



Rigid vs Non-Rigid Face and Head Motion in Phone and Tone Perception

Denis Burnham¹, Jessica Reynolds¹, Guillaume Vignali¹, Sandra Bollwerk¹, Caroline Jones²

¹MARCS Auditory Laboratories, University of Western Sydney, Australia

²School of Education, University of New South Wales

d.burnham@uws.edu.au, 13506894@scholar.uws.edu.au, guillaume@vignali.net, sc.jones@unsw.edu.au

Abstract

There is recent evidence that the visual concomitants, not only of the articulation of phones (consonants & vowels), but also of tones (fundamental frequency variations that signal lexical meaning in tone languages) facilitate speech perception. Analysis of speech production data from a Cantonese speaker suggests that the source of this perceptual information for tones involve rigid motion of the head rather than non-rigid face motion. A perceptual discrimination study was conducted using OPTOTRAK output in which rigid or non-rigid motion of the head could be presented independently, using two conditions: one in which words to be discriminated only differed in tone, and another in which they only differed in phone. The results suggest that non-rigid motion is the critical determinant for successful discrimination of phones, whereas both non-rigid *and* rigid motion are required for the discrimination of tones.

Index Terms: speech perception, auditory-visual speech perception, lexical tone, Cantonese,

1. Introduction

Visual (lip, face, and head and neck motion) information both augments and modifies speech perception. On the one hand, perception accuracy for speech presented in noise is augmented by 40-80% [1]; and on the other, in the McGurk effect dubbing visual [ga] onto auditory [ba] modifies perception to “da” or “tha” [2]. So, both auditory and visual speech information are important in speech perception, and it is argued that this is because convergent information better specifies the speech source [3].

All languages use phones, consonants and vowels, to distinguish lexical items, but in 70% of the world’s languages [6] spoken by over half of the world’s population [7], tone is also used. Tone in tone languages is primarily based on F₀, e.g., Cantonese has 6 tones, as in /fu55/ ‘husband’, /fu33/ ‘rich’, and /fu22/ ‘father’, /fu25/ ‘tiger’, /fu21/ ‘to hold’, and /fu23/ ‘woman’. In the related pitch-accented languages tone is carried between syllables, e.g., Japanese has two pitch-accents, high-low, e.g., ka^hki ‘oyster’, and low-high, e.g., ka^lki ‘persimmon’.

Most auditory-visual speech perception research has been conducted in non-tone languages, and even when in tone languages (Cantonese [4], and Thai [5]), it has mostly concerned phones - consonants and vowels - rather than tones. It may be argued that auditory-visual speech perception operates differently in tone languages: Sekiyama found that Japanese listeners’ McGurk effect perception is less influenced by visual speech than is that of their American counterparts [8], and the effect is further reduced in Chinese

perceivers [9]. Sekiyama reasons that as there are six tones in Cantonese, two 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.

Research on auditory-visual *perception* of tones and on auditory-visual *production* of tones is reviewed here ahead of the presentation of an analytic study of the determinants of visual perception of tones compared with phones.

1.1 Auditory-Visual Perception of Lexical Tone

There are visual correlates of F₀ in speech production, both at the intonation and tone level. Regarding intonation, there are positive correlations between French speakers’ eyebrow motion and intonation in sentences [10], a finding supported by the observation of strong correlations between head motion and F₀ during speech [11], that are continuous and seemingly used in auditory-visual perception [3]. Regarding tone, experiments have now been conducted on the auditory, visual, and auditory-visual perception of tones using Phone Identical, but Tone Differing (PhoneIdent/ToneDif) words. It has been found that Cantonese perceivers’ identification of Cantonese PhoneIdent/ToneDif words from visual information alone is slightly above chance [12], and discrimination in noise of Cantonese PhoneIdent/ToneDif words pairs is augmented in auditory-visual (AV) versus auditory-only (AO) presentations [13]. As this latter study was conducted using a “same”/“different” discrimination task, then non-tone language speakers could also be tested. It was found that augmentation of discrimination in AV versus AO occurs not only for Cantonese perceivers (3.5 proportional increase, AV versus AO) but also for non-Cantonese tone language (Thai) perceivers and even non-tone language (Australian English) perceivers (both 1.7 proportional increase).

