study was that Chinese children with exposure to English would be more influenced by visual cues and show stronger McGurk effect than Chinese children without English exposure. In fact, there was no significant difference in performance between the two Chinese groups. Furthermore, contrary to some previous studies, there was no difference in visual effect between the English and Chinese adult groups.

Finally, in the A-only and V-only conditions, adults also showed overall better performances than child groups, irrespective of their language background. It should be noted that, compared with the study of Sekiyama et al. [7], the number of repetitions was not reduced for child participants relative to adult participants in the current study which provided more reliable results. In addition, based on previous findings by Massaro [5] of a predictive relation between the lip-reading ability and the use of visual information, we also examined this relationship in the current study, and a positive correlation between the total visual effect and the speechreading performance (V-only) was found. Therefore, our results provide further evidence that the smaller influence of visual information in the bimodal condition for children might be accounted for by their less sophisticated lip-reading ability.

In summary, the results of the present study confirm the initial hypotheses that there is a developmental increase in the degree of visual influence regardless of language background. In addition, a positive correlation between the total visual effect and the speechreading performance suggests that there is a general tendency for visual speech processing to increase over age and this speechreading ability might facilitate the acquisition of auditory-visual speech perception. About the second-language learning effect, the original hypothesis that Chinese children with exposure to English would be more influenced by visual cues and show stronger McGurk effect than children without English exposure has not been supported in this study, but there was some evidence of a 'foreign-language' effect for English children and adults at least, who showed much greater visual influence when listening to foreign speakers.

5. References

[1] Sumby, W.H., Pollack, I., "Visual contribution to speech intelligibility in noise", Journal of the Acoustical Society of America, Vol. 26, 1954, p212-215

[2] Reisberg, D., McLean, J., and Goldfield, A, "Easy to Hear but Hard to Understand: A Lip-reading Advantage with Intact Auditory Stimuli", In B. Dodd and R. Campbell (Eds.), Hearing by Eye: The Psychologoy of Lip-reading.(pp.97-113) Lawrence Erlbaum Associates Ltd, 1987

[3] McGurk, H., MacDonald, J. W., "Hearing lips and seeing voices", Nature, Vol. 264, 1976, p746-748

[4] Massaro, D.W., "Children's perception of visual and auditory speech", Child Development, Vol.55, 1984, p1777-1788

[5] Massaro, D.W., Thompson, L.A., Barron, B. and Laren, E., "Developmental changes in visual and auditory contributions to speech perception", Journal of Experimental Child Psychology, Vol. 41, 1986, p93-113

[6] Hockley, N. and Polka, L., "A developmental study of audiovisual speech perception using the McGurk paradigm", Journal of the Acoustical Society of America, Vol. 96, 1994, p3309

[7] Sekiyama, K., Burnham, D., Tam, H. and Erdener, D., "Auditory-Visual Speech Perception Development in Japanese and English Speakers", In Proceedings of the International Conference on Auditory-Visual Speech Processing, St. Jorioz, France, 2003, p61-66

[8] Sekiyama, K., "Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects", Perception & Psychophysics, 59(1), 1997, p73-80

[9] Werker, J. F., Frost, P., & McGurk, H., "Cross-language influences on bimodal speech perception", Canadian Journal of Psychology, Vol. 46, 1992, p551-568