The Perception and Production of Phones and Tones: The Role of Rigid and Non-Rigid Face and Head Motion

Denis Burnham[1], Jessica Reynolds[1], Eric Vatikiotis-Bateson[2], Hani Yehia[3], Valter Ciocca[4], Rua Haszard Morris[1], Harold Hill[5], Guillaume Vignali[1], Sandra Bollwerk[1], Helen Tam[1], Caroline Jones[6]

[1]MARCS Auditory Laboratories, University of Western Sydney, Australia

[2]Department of Linguistics, University of British Columbia

[3]Center for Research on Speech, Acoustics, Language and Music, UFMG, Brazil

[4]Division of Speech and Hearing Sciences, University of Hong Kong

[5]Cognitive Information Sciences Laboratories, ATR, Japan

[6]School of Education, University of New South Wales

d.burnham@uws.edu.au, 13506894@scholar.uws.edu.au, evb@interchange.ubc.ca, hani@cefala.org, vciocca@hkusua.hku.hk, rua@adi.co.nz, hill@atr.jp, guillaume@vignali.net, H.H.Y.TAM@soton.ac.uk, sc.jones@unsw.edu.au

Abstract. *There is evidence, mostly with phones (consonants & vowels), that visual concomitants of articulation facilitate speech perception. Here the visual concomitants of lexical tone are considered. In tone languages fundamental frequency variations signal lexical meaning. In a word identification experiment with auditory-visual words differing only in tone, Cantonese perceivers performed above chance in a Visual Only condition. A subsequent study showed augmentation of word pair discrimination in noise in an Auditory-Visual compared to an Auditory Only condition for Cantonese, tonal Thai speakers, and even non-tone Australian speakers). The source of this perceptual information was sought in an OPTOTRAK production study of a Cantonese speaker. Functional Data Analysis (FDA) and Principal Component (PC) extraction suggests that the salient PCs to distinguish tones involve rigid motion of the head rather than non-rigid face motion. Results of a final perception study using OPTOTRAK output in which rigid or non-rigid motion could be presented independently in tone differing or phone differing conditions, suggests that non-rigid motion is most useful for the discrimination of phones, whereas rigid motion is most useful for the discrimination of tones.*

1. Background

Speech is auditory-visual: whenever visual (lip, face, and head and neck motion) information is available, humans use it to augment, and modify speech perception. Sumby and Pollack, (1954) showed augmentation: accuracy increases of 40-80% when speech presented in a noisy environment is accompanied by the speaker's face; and McGurk and McDonald (1976) showed modification: the McGurk effect, in which auditory [ba] paired with visual [ga] is perceived as "da"

or “tha”. So, both auditory and visual speech information is important in speech perception, and it is argued that this is because convergent information better specifies the speech source (Vatikiotis-Bateson, Kuratate, Munhall, & Yehia, 2000). The McGurk Effect occurs in tone languages - Cantonese (deGelder, Bertelson, Vroomen & Chen, 1995) and Thai (Burnham, 1992), but only insofar as it effects phones - consonants and vowels. So, until recently all the research on auditory-visual speech perception has concerned the perception of phones, with no consideration of visual information for tones.

Tone is primarily based on F_0, e.g., Cantonese has 6 tones, as in /fu55/ ‘husband’, /fu33/ ‘rich’, and /fu22/ ‘father’, /fu25/ ‘tiger’, /fu21/ ‘to hold’, and /fu23/ ‘woman’. In pitch-accented languages tone is carried between syllables, e.g., Japanese has 2 pitch-accents, high-low, e.g., ka⌈ki ‘oyster’, and low-high, e.g., ka⌉ki ‘persimmon’. Auditory-visual speech perception may operate differently in tone languages: Japanese listeners’ McGurk effect perception is less influenced by visual speech than is that of their American counterparts (Sekiyama, 1994), and the effect is further reduced in Chinese perceivers (Sekiyama, 1997). Sekiyama reasons that as there are 6 tones in Cantonese, 2 pitch-accents in Japanese, and none in English, these cross-language auditory-visual effects could result from the relative prevalence of tone, which presumably has few visual concomitants.