These studies show visual face information for tone perception is available and perceptually useful. It appears that there is information in the face that facilitates tone perception and while the better performance by tone language speakers [13] suggests some linguistic influence and/or effects of linguistic learning, the results with non-native and even non-tone language speakers show that the information is not wholly linguistic in nature, and is either readily perceivable or readily learnable in a short period of time. Thus, while the locus of the tone information in the face is not clear, it appears that the information, at least in part, may be relatively low-level in nature. Burnham et al. (2006) used a 4-step process to ascertain the nature of this information:

- Record auditory and visual aspects of tone productions using OPTOTRAK motion capture

- Derive Principal Components (PCs) from face motion of auditory-visual tone productions ,
- Predict tone and phone categories from PCs based on visual speech production,
- Correlate visual speech PCs with Audio F0

ahead of a perception experiment in which non-tone language perceivers are tested for perception of tones based on aspects of face information in speech production.

1.2 Auditory-Visual Production of Lexical Tone

1.2.1 AV Tone Production Recording

A 24-year-old female Cantonese speaker was recorded using the OPTOTRAK face marker apparatus speaking each of 30 words. These were 5 PhoneIdent/ToneDif Cantonese strings /fan/, /fu/, /hau/, /soej/ and /wai/ each spoken with each of the 6 Cantonese tones. Productions were in isolation or in one of 5 sentences constructed for each of the 30 words. Audio data were recorded and visual movement data also recorded using OPTOTRAK active markers - 17 face markers to record non-rigid data (see the small squares in Figure 2), and 4 markers on a head rig to record six parameters of rigid head motion - x y z translation and roll, pitch, and yaw.

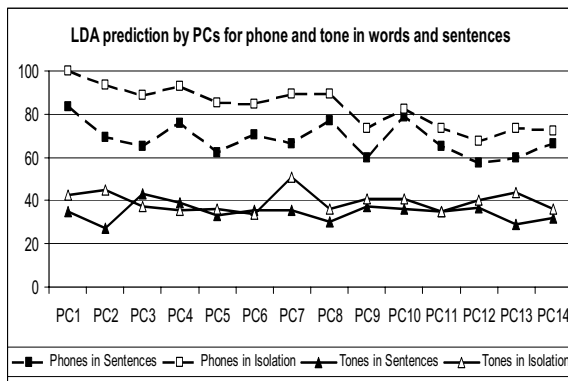


Figure 1: Predictions of Phone and Tone Identity

1.2.1 Principal Component Extraction

Starting from a motion description based on frames and Cartesian coordinates of markers, Principal Component Analysis (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 [14, 15]. Functional Data Analysis [16] 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 tone-specific motion, and mean motion for each tone category was thus derived.

1.2.2 Tone and Phone Category Prediction from PCs based on Visual Speech Production

Using the procedure described in 1.2.1, 14 PCs were sufficient to describe 99% of the variance. Most of these contained both rigid and non-rigid head motion, but often a preponderance of one over the other. For example, PC1 was mainly vertical mouth opening (non-rigid) and jaw motion (rigid), PC2 mainly horizontal mouth opening (non-rigid), PC3 mainly head nodding (rigid), PC4 nodding plus some mouth movement, and PC5 rigid movement of the head forward towards the camera. These 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. As there were 5 different context sentences for each of the 30 words, repetitions of words extracted from sentences were used after removing the motion of the average word. (For the repetitions of words in isolation no such correction for context was required.) As can be seen in Figure 1 prediction for words in isolation was generally better than for words in sentences. This was to be expected as in sentences less than perfect citation forms are used, there is co-articulation with surrounding words, and this co-articulation differed over the 5 different sentences used over the 5 sentences. Nevertheless, the fact that prediction was still above chance for the words in sentences suggests that the visual PC information is robust in extracting and predicting the identity of the phone differences or the tone differences in the words.

