Versatility of phonemic pitch in affective iconicity and perceptual reorganisation

Zheng, T.

Chapter 2
Affective iconicity of tonemes in bisyllabic units in Standard Chinese

A version of this chapter has been published as: Zheng, T., Levelt, C. C., & Chen, Y. (2025). The affective iconicity of lexical tone: Evidence from Standard Chinese. *The Journal of the Acoustical Society of America*, 157(1), 396–408. https://doi.org/10.1121/10.0034863

Abstract: Previous studies suggested that pitch characteristics of lexical tones in Standard Chinese influence various sensory perceptions, but whether they iconically bias emotional experience remained unclear. We analysed the arousal and valence ratings of bisyllabic words in two corpora (Study 1) and conducted an affect rating experiment using a carefully designed corpus of bisyllabic words (Study 2). Two-alternative forced-choice tasks further tested the robustness of lexical tones' affective iconicity in an auditory nonce word context (Study 3). Hierarchical linear models, generalised linear mixed models, and cross-validation were employed to understand the relationship between lexical tones and the emotional responses of tone-carrying words. Results consistently indicated that words with a falling-falling tonal sequence, both real and nonce words, received higher arousal ratings than those with rising-rising and rising-low tones. Only in nonce words, the high-high sequence was more likely to be associated with the low-arousal option; the falling-falling tone sequence was more often linked to negative-valence choice, while high-high and rising-rising tones with positive-valence. These findings, though subtle, suggest that the use of pitch in lexical tones influences emotional responses during the processing of tone-carrying words, pointing to an inherent iconic quality in lexical tones that may subtly shape speakers' emotional experiences.

Keywords: Affective iconicity, Standard Chinese, Emotional arousal, Emotional valence, Bisyllabic tonal sequence

2.1 Introduction

While lexical items have been conventionally characterised by the seemingly arbitrary association of sounds with meanings (Hockett, 1958), recent research has brought to light a more systematic sound-meaning resemblance known as sound symbolism or iconicity (see a review in Winter et al., 2023). Notably, research has observed emotional sound symbolism across many languages. For instance, Adelman et al. (2018) find that individual phonemes of words in Dutch, German, Spanish, English, and Polish predict their emotional valence ratings, with the first phoneme of a word predicting its valence better than subsequent phonemes. Similarly, a nasal consonant as the first phoneme of a word has been reported to help predict word valence in English, German, Dutch, and Chinese (Louwerse & Qu, 2017). Aryani et al. (2018) extend their examination of the predictive effect of phonetic features in German to arousal ratings, and conclude that phonemes at the segmental level contribute to both the emotional arousal and valence ratings of words, with a more pronounced effect on emotional arousal.

It is important to note that about half of the world's languages, if not more, have lexical tones (Maddieson, 2023; Yip, 2002), where words are distinguished not only via segmental phonemes but also through supra-segmental cues, especially pitch variation (Hyman, 2014). For example, in Standard Chinese, the segmental syllable *ma* with a high-level pitch means "mother," but with a falling-rising pitch contour, it means "horse." Little is known about the potential connections between lexical tone and the affective arousal and valence responses of tone-bearing words. This study aimed to explore such connections in Standard Chinese to better understand the relationship

between speech sounds and affective experiences in human language. Although lexical tones are also signalled via other acoustic cues such as duration and intensity (Whalen & Xu, 1992), our primary focus is their pitch characteristics, the primary acoustic correlate of which is fundamental frequency changes (*f0*).

There has been research on the general relationship between pitch and emotional arousal and valence in quite a few languages (e.g., Banse & Scherer, 1996; and Bänziger & Scherer, 2005 on German; Liu & Pell, 2012 on Standard Chinese; see also reviews in Pakosz, 1983; and Scherer, 1979, 1989; Frick, 1985). High-arousing emotions generally correlate with a higher pitch level, higher average pitch height, wider pitch range, and steeper pitch slope, in comparison to low-arousing emotions (Bänziger & Scherer, 2005; Laukka et al., 2005; Scherer et al., 2003). Research on the pitch correlate of valence, however, remains inconclusive. While some report a lower average pitch height for positive valence (Scherer & Oshinsky, 1977), the more prevailing view suggests a trend towards a higher pitch level and average pitch height, and wider pitch range (Belyk & Brown, 2014; Kamiloğlu et al., 2020; Laukka et al., 2005). The variability in findings underscores the complexity of the pitch-valence relationship.

Given the existing findings concerning the (possible) relationship between pitch and emotional arousal and valence, and the relationship between segmental phonemes and their potential emotional attributes, the specific question we addressed in this project is: Does pitch, a fundamental cue for emotional prosody, influence emotional responses in lexical tone languages like Standard Chinese, in a way similar to the influence of segmental phonemes on arousal and valence ratings in non-tonal languages?

Pitch has received considerable attention in iconicity research (see Appendix A, Table A1 for a summary of related studies). The connection of nonlinguistic pure tone with various sensory modalities has been well-established (see an overview in Deroy & Spence, 2013; Spence, 2011). We now know that the perceived pitch height of a sound is linked to a range of sensory attributes, including size, shape, lightness, weight, colour, movement direction, taste, and touch. In linguistic research, the concept of pitch iconicity is often discussed from the perspective of the "frequency code" (Gussenhoven, 2016; Ohala, 1984), which relates higher and/or rising *fo* to socio-cognitive interpretations such as submission and politeness, in contrast to low and/or falling *fo* for dominance and aggression (Perlman, Clark, et al., 2015 in English; Rojczyk 2011 in Polish; Stel et al. 2012 in Dutch; cf. Winter et al. 2021, which shows cultural variations in pitch encoding of politeness).

To our knowledge, Ohala (1984, 1997) is the first to illustrate pitch iconicity in lexical tones, showing the associations of high tone with connotations of smallness, diminutiveness, and familiarity, while low tone with largeness in three lexical tone languages (i.e., Ewe, Yoruba, Cantonese). Subsequent studies confirm the iconicity of lexical tones for various dimensions such as size, shape, gender, and storybook character features (Chang et al., 2021; Lapolla, 1995; Shang & Styles, 2017; X. Wang, 2021; Wong & Kang, 2019). Thompson (2018) reports a distribution bias for high-level tone but only in onomatopoeic words in Mandarin, Taiwanese Southern Min, and Hong Kong Cantonese. Yao et al. (2013) suggest a direct link between tonal sequences of emotion words and specific emotion types, observing a distribution bias for falling tones in anger and joy words, high-level tones in sadness words, and rising and low-dipping tones in fear words. Similarly, Yap

et al. (2014) propose that the direction of tonal pitch contours in Chinese words correlates with the emotional valence of their English translation equivalents, with rising contours perceived as the most positive and falling contours as the most negative.

Despite findings on 1) the association between pitch characteristics and emotional experiences and 2) the iconic significance of pitch across different sensory domains in tonal languages, it remains unclear whether the phonemic use of pitch for lexical tone in these languages can iconically influence or bias emotional experiences. To address this knowledge gap, we set out to examine whether speakers perceive the phonemic pitch of lexical tones as affectively iconic, particularly concerning emotional dimensions such as arousal and valence in a tonal language.