There are visual correlates of F_0 in speech production. Cavé, Guaïtella, Bertrand, et al. (1998) showed correlation between French speakers’ eyebrow motion and intonation in sentences, and there are strong correlations between head motion and F_0 during speech (Yehia, Kuratate, & Vatikiotis-Bateson, 2002), which are continuous and seem to be used in auditory-visual perception (Vatiktiotis-Bateson et al., 2000). However, studies of the visual concomitants of tone are lacking.

Auditory-visual perception and production of Cantonese tone is investigated here in two perception studies of Cantonese perceivers’ identification, and Cantonese, Thai, and Australian perceivers’ discrimination of tone in auditory-only(AO), visual-only(VO), and auditory-visual(AV) modes; and in a production study, measuring auditory and visual concomitants of phone and tone production, leading to the phone/non-rigid, tone/rigid hypothesis, which was tested in a final discrimination study with words presented with rigid, non-rigid or combined motion in AO, VO, or AV conditions.

2. AV Perception Tone: Preliminary Studies

Brief versions of two experiments are given here. For further details see Burnham, Ciocca, and Stokes (2001), and Burnham, Lau, Tam, Schoknecht, C. (2001), respectively.

2.1 AV Tone Identification - Cantonese Perceivers

Method: A 2 x 2 x 2 x (3 x 6 x 4 x 2) design was employed. The first three factors were group manipulations: *phonetic background* - participants with or without prior phonetic training; *word presentation* - isolated words / words in sentences; and *feedback,* feedback for correct responses / no feedback. The remaining within-subjects factors were *presentation mode* – AO, VO, or AV; tone – the *6 Cantonese tones*, high (5-5), low-mid/high-rising (2-5), mid (3-3), low-mid/low-falling (2-1), low-mid/mid-rising (2-3), and low-mid (2-2); phonemic strings - *4 Cantonese phonemic strings* on which tones were carried – 2 with monophthongal vowels, /fu/ and /fan/, and two with diphthongs, /soej/ and /hau/; and *repetitions* – each of the above 72 combinations were presented twice.

Forty-eight adult native Cantonese speakers were tested, 24 trained phoneticians, and 24 non-phoneticians with appropriate group counterbalancing. Stimuli were the 24 Cantonese words (/fu/, /fan/, /soej/ and /hau/ x 6 tones) spoken by a 23-year-old native Cantonese female, and recorded on

a digital video-recorder. Stimuli were presented in two contexts: isolated Cantonese words (Mean = 57.34dB), or in a semantically-neutral Cantonese carrier sentence (Mean = 55.17dB). Participants were tested individually in a sound-attenuated room, approximately 50 cm from the screen of a PC running the DMDX experimental software. After each audio stimulus, orthographic forms of the correct word, along with 5 distracters from the same PI-TD sextuplet were presented, labeled 1-6. Participants were instructed to identify the word said by pressing the corresponding number key, as quickly and as accurately as possible.

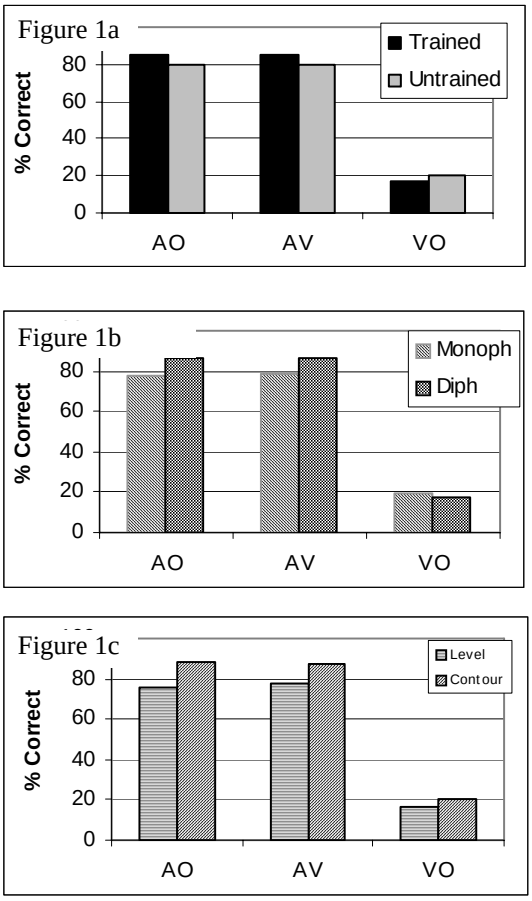


Figure 1: Identification: effect of (a) Phonetic Training (b) Vowel Type, and (c) Tone Type

