Spatially Upward and Emotionally Uncertain: A Pilot Study on Mental Representations of Lexical Tones

Feier Gao (feiergao@seu.edu.cn), Southeast University, Nanjing, China
Chun Hau Ngai (cngai059@uottawa.ca), University of Ottawa, Ottawa, Canada
He Zhou (he.zhou@polyu.edu.hk), Hong Kong Polytechnic University, Hong Kong, SAR

Abstract

The present study investigated cross-modal correspondences between Cantonese tones and two dimensions: (1) spatial motion and (2) emotional valence, via a forced-choice mapping task on Hong Kong native speakers. Results show that the two contour tones (rising and falling contours) could be reliably matched with motions (upward and downward) that are congruent with pitch trajectories; and in contrast, the correspondence between contour tones and emotion valence (rising tone is positive and falling tone is negative) was less robust and limited for selective vowels. In summary, our findings indicate that beyond arbitrary form-meaning correspondence, vertical spatial information, both concrete (motion) and abstract (valence), are also encoded in lexical tones of Hong Kong Cantonese, but that the relative strengths of the two types of correspondences were not equivalent.

Keywords: cross-modal correspondence; lexical tones; conceptual metaphor; spatial motion; emotional valence;

Introduction

The arbitrary pairing between words' forms and meanings has been widely recognized as a prominent assumption of structuralism (de Saussure, 1916). That is, linguistic sign is not directly connected with referents in the real world but instead forms an arbitrary form-to-meaning pairing. This notion has been challenged by the robust sound symbolism across languages (see Lockwood & Dingemanse, 2015, among others). Iconicity constitutes an essential part of non-arbitrariness, referring to the phenomenon that sounds are cross-modally linked with referents in the dimensions of visual, tactile, proprioceptive or many other sensory percepts. One of the well-documented examples is the *bouba-kiki* phenomenon (Maurer et al., 2006), referring to the iconic resemblance between articulatory gestures (i.e., a round vs. unrounded vowel) and shapes (i.e., a round vs. angular).

Beyond segmental elements, little is known about the role of suprasegments, such as tones, in the demonstration of non-arbitrariness. Unlike vowels and consonants, production of tones itself is not always characterized by articulatory movements (Shaw et al., 2016), but more by the change of pitch trajectories. This distinctive feature provides innovative window into the symbolic potentials of tones. As hypothesized in Thompson (2018), it is possible that a falling pitch contour adds an imitative dimension to iconicity by resembling 'a physical or emotional drop'. Dynamic pitch trajectories make spatially-relevant iconicity plausible for tones, especially the ones involving pitch contours.

Audiospatial binding has been commonly identified in the production and perception of tones. High- and low-frequency sounds are often perceived as coming from higher and lower space, respectively (Parise et al., 2014). In languages where tones are used phonemically, such as Mandarin Chinese, facial cues such as head, eyebrows, and lips move in the way that aligns with the pitch trajectories of each lexical tone in terms of spatial and temporal dynamics during speech production (Garg et al. 2019); and movements of neck, head, and lips that are relevant to tone articulation, as well as hand gestures that spatially resemble pitch changes, are attested to improve visual tone identification (Chen and Massaro 2008; Hannah et al. 2017; Morett et al., 2022). These findings indicate that conceptualization of tones is likely to be spatially relevant, as spatial motions of objects in real world can be encoded in lexical contour tones (e.g., rising and falling).

This therefore raises question of whether people can activate spatio-motor representations when hearing pitch contours. In addition, since many abstract concepts such as emotion (e.g., "high spirit", "cheer up", "feeling down") are grounded in vertical conceptual metaphors related to space and motion (Lakoff & Johnson, 1980), we ask whether abstract notions such as emotional valence also plays a role in the conceptualization of pitch contours.

The Present Study

Previous studies have demonstrated the correspondence between size/shape dimensions with tones—both nonlinguistic pitch (Gallace & Spence, 2006; Marks, 1987; O'Boyle & Tarte, 1980) and phonemic tones (Chang et al., 2021; Shang & Styles, 2017, 2023). However, these studies have revealed inconsistent findings in terms of how different lexical tones are mapped with object shapes and sizes. This raises question of where the core symbolic potential of tones lies, especially for languages with a complex tonal inventory such as Mandarin and Cantonese. As previous work have revealed a robust audiospatial binding in the production and perception of phonemic tones (e.g., Connell et al., 2013; Garg et al., 2019), it is plausible for us to hypothesize that space-related concepts, relative to size or shape, serves a more important role in the conceptualization of pitch contours.

