Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Can English perceivers match Cantonese auditory and visual prosody?

Sonya Karisma Prasad[1], *Jeesun Kim*[2], *Chris Davis*[3]

[123] The MARCS Institute, Western Sydney University

sonya.prasad@westernsydney.edu.au, j.kim@westernsydney.edu.au, chris.davis@westernsydney.edu.au

Abstract

The prosody of an utterance can be varied by changing F0, duration and amplitude. Such changes are typically accompanied by variation in the talker's face/head motion (visual prosody). For native language utterances, people can match auditory and visual prosody accurately. We tested whether English perceivers can do this with an unfamiliar language, Cantonese, which differs from English specifically with regard to suprasegmental properties (e.g., different rhythm type; use of lexical tone). These differences may make extraction of prosody difficult, because they distract English perceivers and/or because they affect the way prosody is realized. However, AV cues for prosody may be similar across languages and sufficiently salient to overcome the suprasegmental differences. We tested native Australian-English participants (N=27) with 50 Cantonese sentences spoken as questions, narrowly focused or broad focused utterances by two native Cantonese talkers. Participants completed a same-different matching task for auditory-auditory (AA); visual-visual (VV) and auditory-visual (AV) pairs. Each pair type consisted of the same sentence and talker, but different tokens. Matching performance was above chance for all conditions: AA > AV = VV. Results are discussed in terms of how auditory and visual prosody is conveyed and how this may be affected by language properties.

Index Terms: auditory prosody, visual prosody, language

1. Introduction

Speech conveys information via segmental sounds (consonants and vowels) and also via prosody, that is, variations in fundamental frequency (F0), intensity, duration, and voice quality. For example, prosody can be used to change a statement to a question, or to emphasize (focus on) an element of an utterance. That is, questions can be signaled by a high peak in F0 or by a terminal rise (compared to statement with a low peak or terminal fall) [1]; focus (emphasis) by varying the duration, F0, and the intensity of the focused syllable(s) [2].

Prosody can also be realized by variation in the talker's face/head motion (visual prosody) [3]. For example, compared to broad focused utterances, narrow focused elements tend to be accompanied by more head and eyebrow motion (although there is considerable variation in how and whether this is realized) [4].

Not only is prosodic information conveyed by auditory and visual cues, but it seems that people are sensitive to whether they match or not (at least when tested in their native language) [5]. That is, studies have shown that perceivers can accurately match native language prosodic expressions (i.e., questions, narrowly focused or broad focused utterances) within and across auditory and visual modalities [6].

The current study builds on this native language auditory and visual prosody matching result by examining whether people can match prosody cues (both within and across modalities) in an unfamiliar language. The rationale for this extension to an unfamiliar language is that the level of performance will provide a novel window into native language prosodic bias. That is, we presume that when a language is unfamiliar, people will fall back on the cues that they use in their native one. This should lead to reasonably good performance when the languages use similar cues and poorer performance when they do not. Clearly, here the choice of unfamiliar language is the key for revealing an interesting pattern of results.

We chose Cantonese as the unfamiliar language for English perceivers based on the results of acoustic studies that indicate that final questions are marked in a similar fashion to English; whereas focus may not be. That is, studies suggest that similar to English, final questions in Cantonese are marked by a pitch rise. Indeed, it has been suggested that final rising is an essential cue for Cantonese questions [7]; that all six tones have rising contours for a question in final position, and that for statements the tone was lowered in the final position [8]. Prosodic focus, on the other hand, appears to be differently marked in Cantonese and English. For instance, an important acoustic correlate of prosodic focus is post focus compression (PFC) [9]. Indeed, PFC has been reported for many languages including English [10]. However, it appears that PFC for pitch and intensity is absent in Cantonese [11].

With respect to visual cues to prosody, there are no studies to our knowledge that have examined the degree to which visual cues to prosody are similar across Cantonese and English. Work by [12] suggests that visual speech information (especially lip dynamics) conveys prosodic information about prosodic focus and that this cue may be similar for English, Swedish, and French. Likewise, a study of the relationship between head motion (nods) and prosodic focus in Swedish [13] bears some similarity with that found for English [4]. These studies suggest that at least for some languages, it may be quite possible to detect an across language match between visual cues to prosody.

In summary, the aim of the current study was to examine the degree to which native Australian-English perceivers can match within and across modality cues to prosody in Cantonese. We suspect that this may be a challenging task due to general differences between English and Cantonese (e.g., the use of lexical tone in Cantonese and its syllable-timed rhythm), which may interfere with extracting prosodic cues.

doi: 10.21437/SpeechProsody.2016-214

However, it may be the case that the auditory and visual cues for some prosodic contrasts are sufficiently similar and salient across languages to overcome other differences.