Results and Discussion: Preliminary analyses revealed no feedback or repetitions effects. Results collapsed over feedback, repetitions, and words/ sentences, are given in Figure 1. Performance was generally better for words in sentences than in isolation, $F(1,40)=13.773$, $p<.005$. AO and AV performance was good and statistically equivalent, but VO performance overall did not differ from chance (1/6=16.67%). However, there were significant interactions of mode and:

- phonetic training, $F(1,40) = 6.80$, $p< .05$: those *with* phonetic training did better in AO and AV, whereas those *without* phonetic training did better in the VO) condition (Fig. 1a);
- vowel type, $F(1,40) = 31.31$, $p <.001$: better performance for diphthongs than monophthongs (M = 78.13) in AO and AV, but for monophthongs in VO;
- tone type, $F(1,40) = 11.10$, $p <.005$: more augmentation for contour tones in the AO and AV than in VO.

The differences are underlined by VO tests against chance (Table 1). Together the results suggest that as the vast majority of the population is phonetically *un*trained, visual perception of tone slightly, but significantly above chance is the norm, not the exception, and that visual perception of tone is better under some stimulus conditions than others.

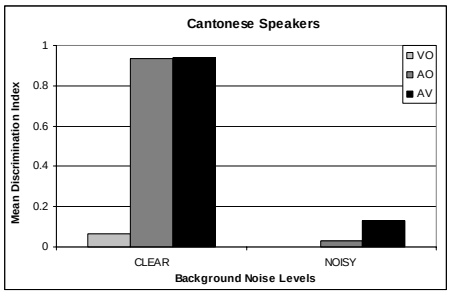
Table 1: Visual Only Tests Against Chance Level – 16.67% (Mean, *SE*), alpha = .05

Factor	***Significantly > Chance***	***Not Significantly > Chance***
Phonetic Training	Without (**20.57** *1.27*)	With (**16.66** *1.18*)
Vowel Type	Monophthong (**19.88** *1.28*)	Diphthong (**17.36** *1.17*)
Word, Tone Type	Word, Contour (**20.31** *1.68*)	Word, Level (**15.63** *1.61*)
Sentence, Tone Type	Sentence, Contour (**20.66** *1.90*)	Sentence, Level (**17.88** *1.74*)

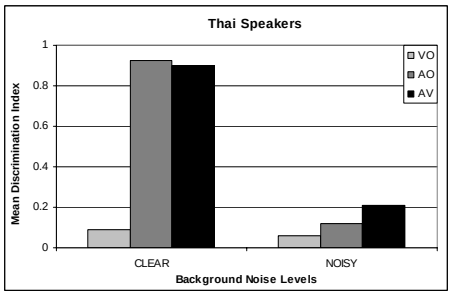
2.2 AV Tone Discrimination - Cantonese, Thai, Australian Perceivers

Is the visual perception of tone due to language-specific learned associations, or language-general principles? To test this Thai speakers (familiar with lexical tone, unfamiliar with Cantonese), and Australian English speakers (unfamiliar with lexical tone and Cantonese) were tested in an AX discrimination paradigm. With respect to the above identification experiment in 2.1: only

phonetically-naïve participants were tested; just two of the words from 2.1 were used (mean = 57.13dB); all 6 Cantonese tones, were used; only words in isolation were used; and an acoustic noise (multi-talker English language babble) condition was introduced (Mean = 63.9dB) to allow any augmentation of tone perception by visual tone information to be shown.

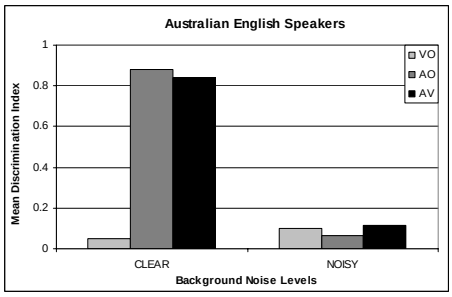


Method: In a 2 x 2 x 2 x (3 x 15 x 4) design, between-subjects factor were presence/absence of auditory *noise* (multi-talker babble); *vowel type* in stimulus word, in the minimal tone sextuplets of the monophthong, /fan/, or diphthong, /soej/; and inter-stimulus interval, 500ms/1500ms. Within-subjects variables were *presentation mode* - AO, VO, AV; *tone pair* - the 15 possible pairings of the six Cantonese tones; and *repetitions* – 4 of each possible pairings, two same and 2 different trials. From these, the dependent variable, a *hits - false positives* discrimination index (DI) was calculated for each of the 15 tone pairs yielding a maximum of +1, a minimum of –1, with a chance level of 0.



