LANGUAGE EFFECTS ON THE DEGREE OF VISUAL INFLUENCE IN AUDIOVISUAL SPEECH PERCEPTION

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 language factors in the use of visual information in auditory-visual speech perception in Mandarin-Chinese, Thai, Japanese and English, four languages differing in the degree to which they use tone information. Adults from these language backgrounds were presented with stimuli consisting of /ba/, /da/, /ga/ spoken by two English and two Mandarin-Chinese speakers. A syllable identification task was used, in auditory-only, visual-only and audiovisual (congruent and incongruent) conditions in clear and in noise. Chinese listeners used visual information in their audiovisual speech processing to the same extent as English listeners, and the magnitude of the McGurk effect was the same between Chinese and English listeners in the noisy condition. The two additional groups (Japanese and Thai) showed a stronger McGurk effect in clear condition but this might be caused by the foreign-language effect, as all four speakers were non-native for them. The hypothesis that a lower reliance on visual cues is found for tone languages is not supported by these results.

Keywords: Audiovisual perception, McGurk effect, cross-language speech perception.

1. INTRODUCTION

It is well accepted that people use visual (speechreading) information in face-to-face communication. Visual speech facilitates comprehension not only in the presence of background noise [11] but also when auditory signals are clear and intact [6]. Moreover, visual speech appears to have an important influence when it is discrepant with auditory speech as shown in the “McGurk effect” [4]: when a visual /ga/ syllable is combined with an auditory /ba/ syllable, listeners report a response (typically /da/) that provides the best fit to the conflicting information. Since the original report of the McGurk effect, the visual biasing effect on speech perception has been established as a robust effect in English-speaking cultures but there is conflicting information regarding the extent to which this process of early audiovisual integration is universal or language/culture-specific. Using synthetic audiovisual stimuli, Massaro et al. [5] showed that the influence of visible speech appeared to be of the same magnitude for native speakers of Japanese, Spanish and English. However, in a study using natural audiovisual stimuli, Japanese listeners hardly showed the McGurk effect when listening to clear Japanese speech, but a highly increased effect when noise was added [7]. In a subsequent cross-language study, both Japanese and English participants reported a stronger McGurk effect with non-native speech stimuli but Japanese participants showed a weaker McGurk effect overall [8].

Sekiyama [10] proposed that the cultural norm of not staring speakers in the face, and the simpler phonological inventory of Japanese might lead Japanese speakers to using a more auditory-dependent type of speech processing, only using visual information when the auditory speech is indistinct. In support of the face avoidance hypothesis was a finding that Chinese participants, who also show face avoidance, showed a weaker visual effect than American participants and similar effect to Japanese participants [2]. However, a weaker McGurk effect in Chinese listeners might also be due to an auditory bias linked to the tonal characteristics of the Chinese language. In this study, in order to investigate the linguistic and second language factors in auditory-visual speech processing, the experimental procedures of the study in [10] were partly replicated. The aim of the project was to investigate cross-language differences in visual influence in listeners of tonal and non-tonal languages to check the hypothesis that speakers of tone languages would show greater auditory bias in perception. A second aim was to investigate the foreign-language effect by testing for differences in visual influence according to whether the speakers were native or non-native.

2. METHOD

2.1. Participants

Participants included 22 Mandarin-Chinese, 18 English, 10 Japanese and 10 Thai adult listeners. English participants were tested in London and other groups were tested in Taiwan. The Japanese and Thai participants had lived in Taiwan for no more than 1.5 years and spoke fairly elementary Chinese. The Chinese, Japanese and Thai listeners had learnt English as a foreign language from junior high school (about 12 years old), mostly focused on reading and writing. All participants were aged between 20 and 54 years.

2.2. Stimuli

The stimuli consisted of the syllables /ba/, /da/, /ga/ uttered by four speakers (two Chinese and two 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 all stimuli was normalised to a fixed level.

Three kinds of stimuli were prepared: Visual only (V), auditory only (A) and audiovisual (AV) stimuli. Half of the AV stimuli were audiovisually congruent, and the other half were audiovisually incongruent. To construct the incongruent stimuli, tokens of /ba/, /da/ and /ga/ were selected that were most similar in terms of their duration, intonation contours and facial movements. The A and V channels were aligned to ensure that there was auditory-visual coincidence at consonant release. 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 V stimuli were created by cutting out the audio track and in the A stimuli, listeners saw a still face of the talker with mouth closed. Stimuli were down-sampled post-editing (250*300 pixels, 25 f/s, audio sampling rate 22.05 kHz). In order to see whether visual influence is greater when the auditory signal is less clear, 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), with four repetitions of each stimulus per block. Thus, participants received 48 trials 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 per participant was 336.