Over and above this generally good prediction, what is of interest here is the differential prediction of phone and tone differences. Prediction of phones was robust and well above chance ($1/5 = 20\%$) across all PCs, with the expected superiority of lower ranked PCs (as these account for more variance). In addition PC1, PC2 and PC4 all involving non-rigid motion (see descriptions above) appear to be particularly good predictors. Prediction of tones was less robust than for phones, but still above chance ($1/6 = 16.67\%$) 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; and troughs, e.g., for tones in sentences at

Table 1: % correct LDA prediction of tone category from F0 contours

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

Table 2: Correlation coefficients between visual PCs and auditory F0 over time. Those in bold are significant at $p < .05$.

Principal Component	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Correlation with F0	-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

PC2 (horizontal mouth opening) and PC8 (slight nodding plus mouth widening) and PC13 (slight lateral turning of the head and slight mouth movement).

1.2.3 Correlation of Visual Speech PCs with Audio F0

Given that visual speech production PCs predict tone category membership (see Figure 1) prediction of tone category membership from auditory information alone was investigated. Audio F0, determined for words in sentences and in isolation using Praat scripts and additional decision rules [17], was used to predict tone category membership via LDA and results are shown in Table 2. Predictions were generally better for words in isolation (for reasons explained in 1.2.2), so only those were considered in subsequent analyses. Table 2 shows auditory-visual correlations - between auditory F0 and the 14 visual motion PCs. Three correlations are significant, PC1, PC9, and PC14. The significance of PC1 shows the involvement of mouth opening and jaw motion. The strongest correlations are for PC9 (corresponding to a pure rigid head rotation backwards); and PC14 (corresponding to a small tilt of the head on the right side). These correlations suggest 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), it appears that the visual concomitants of variations in tone involve rigid head motion more so than non-rigid face motion. It remains to be seen if such speculations are confirmed in perceptual tests independently manipulating rigid and non-rigid motion.

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

Based on the above (1.2), our working hypothesis is that visual information for (a) tones is contained mainly in rigid head motion, and (b) for phones is contained mainly in non-rigid face motion; and that perceivers use visual information to discriminate within tone categories, and phone categories. As the studies reviewed in 1.1 suggest that the visual information for tone may be quite low level and language-independent, reduced visual display was used (see Figure 2), and non-tone language perceivers were tested to investigate if the visual tone information is generally available.

2.1 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 PhoneIdent/ToneDif stimuli, requiring decisions based on tone, and another with Tonically Identical, but Phone Differing (ToneIdent/PhoneDif) 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 [14], which allows display of rigid only, non-rigid-only, or both motion types. Audio files were mixed with noise (-3.3 dB SNR) in order to provide some focus on the visual cues [13]. The visual display is shown in Figure 2.

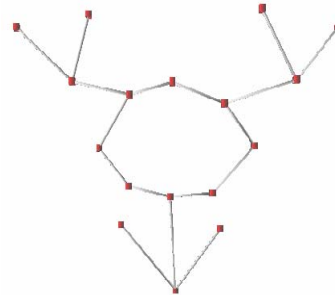


Figure 2: Static view of visual stimulus. The small square points show OPTOTRAK face markers. For the perceptual test the lines between the points were added by OptoPCMarker2 to provide links between the points.