To examine whether native Australian-English perceivers can match auditory and visual prosody accurately for linguistic expressions in Cantonese we used a same-different matching task. Here, perceivers were presented with either the same (narrow focused-narrow focused, question-question) or different (narrow focused-broad focused, question-broad focused) linguistic prosodic expressions. These linguistic expressions were presented in three different modalities: auditory-auditory (AA); visual-visual (VV); and auditory-visual (AV).

Given that acoustic cues for prosody are likely to be more reliable than visual ones [6], we would predict that overall, auditory matching performance should be better than visual. Furthermore, within the auditory modality, we would predict that matching performance for questions should be better than for focus, given the evidence that Cantonese and English both use F0 final rise/fall as a cue to question/statements, but may use different acoustic cues to focus (see above). Across modal matching requires that perceivers extract the cues to prosody from the auditory modality and then match these to the cues in the visual one. Given this dual requirement, we predicted that AV performance will be the poorest. The interesting issue will be whether it is above chance performance or not.

2. Method

2.1. Participants

Twenty-seven undergraduate students (10 males, *Mage* = 19 years, *SD* = .88) from Western Sydney University participated in this study for course credit. All were native Australian-English speakers and none of them had experience with Cantonese. All reported normal or corrected-to-normal vision with no hearing problems.

2.2. Stimuli

The stimuli consisted of auditory and visual recordings of fifty Cantonese spoken sentences selected from the MARCS laboratory database. These sentences are semantically neutral and have a selection of different tones in initial, middle, and final sentence position. The sentences consist of ten characters per sentence and were spoken by two native talkers of Cantonese (1 male, Mage = 28.5 years, SD = 2.12). The sentences were produced in three different prosodic styles: broad focused, narrow focused and echoic questioning.

The video and auditory recordings of the two talkers were captured individually in a well-lit, sound attenuated booth against a neutral coloured background using a video camera (Sony NCCAM HXR-NX30p). The video camera was situated directly in front of the talker and captured video at 1920 x 1080 full HD resolution at 50 frames per second. The talkers were instructed to look into the camera as they uttered each sentence. The audio recording was captured using a microphone (AT 4033a Transformerless Capacitor Studio Microphone) which was placed approximately 20 cm away from the talkers' mouth out of the cameras view. The recordings were interfaced to a PC running CueMix FX digital mixer via a Motu Ultralite mk3 audio interface with a FireWire connection and ported to Adobe Audition.

Each recording session of the 50 sentences was blocked by the type of expression (broad focused, narrow focused, echoic questioning). The talkers were provided with a printed list of 50 sentences and instructed as to which linguistic expression was required before each block. For broad focused statements, talkers said aloud each sentence after first reading it silently. A dialogue exchange task was used to elicit the focus and question expressions. In this task, the talker interacted with an interlocutor by making a correction to an error made by the interlocutor (a narrow corrective focused utterance), or she/he questioned an item that was emphasised by the interlocutor (an echoic question).The interlocutor was seated in front of the talker, face to face, at a level below the video camera. The whole recording session was repeated on the following day. So that overall, each talker was recorded for a total of 300 sentences (50 sentences x 3 expression types x 2 repetitions).

Audio and video recordings were segmented into each sentence using MATLAB. Video recordings were stripped of audio and were cropped to include just the head area (Figure 1). Audio recordings were normalised so that the intensity of all sound files was equal to 60 decibels.

Figure 1: *An example of the head shots of the two Cantonese talkers.*

2.3. Experimental design

The experiment used a same-different matching task as in [6]. In each trial, a pair of sentences were presented, which had the same segmental content but were produced with either the same (i.e., narrow focused and narrow focused; question and question) or different (narrow focused and broad focused; question and broad focused) linguistic expression. The stimulus pair was presented in auditory-auditory (AA), visual-visual (VV) or auditory-visual (AV). Within each pair, the first sentence was always taken from the first recording session and the second sentence from the second session. This was done to rule out instance-specific matching strategies in the trials with same linguistic expressions.

Three sets of 14 different sentence pairs spoken by both talkers were presented blocked by presentation modality: AA, VV and AV. The presentation order of these blocks was counterbalanced across participants. The male talker was always presented first in a block followed by the female talker in a block. The presentation order of the talkers was not counterbalanced since any across talker effects was irrelevant to the aim of the study. In each talker block, the order of stimuli was randomized.

Overall, each participant was presented 336 experimental trials (14 sentences x 2 talkers x 3 presentation modalities x 2 expression contrasts (narrow focused or questioning vs. broad focused) in same or different pairings. Three versions of the experiment were created so that the three sets of 14 sentences appeared in all of the three modality blocks across the versions. Participants were allocated to one of the versions. Eight sentences, which were not used in the experimental trials, were used as practice trials.

