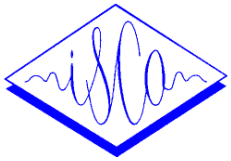
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
(AVSP'99)
Santa Cruz, CA, USA
August 7-10, 1999

THE INTEGRATION OF AUDITORY AND VISUAL SPEECH INFORMATION WITH FOREIGN SPEAKERS: THE ROLE OF EXPECTANCY

Denis Burnham and Susanna Lau

Macarthur Auditory Research Centre, University of Western Sydney, Australia

ABSTRACT

It has been found that auditory-visual integration in the McGurk effect is affected by the relationship between the language of the speaker and that of the perceiver: for a foreign speaker, perceivers tend to incorporate visual information to a greater extent. We investigated whether this is due to the perceivers' detection of the speech sounds as foreign or to an expectancy based upon the appearance of the speaker. English, Japanese, Cantonese, and Thai subjects were presented with English, Japanese, Cantonese, and Thai speakers in a condition in which an expectancy was set (trials blocked by speaker language) or in a random trials condition. There were indeed foreign language effects, albeit in the opposite direction to that expected, and these occurred mainly as the results of expectancies based on the appearance of the speaker rather than the perceived deviation of foreign speech sounds from native language prototypes.

1. INTRODUCTION

In the traditional McGurk effect auditory [ba] dubbed onto visual [ga] is perceived as "da" or "tha" [1]. Recently this effect has been studied in cross-linguistic contexts for a variety of research purposes. These studies can be considered in three categories: *cross-cultural* studies, in which the incidence of the McGurk effect in different languages is examined to ascertain whether there are cross-cultural differences in cue weighting for auditory and visual speech components; *cross-language* studies, in which differences between the phonologies of languages are used to investigate issues such as the manner of auditory-visual integration; and *inter-lingual* studies, in which the effect of being presented with auditory-visual speech by speakers of a foreign language compared with one's native language is studied [2].

This paper is concerned specifically with interlingual effects in auditory-visual integration. Sekiyama found that Japanese listeners incorporate visual information more when they listen to a foreign speaker. In one study American and Japanese listeners' perception of McGurk stimuli spoken by American and Japanese speakers was investigated [3]. As had been found in other studies, Japanese listeners generally showed less visual influence, but over and above this, both Japanese and American listeners demonstrated greater visual influence when presented with stimuli in their non-native language. Sekiyama and Tohkura [3] explained this in terms of auditory ambiguity: when subjects hear phones that are phonemically relevant in their language but acoustically deviant, then any extra information that can be used will be used. Thus visual information is incorporated to a greater extent by Americans listening to Japanese speech and even by Japanese (who tend to use visual information less than English speakers [3]) when listening to American English. Kuhl and her colleagues [4] also found an effect of non-native language upon visual reliance by Japanese and American listeners and explained it in a similar fashion: although the non-native sounds fall within a familiar phone class as in their native language, listeners detect the deviation of non-native speech tokens from the native language prototypes that they have built up through their language experience. Thus there is increased reliance on the visual signal in processing these discrepant auditory speech stimuli.

An alternative to this auditory deviation / visual reliance explanation is that the perceiver sets up expectancies based upon seeing a foreigner's face. When the perceivers see a foreigner's face they may pay greater attention to the visual information because they expect it to be more difficult to understand a foreigner. These alternatives have not been teased apart in previous studies [3,4]. Here we tested English, Japanese, Cantonese, and Thai language subjects in a fully crossed design for their perception of the McGurk effect spoken by speakers from each of the same four languages – English, Japanese, Cantonese, and Thai. Thus for all subjects, one of the speakers spoke their native language while the other three spoke a foreign language. Subjects were given four blocks under one of two conditions: Blocked, in which each speaker presented all trials of a particular block; and Random, in which the order of speakers was random within each block.

In each trial the speaker's face was presented for 1 sec before the speech began. Thus in the random order condition subjects could only ascertain the presence of non-native prototypes *after* the presentation of each stimulus. However, in the blocked condition subjects could expect non-native speech sounds on each trial in the three (of four) blocks on which a foreign speaker was presented. It was expected that if the critical factor in the interlingual effect is the detection of foreign speech sounds discrepant from native speech sound prototypes, then there should be greater visual reliance in the blocked rather than the random condition. On the other hand, if interlingual effects are due to expectations derived from the visual appearance of the speaker, then

visual reliance should be equivalent in the blocked and random conditions.

