# Versatility of phonemic pitch

## In affective iconicity and perceptual reorganisation

Published by
LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461
e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration: Concept by Tingting Zheng, painting by M.Y. Priscilla Lam, and editing by M.Y. Priscilla Lam and Tingting Zheng.

ISBN: 978-94-6093-487-2
DOI: https://dx.medra.org/10.48273/LOT0702
NUR: 616

# Versatility of phonemic pitch

## In affective iconicity and perceptual reorganisation

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op woensdag 19 november 2025

klokke 10.00 uur

door

**Tingting Zheng**

geboren te Shanxi, China

in 1992

**Promotores**: Prof. dr. Claartje Levelt
Prof. dr. Yiya Chen

**Promotiecommissie**: Dr. Victoria Nyst
Prof. dr. Nancy Kula
Prof. dr. Asli Özyürek
(Max Planck Institute for Psycholinguistics)
Dr. Caicai Zhang
(The Hong Kong Polytechnic University)

To my family
Aan mijn familie
致我挚爱的家人

Two roads diverged in a wood, and I,
I took the one less travelled by,
And that has made all the difference.
— Robert Frost, *"The Road Not Taken"* (1915)

# Table of contents

# Acknowledgements

The completion of this dissertation has been a journey full of twists and turns, made possible by the incredible people who walked alongside me.

First and foremost, my deepest gratitude goes to my doctoral promotors, Prof. dr. Claartje Levelt and Prof. dr. Yiya Chen. Your patience, encouragement, and unwavering belief in me have not only shaped my academic growth but also taught me resilience, for which I am eternally grateful. Heartfelt thanks to my master supervisor, Prof. dr. Xuhai Chen, for your early guidance and support, which laid the groundwork for everything that followed.

Being part of such a vibrant and supportive group, including Min Liu, Menghui Shi, Qing Yang, Xinyi Wen, Hana Nurul Hasanah, Matthew Sung, Yiran Ding, Xin Li, Priscilla Lam, Charlie Xu, Katrina Li, Qi Zhang, and Yingyang Wang, was a privilege. To my PhD fellows, Zhen Li, Ami Okabe, Andrew Wigman, Cesko Voeten, Rasmus Puggaard-Rode, Sarah von Grebmer zu Wolfsthurn, Zhuoyi Luo, Zhaole Yang, Ruixue Wu, Jin Wang, Natasja Delbar, Bente De Graeve and many others, thank you for making this journey less lonely and far more fun. Special thanks to Dan Yuan for your research insights and support, and to Min, Han Hu, and Menghui for helping me settle into life in Leiden. I owe so much to my dear friends and colleagues, Qing, Hang, and Jiang. Thank you for being my steady ground and safe space, for seeing me, caring for me, and standing by me through it all.

A huge thank you to those who helped me navigate the technical challenges of this research. Jos Pacilly, for your patience with my Praat scripts; Anne van der Kant, for guiding me through the NIRx system; Szilvia Biro, for your timely fNIRS advice; Guilia Vigna, Heleen de Lange, Ngaira Salomons, Saskia Galindo Jong, and Andreas P. Wölfle,

for assisting with experiments; and Isabella Fritz, for generously sharing your fNIRS study materials. To Andreea Geambaşu, for guiding me through the Methods course teaching.

My gratitude also extends to my Japanese colleagues: Prof. dr. Reiko Mazuka, for your invaluable mentorship and hospitality; Naoto Yamane, for training me on HITACHI fNIRS; and Rachel Ka-Ying Tsui and Akiko Okuno, for your support and lovely companionship. Many thanks to other significant lab members for making my time in Japan special because of your kindness and support.

To my officemate, Dr. Evelyn Bosma, thank you for the shared silences and chit-chats. Camilla, you've made life more beautiful in countless ways. Patty, Niels, and Maxi, your kindness, warmth, and boundless joy made Leiden truly feel like home.

