PERCEPTION OF VISUAL INFORMATION FOR CANTONESE TONES

Denis Burnham*, Valter Ciocca**, Cheryl Lauw**, Susanna Lau*, & Stephanie Stokes**

*Macarthur Auditory Research Centre, Sydney (MARCS), University of Western Sydney, Australia

**Department of Speech & Hearing Sciences, University of Hong Kong

ABSTRACT: It is often assumed that there is little if any visual speech information for lexical tone. However, the presence or absence of visual information for tone is yet to be tested empirically. In this study Cantonese speakers are asked to identify spoken words as one of six words differing only in tone. Words are presented in three different modes: auditory-visual (AV), auditory only (AO), and visual only (VO). It is found that performance is equivalent in AO and AV conditions, ie, that there is no augmentation of auditory tone perception when visual information is added. Performance in the VO condition is considerably worse, but under some circumstances it is significantly greater than chance, and interestingly, distinctly different in nature to performance in the auditory conditions. In particular, visual information for tone is evident in the performance of perceivers *without* phonetic training but not in those *with* phonetic training; for tone carried on *monopthongs* but not on *dipthongs*; in *running speech* but not in *citation form*; and *for contour tones* but not for *level tones*. As there is no augmentation for AV over AO, then these results may have little practical implication for hearing in good listening conditions. Nevertheless, as visual information alone raises performance above chance, then the results may have implications for hearing impaired Cantonese language users, or for situations in which auditory input is degraded.

1. INTRODUCTION

Speech perception is an auditory-visual phenomenon; whenever visual speech information is available, perceivers use it. This is demonstrated in the McGurk effect, in which auditory [ba] paired with the lip movements for [ga] is perceived by English language perceivers as "da" [1]. This is also the case for Japanese and Cantonese language perceivers, although it has been found that the influence of visual information is greatest for English, less for Japanese, and least for Cantonese perceivers [2]. Sekiyama [2] has suggested that this is due to the relative degree to which pitch is used to distinguish lexical items in these three languages. Cantonese has six lexical 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). Japanese has two pitch-accents on two-syllable words, high-low, and low-high. In English however, pitch is seldom used to distinguish lexical items, except in stress contrasts, e.g., `project vs pro`ject, which involve pitch to some extent, and also amplitude and duration. From these linguistic realities it could be concluded that in tone and pitch-accented languages there is relatively less reliance upon visual information because of the presence of lexically-relevant pitch variations and because these pitch variations do not have visual correlates.

However, the last premise in this argument, that pitch variations do not have visual correlates, is yet to be evaluated, and it is this issue which is the subject of the current investigation. The motivation for the study stems from the pervasiveness of visual information for segmental information in speech perception. The McGurk effect demonstrates that whenever visual information for segments is available perceivers use it. Such information is therefore useful even in undegraded speech perception situations, but most especially in noisy environments, and for hearing-impaired perceivers. It would be unfortunate if speakers of tonal and pitch-accented languages, both normal hearing and hearing impaired, would not have available to them this rich source of information for a significant proportion of the lexical distinctions in their language.

The purpose of the study is to evaluate whether there is reliable visual information for lexical tone. An identification paradigm was used, in which Cantonese language participants were required to choose the correct word from a selection of six written words when a native Cantonese speaker presented a word from a set of Cantonese tone sextuplets. Words were presented in auditory-only (AO), visual-only (VO), and auditory-visual (AV) modes. A number of variables were manipulated - phonetic

experience of the perceivers, the provision of feedback to participants, whether words were presented in isolation or in sentences, whether the tone was carried on a monopthong or dipthong, the particular tone of the six Cantonese tones, and most importantly, whether the word was presented in AO, AV, or VO modalities. It was hypothesised that tone perception performance would be:

- better for *phonetically-trained* than *untrained* participants, due to the former's greater knowledge of the relevant cues in segment and tone perception;
- better when *feedback* for correct responses was provided than when it was not, due to the beneficial effect of positive reinforcement of decisions based on the appropriate cues;
- better for words in a *sentence context* than for *isolated words*, due to facilitation by dynamic contextual information;
- better for words containing *dipthongs* than those containing *monopthongs*, due the greater degree of dynamic contextual information in dipthongs;
- better for tones containing pitch movement (the *contour tones* - 2-5, 2-1, 2-3) than those containing little or no pitch movement (the *level tones* - 2-2, 3-3, 5-5), again due the greater degree of dynamic contextual information in the former;
- better for *AV* presentations than *AO* presentations due to augmentation of tone perception by visual information;
- better for *AV and AO* presentations than for *VO* presentations, due to the addition of auditory information; and
- better than chance [1/6 = 16.67%] in all conditions including the VO condition.