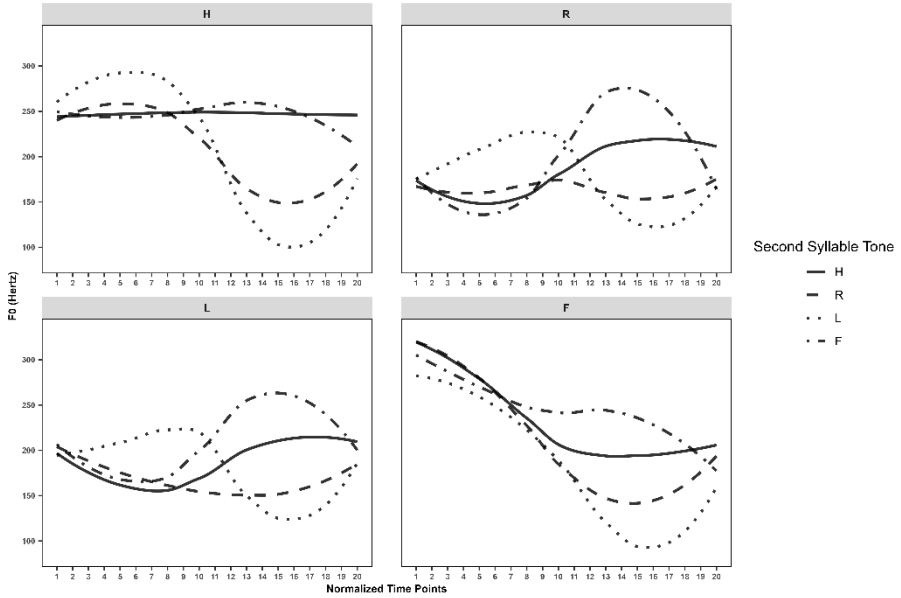
As mentioned earlier, our empirical domain is Standard Chinese, which includes four distinctive lexical tones: high-level (Tone 1, T1, or H), mid-rising (Tone 2, T2, or R), low-dipping (Tone 3, T3, or L) and high-falling (Tone 4, T4, or F). Our investigation centres on bisyllabic words primarily due to the following two reasons. One is that bisyllabic compounds represent the most frequently used basic units in Standard Chinese (C. N. Li & Thompson, 1981). The second consideration, and one of practical significance, is the availability of two databases (Y. Wang et al., 2008; X. Xu et al., 2022) that include arousal and valence ratings. These well-established datasets enabled us to leverage existing, standardised measures of emotional ratings to address our research question on the effects of lexical tone on affective iconicity.

For bisyllabic SC words, the four lexical tones yield 16 tonal sequences: HH, HR, HL, HF, RH, RR, RL, RF, LH, LR, LL, LF, FH, FR, FL, and FF. Figure 2.1 shows the pitch contours of these tonal combinations using the token /mama/ (excluded from any subsequent

analyses) as a reference to visually illustrate commonly observed pitch patterns in bisyllabic words. For further details on contextual tonal variations, readers are referred to Xu (1997) and Chen & Gussenhoven (2008).

Figure 2.1

Illustrated pitch contours of tonal sequences for /mama/ in SC. Tones are denoted as H (T1: high-level), R (T2: rising), L (T3: low-dipping), and F (T4: falling). The titles of each panel indicate the tone category of the first syllable. The pitch contour for "LL" in the "L" panel reflects tone sandhi, resulting in a pitch contour similar to "RL" in the "R" panel.



We aimed to uncover the affective implications of lexical tones by examining how pitch characteristics in these bisyllabic lexical tone sequences may introduce iconic biases in emotional arousal and valence ratings in Standard Chinese. It is important to note that delving

into the affective iconicity of lexical tone does not imply any inherent iconicity of lexical tones themselves. Rather, our interest lies in exploring the broader iconicity effect associated with pitch, the primary conveyor of lexical tone category information, and its impact on the perceived emotional qualities of linguistic expressions featuring specific lexical tone sequences.

Drawing on the aforementioned findings concerning pitch, iconicity, and emotional prosody, we hypothesised that possible pitch iconic effects of lexical tonal sequences would mirror phenomena observed in (lexicalised) emotional prosody. Specifically, we predicted that pitch level, average pitch height, pitch range, pitch slope, and pitch contour direction are important features that could influence the emotional arousal and valence ratings of tone-carrying words. Higher overall pitch level, higher average pitch height, wider pitch range, and steeper pitch slope are expected to be associated with higher arousal (Bänziger & Scherer, 2005; Scherer et al., 2003; Thompson, 2018). Furthermore, higher pitch level, wider pitch range, and upward pitch contour direction are likely associated with positive valence (Belyk & Brown, 2014; Kamiloğlu et al., 2020; Yap et al., 2014).

We conducted three studies to explore the predictive role of lexical tonal pitch patterns for emotional arousal and valence ratings and choices. Study 1 leveraged existing corpus data and analysed data from two existing corpora of emotional ratings for written bisyllabic words. Study 2 employed an online rating task for written bisyllabic words with an enhanced design, balanced stimuli, and advanced statistical modelling, compared to the corpus data. These two studies utilised scaled data (Likert/Slider) to allow for nuanced judgments for real-word stimuli (e.g., Warriner et al., 2013). Study 3 moved onto the auditory domain and further explored the predictive effect of tonal

pitch characteristics, using spoken bisyllabic nonce words (devoid of semantic meaning) in a two-alternative forced-choice task, which is widely used in studies for unfamiliar stimuli to facilitate clear decision without overcomplicating the judgment process (e.g., Monaghan et al., 2012).

Given that the stimuli in Studies 1 and 2 were real words, we controlled for three key factors—word frequency, part-of-speech, and onset consonant type—that could potentially affect the arousal and valence ratings. Word frequency is known to play a significant role in language processing (Brysbaert et al., 2018). Regarding parts of speech, prior research indicates that adjectives and verbs (for English and Spanish) tend to exhibit stronger iconic effects than nouns (Perry et al., 2015). Additionally, consonants have been shown to play a more prominent role than vowels in the classic bouba–kiki iconicity effect (Fort et al., 2015), and nasal onsets are linked to positive valence ratings in Chinese (Louwerse & Qu, 2017). We incorporated these factors into our model analyses. In Study 2, we also balanced tonal sequences across bisyllabic words to ensure similar distributions of lexical tone sequences across categories of onset consonants and parts of speech, with word frequency controlled across categories.

Study 3 utilised nonce words constructed from consonants and vowels commonly used in iconicity research and focused on four lexical tonal sequences to further verify the effects of pitch variation on affective response.

2.1.1 Ethical approval

The studies on affective iconicity in Chapter 2, 3, and 4 adhered to the guidelines set forth by the World Medical Association (Declaration of Helsinki) and obtained approval (2020/16) from the Ethics Committee

of the Faculties of Humanities and Archaeology at Leiden University. Participants willingly registered and provided their consent by acknowledging an informed consent form before taking part in the study. Compensations were provided to all participants upon the complement of the tasks.

2.2 Study 1. Corpus analysis of Chinese bisyllabic words

Study 1 utilised the Chinese Affective Words System (CAWS) corpus (Y. Wang et al., 2008) and the Affective Norms for Chinese Words (NORM) corpus (X. Xu et al., 2022) to investigate the predictive relationship between lexical tonal sequences and the affective ratings of Standard Chinese bisyllabic words regarding emotional arousal and valence.

2.2.1 Method

2.2.1.1 Corpus datasets

The CAWS corpus (Y. Wang et al., 2008) comprises 500 bisyllabic nouns, 500 bisyllabic verbs, and 500 bisyllabic adjectives. Emotional ratings for arousal and valence were obtained from 64 undergraduate university students and subsequently rerated by an additional 30 undergraduates, with all raters aged 18 to 21 years. Valence ratings were provided on a 9-point Likert scale, ranging from negative emotions (e.g., unhappy or annoyed) to positive emotions (e.g., happy, satisfied, or hopeful). Arousal ratings were also given on a 9-point Likert scale, ranging from low arousal (e.g., peaceful, relaxed, or less attentive) to high arousal (e.g., alert, quite excited, or awake). Only the aggregate data across raters were available in the CAWS corpus. For

our analysis, we annotated all words with their tonal sequences. For instance, the bisyllabic word *langman*, meaning "romantic," consists of two falling tones and is marked as "FF." Eight words were excluded from further analysis due to ambiguity in part-of-speech or lack of word frequency information.