My deepest thanks to all the babies, families, and adult participants in my study; your contributions made this research possible. Special appreciation to my colleagues Jurgen Lingen and Jesse Wichers Schreur, who participated with their little ones, and to Aliza Glasbergen-Plas and Ami Okabe, for lending me your voices. To Natasja Delbar, for reviewing and proofreading my Samenvatting.

To the reviewers of my doctoral work, thank you for your time, insights, and for challenging me to think deeper. This dissertation was funded by the China Scholarship Council, with additional support from the Leiden University Centre for Linguistics and the NWO-VICI project "Melody in Speech."

Last but not least, to my family, especially my parents and brother, your unconditional love and support have been my foundation. This dissertation is as much yours as it is mine. Deep appreciation to my friends Min Xu, Xi Liu, and Mengge Tan for always being there for me, and to Shallway, for walking beside me and lighting up the grey.

# Chapter 1
# General introduction

When we listen to speech, our ultimate aim is to extract meaning from the acoustic input. Many studies on speech perception have investigated how listeners differentiate phonemes, the fundamental building blocks of speech, from one another (Mitterer & Cutler, 2006). This process relies on a complex interplay between bottom-up (e.g., sensory input) and top-down (e.g., prior knowledge, context, or cognitive expectation) information (Massaro, 2015).

Among many acoustic cues in speech, fundamental frequency (*f0*), perceived as pitch, stands out due to its universal and multiple functions in communication (Gussenhoven, 2004). Across almost all spoken languages, pitch is used both grammatically and paralinguistically, e.g., to mark distinctions between questions and statements or to convey emotions. Specifically, rising *f0* contour is often used for questions, while falling *f0* contour is often used for statements (see reviews in Cruttenden, 1997; Ohala, 1983, 1994). High-arousing emotions generally correlate with a higher overall pitch level (i.e., a more stable contour at a higher frequency), 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). While some report a correlation between lower average pitch height and positive valence (Scherer & Oshinsky, 1977), the more prevailing view suggests a trend towards a higher pitch level, higher average pitch height, and wider pitch range (Belyk & Brown, 2014; Kamiloğlu et al., 2020; Laukka et al., 2005).

In tone languages, however, pitch serves an additional function: it distinguishes lexical meanings (Yip, 2002). Standard Chinese (SC) is a prototypical tone language which features 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). Some languages,

such as Japanese, are not classified as tone languages but still use pitch in a lexically contrastive way. Japanese employs a lexical pitch-accent system, where pitch distinctions are confined to accented syllables within certain words (Kubozono, 2012; Pike, 1948). As Ota et al. (2018) describe, in Tokyo Japanese, words are categorised as either accented, characterised by a phrase-initial rise (realised as %LH), or unaccented, marked by a lexical falling pitch (realised as H*L).

Pitch, despite its diverse functions in discourse, emotional expression, and lexical meaning, demonstrates remarkable versatility in other contexts. For example, it can convey information through pitch form-meaning iconicity, shaped by its perceptual properties and shared linguistic or social conventions among speakers. In addition, distinct patterns of hemispheric lateralisation across languages are related to the perceptual properties of pitch, and its linguistic function. Building on these specific aspects, this dissertation investigates the affective iconicity of linguistic pitch in lexical tones among adult Standard Chinese speakers and the developmental hemispheric lateralisation of lexical-level pitch processing during perceptual reorganisation in Dutch- and Japanese-learning infants. In both cases, pitch perception results from the interplay of bottom-up acoustic information and top-down cognitive, contextual, and experiential influences. By examining these interactions, this work sheds light on the multifaceted role of pitch in speech perception across different contexts, languages, and developmental stages. The following sections of this dissertation will discuss the roles of pitch in affective iconicity and perceptual reorganisation.