2.3. Procedure

The V stimuli were presented in a 3 by 3 inch frame on the colour monitor of a laptop, and the A stimuli were presented to both ears at a comfortable listening level through headphones. All participants were tested with the same apparatus individually in a quiet room in Taiwan or England. Conditions were presented in the following order: AV, AV-noisy (AVn), A, A-noisy (An) and V blocks. The V condition was set in the last because of its difficulty and the A condition was interposed to avoid a transfer of visual effect from AV to V. Within each block, stimuli were presented in random order.

Instructions were given in the participant’s native language prior to every experimental block and there were five AV practice trials 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. The classical /θa/ choice was not included since /θ/ is not eligible in Chinese. 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. The entire experiment, including instructions and breaks, lasted about 25 minutes.

3. RESULTS

3.1. Auditory and visual alone conditions

The percentages of correct responses were calculated in the A and V conditions (See Table 1). Repeated-measures ANOVAS showed that higher scores were obtained in the clear than noisy conditions (F[1,56]=979.95, $p<0.001$). There was also a significant group effect (F[3,56]=5.502, $p<0.005$) and Bonferroni-adjusted pairwise comparisons suggest that Chinese listeners scored significantly higher in the A-noise condition than English and Thai listeners.

Table 1: Percent correct scores in the A (in clear and in noise) and V conditions for the four listener groups. Standard deviation measures are given in parentheses.

Listener Group	A clear	A noise	V
English	99.5	65.3	72.6
(n=18)	(1.1)	(8.1)	(9.1)
Chinese	99.3	73.2	75.3
(n=22)	(1.6)	(6.7)	(10.2)
Japanese	99.0	69.8	72.7
(n=10)	(2.6)	(7.7)	(6.8)
Thai	94.8	65.5	77.1
(n=10)	(2.6)	(8.0)	(7.8)
ALL	98.6	69.0	74.4
(n=60)	(2.5)	(8.2)	(8.9)

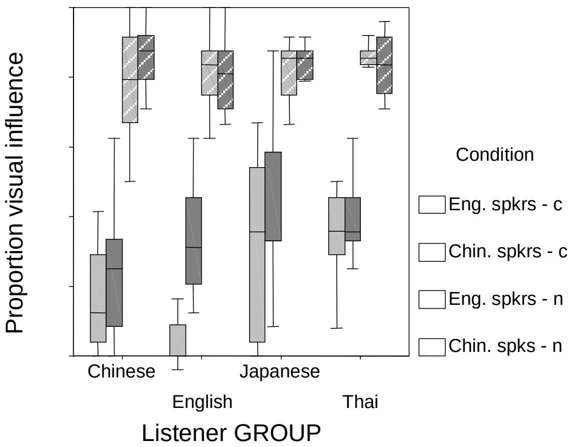
In the V condition, the group effect was not significant but there was a significant listener group by speaker language interaction (F[3,56]=5.09, p<0.005) which appears to be due to English listeners' lower performance in lipreading the Chinese speakers.

3.2. Degree of visual influence

As in [8], the positive effect of visual information was described as the difference in auditory accuracy between the congruent AV and A stimuli, and negative effect as the difference in auditory accuracy between the incongruent AV and A stimuli. The total visual effect was measured by combining these two effects, and this measure was used in the analyses (See Fig. 1).

Group differences in visual effect were analysed via a repeated-measure ANOVA, with participant language as between-subject factor and noise condition and speaker language as within-subject factors. As expected, the degree of visual influence was significantly greater in noise across all listener groups (F[1,56]=362.81, p<.0001). Overall, the effect of listener group was just significant (F[3,56]=3.85, p<.02) but visual influence when responding to Chinese speakers was greater than that for the English speakers (F[1,56]=33.66, p<0.001). The interaction between speaker language, participant language and noise condition was significant: English and Chinese participants showed greater visual influence for Chinese speakers than for English speakers, whereas the Thai and Japanese listeners showed similar visual influence for both groups. This is likely to be due to a 'foreign language' effect as both set of speakers were foreign to the Thai and Japanese listeners.

Figure 1: Boxplot showing the proportion of visual influence in the AV condition in clear (c) and in noise (n) for the four listener groups for English and Chinese speakers.



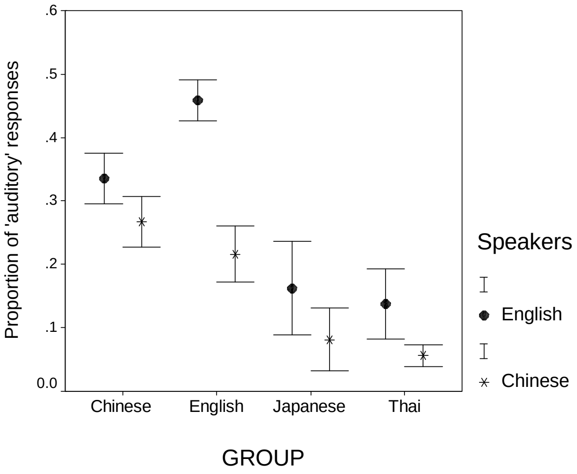
If we focus on comparing the degree of visual influence for the English and Chinese participants, who both heard native and non-native speakers, we see that the main effect of participant group was not significant. There was therefore no evidence of the tone-language listeners showing greater auditory bias in perception than English listeners. However, the significant interaction between participant group, speaker language and noise condition suggests that Chinese listeners showed a less pronounced foreign-language effect than English listeners (See Figure 1), which might be due to their greater exposure to English speakers.