The NORM dataset collected valence and arousal ratings for 11,310 simplified Chinese words, out of which we selected 9,573 bisyllabic words for our purpose. In total, valence ratings were obtained by 1,232 participants, and arousal ratings by 1,189 participants. All participants reported falling within the age range of 18 to 62 years, having education backgrounds that spanned from middle school to graduate school. Notably, 96.9% held a college-level education or higher. Moreover, the rating scales differed from those of the CAWS corpus. A 5-point Likert scale was utilised for arousal, ranging from 0 to 4, and a 7-point Likert scale was employed for valence, ranging from -3 to +3. We followed the same annotation scheme for lexical tones and run HLM for the NORM dataset.

2.2.2 Data analysis

We conducted a Hierarchical Linear Regression (HLM) analysis using R (R Core Team, 2023), comparing models with independent variables (first all control variables and then the predictor variable) entered in successive steps. In the first model, we included the control variables, i.e., word frequency *(log)*, part-of-speech, and onset consonant type, to account for potential confounding effects on emotional ratings. In the second model, we added lexical tonal sequence as the main predictor, allowing us to assess the unique predictive impact of lexical tonal sequences on emotional ratings above and beyond the control variables.

By comparing the coefficient of determination (R^2) between the two HLM models, we gain insights into how much variance in emotional ratings can be attributed to lexical tonal sequences beyond the influence of word frequency, onset consonant type, and part-of-speech category. To delve deeper into tonal sequence distinctions, pairwise multiple comparisons were performed using the *emmeans* package (Lenth, 2023), with Bonferroni correction applied to minimise false positive rate.

2.2.3 Results

The regression diagnostics confirmed that the assumptions of linearity, homoscedasticity, and normality were met, and no collinearity issues were identified. As the dependent variable represented average scores across participants, there were no independence violations.

The first model, including word frequency, part-of-speech, and consonant type, produced a statistically significant model (F = 30.22, p < 0.001), explaining 14.17% (R^2) of the variation in emotional arousal ratings. The second model introduced the lexical tonal sequence, the main factor of our interest, in addition to the control variables. Model comparison between the two models favoured the inclusion of lexical tonal sequence (F = 2.57, p < 0.05), which explained a significant amount of unique variance in the arousal ratings (ΔR^2 = 2.22%, p < .001). Pairwise multiple comparisons showed that the FF (falling-falling, M = 5.84), HF (high-falling, M = 5.71), and HL (high-low, M = 5.82) tonal sequences predicted higher arousal ratings compared to RR (rising-rising, M = 5.31; p < .001, p < .05, p < .05, respectively). However, no significant differences were found between other tonal sequence pairs. Notably, lexical tonal sequences did not significantly contribute to the emotional valence ratings.

Results from the NORM dataset largely aligned with those of the CAWS dataset. Similarly, the first model included three variables: word frequency, part-of-speech, and onset consonant type. The results revealed a significant influence of the three variables on the arousal ratings (F = 45.06, p < 0.001), and the R^2 value showed that the three variables accounted for 10.02% of the variance in emotional arousal. Subsequently, in the second model, the variable lexical tonal sequence was added. Model comparison favoured the second model (F = 3.05, p < 0.001). The lexical tonal sequence collectively explained a significant amount of unique variance in the arousal ratings (ΔR^2 = 0.44%, p < .001). Pairwise multiple comparisons indicated that FF (M = 2.09), HF (M = 2.08), and LF (M = 2.07) tonal sequences predicted higher arousal ratings compared to RR (M = 1.96; p < .001, p < .05, p < .05, respectively). Similar to the CAWS corpus, the HLM for the NORM corpus did not show a significant effect of lexical tonal sequence on the emotional valence ratings. Please refer to Figure 2.2 (CAWS) and Figure 2.3 (NORM) for visualisations.

Figure 2.2

Emotional arousal and valence rating scores for each lexical tonal sequence in the CAWS corpus.

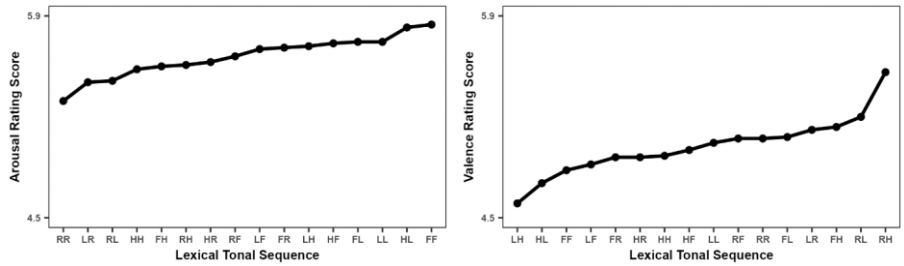
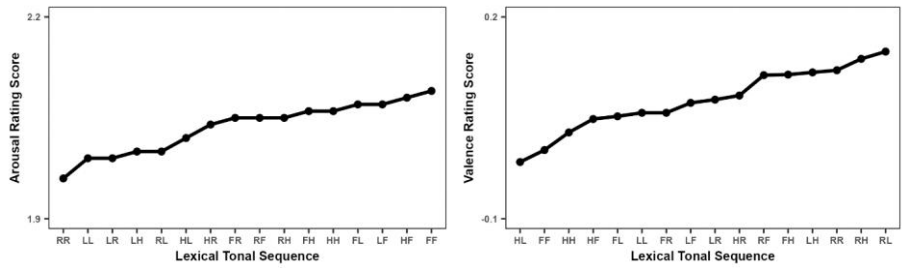


Figure 2.3

Emotional arousal and valence rating scores for each lexical tonal sequence in the NORM corpus.



Moreover, the HL (high-low, in the CAWS corpus) and LF (low-falling, in the NORM corpus) tonal sequences also demonstrated a stronger predictability effect than RR, but the effects of these two contrasts were less consistent, given that we only found a significant difference in one of the two corpora.

2.2.4 Discussion

Using the affective norm data in the corpora and considering relevant lexical properties, our regression analyses showed a significant effect of lexical tonal sequence on the emotional arousal ratings for bisyllabic words. Words with FF and HF tonal sequences were associated with higher arousal values than those with the RR tonal sequence across the CAWS and NORM corpora.

The effect sizes of these contrasts were generally small. However, this finding is well aligned with the small but significant effects of segments. Adelman et al. (2018) found that phonemes collectively accounted for a significant amount of unique variance in valence ratings across five European languages (i.e., English, Spanish, Dutch, German, and Polish; ΔR^2 = 1.40% ~ 4.32%).

The results of Study 1 supported the hypothesis that tonal sequences are likely to affect the emotional arousal ratings of the tone-carrying expressions due to their characteristic pitch levels, average pitch heights, pitch ranges, and pitch slopes. The trend of bias echoes the pitch variations of emotional prosody reported in previous literature (Bänziger & Scherer, 2005; Laukka et al., 2005). Although our results are based on phonological tonal sequences, we believe it is the distinct phonetic pitch patterns that are associated with the perceived varying arousal levels. Specifically, in the FF tonal sequence, increased arousal was linked to a higher average pitch height, a broader pitch range, and a steeper pitch slope. Similarly, in the HF tonal sequence, heightened arousal was associated with both a high pitch level and a higher average pitch height. In contrast, reduced arousal in the RR tonal sequence was associated with a lower pitch level, a lower average pitch height, a narrower pitch range, and a more gradual pitch slope. No significant effect of lexical tonal sequence on the emotional valence ratings was found.