2. METHOD

2.1. Design

A 4 x 2 x (4) language group (Australian-English / Japanese / Cantonese / Thai) x presentation condition (blocked / random) x speaker (English / Japanese / Cantonese / Thai) with repeated measures on the last factor was employed. Native Australian-English, native Japanese, native Cantoneses and native Thai subjects were randomly assigned to either the blocked or the random test groups. All four language backgrounds x blocked / random subgroups were presented with the four speakers -- English, Japanese, Cantonese, and Thai, with order of block presentation counterbalanced between subjects.

2.2. Subjects

A total of 96 adult subjects were tested. Of these 72 were from various language education institutes in Sydney-- 24 native Japanese, 24 native Cantonese, and 24 native Thai speakers; 24 were native Australian-English students and staff members from the psychology department of the University of Western Sydney Macarthur. Half (12) of the subjects in each language group were assigned to the Blocked group, and the other half to the Random group.

2.3. Stimulus Materials

For both the Blocked and Random groups, stimuli consisted of the syllables [bi] and [gi] as auditory-only (Aud-Only), visual-only (Vis-Only), or auditory-visual (AV) presentations, the latter consisting of either matching auditory-visual presentations or mismatching auditory-visual presentations. Native speakers of each of the four languages -- English, Japanese, Cantonese and Thai -- were videotaped producing [bi] syllables and [gi] syllables under controlled conditions.

Trial Types and Stimulus Construction The stimuli were edited and dubbed online in order to produce Aud-Only, Vis-Only, and either matching or mismatching AV stimuli. The visual components originated from the videotapes, while the auditory components were digitized (using the Kay CSL 4500 package) from the original videotapes and stored on disk. Three exemplars of each visual stimulus and three exemplars of each auditory stimulus were used to ensure acoustic/optic variability and phonemic invariance. The CAVE (Computerized Auditory-Visual Experiment) package [5] was used to create test videotapes. In the CAVE, auditory-visual stimuli are created on-line. Based on pre-programmed CAVE software, the sound played from disk on a particular trial is triggered by the original sound from the audio channel of the videotape. In this manner mismatching AV trials can be produced. For matching AV trials the same dubbing procedure is used to ensure uniformity. On Aud-Only trials the stimulus person's face is motionless, and an auditory stimulus is triggered from disk by a tone pre-recorded on the second audio channel of the videotape. On Vis-Only trials the auditory stimulus from the videotape cues the computer to play 'silence' from disk and only the stimulus person's face and lip movements without sound are presented.

CAVE was used to produce stimuli consisting of Aud-Only, Vis-Only, and matching AV presentations of the syllables [bi] and [gi] and mismatching presentations of A[bi]V[gi] and A[gi]V[bi]. Each trial lasted 4 secs, with 1 sec of black background intervening between trials. For Vis-Only and AV trials this consisted of 1 sec of a motionless face with a neutral expression, about 1 sec of articulation, and then another 2 seconds of neutral expression. For the Aud-Only trials, the speaker's motionless face was presented for 4 seconds overdubbed with a speech sound beginning 1 sec into the presentation of the face. Finally, the whole experiment was stored in the form of a videotape, with the dubbings specified in a file created by MAKETAPE (a part of the CAVE package).

2.4. Stimulus Presentations

Practice Trials Practice trials were included to allow the subjects to become familiar with the testing procedure. For each condition there were 24 practice trials (1 presentation each of [bi] and [gi] by 4 speakers x 3 trial types – Aud-Only, Vis-Only, and matching AV). The only difference between conditions was the presentation order of the speakers. For the Blocked groups, there were six consecutive trials of one speaker (Aud-Only, Vis-Only, and A-V presentations of [bi] and [gi]) followed by six consecutive trials for each of the other speakers. In the Random condition, the presentation order of speakers was mixed randomly across the 24 practice trials.