To further our understanding of the mental representation of auditory tones, the present study investigates the cross-modal correspondence between tones and two interrelated dimensions—motion and valence (Casasanto & De Bruin, 2019; Casasanto & Dijkstra, 2010). We specifically address three research questions: (1) do pitch contours activate the

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

representation of a spatially congruent motion (e.g., a rising contour denotes climbing), and if it does, (2) whether such correspondence also exists in a more abstract dimension like emotion (e.g., a rising contour matches with happiness), and (3) are the correspondences demonstrated by lexical tones innate or shaped by language-specific experience.

This study focuses on the correspondence demonstrated by two contour tones in Cantonese—the high-rising tone (Tone 2) and the low-falling tone (Tone 4) (Bauer & Benedict, 1997; Matthews & Yip, 2013). The contrast on pitch trajectories allows us to probe the correspondence demonstrated by dynamic pitch contours. Specifically, native speakers with a Hong Kong origin were chosen as targeted participants of the current study. The rationale that we selected Hong Kong Cantonese, rather than Mandarin varieties, was to tease apart the interference of explicit usage of verbal metaphors during language acquisition. In Mandarin, spatially equivalent verbal metaphors and *pinyin* marks[1], as well as co-speech gestures, are commonly used to describe lexical tones, especially in the context of literacy instruction. The involvement of those visual depictions and embodied actions (Chen & Massaro, 2008; Morett et al., 2022) likely shape the way language users conceptualize lexical tones. Unlike Mandarin, Cantonese, especially the Hong Kong variety spoken outside of mainland China, does not use verbal metaphors or *pinyin* in oral or written communication (Zhang & McBride-Chang, 2011), making it possible for us to probe whether the motion and valence correspondence, if there is any, is independent of language-specific experience.

Methods

This study used a forced-choice mapping paradigm (Cuskley, 2013; Imai et al., 2008), in which listeners match audio clips with words in the way that they were most congruent with each other. The task was conducted in the form of word-meaning guessing game based on an "alien" language.

Participants

Sixty native speakers of Hong Kong Cantonese participated in the online study for payment. 49 of them were recruited via Prolific and 11 were recruited through personal network of the authors. All participants self-reported that they were raised in Hong Kong and have never lived outside of the region for more than two years before the age of 18.

Stimuli

Visual Stimuli For each type of correspondence, seven pairs of Chinese words were used as visual options (total N = 14). Within each pair, one word was "upward" and the other was "downward" for the tone-motion mapping; likewise, "positive" and "negative" for the tone-valence mapping (cf. Table 1). A pre-norming task was conducted to balance intensities and directionalities between two words within each pair. Each word pair was displayed twice (14 × 2=28). Additional 56 pairs of irrelevant visual words were used as fillers.

Table 1: Sample word stimuli.

Type	Directionality	Word	Translation
motion	upward	爬山	"hiking"
	downward	潛水	"diving"
valence	positive	開心	"cheerful"
	negative	害怕	"terrified"

Auditory Stimuli Seven Cantonese vowels /i y ɛ œ a ɔ u/ were used as auditory stimuli. Each vowel was produced in the Cantonese high-rising tone (Tone 2) and the low-falling tone (Tone 4). All stimuli were recorded by a female and a male native speaker of Hong Kong Cantonese, hence giving 28 clips in total (Figure 1). Each stimulus was associated with a word pair, counterbalanced across subjects. Additional 56 auditory monosyllables carrying the other four Cantonese tones, each associated with a filler word pair, were used as filler stimuli. All audio clips were scaled to 74dB.

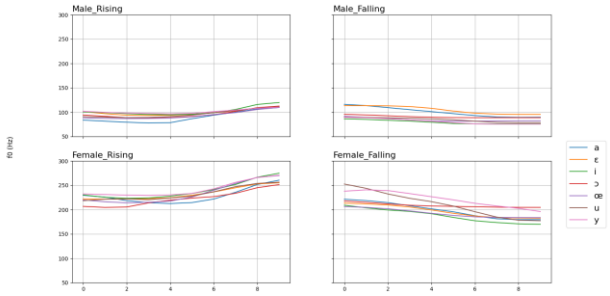

Figure 1: Visualization of auditory stimuli.