It is important to acknowledge certain limitations of our corpus data. Firstly, the stimuli were unevenly distributed across different tonal sequences. For example, in the CAWS corpus, the number of words with FF tonal sequence (N = 204) was twice as large as the number of words with RR tonal sequence (N = 96). This imbalance may have affected the statistical power and raised concerns about the robustness of the findings regarding emotional arousal, as well as the lack of significant effects concerning emotional valence. Secondly, as participants performed the rating task upon written stimuli, factors like visual complexity (e.g., number of strokes; Peng & Wang, 1997) should be considered, as they can impact processing costs in character reading. For example, Peng and Wang (1997) found that characters

with more strokes require greater processing effort. Additionally, the CAWS and NORM corpora include compounds with complex morphological structures that can affect word recognition (Zhou & Marslen-Wilson, 1995) and may introduce confounding effects. Examples include reduplicated words, which may carry iconic motivation (e.g., D. Xu, 2012); words with neutral tones, which have limited pitch variation; and loanwords, which may carry cultural and semantic associations. Lastly, although aggregate data are commonly used in the investigation of iconicity, they are limited by the so-called "ecological fallacy", which compromises the validity of conclusions and makes it challenging to generalise the observed patterns at the individual level (Pollet et al., 2015). These limitations motivated the design of the rating task in Study 2, which aimed to address the aforementioned drawbacks.

2.3 Study 2. Online assessment of Chinese bisyllabic words

Study 2 adopted a designed experiment approach to scrutinise the effect of lexical tonal sequences on affective ratings through an online rating task. Great care was taken in selecting and validating the stimuli used in this task. Furthermore, individual rating scores of the stimuli were collected, which enabled us to analyse the data using both hierarchical linear regression (HLM) analysis (as in Study 1) and Generalised Linear Mixed Model (GLMM). In addition, we included thorough checks on the inter-rater reliability of the rating data and the application of Leave-One-Out Cross-Validation (LOOCV) for the robustness of the results. In this way, we aimed to validate and extend the findings observed in Study 1.

2.3.1 Method

2.3.1.1 Stimuli: Developed-CAWS corpus

The stimuli for Study 2, developed from the CAWS corpus and referred to as the developed CAWS corpus (DCAWS), included 1500 words across ten critical lexical tonal sequences: HH (T1T1, high-high), HL (T1T3, high-low), HF (T1T4, high-falling), RH (T2T1, rising-high), RR (T2T2, rising-rising), RL (T2T3, rising-low), RF (T2T4, rising-falling), LF (T3T4, low-falling), FH (T4T1, falling-high), and FF (T4T4, falling-falling). These sequences were selected based on three criteria. First, we included all sequences that showed significant differences in arousal ratings in Study 1 (i.e., *FF, HF, HL, LF, RR*). Second, additional tonal pairs were included despite their lack of significance in Study 1 to test the possible effects of other pitch features (e.g., average feature and dynamic changes of pitch contours), which the existing literature suggests have an impact on emotion ratings or as a critical cue for emotional prosody and iconicity (Laukka et al., 2005; Stel et al., 2012; Thompson, 2018). For example, HH was included due to its high pitch level and average pitch height within the high-tone sequence, commonly found in Mandarin onomatopoeia. This sequence provided a valuable test case, with RL providing a contrasting low-pitch sequence where the contour drops to a relatively low *fo* level. RH and FH were added to contrast rising-high and falling-high *fo* contours, while RF was added to contrast with HL. Additionally, the contrasts between HF and LF, as well as FF and RR, enabled us to examine the effects of pitch level, average pitch height, pitch range, slope, and contour direction on affective iconicity (Laukka et al., 2005; Shang & Styles, 2017; Yao et al., 2013; Yap et al., 2014).

We took the following steps to finalise the stimulus list. First, we excluded words with specific characteristics (such as reduplication and neutral tone), loan words, words with ambiguous meanings, words with unclassifiable parts of speech, and those exceeding stroke limits (over 21 for a single character or over 35 for both characters). Stroke counts were obtained from the Lists of Frequently Used Characters and Commonly Used Characters in Modern Chinese (State Language Commission, 1988, 2013). To minimise confounding effects, we set a limit of 21 strokes per character to exclude highly complex characters, allowing up to 35 strokes for two-character words. This criterion balances manageable visual complexity with the maintenance of natural linguistic properties. The limits are consistent with prior research (e.g., a typical range of 10.6 ± 2.5 × 4.4 strokes per character and a maximum of 35 strokes for two-character words; Tse et al., 2024). Additional words were then introduced to ensure a consistent quantity of 150 tokens per tonal sequence, with a balanced distribution of lexical tonal sequences across all onset consonant types (e.g., voiceless unaspirated, voiceless aspirated, fricatives, nasals and laterals, approximants, and onset-less) (Fort et al., 2015; Louwerse & Qu, 2017) and parts of speech (Perry et al., 2015). Word frequency was also controlled (Cai & Brysbaert, 2010). As a result, the DCAWS corpus consists of 910 bisyllabic words chosen from the CAWS corpus, along with an additional 590 bisyllabic words sourced from the dictionary *Xiandai Hanyu Cidian* (*7th*).

2.3.1.2 Participants

Fifty-four native Chinese speakers participated in the emotional arousal rating task, and another fifty-five native Chinese speakers

participated in the emotional valence rating task. Both groups rated all 1500 bisyllabic words in the DCAWS.

Five participants were excluded from data analyses due to inconsistencies in their task performance, such as irregularities in their responses and missing data points. The final sample for data analysis of the emotional arousal rating task included fifty-one participants (aged 22.26 ± 4.79 years; 27 females, 24 males), while analysis for the emotional valence rating task included fifty-three participants (aged 21.12 ± 2.43 years; 33 females, 20 males).

2.3.1.3 Procedure

To elicit a more comprehensive and nuanced range of responses, while fostering heightened participant engagement, we utilised a 100-point gradient slider scale. This scale allows participants to smoothly slide a point across a continuum from 1 to 100, with increments of 1 unit. By adopting this numeric response format, we provided a broader spectrum of answers, enhancing precision, dynamism, and interactivity in the data collection process. The results, if consistent with the outcomes of Study 1, will further strengthen the validity of our findings.

The arousal and valence rating tasks were conducted using the Gorilla Experiment Builder (https://gorilla.sc/). The 1500 words were randomly divided into three wordlists, each containing 500 words. Participants were instructed to rate one wordlist at a time and gain access to the following wordlist after a minimum of 24 hours. The third wordlist became available 24 hours after completing the second wordlist. Participants were given a maximum of 120 hours (five days) to complete all the rating tasks.

Each wordlist was presented in five evenly distributed blocks. After completing each block, participants were allowed to take breaks according to their preference. On average, it took approximately 30 minutes to complete each wordlist, including optional breaks.

2.3.2 Data analysis

Inter-rater reliability was assessed using the Intra-Class Correlation (ICC) with a mean-rating, absolute agreement, and two-way random-effects model (ICC 2k, Koo & Li, 2016), implemented in the *psych* package (Revelle, 2023). The model was selected based on three criteria. First, mean rating scores assigned to each item by all raters were used for making inferences. Second, the determination of absolute agreement scores between raters was based on their respective ratings for the same items. Third, the random effect model in ICC was chosen, given that we had the same set of randomly selected raters who rated all items. This decision enabled us to generalise our reliability results to the broader population of Standard Chinese speakers.