Test Trials Following the practice trials, for each group (Blocked, Random) there were four 20-trial test blocks. In each block there were 2 Aud-Only [bi], 2 Aud-Only [gi], 2 Vis-Only [bi], 2 Vis-Only [gi], 2 each of AV matching presentations of [bi] and of [gi], and 4 each of mismatching A[bi]V[gi] (McGurk trials) and A[gi]V[bi] (combination trials). Each group received this same combination of trial types but the distribution of speakers across blocks varied between groups. For the blocked groups one speaker only was presented in each test block, and at the start of each block the language of the speaker, eg, "Japanese Speaker" was displayed for 5 secs. In the random group, the presentation order of speakers was mixed randomly within and between the four test blocks, and the blocks were preceded by 5-sec labels, "Block A" etc.

2.5. Apparatus and Procedure

Subjects were tested either individually or in groups (maximum = 3 per group) in a room, seated in front of a monitor connected to the video-recorder. The stimuli were presented in the form of dubbed videotapes previously created by the CAVE package. Subjects were given a response sheet and asked to circle the sound they heard from 6 options, "bi" "gi", "di", "thi", "bgi", or "gbi". They were required to respond as quickly and accurately as possible with the response which "best matched the syllable spoken by the speaker." In each group subjects were given one block of practice trials followed by four test trial blocks. Testing lasted approximately 20 minutes.

3. RESULTS

Two sets of analyses were conducted, one for the trials in which auditory and visual information did not conflict (Aud-Only, Vis-Only, and A-V [bi] and [gi] trials) and one set for the A[bi]V[gi] (McGurk) and A[gi]V[bi] (combination) trials.

3.1 Auditory, Visual and A-V [bi] and [gi] Trials

The percent of correct responses for Aud-Only, Vis-Only, and A-V [bi] trials were examined in a 2 (blocked/random) x 4 (language background group -- English, Japanese, Cantonese, Thai) x (speaker -- English, Japanese, Cantonese, Thai) x mode of presentation (Aud-Only, Vis-Only, A-V) analysis of variance (ANOVA) with repeated measures on the last two factors. A similar ANOVA was conducted for the [gi] responses. The percent correct responses in the [bi] trials and the [gi] trials are shown in Table 1.

Analysis of the [bi] trials revealed that English perceivers obtained most correct responses, followed by Japanese, $F(1,88) = 19.20$, Cantonese, $F(1,88) = 4.87$, and Thai, $F(1,88) = 8.61$ (see Table 1). As would be expected and as shown in Figure 1, there were more correct responses when full AV information was presented than when either auditory or visual information alone was presented, $F(1,88) = 75.24$. Most importantly, there were more correct responses in the blocked than the random condition, $F(1,88) = 4.88$, and this interacted with the percent correct responses in Aud-Only and Vis-Only trials, $F(1,88) = 7.51$: in general subjects had higher scores for the Aud-Only than the Vis-Only trials in the blocked condition and higher scores for the Vis-Only than the Aud-Only trials in the random condition. These results suggest that the blocked condition allowed subjects to perceive the auditory information for [bi] more veridically than in the random condition, while the visually-distinct information for [bi] was relatively unaffected by the blocked / random manipulation. These results were obtained across all four speakers, so there are no interlingual effects -- it made no difference whether the subjects were presented with a foreign language or a native language speaker.

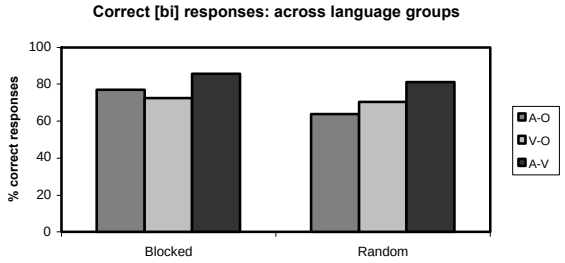


Figure 1. Percent correct AO, VO, AV [bi] responses across language groups and speaker conditions: block vs. random conditions

Analysis of the [gi] trials revealed that English perceivers obtained more correct responses, followed by the Japanese, $F(1,88) = 7.67$, Cantonese, $F(1,88) = 8.15$, and Thai, $F(1,88) = 5.86$ (see Table 1). Again, as would be expected, there were more correct responses when full AV information was presented than when either auditory or visual information alone was presented, $F(1,88) = 28.33$. In addition, auditory information alone resulted in a greater percent of correct responses than visual alone $F(1,88) = 24.50$. For [gi], as opposed to [bi] in the previous analysis) there *was* an effect of native language but only in an interaction with Aud-Only/Vis-Only vs AV trials, $F(1,88) = 7.34$. As can be seen in Figure 2, this is due mainly to the superiority of AV (over Aud-Only and Vis-Only) trials in the blocked conditions being ameliorated in the random conditions but mainly when the subjects view their native language. It seems that the visual information for [gi], especially in non-native presentations, is not so perceptible in the random conditions, for in this

