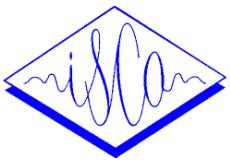

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

The Effect of Tonal Information on Auditory Reliance in the McGurk Effect

Denis Burnham and Susanna Lau

School of Psychology, University of NSW, Sydney, Australia

ABSTRACT

The McGurk effect occurs when conflicting auditory and visual speech information result in an emergent percept. The incidence of the McGurk effect is greater for speakers of English than Japanese, and in turn for speakers of Japanese than Cantonese. Sekiyama postulates that this is because speakers of tonal languages rely more upon auditory than visual information in speech perception. Here this hypothesis is tested by presenting both tonal (Cantonese) and non-tonal (English) language speakers with McGurk stimuli in which the tone on syllables either varied or remained constant across trials. Cantonese perceivers relied more upon auditory information than did Australian perceivers, but over and above this, tone variation affected the relative salience of auditory and visual information. This effect was different for the two syllables [ba] and [ga]. In auditory-visual conflict trials subjects' responses appear to depend on the effect of the tone variation on the specific auditory and visual syllables. The results are discussed in terms of the auditory and visual correlates of tone.

1. INTRODUCTION

The McGurk effect occurs when discrepant auditory and visual information result in an emergent percept, eg, auditory [ba] and visual [ga] are usually perceived as [da] or [ða] [1]. This has been found with the speakers of various languages - English [1], German and Spanish [2], Thai [3], Dutch [4] Japanese [5,6] and Chinese [4,7]. Despite this apparent universality, Sekiyama has uncovered a cross-cultural difference - the "Japanese McGurk", in which Japanese perceivers rely less upon visual information than do English language perceivers. Sekiyama has explained this in terms of the impoliteness of looking directly at the talkers' face in Japanese culture [6, but see 8]; the lack of visually distinct phonemes and consonant clusters in Japanese [6]; and most recently by what could be called the 'tone hypothesis'. As Japanese is a pitch-accented language it is suggested that there is less visual reliance by Japanese perceivers because pitch accent or tones are more auditorally distinct than they are visually distinct [7].

Nevertheless, Japanese language users are not insensitive to visual information, for Sekiyama [8] found that Japanese subjects report incompatibilities between conflicting auditory and visual information, even more than do American subjects. In addition, Sekiyama [6,8] has found that when auditory noise is added to the signal, the Japanese McGurk effect is augmented compared with a no-noise condition, ie, auditory unintelligibility of speech influences Japanese subjects' reliance upon visual information. Thus, Japanese subjects tend not to integrate visual information as long as the auditory modality provides them with sufficient information, but show increased visual reliance when the auditory stimulus is degraded.

This evidence is consistent with the tone hypothesis, but more direct evidence is required for its support. Recently Sekiyama [7] found that Chinese perceivers have an even weaker McGurk effect than Japanese perceivers. Chinese could be considered to be both quantitatively and qualitatively more tonal than Japanese - Chinese has six tones, while Japanese has just two pitch-accents, therefore this evidence is consistent with a strong formulation of the tone hypothesis - that the more tonal the language, the greater the auditory reliance.

Nevertheless, the tone hypothesis still awaits a direct test. It might be considered that if the presence of lexical tones in speech input result in perceivers' increased reliance on auditory input, then it should be possible to witness this by manipulating the degree of tonal input. This is what was done here: two groups of subjects were tested, a Tone Varying group, for whom the lexical tone on the stimuli varied across trials, and a Tone Constant group, for whom the same lexical tone was used in all trials. Within these groups, half of the subjects were native speakers of the tonal language, Cantonese, and half were native speakers of a non-tonal language, English. The speech stimuli were produced by a native Cantonese speaker in one condition, and a native Thai speaker in another, with all subjects being presented with both speaker conditions. For English language users unschooled in Cantonese or Thai, this manipulation should have no differential effect - lexical tones of both speakers would be equally unfamiliar to them. For the Cantonese perceivers, the Cantonese speaker would, of course, be native to them, and the Thai speaker foreign, with the specific tones of Thai, but not the concept of tonal variation on lexical items, being unfamiliar.

If tones do result in greater auditory reliance then this could take two main forms. If the effect of tones on auditory-visual speech perception is independent of the language background of the perceiver, and if the effect is relatively immediate, then participants in the tone varying groups should demonstrate greater auditory reliance than those in the tone constant groups, irrespective of the language background of the speaker (Cantonese or Thai), or the perceiver (Cantonese or Australian English). Alternatively, if the effect is only manifest after intensive linguistic induction, then only the Cantonese perceivers should demonstrate heightened auditory reliance.