Identical steps and fitted models of HLM were employed to replicate the findings of Study 1. Further, we conducted an analysis using GLMM in the *lme4 package* (v1.1-26; Bates et al., 2014) to account for potential variance associated with items and subjects (Baayen et al., 2008). Given the balanced distribution of lexical tonal sequences regarding word frequency, consonant, and part-of-speech categories, we performed a simplified GLMM by focusing solely on testing the effect of lexical tonal sequence.

To account for potential heterogeneity among raters and variations in the relationship between ratings and tonal sequences, we included both by-subject intercepts and by-subject slopes in the model. This

approach effectively controlled for individual differences in baseline levels and individual-specific responses to predictors. However, both arousal and valence models encountered singular fit issues, indicating that one or more random effects did not significantly contribute. As a result, we simplified both models by only keeping the by-subject intercept effect. In addition, by-item intercepts were introduced to the model to accommodate fluctuations in ratings across distinct items due to item-specific characteristics and individual differences among raters. This step also addressed the potential lack of independence between ratings by multiple raters for the same item. In summary, the GLMM consisted of lexical tonal sequence as the fixed effect, and the random effects included by-subject intercepts and by-item intercepts. Pairwise multiple comparisons among tonal sequences were conducted using the *emmeans* package (Lenth, 2023) and Bonferroni correction was applied to adjust the error rate for false-positive decisions in both HLM and GLMM analyses.

The Leave-One-Out Cross-Validation (LOOCV) method was adopted, which calculates statistics based on the retained samples and provides valuable insights into the predictive ability of the models. The entire dataset was divided into ten subsets, and in each iteration, one subset was excluded for analysis, resulting in the examination of nine subsets. In total, we conducted ten iterations to assess the extent to which lexical tonal sequence predicts the emotional arousal and valence ratings of the bisyllabic lexical items in the DCAWS corpus.

2.3.3 Results

2.3.3.1 Inter-rater reliability

There was exceptionally high reliability among all raters included in the analysis for both emotional arousal ratings (*ICC* = 0.836, $p < .001$, *95%CI* = 0.820–0.851) and valence ratings (*ICC* = 0.98, $p < .001$, *95%CI* = 0.983–0.985). The model accounted for systematic differences between raters and random variations in their ratings. The ICC coefficients indicated strong agreement, confirming that the raters provided comparable ratings for the same items.

2.3.3.2 Hierarchical linear regression analysis (HLM)

The aggregated rating scores across subjects were used for the HLM analyses. The lexical tonal sequence explained a significant amount of unique variance in the arousal ratings (ΔR^2 = 1.6%, $p < .01$). Pairwise multiple comparisons showed that the FF tonal sequence (M = 53.24) yielded higher arousal ratings than RR (M = 50.63, $p < .05$) and RL (M = 49.95, $p < .001$). However, no effect of tonal sequences was found for emotional valence ratings.

2.3.3.3 Generalised linear mixed-effects modelling (GLMM)

Individual participants' rating scores were used for the GLMM analyses. Model diagnostics showed no violation of linear regression assumptions. Lexical tonal sequence showed a significant effect on the emotional arousal ratings (F = 33.12, $p < .001$). Pairwise multiple comparisons showed that FF (M = 53.63) had higher arousal ratings than RR (M = 50.74, $p < .01$) and RL (M = 50.00, $p < .001$) tonal sequences. Moreover, HF (M = 52.58, $p < .05$) and HL (M = 52.58, p

< .05) showed higher arousal scores than RL. Again, no significant effect of tonal sequences on emotional valence was observed.

2.3.3.4 Cross-validation: LOOCV (HLM & GLMM)

The results of LOOCV (see details in Appendix, Table A2) supported the adequacy of the dataset size and highlighted specific tonal sequences that had a greater impact than others, as discussed below.

Regarding the HLM analyses, the coefficients of determination across all datasets consistently demonstrated a significant role of lexical tonal sequence in explaining the arousal ratings (ΔR^2 = 1.4% ~ 1.7%, *ps* < .001). Specifically, the FF tonal sequence consistently elicited higher arousal ratings than the RR and RL sequences across all ten datasets.

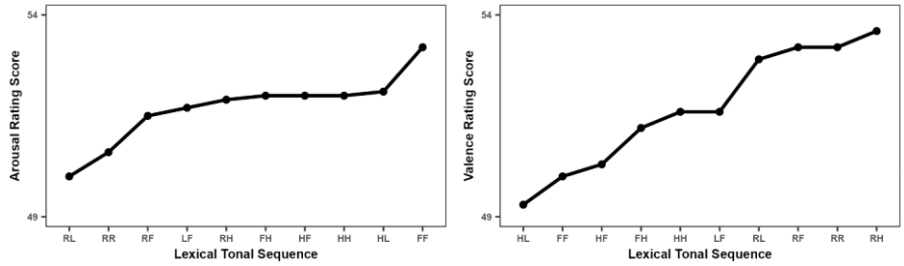
Similarly, in the GLMM analyses, comparisons between null and interest models revealed significant effects of lexical tonal sequence on arousal ratings in all datasets (F = 31.10 ~ 34.35, *ps* < .001). Moreover, FF tonal sequence consistently yielded higher arousal ratings compared to RR and RL. However, the differences in the arousal ratings between the HF/RL and HL/RL pairs were less robust, with a significant effect observed in only eight subsets.

LOOCV revealed no significant prediction of lexical tonal sequence on emotional valence ratings in both HLM and GLMM analyses.

Figure 2.4 visualises the lexical tonal sequence and emotional arousal and valence rating scores in the DCAWS corpus.

Figure 2.4

Emotional arousal and valence rating scores for each lexical tonal sequence in the DCAWS corpus.



2.3.4 Discussion

Study 2 employed an experimental approach to more effectively evaluate the effects of lexical tonal sequences on emotional arousal and valence ratings, by ensuring better control over potential variables (i.e., word frequency, part-of-speech, consonant type, visual complexity, and morphological structure) that could introduce bias in the rating results. Additionally, various methodological strategies were utilised to enhance the reliability and validity of the online rating outcomes, thereby yielding more robust findings concerning the predicted association between lexical tone sequences and emotional arousal and valence ratings.

Comparable results between Study 1 and Study 2 enhance the reliability of our findings. HLM analyses revealed a significant contrast between the FF and RR tonal sequences, with FF consistently predicting higher arousal ratings. The effect sizes (1.4%~1.7%) align with the typical effect sizes observed in iconicity research (e.g., Adelman et al., 2018; Sidhu et al., 2022). Additionally, a noteworthy observation in Study 2 is that the FF tonal sequence predicted higher arousal ratings relative to the RL tonal sequence.

Parallel GLMM analyses showed that bisyllabic words with the FF tonal sequence, characterised by a high average pitch height, wide pitch range and steep slope, consistently elicited higher arousal ratings than the RR and RL tonal sequences, which are realised with a lower average pitch height, narrower pitch range and more gradual slope. Additionally, contrasts such as HF vs. RL and HL vs. RL were identified, although these effects were less stable according to the LOOCV analysis. These distinctions suggest further research is warranted into how average pitch height, pitch range, and slope collectively impact emotional arousal.

In terms of the emotional valence ratings, however, neither HLM nor GLMM analyses showed a significant effect for tonal sequences. Notably, the expected influence of the HH sequence (with a consistently high average pitch height) reported in earlier studies (e.g., Thompson, 2018) seemed to be absent in the general lexicon.