Table 1. Percent correct [bi] and [gi] responses

Language Group		*English Perceivers*						*Japanese Perceivers*						*Cantonese Perceivers*						*Thai Perceivers*					
speaker	*condition*	*AO*		*VO*		*AV*		*AO*		*VO*		*AV*		*AO*		*VO*		*AV*		*AO*		*VO*		*AV*	
		bi	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*	*bi*	*gi*
English	Blocked	92	92	73	62	100	96	75	71	75	67	88	83	79	63	71	71	71	83	37	54	83	29	58	58
	Random	88	92	71	21	96	83	33	88	58	71	71	75	67	71	71	71	58	67	8	46	46	29	29	38
Japanese	Blocked	88	92	83	29	100	83	87	79	75	42	92	92	75	67	62	38	83	71	37	63	58	38	87	75
	Random	100	75	71	46	96	96	87	83	58	46	92	58	75	63	79	21	96	63	33	50	83	38	88	75
Cantonese	Blocked	100	92	83	87	96	96	86	79	79	87	96	92	75	54	79	63	79	79	62	67	71	50	79	63
	Random	88	92	83	50	92	88	58	67	100	67	83	75	67	63	75	54	71	62	42	63	75	46	86	63
Thai	Blocked	100	21	88	54	96	58	96	33	79	33	83	21	79	42	54	42	96	46	54	8	58	25	71	38
	Random	79	50	75	38	88	42	58	42	54	63	67	46	75	58	67	38	83	58	29	33	71	46	79	17
Total	Blocked	95	74	82	58	98	83	86	66	77	57	90	72	77	57	67	54	82	70	48	48	68	36	74	59
	Random	89	77	75	39	93	77	59	70	68	62	78	64	71	64	73	46	77	63	28	48	69	40	71	48

condition there is a decrease in both AV and Vis-Only percent correct and a corresponding rise for Aud-Only trials.

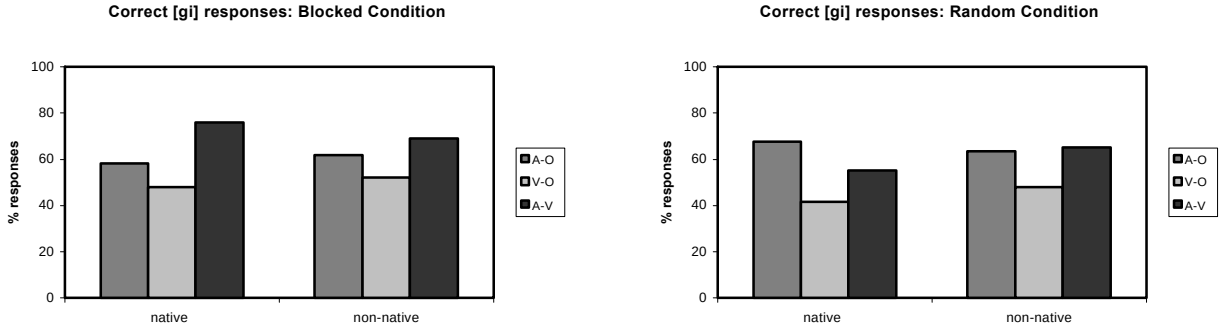


Figure 2: Percent correct Aud-Only, Vis-Only, and AV responses averaged across language groups: blocked / random x native / non-native speaker conditions

To summarize, blocking the four different speakers allowed subjects to make more correct responses based on the auditory information for [bi] and the visual information for [gi], especially when viewing one's native language. Blocking speakers from different language backgrounds thus appears to allow subjects to set up an expectancy that the speech will be foreign and to concentrate upon those aspects of the speech sound which are *less* salient -- auditory information for the visually distinct [bi] and visual information for the more auditorily-distinct but visually-ambiguous [gi]. So blocking here allows *perceptual learning* to occur across trials especially for those features which are the least discriminable. As this occurs both for the auditory and the visual components, these results reinforce the interactive nature of auditory and visual information in auditory-visual speech perception.