2. METHOD

2.1 Design

A 2 x 2 x 2 x (3 x 6 x 4 x 2) design was employed with repeated measures on the last four factors. The first three factors, the group factors, were phonetic background - participants with prior phonetic training / without phonetic training; word presentation condition - isolated words / words in sentences; and feedback - feedback provided for correct responses / no feedback. The within-subjects factors were mode of presentation - auditory-only (AO), visual-only (VO), auditory-visual (AV); tone – the six 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); words - 4 different Cantonese words on which the tones were carried; and repetitions – each of these 3 x 6 x 4 = 72 combinations was presented twice.

2.2 Subjects

A total of 48 adult participants were tested. All were native Cantonese speakers and all were members of the Department of Speech and Hearing Sciences at Hong Kong University. Of the 48, 24 were untrained phoneticians - non-academic staff members or first year students; and 24 were academic staff members or third or fourth year students with intensive phonetic training. Half (12) of the subjects in each phonetic background group were assigned to the Isolated Words group, and the other half to the Sentence Context group. Within each of these subgroups, half of the participants were assigned to a group in which feedback for correct responses was given, and half were assigned to a no-feedback group

2.3 Stimulus Materials

Stimuli consisted of four Cantonese tone sextuplets, two with monopthongs, /fu/ and /fan/, and two with diphthongs, /soej/ and /hau/. Each of these phonetic strings has a lexical meaning for each of the six Cantonese tones, and therefore each comprises a Cantonese tone sextuplet [3]. Each of these 24 words were presented in auditory-only (AO), visual-only (VO), and auditory-visual (AV) modes. For the Isolated Words condition, stimuli were presented in the form of isolated Cantonese words, e.g., /fu55/ 'husband', /fu25/ 'tiger', /fu33/ 'rich', /fu21/ 'to hold', /fu23/ 'woman', /fu22/ 'father', and similarly for the other three tone sextuplets. In the Sentence Context condition, the 24 Cantonese words were presented in a semantically-neutral Cantonese carrier sentence, e.g., /ha6 yat1 go3 ji6 hai6 fu1/ 'The next word is husband'.

For both conditions, there were a total of 144 test trials (AO/VO/AV x 6 tones x 4 words x 2 repetitions) in six 24-trial test blocks. The 6 blocks were made up of 2 AO blocks, 2 VO blocks, 2 AV blocks with

block order counterbalanced between participants. In addition, 24 practice trials were included to allow participants to become familiar with the testing procedure: there was one block of 6 practice trials (2 AO, 2 VO, and 2 AV trials) at the start of the experiment, before test blocks began, clearly labeled as practice trials for the participants; and then at the beginning of each of the 6 test blocks there were another three practice trials (in the appropriate mode, AO, VO, or AV), not labeled as practice trials, so that they served as warm-up trials for participants for each presentation mode.

The stimuli consisted of the 24 Cantonese words (/fu/, /fan/, /soej/ and /hau/ in their 6 tonal representations) spoken by a 23-year-old a native Cantonese speaking female, and recorded on a digital video-recorder. The video was taken face-on resulting in a full front-on view of the face, head and throat of the speaker. Thus there were various possible cues available to the perceivers - eyes, face, larynx, etc., but the source of any visual information for tone could not be determined from the this study. The resulting images were then digitally edited into digital video files (Apple Quicktime) using the video-editing software package Adobe Premiere. The experiment was created and run using the DMDX experimental software [4].

2.4 Procedure

Subjects were tested individually in a sound-attenuated room, on a PC running the DMDX software. The DMDX software presented stimulus materials, and recorded subjects' responses and reaction times via keyboard input. Only the correct response data are presented in this paper.

3. RESULTS

The percent correct results are graphed in Figures 1, 2, and 3. Figure 1 shows the effect of phonetic knowledge on correct responses in AO, VO, and AV presentation modes; Figure 2 shows the effect of type of vowel (monopthong vs dipthong) words on correct responses in AO, VO, and AV presentation modes; and Figure 3 shows the effect of words in isolation and words in contextual sentences for level and contour tones for each of the three presentation modes, AO, VO, and AV. The percent correct responses were analysed using a combination of analyses of variance to fit the 2 x 2 x 2 x (3 x 6 x 4 x 2), phonetic background x word presentation condition x feedback x (mode of presentation x tone x words x repetitions) design. In addition, where there were significant differences between conditions, single sample t-tests against chance (16.67%) were also conducted. This procedure is relatively conservative, because t-tests against chance were not conducted across the board for every possible factorial cell, only when the overall analysis indicated that there was a significant difference between

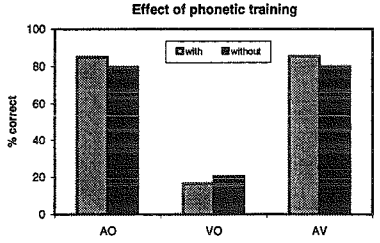

Figure 1. Effect of Phonetic Training.

conditions.