Further investigation is needed to clarify the robustness of tonal influences on emotional arousal and valence ratings. It is particularly important to explore whether the subtle tonal biases observed in affective ratings in Studies 1 & 2 hold in settings where semantic meaning is absent. Despite controls, the inherent meanings of real words might have impacted the ratings by introducing linguistic and contextual cues, potentially contributing to the observed effects. To minimise these influences, Study 3 employed nonce words to better isolate and analyse the role of tonal pitch characteristics in affective iconicity. It is worth emphasising that Studies 1 and 2 relied on text stimuli, offering insights primarily from the characteristic pitch contours of abstract phonological tonal sequences. To complement these findings, Study 3 incorporated spoken nonce words, providing additional evidence based on acoustic data.

2.4 Study 3. Two-alternative forced-choice task using Chinese bisyllabic nonce words

Study 3 utilised nonce words to encourage participants to rely solely on the lexical tone sequences when determining the arousal and valence meanings of each word. These nonce words were presented auditorily to 1) eliminate any potential influence from written characters or associated semantic meanings and 2) further verify the unique contribution of pitch variation linked to lexical tones.

We selected four lexical tonal sequences: HH (high-high), RR (rising-rising), LL (rising-low), and FF (falling-falling). In this way, all four lexical tones were included. Note that in Standard Chinese, LL sequences surface with the first low tone realised with a rising *f0* contour, comparable to the lexical Rising tone (see, e.g., Yuan & Chen, 2014 and references therein). Thus, one may also consider the LL sequence as an RL sequence, given that in this study, participants only heard the auditory signal and were likely to process the sequence as RL. More importantly, the selection of the tonal sequences was based on the significant differences found in Studies 1 and 2 and their contrastive effects on affective ratings that we aimed to verify in this study. Schematically speaking, compared to RR, LL, and FF, the HH sequence exhibits the highest pitch level and average pitch height with little slope, as both syllables maintain the highest average pitch height. The FF sequence shows the greatest pitch range and steepest slope, with both syllables contributing to this dynamic variation. The RR sequence displays an overall upward pitch contour, while the FF sequence presents an overall downward pitch contour, shaped by each syllable's pitch trajectory. Based on their specific pitch characteristics and findings in Studies 1 and 2, we predict that FF will be more likely

to be associated with high arousal option, while HH will be linked to low arousal option compared to the RR and LL (RL) sequences. Furthermore, HH and RR are expected to be more strongly associated with positive valence option than the other two tonal sequences.

2.4.1 Method

2.4.1.1 Stimuli

The auditory stimuli consist of two CVCV nonce words, i.e., /titu/ and /ninu/. The consonants (i.e., /t/, /n/) and vowels (i.e., /i/, /u/) were chosen for their widespread occurrence across languages and their importance noted in iconicity studies (Blasi et al., 2016; Louwerse & Qu, 2017; Styles & Gawne, 2017). The two CVCV nonce words were uttered with four lexical tonal sequences, i.e., HH, RR, RL, and FF. Moreover, we added *baozao* (/pɑʊ//tsɑʊ/T4T4, meaning *irritable*) and *xingfen* (/ɕɪŋ//fən/T1T4, meaning *excited*) as two validation items. Previous studies have consistently shown that *baozao* is associated with high arousal and negative valence, while *xingfen* is linked to high arousal and positive valence (Y. Wang et al., 2008; X. Xu et al., 2022). These items were used to assess participants' engagement in the online tasks, and any incorrect ratings served as an objective measure for excluding participants.

The stimuli were recorded using a Sennheiser MKH416T microphone with a sampling rate of 44.1 kHz and 16-bit resolution at Leiden University's Phonetics Lab. All stimuli were produced by a male native speaker of Standard Chinese from Beijing, who was not familiar with the objectives of the study. He was instructed to read the stimuli one word at a time in a consistent way. Each word can thus be taken as an utterance with a statement intonation without any additional, e.g.,

emotional or attitudinal, implications, and may therefore be considered to have neutral intonation. The stimulus list was read three times, each time in a randomised order. The complete recording was then further processed in Praat to select the best tokens and edited to produce individual sound files, each containing a single token of the stimuli. Token selection was based on perceptual clarity, as assessed by the first author and another native speaker consulted. All stimuli were normalised to an average intensity of 70 dB SPL. Figure 2.5 presents the average time-normalised *f0* contours of the tokens, calculated over 20 equidistant points across the entire word. Additional details on the pitch measurements for each token, including average pitch height, pitch range, and slope (i.e., pitch range/duration) are provided in Table 2.1.

Figure 2.5

Pitch contours of auditory stimuli used in Study 3. The tonal sequences HH, RR, LL, and FF correspond to T1T1 (high-high), T2T2 (rising-rising), T3T3 (rising-low), and T4T4 (falling-falling). The "LL" sequence includes tone sandhi, producing a rising-low pitch contour.

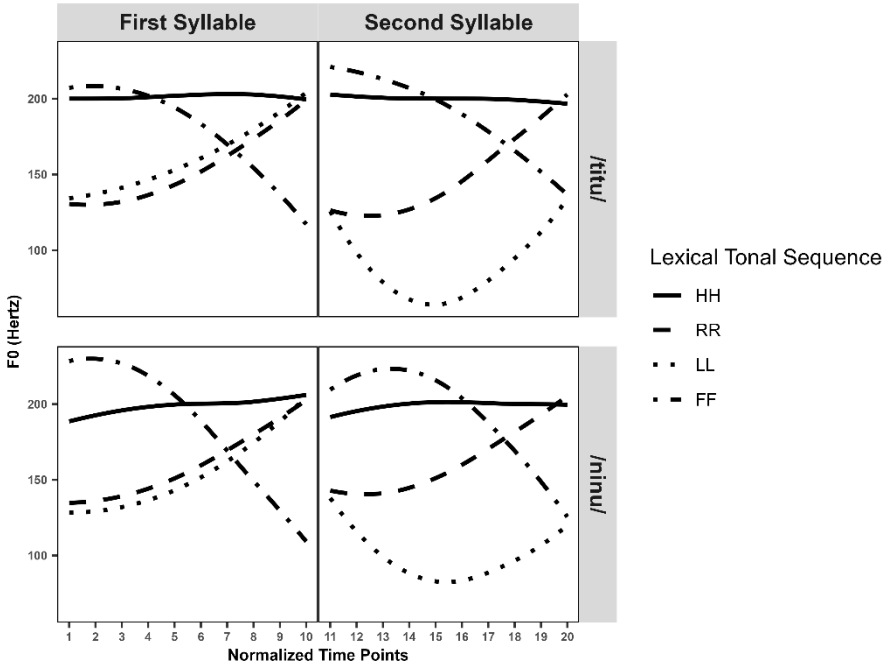


Table 2.1

Pitch measurements of tokens used in Study 3.

Token	Tonal sequence (Syllable position)	Duration (*ms*)	Slope	Average pitch height	Pitch range
/titu/	HH (First)	412	0.03	201	12
/titu/	HH (Second)	418	0.03	200	13
/titu/	RR (First)	356	0.20	154	72
/titu/	RR (Second)	429	0.18	151	75
/titu/	RL (First)	342	0.20	162	69
/titu/	RL (Second)	508	0.15	93	76
/titu/	FF (First)	262	0.33	178	88
/titu/	FF (Second)	224	0.37	188	84
/ninu/	HH (First)	486	0.05	198	25
/ninu/	HH (Second)	415	0.03	199	12
/ninu/	RR (First)	386	0.17	161	64
/ninu/	RR (Second)	476	0.14	163	67
/ninu/	RL (First)	376	0.20	155	74
/ninu/	RL (Second)	508	0.13	103	67
/ninu/	FF (First)	320	0.38	185	121
/ninu/	FF (Second)	350	0.25	192	87

The visual stimuli comprise two pairs of emojis, each representing faces with differing levels of arousal (high vs. low) and valence

(negative vs. positive), chosen from the EmojiGrid (Toet & van Erp, 2019). The emojis are shown in Figure 2.6.