3.2 A[bi]V[gi] and A[gi]V[bi] Trials

The A[bi]V[gi] fusion trials and the A[gi]V[bi] combination trials were analyzed separately, each in a 2 (blocked/random) x 4 (language background group -- English, Japanese, Cantonese, Thai) x (speaker -- English, Japanese, Cantonese, Thai) x response ("bi", "gi", "di", "thi", "bgi", "gbi") ANOVA with repeated measures on the last two factors split into auditory and visually influenced responses as shown in Table 2.

For the fusion trials the analysis revealed that subjects generally made more visually influenced responses than pure auditory responses, $F(1,88) = 71.98$. There were no overall effects of the Blocked vs. Random conditions nor were there any overall effects of whether the perceiver was presented with their native language or a foreign language. However there was a significant interaction of the incidence of auditory and visually-influenced responses with whether the subjects were presented with their native or a foreign language, $F(1,88) = 23.12$. This is shown in the upper 4 graphs of Figure 3. As can be seen, the relative preponderance of visual over auditory responses in subjects' native language is ameliorated when the subjects are presented with a foreign language. This occurs for all four language groups (although the effect is somewhat reduced for the Japanese subjects, $F(1,88) = 4.35$.) Thus it appears that the effect of viewing a foreign speaker presenting A[bi]V[gi] is to increase the perceiver's *auditory* reliance (on [bi]) rather than increase the visual reliance as Sekiyama and Kuhl suggest [3,4].

Furthermore, as there is no difference between the blocked and random conditions in the relative frequency of auditory and visual responses, it would appear that it is the optical stimulus of the person's face (presented for

Table 2. Percent auditory ("bi") vs. visual (gi,di,thi,bgi,gbi) responses in A[bi]V[gi] (Fusion Trials) and A[gi]V[bi] (Combination Trials)

Language Group		*English Perceivers*				*Japanese Perceivers*				*Cantonese Perceivers*				*Thai Perceivers*			
speaker	***condition***	*auditory*		*visual*		*auditory*		*visual*		*auditory*		*visual*		*auditory*		*visual*	
		McGurk	*Combo*	*McGurk*	*Combo*	*McGurk*	*Combo*	*McGurk*	*Combo*	*McGurk*	*Combo*	*McGurk*	*Combo*	*McGurk*	*Combo*	*McGurk*	*Combo*
English	Blocked	4	48	96	52	9	38	91	62	6	36	94	64	0	13	100	87
	Random	2	31	98	69	6	40	94	60	30	27	69	75	15	4	85	96
Japanese	Blocked	12	40	88	60	0	23	100	77	6	23	94	77	0	10	100	90
	Random	13	27	87	73	6	46	94	54	33	21	66	78	47	21	43	79
Cantonese	Blocked	7	19	93	81	4	29	96	71	0	25	98	75	2	10	98	90
	Random	9	15	91	85	10	23	90	77	28	10	72	90	29	2	60	98
Thai	Blocked	33	2	67	98	19	17	81	83	15	8	75	92	0	10	100	90
	Random	19	2	81	98	15	10	85	90	44	8	46	92	21	0	79	100
Total	Blocked	14	27	86	73	8	27	92	73	7	23	90	77	1	11	99	89
	Random	11	19	89	81	9	30	91	70	34	17	63	84	28	7	67	93

1 sec before the syllable is uttered), rather than the particular lip movements or sounds which sets the system to greater auditory reliance.