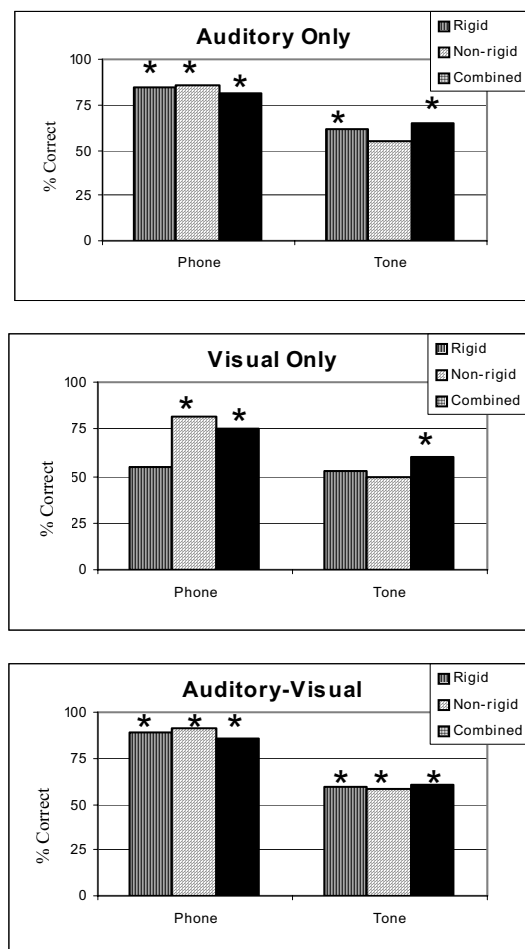


Figure 3: Phone & Tone % correct for rigid, non-rigid, & combined motion in AO, VO, and AV. * means > chance (50%).

2.2 Results

Discrimination performance (Figure 3) 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 this interaction failed to reach significance, $F(1,41) = 3.23$. In general the AO results show better discrimination for phones than for tones, with tone discrimination approaching chance. (Note that for AO the motion manipulations are simply control conditions, as there was no actual visual stimulation involved.)

Of greatest importance is the two visual conditions. 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. This suggests that *non-rigid motion is sufficient and necessary for the visual perception of phones*. In VO for tone perception, only when both rigid and non-rigid motion are available is performance above chance, although in AV all conditions are above chance. This suggests that *both rigid and non-rigid motion are necessary for the visual perception of tones, that some degree of tone perception is possible on the basis of visual information alone*.

3 Conclusions

The results support the phone/non-rigid, and tone/rigid hypothesis. Participants in the final experiment were non-tone language speakers, which suggests that the visual information for tones is low-level and language-general. While conclusions must remain tentative until tone language perceivers are tested, the prognosis for stronger findings with such perceivers in future studies is good. Moreover, given the VO results, stronger results with hearing impaired (especially hearing impaired tone language speakers) would be expected.

Previous studies have shown that there are strong correlations between head motion and sentential intonation [3, 11]. Here the relationship between head and face motion and visual perception of tone was investigated. The results suggest that similar correlations between head motion and tone, and that the rigid head motion associated with tone, while fine-grained, is available perceptually.

4 References

- [1] Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.
- [2] McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.
- [3] Vatikiotis-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.
- [4] 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.
- [5] Burnham, D. K. (1992) Auditory-visual perception of Thai consonants by Thai and Australian listeners. *Pan-Asiatic Linguistics*, Bangkok: Chulalongkorn University Press, 531-545.
- [6] Yip, M (2002). *Tone*. Cambridge, Cambridge University Press.
- [7] Fromkin, V. (1978). *Tone: A linguistic survey*. New York: Academic
- [8] 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.
- [9] Sekiyama K. (1997) Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, 59, 73-80.
- [10] Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., and Espressor, R. (1998) About the relationship between eyebrow movements and F_0 variations. In T. Bunnell & W. Idsardi (Eds) *Fourth International Conference on Spoken Language Processing. Vol 4*, 2175-2178.
- [11] Yehia, H., Kuratate, T., & Vatikiotis-Bateson, E. (2002), Linking facial animation, head motion, and speech acoustics, *Journal of Phonetics*, 30, No.3, 555-568.
- [12] Burnham, D., Ciocca, V., & Stokes, S. (2001) Auditory-visual perception of lexical tone. *Eurospeech Conference 2001, Aalsborg, Denmark*, ISCA, Bonn, Germany, 395-398.
- [13] 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.
- [14] 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>
- [15] 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.
- [16] Ramsay, J.O. & Silverman, B.W. (1997) *Functional Data Analysis*. Springer
- [17] 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).