## 1.1 Iconicity

Although lexical items have conventionally been characterised by the seemingly arbitrary association of sounds with meanings (Hockett, 1958), numerous studies have enlightened a significant sound-meaning resemblance, known as iconicity or sound symbolism. Traditionally, iconicity is used when a specific phonetic feature or sound unit, such as phoneme or syllable, is said to go beyond its linguistic function as a contrastive, nonmeaning-bearing unit, to directly express particular meanings (e.g., Ćwiek et al., 2021; Nuckolls, 1999).

Notably, the concept of iconicity is distinguished from systematicity, which refers to a statistically consistent relationship (i.e., statistical regularities) between arbitrary, potentially language-specific patterns of sounds and groups of words that share those patterns (Dingemanse et al., 2015; Thompson & Do, 2019). As argued by Dingemanse et al. (2015), systematicity is presumed to be language-specific due to phonological regularities that are unique to each language. For instance, the onset cluster *gl-* often appears in English words related to light, such as *glimmer*, *glitter*, *glow*, and *glisten* (Sidhu, 2025). While all phonemes that exhibit such patterns are examples of systematicity, not all of them are iconic (Dingemanse et al., 2015; Sidhu, 2025). Iconicity, in contrast, is fundamentally subjective and relies on interpretative processes (Winter et al., 2023). For example, vowels can be iconically associated with multiple meanings across contexts such as size (Blasi et al., 2016), shape (Godoy & Ananias, 2022), colour (Mok et al., 2019), and gender of the entity referred to by the label containing the specific vowel (Wong & Kang, 2019).

Based on these distinctions, Winter et al. (2023) redefined iconicity as follows:

> *A signal in any medium or modality, such as a word, sign, or gesture, is iconic to the extent that language users produce or perceive it through a sense of resemblance between some aspect of its form and some aspect of its meaning.*

In this dissertation, we adopt this definition but focus specifically on how iconicity is perceived. Here, we define iconicity as: A signal in any modality is iconic to the extent that language users perceive it through a sense of resemblance between some aspect of its form and some aspect of its meaning.

Iconicity is a pervasive phenomenon (Adelman et al., 2018; Aryani et al., 2018; Perniss et al., 2010; Winter et al., 2023; Yap et al., 2014) showing universal tendencies (e.g., Auracher et al., 2011; Y.-C. Chen et al., 2016; Huang et al., 1969; Nikroshkina, 2019) while also exhibiting cultural and cross-linguistic variation (e.g., Winter et al., 2021). A widely studied example is the vowel-shape correspondence, where the vowel /i/ is often associated with sharp shapes, and /u/ with round shapes (Köhler, 1929; Ramachandran & Hubbard, 2001). This phenomenon, known as the bouba–kiki effect, has been consistently observed across diverse languages and writing systems (e.g., German, French, Spanish, Polish, Japanese, Korean, Georgian, Chinese, Zulu, Thai, Turkish, Finnish, etc; Ćwiek et al., 2021), and populations, including remote groups like the Himba of Northern Namibia (Bremner et al., 2013), as well as across developmental stages (e.g., toddlers and adults; Maurer et al., 2006). The effect extends to implicit processing, where congruency between spiky shapes and angular

letters leads to faster response times (De Carolis et al., 2018). Moreover, the activation of the prefrontal cortex during implicit association tasks suggests that bouba–kiki correspondences may involve modulation of executive processes (Peiffer-Smadja & Cohen, 2019). This implies that bouba–kiki effects may be subject to top-down modulation, with attention and cognitive control playing a role in resolving perceptual conflicts between mismatching stimuli.