There were no significant effects of feedback, or the repeat factor repetitions. There were no significant main effects or interactions between the percentage correct responses in AO (mean = 82.2%) and AV (mean = 82.6%) conditions, but in both of these the percentage correct was greater than in the VO condition (mean = 18.6%), F (1,40) AO vs VO = 1281.33, F (1,40) AV vs VO = 1133.02. The lack of difference between AO and AV conditions and their superiority over the VO condition can be clearly seen in Figures 1, 2, and 3. The overall performance level in the VO condition, 18.6% (SE = 1.23), was slightly higher than the 1 in 6 (16.67%) chance level but failed to reach significance, t

(1151) = 1.59, p > .05. However, as will be seen shortly, performance in the VO condition did rise above chance under some interesting conditions.

There was a significant effect of phonetic training of the participants, but only insofar as it interacted with AO vs VO, F (1,40) Phonetic Training x AO vs VO = 6.58, and with AV vs VO, F (1,40) Phonetic Training x AV vs VO = 6.08. The nature of this effect can be seen in Figure 1: while participants with phonetic training performed better than those without training for auditory stimuli (AO or AV), phonetic training hindered performance for non-auditory (VO) stimuli. It is possible that due to its auditory emphasis, phonetic training serves to direct participants' attention towards auditory and away from visual speech information. Thus those participants without phonetic training performed better on VO stimuli (mean = 20.57%) than did those with phonetic training (mean = 16.67%). The significance of this finding is verified by tests against chance level for performance on the VO trials. For participants with phonetic training the level of performance (mean = 16.667%, SE = 1.18) was not significant, t (575) = -.003, p>.05. On the other hand, for participants without phonetic training the level of performance (mean = 20.57%, SE = 1.27) was significant, t (575) = 3.07, p <.01.

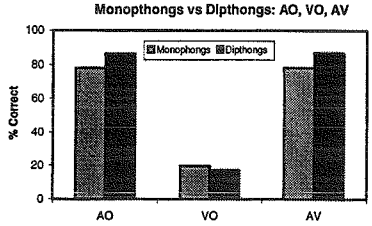

Figure 2. Monopthongs vs. Dipthongs: AO,VO,AV

There was also a significant interaction of type of vowel in the word (monopthong vs dipthong) with mode of presentation, F (1,40) Vowel Type x AO vs VO = 10.18, F (1,40) Vowel Type x AV vs VO = 5.29. As can be seen in Figure 2, participants performed better with dipthongs than monopthongs both in the AV condition (mean dipthong = 86.89%; mean monopthong = 78.30%), and the AO condition (mean dipthong = 86.46%; mean monopthong = 77.95%). Conversely for the VO condition, participants performed better for the monopthongs (mean = 19.88%) than for the dipthongs (mean = 17.36%). In addition, the performance for monopthongs (mean = 19.88%, SE = 1.28) was significantly superior to chance, t (575) = 2.50, p<.05, while for dipthongs performance (mean = 17.36%, SE = 1.17) did not depart from chance, t (575) = 0.59, p>.05.

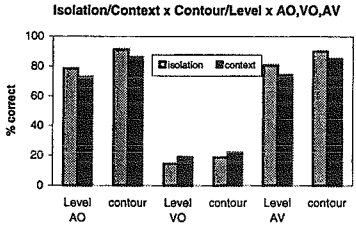

Figure 3. Isolation/Context x Contour/Level x AO, VO, AV

Finally, there were a number of effects associated with words presented in isolation vs in a sentence, and with the type of tone - contour or level. Participants generally performed better with words in isolation (mean = 62.27%), than words in context (mean = 60.01%), F (1,40) = 13.77; and were better

with contour (mean = 65.60%) than level (mean = 56.68%) tones, F (1,40) = 42.98. However most interesting are the three way interactions between isolation vs context, contour vs level tones and mode of presentation, F (1,40) Isolation/Context x Contour/Level x AO vs VO = 13.61, F (1,40) Isolation/Context x Contour/Level x AV vs VO = 11.62, shown in Figure 3. Participants were generally better with contour than level tones in all three modes, AO, AV, and VO, with the degree of superiority being less marked in the VO condition. For the AO and the AV conditions participants performed better with words in isolation, while in the VO condition participants perform better with words in sentences. Tests against chance for the VO trials confirmed the significance of these results. For words in isolation performance (mean = 17.97%, SE = 1.17) did not deviate from chance, t (575) = 1.11, p>.05, while for words in sentences performance (mean = 19.27%, SE = 1.28) was significantly greater than chance, t (575) = 2.02, p<.05.