Figure 2.6

Emojis used in the 2AFC tasks in Study 3. Positive and negative were used for valence task, while high-arousal and low-arousal were used for arousal task.

2.4.1.2 Participants

A total of 179 college students, all native Standard Chinese speakers without hearing, visual, speech, or alexithymia disorders, participated in the study. All participants reported using Standard Chinese for over 80% of their daily communication.

The final analyses included participants who accurately rated the probe items. This comprised 121 participants for the arousal task (aged 21.53 ± 2.11 years; 72 females) and 135 participants for the valence task (aged 21.52 ± 2.13 years; 84 females).

2.4.1.3 Procedure

The experiment was conducted via the Gorilla Experiment Builder (https://gorilla.sc/). Participants were introduced to the concepts of the two emotional dimensions (arousal and valence), along with corresponding emojis. They then completed a quiz, matching emojis to emotional dimensions in real-life scenarios, with feedback provided. Only participants who correctly matched emojis for all scenarios proceeded to the following main task.

In the main task, participants completed two 2AFC tasks to rate arousal and valence, with the task order counterbalanced: 93 participants began with the valence task, and 86 with the arousal task. Within each 2AFC task, auditory stimuli were presented in pseudo-random order, alternating between /titu/ and /ninu/ segments, with identical tonal sequences spaced to avoid consecutive repetitions.

In each trial, a stimulus was played automatically, with an option to replay it up to five times. Simultaneously, two emojis (representing either valence or arousal) were displayed at the same horizontal level on the screen. Participants were instructed to select the emoji that best represented their perception of the sound's arousal or valence.

2.4.2 Data analysis

Two sets of factorial GLMM were fitted using the *lme4* package (Bates et al., 2014) in R to examine the impact of the lexical tonal sequences on the emotional arousal and valence choices, respectively. Based on the binary outcomes, the GLMMs applied logistic regression with a binomial probability distribution and the logit link function. Similarly to the GLMMs in Study 2, we initially included by-subject intercepts and slopes as random effects, but both models encountered convergence issues. Additionally, the valence model faced a singular fit issue with the by-subject intercept as a random effect. Therefore, the final arousal model included a fixed effect for lexical tonal sequence and a by-subject intercept, while the final valence model included only the fixed effect for lexical tonal sequence. Pairwise multiple comparisons were conducted using the *emmeans* package (Lenth, 2023), with Bonferroni correction applied to control type I error.

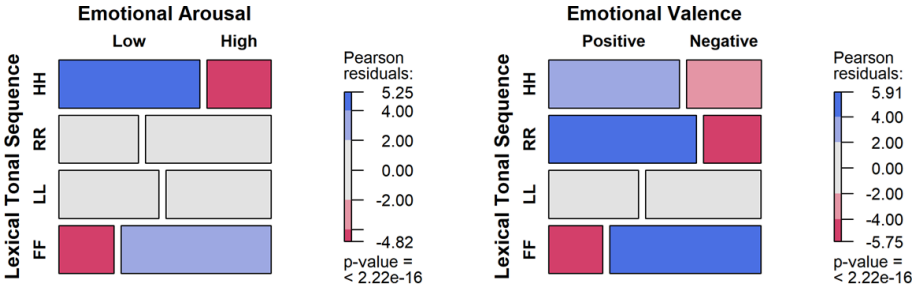
We utilised the Odds Ratio as the main measurement for discussing the outcomes. Odds represent the ratio of the probability of

an event (e.g., high arousal) occurring to the probability of it not occurring (e.g., low arousal). An Odds Ratio (hereafter OR) is a specific type of comparison between two categories, indicating how many times an event (e.g., high arousal) is more likely to occur in one category compared to another (e.g., high arousal in FF compared to RR).

To visually represent the frequency of arousal or valence choices as a function of the lexical tonal sequences, we utilised mosaic plots generated by the *mosaic()* function from the *vcd* package (Friendly & Meyer, 2015). Figure 2.7 displays the observed frequencies of emotional arousal (left panel) and valence (right panel) for each tonal sequence. Tiles were colour-coded and shaded based on standardised residuals, highlighting deviations from the expected frequencies under the assumption of independence (e.g., equal probabilities for high and low arousal in each tonal sequence). Positive residuals (blue-shaded) indicated higher-than-expected frequencies, whereas negative residuals (red-shaded) indicated lower-than-expected frequencies. Unshaded tiles indicate no statistically significant association for a tonal sequence and emotional dimension based on Pearson residuals. Note that Pearson residuals show the strength and direction of associations between tonal sequences and emotional dimensions but do not directly correspond to logistic regression models. This visualisation complements the fitted GLMMs, offering an alternative view by indicating significant deviations from the expected frequency.

Figure 2.7

Arousal and valence frequency variations across lexical tonal sequences. The tonal sequences HH, RR, LL, and FF represent T1T1 (high-high), T2T2 (rising-rising), T3T3 (rising-low), and T4T4 (falling-falling).



2.4.3 Results

2.4.3.1 Arousal

A null model and an alternative model were fitted for the arousal ratings. The comparison of likelihood ratio tests between the two models revealed that lexical tonal sequences predicted the emotional arousal choices of the CVCV nonce words (χ^2 = 95.22, p < .001, R^2 = 7.13%) after accounting for the individual differences (variance = 0.12 ± 0.34). Pairwise multiple comparisons indicated that FF, RR, and RL tonal sequences were 6.25, 3.57, and 2.33 times (OR) more likely to be rated as high arousal than HH, respectively. Additionally, FF is more likely to be rated as high arousal than RL and RR (OR = 2.66 and 1.75, respectively).

2.4.3.2 Valence

Similarly, a null model and an alternative model were fitted for the valence ratings, and a comparison between the two models was conducted. Consequently, the comparison indicated that lexical tonal

sequences significantly predicted the emotional valence judgments of the CVCV nonce words (χ^2 = 140.54, p < .001, R^2 = 9.39%). Random effects were omitted due to a singular fit issue, suggesting that individual differences had minimal influence on valence choices. Pairwise multiple comparisons revealed that FF was more likely to predict negative valence than HH, RL, and RR (OR = 4.92, 2.18, and 7.15, respectively). Additionally, RL was more likely to predict negativity than RR and HH (OR = 3.29 and 2.27, respectively).

2.4.4 Discussion

Study 3 explored the impact of lexical tonal sequences on emotional arousal and valence choices in CVCV nonce words presented auditorily to participants in 2AFC tasks. The findings from Study 3 reaffirmed the hypothesis that pitch characteristics of lexical tones significantly influence emotional arousal and valence responses, with a stronger effect when semantic content is absent. The FF tonal sequence, with extensive pitch variation in both range and slope, was consistently rated as conveying high arousal. In contrast, the HH sequence, despite its high pitch level and average pitch height, was associated with low arousal, likely due to its minimal pitch variation. This suggests that dynamic variations in both pitch range and slope have a greater impact on conveying emotional arousal than high average pitch height (and consistently high pitch level).

Interestingly, while Studies 1 and 2 did not identify an effect of lexical tonal sequences on emotional valence, Study 3 revealed a significant influence. The FF and RL tonal sequences were more likely to be associated with negative valence, whereas the RR and HH sequences tended toward positive valence, emphasising the importance of both pitch level and pitch variation in influencing

emotional valence choices. Additionally, the contrast between FF and RR suggests that overall pitch contour direction may significantly contribute to valence perception. This difference aligns with the "Good is Up, Bad is Down" conceptual metaphor, where upward-pitch contours are associated with positive emotions and downward-pitch contours with negative emotions (Yap et al., 2014). Supporting this association, previous research has shown that rising pitch (compared to falling pitch) enhances a positive emotional state, making people feel happier, with evidence from pure tones and musical sounds (Ley-Flores et al., 2022).