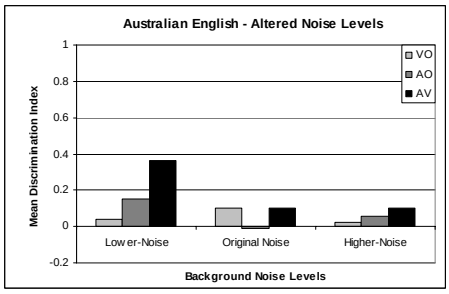
Cantonese Perceivers

48 native Cantonese speakers were tested. As shown in Figure 2a, there was a significant interaction between background noise and performance in AO vs AV, $F(1, 40) = 7.906$, $p<.05$. Discrimination was uniformly better in AV than in AO, but this effect was more apparent in noisy audio than in clear audio. Thus for Cantonese perceivers, in both audio noise and in audio clear conditions, visual face information significantly augments tone discrimination in AV compared with AO.



Thai Perceivers

The data for 48 native Thai speakers are shown in Figure 2b. There was a marginal superiority of AV over AO, $F(1, 40) = 4.35$, $p>.05$. This was qualified by a significant interaction between background noise and AO vs AV modes, $F(1,40) = 15.09$, $p<.001$, showing similar performance in AV ($M = 0.90$) AO and ($M = 0.93$) in clear audio, but in noisy audio better performance in AV ($M = 0.21$) than AO ($M = 0.12$). So visual face information for tone is perceptible even when the target language is foreign, so long as the perceivers are tone language speakers.



Australian Perceivers

Forty-eight 48 native Australian English speakers were tested. As shown in Figure 2c, there was no effect of AV vs AO either across or within each noise condition, suggesting that non-tone language speakers cannot perceive visual face information for tone.

Australian Perceiver Varying Noise Levels Experiment

In Experiment 2c, AE participants complained that the noise was distracting. So in this experiment, another 24 native AE speakers were tested, 12 in a Lower-Noise condition (babble level, 53.25dB),

and 12 in a Higher-Noise condition (babble, 74.6dB), only in the noise condition, and only on the /fan/ vowel level. Their results were compared to the 12 participants in the Original Noise AE sample who were tested in the noise condition with the /fan/ stimulus (babble level, 63.9dB). As shown in Figure 2d, for both the Higher-Noise and the Original Noise groups there were no significant effects of mode AO, VO, or AV - no augmentation due to visual face information. For the Lower Noise group performance in AV was significantly augmented over that in AO, $F(1, 10) = 9.93$, $p<.05$. Thus visual face information for tone is perceptible even for a foreign language, and even when tone languages are unfamiliar.

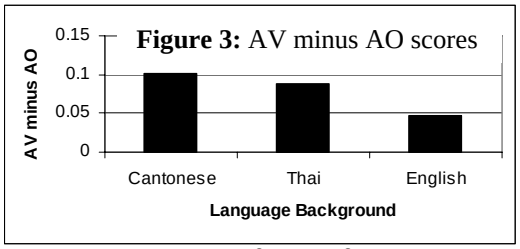


Figure 3: AV minus AO scores

It is interesting that, while visual information for tone is a language-general phenomenon, augmentation is graded in terms of language familiarity, and tone languages in general (Fig. 3).

3. Auditory-Visual Production of Tone

There is visual face information that allows tone perception, even by foreign, non-tone language speakers. We used a 4-step process to ascertain the nature of this information: (i) record auditory-visual productions (ii) derive Principal Components (PCs) from face motion, (iii) predict tone categories from PCs, (iii) correlate PCs with F0.