2.4. Procedure

Participants were tested individually in a quiet room. First each participant was told about the three prosodic types that would be presented in the experiment (broad focused, narrow focused, echoic questioning). They were then told that in each trial they would be presented with a pair of sentences in AA, VV, or AV modality that were made up of the same words and that their task was to judge whether the pair was spoken with the same prosodic expression or not. Participants were informed that half the time the pair was spoken with the same prosody ("yes") and half the time this was not the case ("no"). The participants were also informed about the blocking of presentation modality; that two talkers would be presenting the sentences and that presentations from each would also be blocked.

In each talker block, participants were first presented with 4 practice trials (2 expressions x 2 pair types) followed by 56 experimental trials (2 x 2 x 14). In each trial, participants were presented with the two sentences in sequence and a yes/no response options appeared on the screen for a keyboard response with no time limit. No feedback was given as to the correctness of a response. Participants were given three breaks throughout the experiment. At the conclusion of the experiment, participants were debriefed as to the purpose of the study. Stimulus display and response collection was carried out using the DMDX software [14].

3. Results

We used d-prime (d') scores [15] to take account of potential response biases in the same-different matching task. Figure 2 shows mean d' scores for each prosodic contrast type in each presentation modality condition. As can be seen, all the d' scores were higher than chance level (d' = 0). A series of one-sample t tests were conducted and the values are presented in Table 1. All d' scores were significantly greater than zero.

To compare performance across conditions, the mean d' scores were further analysed in a 3 (modality: AA, VV, AV) x 2 (expression contrast type: narrow focused, question) repeated measures ANOVA. Overall, there was a significant main effect of modality, F(2,52) = 11.36, p < .01, ηp2 = .31. The AA (mean d' = 1.09) presentation modality resulted in significantly better performance than the VV (mean d' =.60) and AV (mean d' =.65) presentation modalities. There was no statistically significant difference between the VV and AV presentation modalities.

Participants showed better performance for the questions (mean d' = .87) than the focused utterances (mean d' =.69), F(1,26) = 4.68, p = .04, ηp2 = .15. The interaction between modality and prosody type was significant, F(2,52) = 5.41, p = .01, ηp2 = .17. The interaction between modality and expression type was further analysed by using a simple effect comparison with a Bonferroni adjusted alpha of .01. Simple comparisons reveal that the participants showed no significant difference when matching questions and focused utterances in the VV modality, F(1,26) = .17, p = .68, ηp2 = .01 and AV modality, F(1,26) = 1.12, p = .30, ηp2 = .04 and that the questions attracted a higher mean d' score than focused utterances in the AA modality, F(1,26) = 10.47, p < .00, ηp2 = .29.

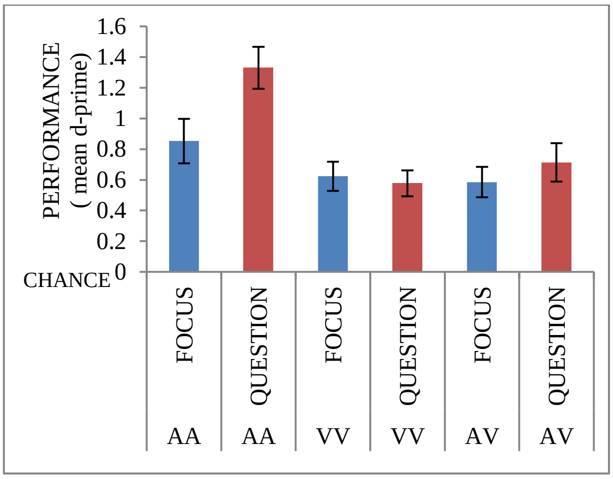

Figure 2: *Mean d-prime scores for each expression type across AA, VV, and AV presentation modalities. The error bars represent standard error.*

Table 1. *One-sample t test results for each modality and expression type.*

Conditions	*T*	*df*	Sig.
AA Focus	5.90	26	.00
AA Question	9.74	26	.00
VV Focus	6.63	26	.00
VV Question	6.85	26	.00
AV Focus	5.92	26	.00
AV Question	5.69	26	.00

4. Discussion

The aim of the current study was to examine the degree to which native Australian-English perceivers can match within and across-modality prosody in Cantonese, an unfamiliar language. More specifically, our interest was to assess whether the matching performance was above chance or not, whether the auditory modality afforded better matching performance than the visual one, and whether questions were matched better than focused utterances.