2. METHOD

2.1. Design

A 2 x 2 x (2) Language group (Cantonese / Australian) x Tone Condition (varying / constant) x Speaker (Cantonese / Thai), with repeated measures on the last factor was employed. Native Cantonese and native Australian-English speakers were randomly assigned to either the tone varying or tone constant condition, but all four subgroups were presented with a Cantonese speaker and a Thai speaker, with order of presentation counterbalanced. Subjects' perceptual decisions and their reaction times (RTs) were measured for each trial, but only the former are analysed here.

2.2. Subjects

A total of 48 adult subjects were recruited from the introductory psychology pool at the University of NSW. Of these 24 were native Cantonese speakers with no knowledge of Thai, and 24 were native Australian-English speakers with no knowledge of Thai or Cantonese. Half (12) of the subjects in each language group were assigned to the Tone Varying group, and the other half to a Tone Constant group. Within each of these four Language Background x Tone Condition sub-groups, subjects were presented with two separate blocks of trials one with a Cantonese-speaker and the other with a Thai speaker with the order of presentation of speakers counterbalanced across subjects.

2.2. Stimuli

For both the tone varying and tone constant groups, stimuli consisted of the syllables [ba] and [ga] presented as auditory-only, visual-only, or auditory-visual presentations, the latter being either matching or mismatching. The syllables [ba] and [ga] are both used in both Cantonese and Thai. A native Central Thai speaker and a native Cantonese speaker were videotaped under controlled conditions producing these CV syllables using all different tones in the language - six in Cantonese and five in Central Thai.

Lexical Tone Selection: For the tone constant groups for each language a relatively neutral lexical tone was chosen as the tone in which all stimuli would be presented. For Thai this was the mid tone, and for the Cantonese condition this was the low tone. For the tone varying groups, three tones were used in each language condition. For Thai, the three tones, high, low, and falling were used. These were chosen on the basis of the results of a perception study on the discriminability of the five Thai tones [9]. No similar data are available for Cantonese, however on the basis of inspection of F_0 contours and informal perception experiments with four members of our laboratory, the high-falling, high-mid and low rising tone were chosen for Cantonese. So tone constant groups subjects were presented with the various syllables always in the same tone, a condition similar to normal conditions in a non-tonal language, whereas for tone varying groups trials randomly varied between the three tones, a condition similar to normal conditions in the tone languages in question.

Trial Types and Stimulus Construction: The stimuli were edited and dubbed on-line in order to produce auditory-only (Aud-Only), visual-only (Vis-Only), and either matching or mismatching auditory-visual stimuli (AV). The visual components were produced from the original videotapes, while the auditory components were digitised (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 variability and phonetic invariance. The CAVE (Computerised Auditory-Visual Experiment) package, developed by the first author [9], was used for presenting stimuli and collecting speeded forced choice responses. In this, auditory-visual stimuli are created on-line at the time of testing each subject. 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 so just the stimulus person's face and lip movements without sound are presented.

This CAVE package was used to produce stimuli consisting of Aud-Only, Vis-Only, and matching AV presentations of the syllables [ba] and [ga] and mismatching presentations of A[ba]V[ga] and A[ga]V[ba]. 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, about 1 sec of articulation, and 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.

2.3. Stimulus Presentations

Each speaker condition (Cantonese or Thai) was presented separately. In each there was a practice and a test trial phase as described hereunder.

Practice Trials: For each speaker condition (Thai, Cantonese) there were 18 practice trials. The only difference between conditions was the language background of the speaker. For the tone-varying groups these practice trials comprised of one presentation of each of the three tones on [ba], and each of the three tones on [ga] in each of the three modes, Aud-Only, Vis-Only, and AV. In the tone-constant condition, since stimuli were presented using the same lexical tone, these 18 practice trials were then comprised of three instead of one presentation of [ba] and [ga] in each of the three different modes. Practice trials were included to allow the subjects to become familiar with the testing procedure. The practice trial results were used only to eliminate the data of subjects who responded too slowly or inaccurately.