Recordings of Cantonese Speech: A 24-year-old female Cantonese speaker was recorded speaking each of the 5 PI-TD Cantonese words /fan/, /fu/, /hau/, /soej/ and /wai/ with each of the 6 Cantonese tones, in isolation or in one of 5 sentences constructed for each of the 5x6=30 words. Audio data; six parameters of head motion - x y z translation and roll, pitch, yaw from 4 rigid body markers; and movement from 17 face markers in 3D space (see Figure 4) using the OPTOTRAK apparatus.

Principal Components Analysis (PCA) on Face Motion: A method to work with 3-dimensional motion data preserving the time varying aspect was developed to analyse the visual component of head and face motion (Vignali, 2005a, 2005b). Starting from a motion description based on frames and Cartesian coordinates of markers, PCA was used to build a meaningful coordinate system, and a tool, OptoPCMarkerView2 was developed to display and independently manipulate PCs, and rigid and non-rigid motion. Functional Data Analysis (Ramsay & Silverman, 1997) assisted in the description of the motion on the basis of B-splines to capture the time dependency of frames. Time warping was applied to time-align similar features (e.g., jaw motion), and compute average curves to highlight motion differences. Deviation from the average was considered to characterize the tone-specific motion. The method was applied to the Cantonese tone OPTOTRAK data to give an average motion for each tone category. 14 PCs were sufficient to describe 99% of the variance. PC1 concerns jaw opening, PC2 lip aperture with jaw fixed, and PC3 rigid head nodding.

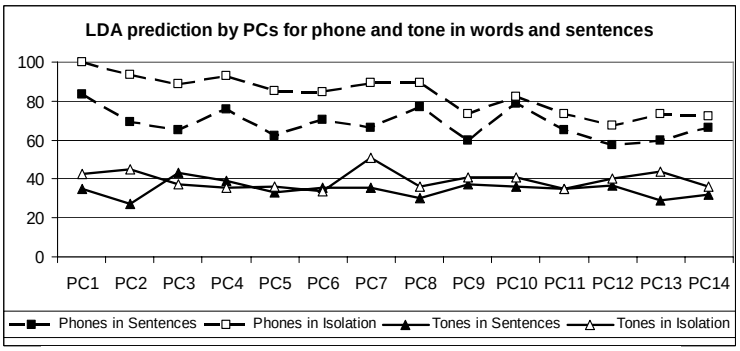


Figure 4: Predictions of phone & tone

Tone Category Prediction: The 14 PCs

were used in Linear Discrimination Analysis (LDA) to predict (a) membership of the 6 Cantonese tone categories, and (b) membership of the 5 phonetic string (word) categories. Repetitions of words extracted from sentences were used after removing the motion of the average word. As shown in Figure 4, prediction of phones is robust and well above chance (20%) across all PCs, with the expected superiority of lower ranked PCs. In contrast, prediction of tones is less robust compared to chance (16.67%) and there are some noticeable peaks – at PC3 (rigid head nodding) for words in sentences; and for PC7 (a forward downward movement of the head with a small non-rigid component) for words in isolation.

Correlation of PCs with Audio F0: F0, determined for words in sentences and in isolation using Praat scripts and additional decision rules (Vignali, 2005c), was used to predict tone category membership via LDA. Table 2 shows predictions were generally better for words in isolation, so only those are considered in the subsequent analyses. Table 3 shows correlations between auditory F0 and the 14 visual motion PCs. The highest are with PC 1, concerned with jaw opening only, without rigid head motion; PC9, corresponding to a pure rigid head motion, a rotation toward the back (as opposed to PC3 - a nodding rotation around the head centre); and PC14, which corresponds to a small tilt of the head on the right side.

Table 2: % correct LDA prediction of tone category from F0 contours

Words in -	Tone 55	Tone 25	Tone 33	Tone 21	Tone 23	Tone 22	Mean
Sentences	59	56	24	61	56	24	46.67
Isolation	92	96	60	96	100	60	84.00

Table 3: Correlation coefficients between visual PCs, and auditory F0 over time.