We found that the Australian-English perceivers were able to match Cantonese prosody better than chance level for all conditions. Since perceivers most likely based their judgments on cues appropriate to their native language, it suggests that auditory and visual cues to prosody in Cantonese and English are somewhat similar. As mentioned above, we predicted that overall, auditory matching performance will be better than visual matching performance. Consistent with this prediction, we found that the Australian-English perceivers' matching performance for the auditory stimuli was significantly better than for the visual ones. These results indicate that the auditory cues to linguistic prosody are most likely more prominent/reliable than visual cues. This is consistent with the results of [6] that found that acoustic cues for prosody were more reliably realized than visual ones.

We also predicted that within the auditory modality, matching performance for questions would be better than that for the focused utterances. This prediction was based on suggestions that both Cantonese and English employ an F0 rise to mark a question, whereas cues for focus may differ across the languages. This prediction was borne out by the results, as matching performance was significantly better for questions than focused utterances. It is worth noting though that even though matching performance for the focus stimuli was poorer than the question ones, it was still better than chance. This indicates that there may be common cues for focus (such as the relative greater amplitude of mouth opening for focused constituents) across the languages.

Our final prediction was that matching performance will be the poorest across auditory-visual modality. As predicted, we found this to be the case, with performance for across modality matching significantly poorer than that for within auditory modality matching. However, across modality matching performance did not differ significantly from that for within visual modality matching.

The perceivers' difficulty in matching performance in the visual modality might be due to the poorer reliability of visual prosody cues overall. Furthermore, it could be that seeing a foreign talker's face may have distracted the perceivers from performing the task [16]. In this regard, a future study could examine the perceivers' eye gaze patterns for native and non-native visual speech to determine which face regions the perceivers look at. It may also be that the cues to visual prosody in Cantonese are not the same as English. An interesting future study would be to examine how well native speakers of Cantonese can match auditory and visual prosody in their native language and in English. This will help to ascertain how similar the auditory and visual cues to linguistic prosody might be between these languages.

5. References

[1] M. Studdert-Kennedy and K. Hadding, "Auditory and linguistic processes in the perception of intonation contours," *Language and Speech*, vol. 16, no. 4, pp. 293-313, 1973.

[2] S. J. Eady, W. E. Cooper, G. V. Klouda, P. R. Mueller, and D. W. Lotts, "Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments," *Language and speech*, vol. 29, no. 3, pp. 233-251, 1986.

[3] M. Swerts and E. Krahmer, "Facial expression and prosodic prominence: Effects of modality and facial area," Journal of Phonetics, vol. 36, pp. 219-238, 2008.

[4] J. Kim, E. Cvejic, and C. Davis, "Tracking eyebrows and head gestures associated with spoken prosody," *Speech Communication*, vol. 57, pp. 317-330, 2014.

[5] E. Cvejic, J. Kim, and C. Davis, "Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody," *Cognition*, vol. 122, no. 3, pp. 442-453, 2012.

[6] S. Simonetti, J. Kim and C. Davis, "Cross-modality matching of linguistic and emotional prosody," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, 2015.

[7] B. R. Xu and P. Mok, "Final rising and global raising in Cantonese intonation," in *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong*, 2011, pp. 2173–2176.

[8] J. K. Y. Ma and T. L. Whitehill, "The effects of intonation patterns on lexical tone production in Cantonese," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 2004.

[9] Y. Xu, "Post-focus compression: Cross-linguistic distribution and historical origin," in *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong*, 2011, pp. 152-155.

[10] W. E. Cooper, S. J. Eady, and P. R. Mueller, "Acoustical aspects of contrastive stress in question answer contexts," *Journal of the Acoustical Society of America*, vol. 77, pp. 2142-2156, 1985.

[11] W. L. Wu and Y. Xu, "Prosodic focus in Hong Kong Cantonese without post-focus compression," in *Proceedings of Speech Prosody, Chicago*, 2010.

[12] M. Dohen, H. Loevenbruck, and H. Hill, "Recognizing prosody from the lips: Is it possible to extract prosodic focus from lip features?" in *Visual Speech Recognition: Lip Segmentation and Mapping*, 2009, pp. 416-438.

[13] S. Alexanderson, D. House, and J. Beskow, "Aspects of co-occurring syllables and head nods in spontaneous dialogue," in *Proceedings of 12th International Conference on Auditory-Visual Speech Processing*, 2013.

[14] K. I. Forster and J. C. Forster, "DMDX: A Windows display program with millisecond accuracy," *Behavior Research Methods, Instruments, and Computers*, vol. 35, no. 1, pp. 116-124, 2003.

[15] H. Stanislaw and N. Todrov, "Calculation of signal detection theory measures," *Behavior Research Methods, Instruments, & Computers*, vol. 31, no. 1, pp. 137-149.

[16] M. Babel and J. Russell, "Expectations and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 2823-2833, 2015.