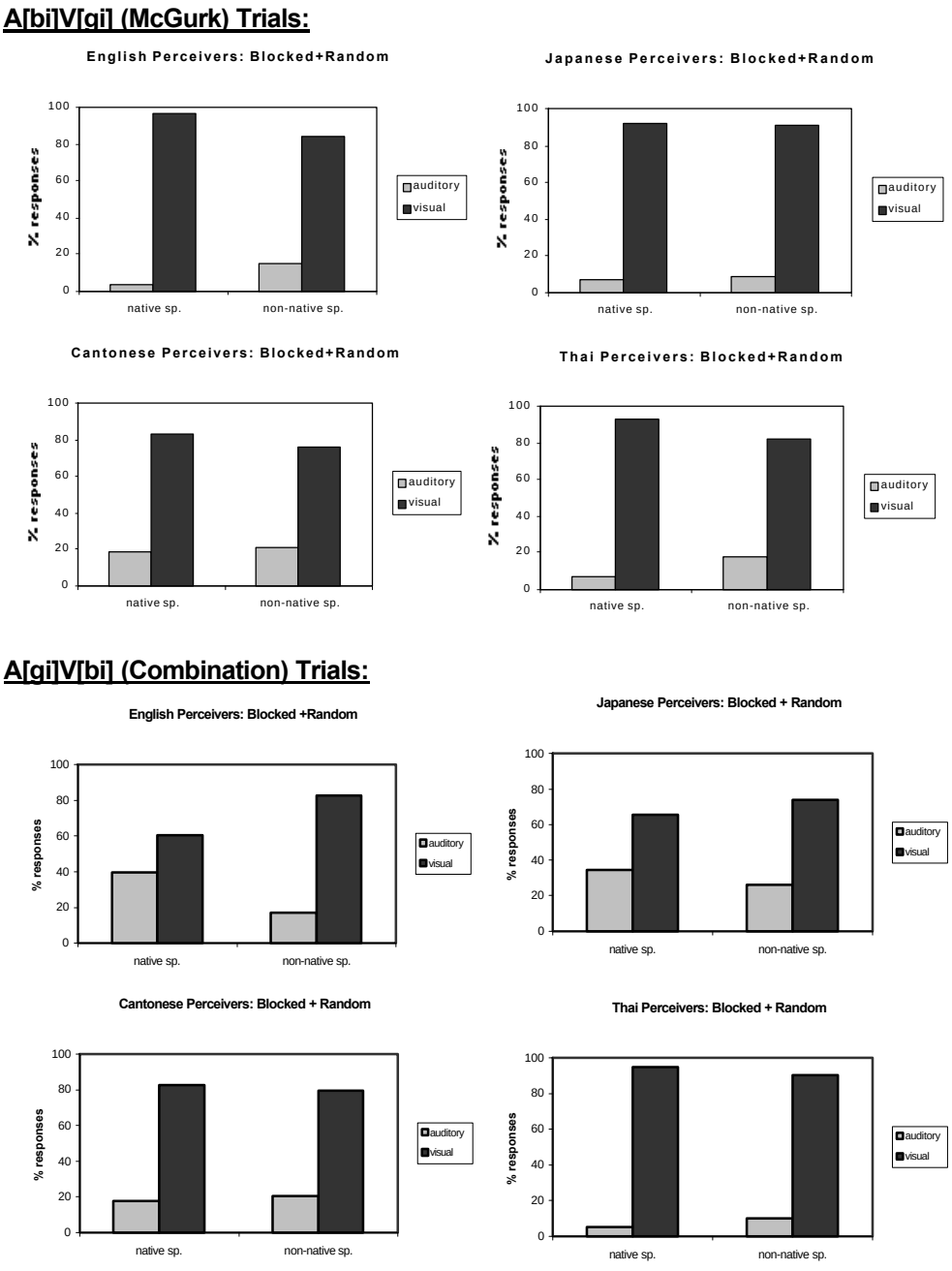


Figure 3. Percent auditory ("bi") vs. visual ("gi, di, thi, bgi, gbi") responses x native vs. non-native speakers averaged across blocked and random conditions in A[bi] V[gi] (McGurk) and A[gi]V[bi] Trials

being blocked or random. Subjects generally made similar numbers of auditory ("gi") responses and visually-influenced responses ("bi", "di", "thi", "bgi", and "gbi") but there were comparatively more auditory than visual response made by Japanese than Cantonese or Thai subjects, $F(1,88) = 9.39$. As for the fusion trials, the proportion of auditory and visually-influenced responses interacted with native vs. non-native language, $F(1,88)= 4.45$. As can be seen in the lower four graphs of Figure 3, the predominance of

For the A[gi]V[bi] trials, as was the case in the McGurk trials, there were no significant effects due to the trials

visual over auditory responses increased dramatically when subjects were presented with non-native as opposed to native speech. This was particularly so for the English-speaking subjects, $F(1,88) = 12.20$.