Tone	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*	*PC7*	*PC8*	*PC9*	*PC10*	*PC11*	*PC12*	*PC13*	*PC14*
Fan1-6	**-0.4**	**-0.39**	**-0.35**	-0.04	-0.18	-0.02	0.04	0.24	**0.28**	**-0.26**	0.21	-0.15	0.06	**-0.61**
Fu1-6	-0.21	0.11	-0.13	0.1	**0.38**	-0.18	0.23	0	0.22	0.04	**0.32**	0	**0.54**	**-0.34**
Hau1-6	-0.05	-0.22	0.1	**-0.36**	0.05	0.09	-0.19	-0.13	**0.32**	**0.32**	0.04	0.03	0.14	**-0.38**
Soej1-6	**-0.39**	0.09	**0.4**	**-0.45**	0.15	**0.38**	-0.09	**-0.48**	**0.28**	**0.5**	**0.35**	0.11	**0.69**	-0.23
Wai1-6	**-0.39**	-0.13	**0.48**	**-0.31**	-0.13	0.04	**-0.26**	-0.11	**0.6**	**-0.36**	0.21	0.23	-0.15	-0.24
Mean	**-0.29**	-0.11	0.1	-0.21	0.05	0.06	-0.05	-0.09	**0.34**	0.05	0.22	0.04	0.25	**-0.36**

Summary: In the correlation of PCs with audio data, while there is evidence from PC1 for the involvement of articulatory jaw motion, the strongest evidence comes from PC9 and PC14, suggesting that rigid head motion is especially important in distinguishing between the production of different Cantonese tones. Together with data from the prediction of the tone category from visual PCs suggesting the involvement of rigid head motion (PC3 and PC7, in isolation and in sentences respectively), it appears that the visual concomitants of variations in tone involve rigid motion of the head more so than non-rigid motion of the face.

4. Perceptual Test of the Phone/Non-Rigid, Tone/Rigid Hypothesis

Based on the analyses in section 3, underpinnings for the perceptual results in section 2 can be suggested. As a working hypothesis it is predicted that (1) visual information for (a) tones is contained mainly in rigid head motion, and (b) for phones is contained more in the non-rigid face motion, and that (2) this information can be used by perceivers to

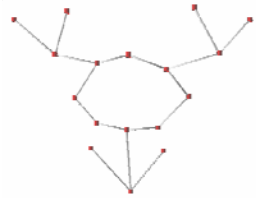


Figure 5: Static view of visual from OptoPCMarkerView

discriminate tones and phones.

Method: A speech segment (phone/tone) x motion type (rigid/non-rigid/combined) x modality (AO/VO/AV) design was used with repeated measures on all factors. An AXB paradigm in DMDX was used – participants decided whether the first or the last stimulus in a triad was most similar to the second. There were two test phases: one with PI-TD stimuli, requiring decisions based on tone, and another with Tonetically Identical, but Phone Differing (TI-PD) stimuli requiring decisions based on phone. Each phase included an AO, a VO, and AV block of trials, with 3 8-trial sets, one each for rigid, non-rigid, and combined motion. Tests sessions, trial sets and blocks were counterbalanced across the 42 adult monolingual English speakers (27 females, 17 males, mean age, 27.45 years, range, 18-63 years). They were tested with 30 Cantonese words - 'fan', 'fu', 'hau', 'soej' and 'wai' with each of the 6 Cantonese tones. 5 auditory-visual 5 recordings of each of the 30 words were captured from OptoPCMarkerView2 (Vignali, 2005a), which allows display of rigid only, non-rigid-only, or both motion types. Audio files were mixed with noise (-3.3 dB SNR). The visual display is shown in Figure 5.

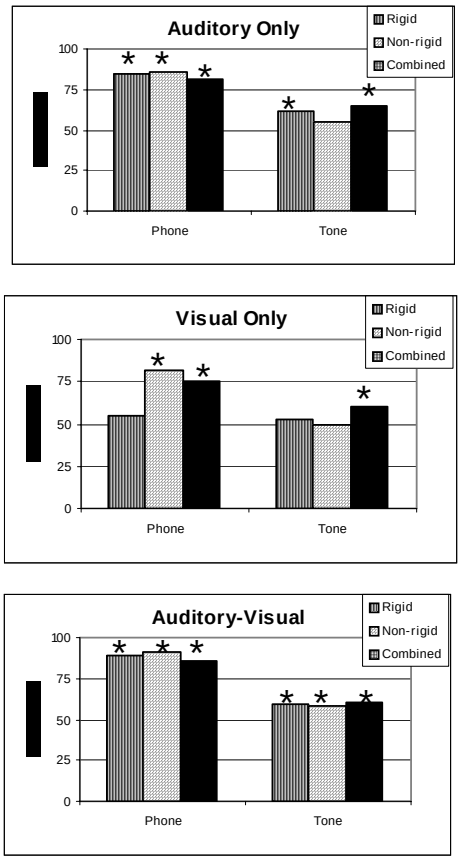


Figure 6: Phone & Tone % correct for rigid, non-rigid, & combined motion in AO, VO, and AV. * means > chance (50%).