4. DISCUSSION AND CONCLUSIONS

In relation to the hypotheses put forward earlier, performance was indeed better for phonetically trained participants, but only when auditory information was available in the AO and the AV conditions. When only visual information was available, phonetically-trained participants performed *worse* than untrained naïve perceivers. Moreover, untrained perceivers performed better than chance, while trined phoneticians did not. As it is fair to assert that a far greater proportion of the Cantonese population (or any language population, for that matter) is untrained phonetically, these results suggest that there is perceptually-salient visual information for Cantonese tone, but that the information can be ignored if the perceiver has been trained to concentrate on other, in this case auditory, cues.

These results suggest an effect of training on the auditory-visual perception of speech. The effect of training must, it appears, occur over a relatively extended period of time, as the hypothesis that feedback would enhance performance was not upheld in any of the three modalities. The veracity of this explanation and the generality of the present results could be tested by presenting conflicting auditory and visual information, i.e., McGurk effect stimuli, to those with and without phonetic training, providing feedback or no feedback. To the extent that the integration of auditory and visual speech information is affected by long-term and not short-term training, it would be expected that those without phonetic training should show stronger McGurk effect performance than those with phonetic training, and that feedback should not significantly affect performance.

With respect to the sentence versus word context, results were again significant, but again they differed between the auditory (AO & AV) and the visual (VO) condition in an unexpected manner. Contrary to expectations, performance was better in isolated words for the auditory conditions. On the other hand performance for the VO condition was better in the sentence context. There are two possible reasons for this - either visual information for tone is more perceptible in sentences, or there is more, or a different kind of, visual information for tone produced by speakers in sentences. These are not, of course mutually exclusive, and further studies are required to investigate these perception and production aspects.

Consistent with expectations, performance was better for tones carried on dipthongs than monopthongs, but this was only so for the auditory conditions; for the VO condition, tones carried on monopthongs were perceived more clearly than those carried on dipthongs. It is possible that the dynamic lip and face movement information contained in sentences facilitates the extraction visual cues for lexical tone. However, if this were the case then one would expect that the presumably greater dynamic information in dipthongs than monopthongs should also facilitate visual processing of tone.

Another possible explanation for this set of results could involve the temporal aspects of speech perception. The duration of the dipthong words was greater than that of the monopthong words. Similarly, although it has not yet been measured systematically, the duration of the words in isolation was probably greater than when the same words were presented in sentences. Thus it may be the case that acoustic-based judgements are better when based upon information spread over time, while optically-based judgements are better when information is presented over a shorter time span. This explanation, while plausible, lacks generality, because it was found here, consistent with predictions, that contour tones were perceived more veridically than level tones in all three modalities, AO, AV, and VO. So there is visual information for Cantonese tone, and under a number of conditions

this information raises performance above chance, but this visual information cannot be reduced merely to durational information.

Finally, with respect to the relationship between the three modality conditions, performance was not better for AV presentations than AO presentations, ie, visual information did not augment tone perception over and above the auditory alone presentations. Nevertheless, performance in the VO condition was significantly greater than chance in a number of interesting situations. Performance was significantly greater than chance in the VO condition when participants did not have phonetic training, when tones were carried on monopthongs, and when the target words were embedded in sentence context. As there is no augmentation of AO tone perception by the addition of visual information, the relevance of visual information for tone in the hearing population in undegraded acoustic conditions is questionable. However, given that there is information for tone in the visual signal alone, it is possible that in tonal languages those who are deaf or hearing impaired perceive lexical tone on the basis of visual information alone, and may do so irrespective of top-down linguistic context information. Thus auditory-visual speech perception studies with hearing impaired speakers of tonal languages are urgently required. Furthermore, even hearing language users may use visual information for tone in degraded acoustic conditions. To test this it would be necessary to repeat the current experiment with noise added to the auditory signal in the AV condition to discover whether there is augmentation of auditory-visual speech perception by visual information when the auditory signal is degraded.

5. REFERENCES

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

[2] Hayashi, Y., & Sekiyama, K. (1998). Native-foreign language effect in the McGurk effect: a test with Chinese and Japanese. Proceedings of the International Conference on Auditory-Visual Speech Processing. Sydney, 1998, 61-66.

6. NOTES

[3] This is true for all but the 5th tone for /hau/, in which the phonetic string is /a:o/ 'to bite', rather than /hau/.

[4] To download DMDX go to http://www.u.arizona.edu/~jforster/dmdx.htm