Test Trials: Following the practice trials, in each speaker condition (Thai, Cantonese), there were two 36-trial test blocks. Each block consisted of exactly the same trial types with trial presentation order varied between blocks, and test block sequence counterbalanced between subjects. In each block there were 3 Aud-Only [ba], 3 Aud-Only [ga], 3 Vis-Only [ba], 3 Vis-Only [ga], 3 AV matching presentations of [ba], 3 AV matching presentations of [ga], and 9 each of mismatching A[ba]V[ga] (McGurk stimulus), and A[ga]V[ba] (combination stimulus). For the tone constant groups all stimuli were presented using the same lexical tone (low for Cantonese and mid for Thai). In the tone varying group, one of the 3 repetitions of the Aud-Only, Vis-Only, and AV matching trials was in each of the three tones, and for the 9 A[ba]V[ga] and 9 A[ga]V[ba] trials 3 in each were of each of the three tones.

Trials were presented in pseudo-random order with no more than three of the same trial type (Aud-Only, Vis-Only, AV), or the same syllable ([ba], [ga]) occurring consecutively. In the tone varying groups this also applied to the tone type, no more than three exemplars of the same tone occurring consecutively.

2.4. Apparatus and Procedure

Subjects were tested individually in a sound-attenuated room, seated in front of a monitor connected to the videorecorder and computer in the control room. A response pad placed in front of the monitor had a central 'ready' key with six response buttons (labelled 'ba', 'ga', 'da', 'tha', 'bga' & 'gba') arranged in a semicircle around it. A reward light output from the computer was attached to the left side of the monitor. This flashed when subjects responded correctly (only during practice trials). A computer-controlled error buzzer sounded to inform subjects and experimenter of any failure to respond appropriately, eg, if response times were too long.

In each condition there was one block of practice trials followed by two test trial blocks. Subjects pressed the ready key to start each trial. On stimulus presentation the subject was required to respond as quickly and accurately as possible, by pressing the response button which "best matched the syllable the speaker used". If the subject failed to respond within 3000msec, or took their finger from the button prior to the onset of the sound, the error buzzer sounded and a null trial was recorded. When subjects finished the first block of trials (Thai or Cantonese speaker depending on counterbalancing) they were given a few minutes rest before beginning the second condition. Results were recorded on-line and collated by the CAVE package. Testing lasted approximately 30 minutes.

3. RESULTS

3.1. Aud-Only, Vis-Only, and AV Trials

The first stage of analysis concerned the effect of tone condition (varying or constant) on auditory and visual information. The percent correct responses in Aud-Only, Vis-Only, and matching AV trials were analysed in two analyses of variance (ANOVA), one for [ba] stimuli, and another for [ga] stimuli. In each the design was language background of the perceivers (Cantonese / Australian) x tone condition (varying / constant) x (language of the speaker, Cantonese / Thai) x presentation mode (Aud-Only / Vis-Only / AV stimuli) with repeated measures on the last two factors. The results, split for Cantonese and Australian perceivers, but collapsed over the Cantonese and Thai speakers are shown in Figure 1.

If tone variation induces greater auditory reliance, then it might be expected that percent correct should be greater for the Aud-Only trials in the tone varying condition than in the tone constant condition, and that on Vis-Only trials percent

correct should be greater for tone constant than the tone varying condition. However, as can be seen in Figure 1, the results appear to depend upon the syllable, [ba] or [ga].

For the [ba] trials there were significantly more correct responses in the Aud-Only than the Vis-Only trials, F (1,44) = 8.22, and this trend was more pronounced for the Australian perceivers, F (1,44) = 12.28. Most interestingly, there was an interaction of Aud-Only / Vis-Only with tone varying / tone constant: Unexpectedly, in the Aud-Only [ba] trials, there were slightly more correct responses when the tone remained constant than when it varied, and in the Vis-Only trials there were more correct responses in the tone varying than the tone constant condition.