Results: Discrimination performance (Fig. 6) was generally better for phones than tones, $F(1,41) = 194.93$; and non-rigid than rigid motion, $F(1,41) = 4.28$. In general and in accord with the phone/non-rigid, and tone/rigid hypothesis, non-rigid motion allows better discrimination for phones, and rigid motion for tones, $F(1,41) = 31.15$, and this difference appears to be greater in the visual conditions (VO & AV) than in AO, though it failed to reach significance, $F(1,41) = 3.23$. In AV, responses in all conditions were above chance, there was a clear phone>tone advantage, and little effect of motion type. In VO, phone perception is above chance in the non-rigid and combined conditions but dramatically drops to chance when only rigid motion is available showing respectively that *non-rigid motion is sufficient and necessary for the visual perception of phones*. However, in VO for tone perception, only when both rigid and non-rigid motion are available is performance above chance, suggesting that *both rigid and non-rigid motion are necessary, and together are sufficient for the visual perception of tones.*

5. Conclusions

The results support the phone/non-rigid, and tone/rigid hypothesis. Participants in the final experiment were non-tone language speakers, so while conclusions must remain tentative, the prognosis for stronger findings with tone language speakers in future studies is good; and given the VO results, stronger results with hearing impaired (especially hearing impaired tone language speakers). There are strong correlations between head motion and sentential intonation (Vatiktiotis-Bateson, 2000; Yehia et al., 2002), and the results here suggest that similar correlations occur for tone, and that the associated rigid head motion, while fine-grained, is available perceptually.

6. References

Burnham, D. K. (1992) Auditory-visual perception of Thai consonants by Thai and Australian listeners. *Pan-Asiatic Linguistics*, Bangkok: Chulalongkorn University Press, 531-545.

Burnham, D., Ciocca, V., & Stokes, S. (2001) Auditory-visual perception of lexical tone. *Eurospeech Conference 2001, Aalsborg, Denmark*, ISCA, Bonn, Germany, 395-398.

Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001) Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. *Auditory-Visual Speech Perception Conference 2001*, Causal Productions, www.causal.on.net, 155-160.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espressor, R. (1998) About the relationship between eyebrow movements and F_0 variations. In In T. Bunnell & W. Idsardi (Eds) *Fourth International Conference on Spoken Language Processing. Vol 4*, 2175-2178.

de Gelder, B., Bertelson, P., Vroomen, J., & 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* (1699-1702). Madrid.

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

Ramsay, J.O. & Silverman, B.W. (1997) *Functional Data Analysis*. Springer

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

Sekiyama, K. (1994) Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *J. Acoust. Soc. Japan*, *15*, 143-158.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, 26, 212-215.

Vatiktiotis-Bateson, E., Kroos, C., Kuratate, T., Munhall, K. G., & Pitermann, M. (2000). Task constraints on robot realism: The case of talking heads. In K. Kamejima (Ed.), *9th IEEE Internat. Workshop on Robot & Human Interactive Comm. (RO-MAN 2000)*, (352-357). Osaka: IEEE.

Vignali, G. (2005a) Analysis of 3D multivariable data of expressive speech motion. Symposium, Cross-Modal Processing, Faces & Voices, ATR, Japan. Also http://www.vignali.net/~guillaume)

Vignali, G. (2005b) Study of the visual component of tone in Cantonese and Mandarin, and stress in English and Japanese. Report for MARCS Auditory Labs, April, 2005.

Vignali, G. (2005c) Relation between voice pitch and rigid and nonrigid head motion in Cantonese and Mandarin. Report for MARCS Auditory Labs (at CRSLP, Chulalongkorn Univ., Thailand).

Yehia, H., Kuratate, T., & Vatikiotis-Bateson, E. (2002), Linking facial animation, head motion, and speech acoustics, *Journal of Phonetics*, *30*, No.3, 555–568.