Procedure

Participants completed the online experiment using their own Internet browser. The experiment was administered online and programmed using the PennController for Ibex (Zehr & Schwarz, 2018). Participants provided their consent in the beginning and were introduced with some basic information of an "alien" planet. They were then asked to take part in a vocabulary game based on an "alien" language. Each page contained only one trial. For each trial, the audio clip appeared on the top of the page, and two visual words were presented below the audio side by side (randomized position). Participants clicked the button to listen to the audio clip only once and then selected the word meaning from two options. Each participant completed a total of 84 trials (critical N = 28, filler N = 56) in a fully randomized order. In the end, participants

[1] *píng* 'flat' for the high-level tone, *yáng* 'rise' for the rising tone, *guǎi-wān* 'turn' for the low-dipping tone and *jiàng* 'descend' for the falling tone.

completed a background questionnaire and answered a validation question asking which audio file has never appeared in the experiment. The task was demonstrated to participants in colloquial Cantonese using traditional Chinese characters.

Data Screening

Six participants were excluded for clicking wrong answers to the validation question, leaving 54 valid subjects for analysis.

Results

A series of logistic regression models, using the lme4 packages 1.1-26 (Bates et al., 2015) in R (R Core Team, 2022), were constructed to examine the effects of *tone* (rising vs. falling) and *vowel* (/i/, /y/, /ɛ/, /œ/, /a/, /ɔ/ vs. /u/) as well as their interaction on the matching judgements. For both tone-motion and tone-valence associations, *selections* were treated as a dependent binomial variable (upward vs. downward for the motion mapping; positive vs. negative for the valence mapping), and *subjects* as a random variable. A maximal model[2] was constructed first and sequentially pruned based on results of model comparisons.

Tone-Motion Correspondence

For the mapping between tones and spatial motions, results of model comparisons showed a similar goodness of fit [$\chi^2(6) = 9.836$, $p = 0.132$] between models with and without the *tone* × *vowel* interaction. The best-fitting model[3] indicated that the two factors did not have interactional effect on the matching judgements.

Table 2: Selection summary (tone-motion mapping).

selection	rising	falling	Sum
"upward"	207	118	325
"downward"	171	260	431
Sum	378	378	756

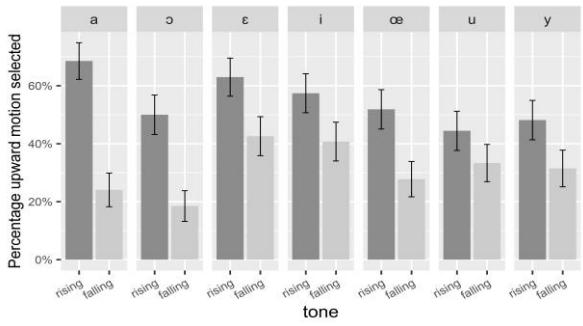

Figure 2: Observed percentages of "upward" motions selected (tone-motion correspondence).

Effects of *tone* and *vowel* were examined separately by pairwise comparisons, using the emmeans package (Lenth et al., 2021) in R with a Bonferroni correction for the *p* values on each factor. Results show that the rising tone and the falling tone demonstrated significantly different motion mapping judgements ($\beta = 1.033$, $p < .001$): the falling contour was better matched with a downward motion (average = 60.3%) than the rising contour, and the rising contour with an upward motion (average = 63.7%) than the falling contour. Also, pairwise comparisons on *vowel* showed that there was no significant difference across the vowel categories in terms of motion mapping (*p*s > .05).

These results indicate a robust tone-motion correspondence demonstrated by two contour tones in Hong Kong Cantonese. Specifically, native speakers could associate tones with spatial motions in the way that pitch trajectories are congruent with movement directions, i.e., the rising tone is upward and the falling tone is downward (Gao, 2024). Moreover, such correspondence was *not* modulated by vowels. The results indicated that vowel category did not significantly affect native speakers' tone-motion matching judgements.

Tone-Valence Correspondence

Model comparison for the tone-valence mapping followed the same procedure. Results showed that the more complex model was a better fit than the one without the *tone* × *vowel* interaction [$\chi^2(6) = 24.041$, $p < .001$]. The best-fitting model[4] therefore indicates that there is a significant interactional effect on the matching judgements.

Table 3: Selection summary (tone-valence mapping).

selection	rising	falling	Sum
"positive"	180	119	299
"negative"	198	259	457
Sum	378	378	756

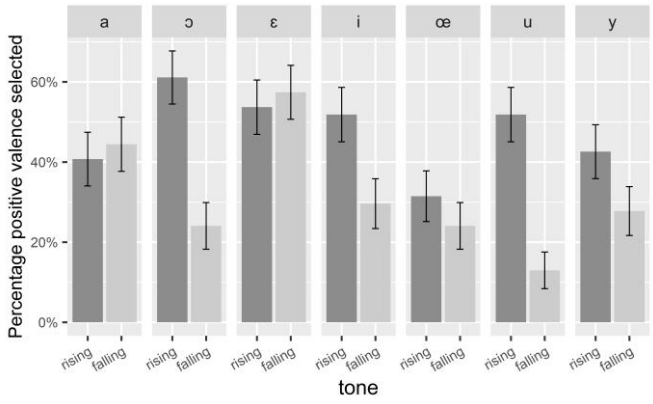

Figure 3: Observed percentages of "positive" emotions selected (tone-valence correspondence).