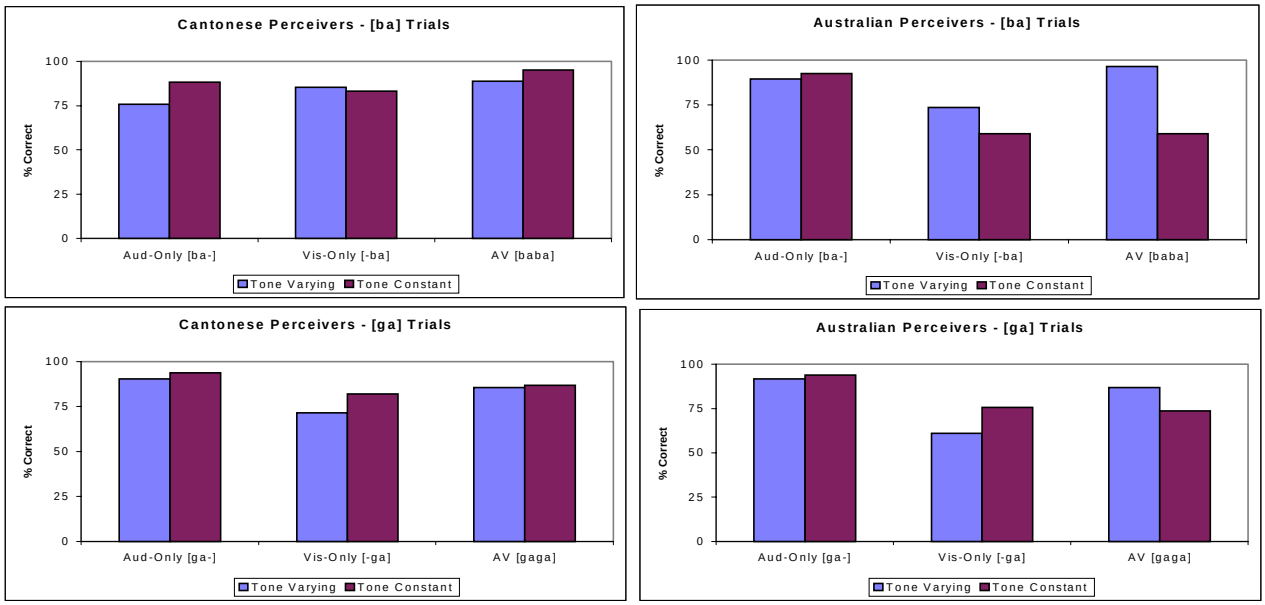

Figure 1: Percent Correct Responses on Aud-Only, Vis-Only, and AV [ba] and [ga] Trials

For the [ga] trials, again there were more correct responses for Aud-Only than Vis-Only presentations, F (1,44) = 6.08. There was also an interaction with tone condition and the language of the perceiver, F (1,44) = 19.32: while there were generally less correct visual only responses for the Australian perceivers, for both the Cantonese and Australian perceivers, there were more correct responses for Vis-Only [ga] when tone was constant than when it varied, and for Aud-Only [ga], only slightly more correct responses for the tone constant condition.

To summarise, the correct responses in the tone conditions showed the following pattern.
Aud-Only [ba]: Tone Constant > Tone Varying
Vis-Only [ba]: Tone Varying > Tone Constant
Aud-Only [ga]: Tone Varying > Tone Constant (but very slight).
Vis-Only [ga]: Tone Constant > Tone Varying.

Thus the results for [ba] are the opposite of what was expected, and for [ga], the expected result only occurred for the Vis-Only trials. There appears to be a definite effect of the type of syllable here. If this is simply idiosyncratic for these two syllables, then this is relatively uninteresting. However, if some general principle can be extracted then this finding may be theoretically significant. As [ba] is more visually distinct than [ga], it may be the case that tone variability on [ba] actually accentuates visual information for [ba], but that on the visually less distinct [ga] the tonal variation is not so salient. This of course would imply that there is visually distinct information for different tones, a proposition that has not yet been empirically tested.

Irrespective of the differences in the Aud-Only and Vis-Only trials, it is interesting to note for the AV trials there is a slight tendency for Cantonese perceivers in the tone constant condition to perform more accurately, and a clear tendency for Australian perceivers in the tone varying condition to perform more accurately. This may be a novelty effect, that is, for an unfamiliar situation (tone constant for Cantonese perceivers and tone varying for the Australian perceivers), participants may invest greater attention and thus identify more trials correctly. Again this is a proposition that requires testing.

3.2. A[ba]V[ga] Trials

In the second stage of the analysis, responses in auditory-visual conflict trials were tested in two ANOVAs, one for the A[ba]V[ga] (McGurk), and one for the A[ga]V[ba] (Combination) trials. In both a perceiver language background (Cantonese / Australian) x tone condition (varying / constant) x language of the speaker (Cantonese / Thai) x response ("ba", "ga", "da", "tha", "bga", "gba") design with repeated measures on the last two factors was employed.