Overall, Study 3 not only replicated the association between lexical tonal sequences and emotional arousal ratings observed in our Studies 1 and 2 but also revealed a clear effect of tonal pitch variation on emotional valence iconicity. The valence effect, which may be obscured by semantic meanings in the general lexicon, became evident in the nonsense auditory context. Additionally, the 2AFC results highlighted the distinct propensity of FF and HH tonal sequences on emotional arousal and of FF, HH, and RR tonal sequences on emotional valence.

2.5 General discussion

With three studies, we investigated the potential pitch iconic effects of lexical tones on affective interpretations in Standard Chinese. Our findings present convincing evidence that affective iconicity in lexical tone systematically permeates both the general lexicon and nonce words of Standard Chinese. While lexical meaning unquestionably impacts these emotional ratings and choices, the pitch characteristics of lexical tonal sequences over bisyllabic linguistic items make a subtle, but systematic and significant contribution to emotional arousal and valence perception. This suggests that even in a lexical tone language

where pitch variation is used phonemically for lexical distinctions, the pitch characteristics of tones retain an iconic emotional significance.

Consistent with our hypotheses, the FF tonal sequence, marked by extensive pitch variation in range and slope, was consistently associated with higher arousal than RR, across both written words (Studies 1 and 2) and spoken nonce words (Study 3). This finding aligns with previous research on pitch and emotional prosody (Bänziger & Scherer, 2005; Laukka et al., 2005; Scherer et al., 2003), which suggests that high-arousing emotions are typically associated with a higher pitch (level and/or height), wider pitch range, and steeper pitch slope. These pitch characteristics, even when embedded in phonemic tones, appear to iconically influence emotional perception. Additionally, the HH tonal sequence, with the least pitch variation, was significantly associated with low arousal, though only in spoken nonce words (Study 3). This pattern may relate to the "frequency code" hypothesis (Gussenhoven, 2016; Ohala, 1984), which posits that a higher *fo* conveys meanings related to smallness, submissiveness, and politeness. In this context, we speculate that the HH tonal sequence may evoke the perception of low emotional intensity, contributing to a sense of stability and calmness that aligns with submissive or polite interactions.

The associations of FF and RL tonal sequences with negative valence and RR and HH sequences with positive valence align with previous findings on the relationship between pitch characteristics and emotional prosody (Belyk & Brown, 2014; Kamiloğlu et al., 2020) as well as valence iconicity (Yap et al., 2014). These studies have demonstrated that high pitch (levels) and rising contours are more likely to convey positive valence, while falling contours and lower pitch (levels) are more likely to convey negative valence. Given the

complexity of the pitch-valence relationship, our findings highlight additional influential factors in valence perception, namely, semantic meaning.

These findings point to distinct roles for tone–arousal and tone–valence iconicity within the context of lexical tones, with interesting implications for the relationship between lexical tone and iconicity. The observation that both tone–arousal and tone–valence associations occur in nonce words supports the notion that iconicity is a pre-semantic phenomenon at the acoustic level (rather than relying on semantic content), which is possibly a foundational mechanism in early human communication systems (Sučević et al., 2015; Westbury, 2005). In this context, the tone's acoustic properties, such as pitch level, average pitch height, pitch range, slope, and contour, likely play a direct role in shaping emotional responses.

Moreover, the presence of tone–arousal associations in both general lexicon and nonce words, contrasted with tone–valence association observed only in nonce words, implies that while both associations may be pre-semantic, arousal is more pronounced than valence for iconicity at the lexical tone level (see their different effect sizes in Appendix A, Table A3). This aligns with findings by Aryani et al. (2018), which demonstrated that phonemic segments more strongly influence arousal than valence. This distinction may arise from their unique affective attributes: arousal, an ancient and innate emotional dimension, is tied to physiological responses and reflected in salient acoustic features like sharp pitch rises or falls. In contrast, valence is more abstract and susceptible to cognitive influences such as semantic meanings and cultural interpretation (Darwin, 1998; Russell, 2003). Meanwhile, the consistent findings of tone–arousal associations across both silent reading and active listening contexts reinforce that the

acoustic profiles of words provide implicit affective cues, which language users may use in constructing emotional meaning (Aryani et al., 2018).

It is noteworthy that the development of lexical tones in languages like Standard Chinese is believed to have occurred after the establishment of segmental features (see review in Michaud & Sands, 2020). Following this view, lexical tones evolved to distinguish lexical meanings due to sound changes in the segmental syllables instead of serving to express emotions. However, our findings suggest that this does not preclude the possibility that 1) emotional prosody exerted an influence on the formation of lexical tone sequences in certain linguistic expressions; 2) the pitch characteristics of lexical tones retain their iconicity effect for affective expression beyond lexical meanings. Cross-linguistic studies can explore further whether, and if so, to what extent, the affective iconicity of lexical tonal sequences, rooted in pitch variations, holds true universally.

Compared to previous studies (Adelman et al., 2018; Aryani et al., 2018; Thompson, 2018; Yao et al., 2013; Yap et al., 2014), our research extended the investigation of affective iconicity from phoneme to toneme, from one dimension (valence) to two dimensions (arousal and valence), from real words to nonce words, from nontonal to tonal languages. To capture the intricacies of human emotions, we adopted the emotional dimension theory, which transcends basic emotional categories (Posner et al., 2005; Russell, 1980). To ensure precision and reliability in emotional rating data, we developed a gradient scale for emotional meaning ratings, minimising measurement variability and capturing finer nuances of affective variations. Additionally, our study employed a rigorous methodology, combining hierarchical linear models, generalised linear mixed models, and sound analytical

techniques (i.e., inter-rater reliability and cross-validation procedures) to enhance the validity and robustness of our findings. Last but not least, the use of the 2AFC task extends findings from written words to auditory sounds and reveals the nuanced interplay between lexical tones and affective meanings in multiple modalities and contexts.

It is important to acknowledge the limitations of our research. For example, the size of our datasets could be increased in future studies. Although our study involved a larger pool of raters than previous attempts, a larger dataset (or nonce tokens) would provide a more comprehensive overview. Additionally, despite our preliminary explorations of how pitch characteristics influence affective iconicity, more research designs are needed to elucidate their interplay and significance in biasing the arousal and valence ratings, respectively. Whereas our current investigations have controlled phonemes to isolate the effects of lexical tone on arousal and valence ratings, future studies are encouraged to explore further the interplay between phonemes and lexical tone in predicting arousal and valence ratings or choices. Finally, the progression from Studies 1 and 2 to Study 3, along with the converging results, suggests that the affective iconicity of lexical tones is not linked to the phonological tonal categories but is instead acoustically associated with the tonal pitch characteristics. However, further large-scale investigations are needed to confirm and consolidate these findings.

To summarise, our findings contribute to a growing body of evidence that challenges the conventional view of arbitrariness in the relationship between word meaning and form in language (Hinton et al., 2006; Monaghan et al., 2014; Nuckolls, 1999; Ohala, 1997). The novel evidence of affective iconicity in the general lexicon and nonce words of Standard Chinese indicates that language may not be as

arbitrary as previously believed and that sound can convey meaningful information beyond its primary function. Further research is needed to fully comprehend the underlying mechanisms and generalisability of the lexical tone effects on affective iconicity. This research has unveiled new avenues for investigating the role of lexical tonal sequences in understanding the relationship between pitch iconicity, emotional experiences, and, more broadly, the interplay of sound and meaning in languages with typologically diverse lexical tone systems.