Post-Hoc tests were conducted to unpack the interaction, using the same package in R. A pairwise comparison (with

[2] selection ~ vowel : tone + vowel + tone + (1 + vowel : tone + vowel + tone |subject)

[3] selection ~ vowel + tone + (1 + tone |subject)

[4] selection ~ vowel : tone + vowel + tone + (1|subject)

Table 4: Post-Hoc Analysis on tone-valence correspondence (*p* values corrected; Partial results are provided).

Factors		***Est.***	**SE**	***z***	***p*-values**
Comparison = *Vowel*					
rising	/ɔ/ - /œ/	-1.31e+00	0.420	-3.119	0.038 *
falling	/ɛ/ - /ɔ/	-1.447	0.421	-3.439	0.012 *
	/ɛ/ - /œ/	-1.447	0.421	-3.439	0.012 *
	/ɛ/ - /u/	-2.2027	0.490	-4.498	< 0.001 ***
	/ɛ/ - /y/	-1.254	0.410	-3.059	0.047 *
Comparison = *Tone*					
/a/	rising - falling	-0.152	0.390	-0.389	0.697
/ɛ/		-0.170	0.413	-0.412	0.681
/i/		0.973	0.427	2.277	0.023 *
/ɔ/		2.200	0.677	3.248	0.001 **
/œ/		0.427	0.468	0.911	0.362
/u/		15.012	2.934	5.116	< 0.001 ***
/y/		0.678	0.424	1.600	0.110

Bonferroni correction for the *p* values) was conducted. Factors were held constant to examine the contribution of individual factors. Partial statistical results are shown in Table 4. Results showed that the mapping between tone and emotional valence is modulated by vowel categories. First, certain tone-vowel combinations demonstrate significantly different valence association than others. When paired with a rising contour, /ɔ/, compared with /œ/, was more likely to be associated with a positive emotion; Similarly, when paired with a falling contour, /ɛ/, compared with /ɔ/, /œ/, /u/ and /y/, is more likely to be mapped onto a positive emotion. Second, the contrast between two tones in terms of valence mapping (i.e., the rising tone is positive and the falling tone is negative) was only found to be significant for /i/, /ɔ/ and /u/, but not the other four vowels.

As opposed to the tone-motion mapping, matching judgements obtained from the forced-choice mapping task indicate a less robust association between lexical tones and emotional valence. Specifically, the tone-motion correspondence was only robust for selective vowels, and the mechanism underlying those patterns is pending further investigation.

Discussion

The present study investigated two types of crossmodal correspondences (spatial motion and emotional valence) demonstrated by lexical tones (rising and falling contours) in Cantonese. Our results indicated that native speakers could reliably match the rising tone with an upward motion and the falling tone with a downward motion, and this pattern is consistent across vowel categories. Also, we provided novel evidence that the symbolic potential of tones cannot be fully extended to a less concrete domain, such that the rising tone could be matched with a positive emotion and the falling tone with a negative one. Unlike the tone-motion correspondence, the association between pitch contours and emotional valence was less robust, as the mappings were significantly modulated by vowels.

The distinctive patterns can be explained by the different mechanisms underlying the two types of correspondences. The conceptualization of pitch contours is considered more "iconic" in their use of vertical space. Pitch trajectory of a rising tone goes "up" and that of a falling tone goes "down", resembling spatial movements in an iconic manner. Matching judgements obtained in the tone-motion mappings therefore indicate that subtle acoustic details can contribute to the mental representation of lexical tones. However, the correspondence between (rising/falling) contours and (positive/negative) emotions are relatively more abstract, which is mostly understood in terms of conceptual metaphor (Lakoff & Johnson, 1980; Morett et al., 2022). The less consistent tone-valence mappings likely indicate that correspondence driven by "metaphorical" analogy is inherently weaker than the one that is based on "iconic" resemblance.

It is also noteworthy that whether conceptual metaphors are explicitly described or not in the language may also serve an important role in the mental representation of lexical tones. Although spatially relevant descriptions and *pinyin* marks are explicitly used in Standard Mandarin, they are absent in literacy instruction in Hong Kong. Selecting Hong Kong native speakers as target population, the current study allows us to tease apart the effects of spatial nature of tones and language experience on the tone-based correspondence. The relatively weaker association found in Hong Kong Cantonese as opposed to Standard Mandarin indeed provides evidence that language experience can shape how we understand and represent lexical tones, and implicit use of vertical conceptual metaphor may lead to a relatively loose vertical mapping between tone and space/motion, whereas explicit usage may promote such correspondence.