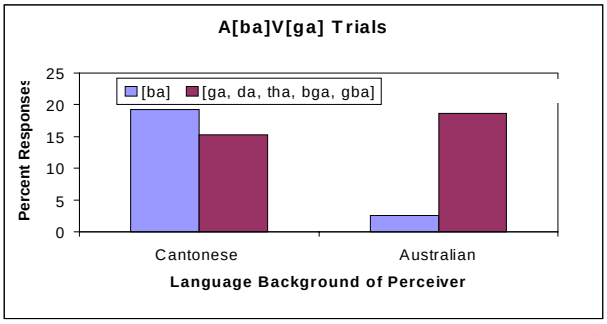

Figure 2: McGurk Trials - Cantonese/English x Auditory/Visual Interaction

There were three questions of interest: whether the tonal language (Cantonese) speakers performed differently than non-tonal (English speakers; whether the presence of tonal variation vs tonal constancy affected subjects' responses; and whether the tonal language (Cantonese) speakers performed differently for their native tonal language (Cantonese) vs a foreign tonal language (Thai). The results did not uphold the notion that any effects were simply native language effects, ie., the effects were not specific to Cantonese perceivers responding to a Cantonese speaker. This factor is not discussed further here. With respect to Cantonese vs English perceivers, there was an interaction of language background and the incidence of auditory-based, "ba", responses vs all visual-based responses ("ga, da, tha, bga, gba"), $F(1,44) = 6.97$, and similarly for the incidence of auditory-based, "ba", responses vs visual-based "ga" responses, $F(1,44) = 6.15$. The former interaction is shown in Figure 2.

The effects for tone variation can be seen in Figure 3. There was a significant interaction of tone varying vs tone constant with fusion responses ("da, tha") vs combination responses ("bga, gba"), $F(1,44) = 6.14$; and also an interaction which just failed to reach significance, between single component, "ba" or "ga" responses, vs compound responses with the tone varying vs tone constant conditions, $F(1,44) = 3.90$.

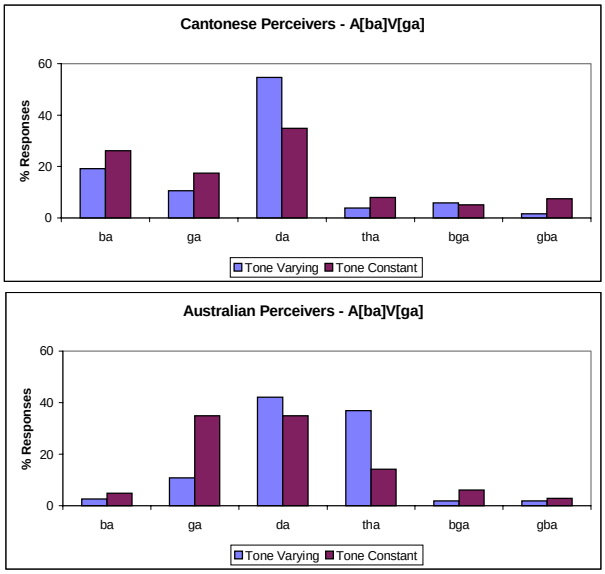

Figure 3: Responses in A[ba]V[ga] Trials

Essentially these suggest that, as might be expected on the basis of the results for the non-conflict trials (see Figure 1, and accompanying summary), there were more visual [ga], and more auditory [ba] responses in the tone constant condition. Correspondingly, there were more compound (especially fusion) responses in the tone varying condition. Thus the effect of tonal variation, irrespective of language background of speaker or perceiver, appears to be to reduce reliance on either the auditory or the visual input alone, and to increase the probability that both auditory and visual information will be combined.

3.3. A[ga]V[ba] Trials

For responses to A[ga]V[ba] there was a significant interaction of language background of the perceiver (Cantonese vs English) with the auditory-based, "ga", responses vs all the other visually influenced responses ("ba, da, tha, bga, gba"), $F(1,44) = 11.20$,

and also between language background and the auditory-based, "ga", responses vs the visual-based,

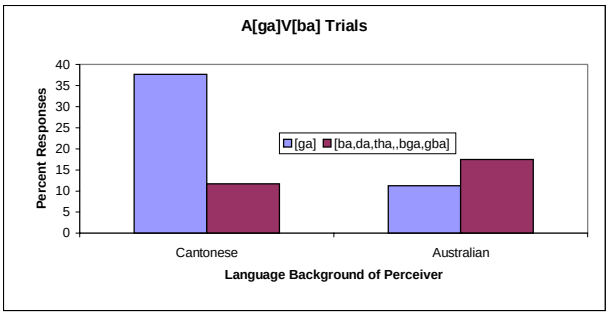

Figure 4: Combination Trials - Cantonese/English x Auditory/Visual Interaction

"ba" responses, F(1,44) = 10.62. The former interaction is shown in Figure 4. As can be seen the result is similar to that for the A[ba]V[ga] trials: when auditory and visual information are in conflict, Cantonese language background subjects appear to rely more upon the auditory, whereas those with an English language background appear to rely more upon the visual information.