For the Fusion trials the effect of a non-native speaker was to increase the proportion of *auditory* responses, while conversely for the combination trials, the effect of a non-native speaker was to increase the proportion of *visually-influenced* responses. Consideration of the make-up of these trials, A[bi]V[gi] and A[gi]V[bi], reveals that non-native speakers concentrated more on the [bi] component whether this was presented auditorily or visually.

4. DISCUSSION

When fusion stimuli, A[bi]V[gi], were presented, subjects showed definite interlingual effects though not in the expected direction. The appearance of a foreign speaker on the screen set the perceiver to allot greater attention to the *auditory* stimulus as opposed to the *visual* stimulus found by Sekiyama and Kuhl [3,4].

It is possible that this difference occurs due to the different vowel contexts used in the two experiments. Sekiyama used the [a] vowel while we used [i]. It has been found that the [i] vowel results in a greater incidence of "di" as opposed to "thi" responses, while [a] results in more "tha" than "da" responses [6]. As [ð] is not phonemic in Japanese, Cantonese, or Thai, the use of [i] here allows these perceivers to report more visually-based "di" fusion responses. So for these speakers there is relatively less visual ambiguity in the [i] than the [a] vowel case, and when a foreign speaker is presented more attention is directed towards the auditory component. This explanation is tenuous at best, but it does raise the possibility that the effect of a foreign speaker is not uniform and that it depends upon the phonetic characteristics of the phones being presented.

This explanation is given some support by the results for the A[gi]V[bi] trials. For these the interlingual effect was in the opposite direction: for non-native speakers there was an increase in *visual* reliance. The common element in the A[bi]V[gi] and the A[gi]V[bi] results is that viewing a non-native speaker increased reliance on the bilabial [b] at the expense of the velar [g]. It is difficult to posit a general principle to explain this, but it seems that the information for [b] does become the focus of attention in AV conflict trials when a non-native speaker is presented.

As these interlingual effects for fusion and combination trials were obtained in both blocked and random presentations, it appears that the bias occurs as a result of an expectancy based on the appearance of a foreign face rather than the divergence of the presented phones (either their auditory or visual manifestation) from stored native language prototypes as Sekiyama and Kuhl have suggested [3,4]. Nevertheless, the increased concentration upon the bilabial in non-native speech in both conditions suggests that there is at least some intrusion of phonetic factors in this interlingual effect. In any event, it appears that the effect of non-native faces and speech on subjects' integration of auditory and visual speech is not straightforward. Insofar as there is not a constant increase in visual *or* auditory reliance upon the presentation of a non-native speaker, these results suggest that subjects' attention may be directed by visual or auditory information depending on the situation.

It is interesting to note that in non-conflict trials subjects were able to make use of the blocked trials to learn to differentiate just those aspects of the auditory and visual components that were difficult to perceive (auditory information for [bi] and visual for [gi]). The bias from perceiving a foreign face must also have been operating in these trials, but perhaps the rather unusual unimodal presentation of the components alone in Aud-Only and Vis-Only trials allowed (or required) perceptual learning to take place. Conversely, the blocked versus random manipulation did not have any effect in the mismatching AV trials (fusion or combination), so the effects in AV trials appear to be more immediate.

It may be the case that the relative weighting of auditory and visual information in non-native presentation depends upon the relative difficulty of incorporating the foreign auditory or visual information into a coherent percept. This issue awaits investigation via an appropriate methodological vehicle. Rather than the blocked / random manipulation, it may be necessary in future experiments to dub speech sounds across speakers such that subjects see a non-native face "producing" native speech and vice-versa.

5. REFERENCES

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

2. Burnham, D. (1998) Language specificity in the development of auditory-visual speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds) *Hearing by Eye II: The Psychology of Speechreading*. London: Psychology Press (27-60).

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

4. Kuhl, P.K. Tsuzaki, M., Tohkura, Y., & Meltzoff, A. (1994) Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces. *Proceedings of the International Conference on Spoken Language Processing*, Tokyo, 1994, S11-4.1.

5. Burnham, D., Fowler, J., & Nicol, M. (1997) An online procedure for creating and running auditory-visual

experiments: hardware, software, and advantages. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (eds) *Eurospeech '97 Proceedings: ESCA 5th European Conference on Speech Communication and Technology, 3*, 1683-1686.

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

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

8. 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. *Journal of the Acoustical Society of America, 84,* S155.