While previous work mostly focused on association between tones and size/shape (e.g., Chang et al., 2021; Shang & Styles, 2017, 2023), the present study provides novel evidence regarding the vertical mapping of pitch contour—both in an iconic and a metaphorical way, in a language where tones are used phonemically. These findings provide further

insights into sound symbolic potentials of tones, showing that the form-meaning correspondence demonstrated by lexical tones is *not* wholly arbitrary. Instead, vertical spatial information is encoded in the pitch contours and can be activated during speech perception.

Acknowledgments

We would like to thank Siying Chen, Yu-Fu Chien, Yanting Li and Ka Fai Yip for their help in stimuli creation. We also benefited a lot from discussions with Stephen Matthews and Marc Brunelle, and several anonymous reviewers of CogSci 2025, whose comments have greatly improved our paper.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), Article 1.

Bauer, R. S., & Benedict, P. K. (1997). Modern Cantonese Phonology. In *Modern Cantonese Phonology*. Mouton de Gruyter.

Casasanto, D., & De Bruin, A. (2019). Metaphors we learn by: Directed motor action improves word learning. *Cognition*, *182*, 177–183.

Casasanto, D., & Dijkstra, K. (2010). Motor action and emotional memory. *Cognition*, *115*(1), 179–185.

Chang, Y.-H., Zhao, M., Chen, Y.-C., & Huang, P.-C. (2021). The effects of Mandarin Chinese lexical tones in sound–shape and sound–size correspondences. *Multisensory Research*, *35*(3), 243–257.

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *The Journal of the Acoustical Society of America*, *123*(4), 2356–2366.

Connell, L., Cai, Z. G., & Holler, J. (2013). Do you see what I'm singing? Visuospatial movement biases pitch perception. *Brain and Cognition*, *81*(1), 124–130.

Cuskley, C. (2013). Mappings between linguistic sound and motion. *Public Journal of Semiotics*, *5*(1), Article 1.

de Saussure, F. (1916). *Cours de linguistique générale*. Payot.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*(7), 1191–1203.

Gao, F. (2024). Crossmodal correspondence between lexical tones and visual motions: A forced-choice mapping task on Mandarin Chinese. *Linguistics Vanguard*, *10*(1), 721–729.

Garg, S., Hamarneh, G., Jongman, A., Sereno, J. A., & Wang, Y. (2019). Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. *Speech Communication*, *113*, 47–62.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By* (First Edition). University of Chicago Press.

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Riebl, H., & Singmann, H. (2021). *Emmeans: Estimated marginal means, aka least-squares means* (Version 1.6.3) [Computer software]. https://CRAN.R-project.org/package=emmeans

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, *6*.

Marks, L. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 384–394.

Matthews, S., & Yip, V. (2013). *Cantonese: A Comprehensive Grammar* (2nd ed.). Routledge.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science*, *9*(3), 316–322.

Morett, L. M., Feiler, J. B., & Getz, L. M. (2022). Elucidating the influences of embodiment and conceptual metaphor on lexical and non-speech tone learning. *Cognition*, *222*, 105014.

O'Boyle, M. W., & Tarte, R. D. (1980). Implications for phonetic symbolism: The relationship between pure tones and geometric figures. *Journal of Psycholinguistic Research*, *9*(6), 535–544.

Parise, C. V., Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, *111*(16), 6104–6108.

R Core Team. (2022). *R: A language and environment for statistical computing*. https://www.r-project.org/

Shang, N., & Styles, S. (2017). Is a high tone pointy? Speakers of different languages match Mandarin Chinese tones to visual shapes differently. *Frontiers in Psychology*, *8*.

Shang, N., & Styles, S. (2023). Implicit Association Test (IAT) studies investigating pitch-shape audiovisual cross-modal associations across language groups. *Cognitive Science*, *47*, 13221.

Shaw, J. A., Chen, W.-R., Proctor, M. I., & Derrick, D. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research*, *59*(6), S1566–S1574.

Thompson, A. L. (2018). Are tones in the expressive lexicon iconic? Evidence from three Chinese languages. *PLOS ONE*, *13*(12), e0204270.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)* [Computer software]. https://www.pcibex.net/

Zhang, J., & McBride-Chang, C. (2011). Diversity in Chinese Literacy Acquisition. *Writing Systems Research*, *3*(1), 87–102.