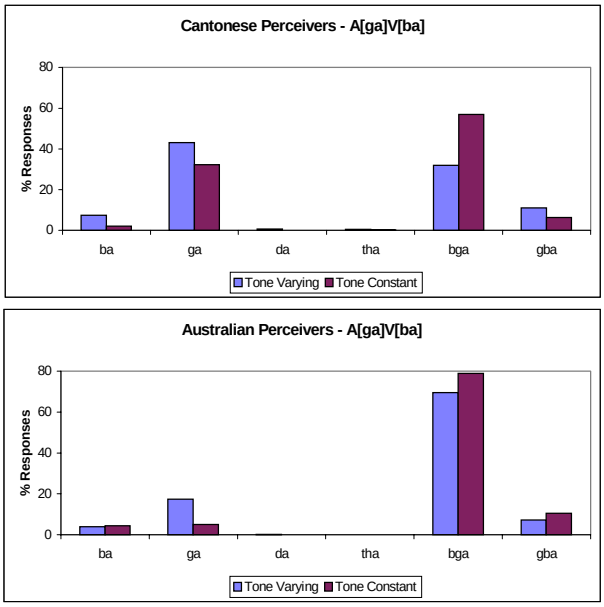

Figure 5: Responses in A[ga]V[ba] Trials

Graphs of the effect of tone variation are given in Figure 5. There are more compound responses than single component (auditory "ga" or visual "ba" responses) for the Australian but not the Cantonese perceivers, F(1,44) = 10.43, and for both language groups, more combination ("bga" & "gba") than fusion ("da & tha") responses, F(1,44) = 10.62. However, none of these effects interact significantly with the tone conditions, even though it may appear from the graphs that they might. For example, it is interesting to note that there are more auditory "ga" responses for the tone varying groups and correspondingly more "bga" responses for the tone constant groups, but this effect is not significant.

4. CONCLUSIONS

In support of Sekiyama's studies, when auditory and visual information conflict, Cantonese subjects make more responses based on auditory information alone, whereas Australian subjects make more visually-influenced responses. Whether this is due to the tonal nature of Cantonese cannot be determined here.

Despite this language background difference, there are effects of tone variation vs constancy, although not in the expected direction - that tonal variation results in greater auditory reliance. There appears to be significant involvement of the type of consonant, and perhaps whether it is visually distinct. Speakers of a tonal language, Cantonese, *and* speakers of a non-tonal language, English, are both sensitive to these manipulations of tonal variation.

5. REFERENCES

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

2. Fuster-Duran, A. (1995). Perception of conflicting audio-visual speech: an examination across Spanish and German. In D.G. Stork & M.E. Hennecke (Eds), Speechreading by Humans and Machines. Berlin: Springer-Verlag.

3. Burnham, D., & Dodd, B. (1996). Auditory-visual speech perception as an direct process: the McGurk effect in human infants and across language. In D.G. Stork & M.E. Hennecke (Eds.), Speechreading by Humans and Machines. Berlin: Springer-Verlag

4. de Gelder, B., Bertelson, P., Vroomen, J. 7 Chen, H. C. (1995). Interlanguage differences in the McGurk effect for Dutch and Cantonese listeners. Proceedings of the Fourth European Conference on Speech Communication and Technology (pp. 1699-1702). Madrid.

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

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

7. Sekiyama, K. (1996). Cultural and linguistic factors influencing audiovisual integration of speech information: The McGurk effect in Chinese subjects. Perception & Psychophysics, 59, 73.

8. Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. Journal of Acoustical Society of Japan, 15(3), 143-158.

9. Burnham, D., & Francis, E. (1997). The role of linguistic experience in the perception of Thai tones. In T.Thongkum (Ed) SouthEast Asian Linguistic Studies in Honour of Vichin Panupong (Science of Language Vol. 8). Bangkok: Chulalongkorn UP.

10. Burnham, D., Fowler, J. & Nicol, M. (1997) CAVE: An on-line procedure for creating and running auditory-visual speech perception experiments - Hardware, software and advantages. In *Proceedings of the 5th ECSCT*, Rhodes.