3.3. 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' stimuli, which were most likely to give the McGurk effect, was compared across the four participant groups (see Figure 2).

Overall, the effect of listener group was significant (F[3,56]=8.447, p<0.001) with the English and Chinese groups showing lower McGurk effects than the Japanese and Thai group. To investigate whether this was due to the 'foreign language' effect, the effect of speaker language was examined. There was a significant group by speaker-language interaction (F[3,56]=4.247, p<0.01): English listeners showed less of a McGurk effect for English stimuli than Chinese stimuli, but there was no difference across speakers for the Japanese, Chinese and Thai groups.

Figure 2: Mean proportion of BA responses given for V[ga]A[ba] stimuli (Error bars show 1 s.e.) averaged over the noise and clear conditions. The lower the proportion of BA responses, the greater the McGurk effect.



4. DISCUSSION

This study confirms previous findings of a much stronger visual influence in speech perception when listening to non-native speakers: Thai and Japanese listeners, for whom all speakers were non-native showed the strongest visual influence overall, thus overriding any auditory bias that may come from their being speakers of tone or pitch-accent languages. English listeners showed the greatest discrepancy in visual influence between native and non-native speakers, maybe because their familiarity with Chinese speakers was generally lower than the familiarity that Thai, Japanese and Taiwanese listeners have with English speakers.

It was expected that speakers of tone languages would show a lower degree of visual influence as auditory cues are more informative for tone perception than visual cues. A direct comparison of results for Chinese and English listeners is most appropriate here as both languages have rich phonetic inventories, and listeners from both groups heard both native and non-native speakers. When listening to native speakers of their language, contrary to expectations, Chinese listeners showed a greater degree of visual influence than did English listeners. Both groups also scored similarly in the A and V alone conditions. This result is inconsistent with the findings in Sekiyama's study [8] in which both the Japanese and Chinese were less susceptible to the McGurk effect and used less visual information in speech perception than Americans. Given that McGurk studies have typically used small numbers of speakers, discrepancies across studies may be due to differences in the visual clarity of individual speakers. Therefore, it should be important in future studies to assess the degree of visual clarity of a given speaker. Also, it could be argued that the face avoidance explanation may be becoming less valid as this cultural habit is less observed in the younger generation [3]. Finally, recent studies have shown that listeners may be able to distinguish tones to some extent based on visual information alone even when their native language is not tonal (Australian English speakers) [1], thus countering the hypothesis of auditory dominance in tone language speakers.

5. REFERENCES

[1] Burnham, D., Lau, S., Tam, H. and Schoknecht, C. 2001. Visual Discrimination of Cantonese Tone by Tonal but non-Cantonese Speakers, and by non-Tonal Language speakers. In *Proc. Int. Conf. Auditory-Visual Speech Processing*, Sydney, 155-160

[2] Hayashi, Y.and Sekiyama, K.(1998). Native-foreign language effect in the McGurk effect: a test with Chinese and Japanese. In *Proc. Int. Conf. Auditory-Visual Speech Processing*, Sydney, 61-66

[3] Issei-Jaakola, T. 2006. Cognition and Physio-acoustic Correlates—Audio and Audio-visual Effects of a Short English Emotional Statement: On JL2, FL2 and EL1. In T. Salakoski et al.(Eds.), *FinTAL 2006, LNAI 4139* (pp. 161-173), Springer-Verlag, Berlin Heidelberg

[4] McGurk, H., MacDonald, J. W. 1976. Hearing lips and seeing voices. *Nature*, 264, 746-748

[5] Massaro, D.W., Cohen, M.M., Gesi, A., Heredia, R. and Tsuzaki, M. 1993. Bimodal speech perception: an examination across languages. *J. Phon.* 21, 445-478

[6] Riesberg, D., McLean, J., and Goldfield, A. 1987. 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

[7] Sekiyama, K. and 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. Am.* 90, 1797-1805

[8] Sekiyama, K. and Tohkura, Y. 1993. Inter-language differences in the influence of visual cues in speech perception. *J. Phon.* 21, 427-444

[9] Sekiyama, K. 1997. Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, 59(1), 73-80

[10] Sekiyama, K., Burnham, D., Tam, H. and Erdener, D. (2003). Auditory-Visual Speech Perception Development in Japanese and English Speakers. In *Proc. Int. Conf. Auditory-Visual Speech Processing*, St. Jorioz, 61-66

[11] Sumby, W.H., Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212-215