Despite its universality, iconicity is influenced by cultural and perceptual experiences. For instance, cultural differences influence shape–sound associations (e.g., Taiwan vs. US participants; Y.-C. Chen et al., 2016), while phonemes (Fort et al., 2015), acoustic cues (Fort & Schwartz, 2022; Knoeferle et al., 2017), phonetic features (D'Onofrio, 2014), and orthography and phonology (Cuskley et al., 2017) further refine these associations. Additionally, the bouba–kiki effect extends to other domains: vowels in names can influence perceptions of facial roundness (Barton & Halberstadt, 2018), and tactile ratings show that /u/ is linked to positive sensations of the materials (e.g., comfort, smoothness, and warmth), whereas /i/ is associated with negative ones (Sakamoto & Watanabe, 2018). Furthermore, the vowel-shape association persists in production tasks, with participants consistently naming shapes using phonemes that align with perceptual findings. For example, spiky shapes were more frequently named with high/front vowels and voiceless stops, whereas round shapes were more often named with back/rounded vowels and lateral or nasal consonants (Godoy & Ananias, 2022).

Among the world's languages, Standard Chinese (SC) stands out as particularly intriguing in the study of iconicity. SC is a tone language where characters correspond to syllables, each carrying distinct phonetic and semantic information. Some studies have explored the

character-iconicity in SC, uncovering links between the visual structure of characters and their conceptual or semantic properties (e.g., Koriat & Levy, 1979; F. Li, 2017; V. P. H. Li, 1986; Luk & Bialystok, 2005; Pan & Schmitt, 1996; Xiao & Treiman, 2012), revealing that a small degree of iconicity has been retained in simple characters.

A more intriguing question arises: can the tonal dimension of SC, specifically its lexical tones, exhibit iconicity? Lexical tones in SC involve pitch variations that distinguish word meanings, and their potential for iconic representation may align with the broader phenomenon of pitch iconicity. This question is explored in detail in the following section.

## 1.2 Iconicity of pitch

Ohala (1984, 1997) is among the first to illustrate pitch iconicity in lexical tones. He demonstrated, based on observations in three tonal languages: Ewe, Yoruba, and Cantonese, that high tones are often associated with connotations of smallness, diminutiveness, and familiarity, while low tones are linked to largeness. Subsequent studies have confirmed the iconicity of lexical tones in Chinese and Cantonese across a range of dimensions, including size, shape, gender, and the personality traits in storybook character naming (Chang et al., 2021; Lapolla, 1995; Shang & Styles, 2017; X. Wang, 2021; Wong & Kang, 2019). For example, in SC, the T4 and vowel /i/ are associated with pointy shapes, while the T1, T2 and vowels /a/ and /u/ are associated with round shapes (Chang et al., 2021; Shang & Styles, 2017). Moreover, Thompson (2018) reports a distributional bias for high-level tones, although only in onomatopoeic words, in Mandarin, Taiwanese Southern Min, and Hong Kong Cantonese. Collectively, these studies reveal a consistent pattern of pitch iconicity, with the pitch

characteristics of lexical tones, particularly pitch height, being associated with diverse sensory and semantic dimensions.

An important point is that the development of lexical tones in languages like Standard Chinese is believed to have occurred after the establishment of segmental features (see the review by Michaud & Sands, 2020 and references therein). For example, it is widely acknowledged that lexical tones are derived from the consonants at both ends of a tone-bearing syllable, with prevocalic consonants determining the pitch height and postvocalic consonants determining the pitch contour (e.g., Haudricourt, 1954; Pittayaporn & Kirby, 2017). Thus, research on the iconicity of lexical tones is typically framed in terms of pitch iconicity, since lexical tones are primarily distinguished by pitch variations.

This leads to the broader concept of pitch iconicity, which has been extensively studied and is frequently discussed in relation to the "frequency code" hypothesis (Gussenhoven, 2016; Ohala, 1984). According to this theory, higher and/or rising *fo* signifies smallness, or is associated with socio-cognitive interpretations such as submission and politeness, while lower and/or falling *fo* conveys largeness, dominance or aggression. These cross-modal associations appear robust across languages, including English (Perlman, Clark, et al., 2015), Polish (Rojczyk, 2011), and Dutch (Stel et al., 2012). A comprehensive list of studies on pitch iconicity is presented in the Appendix A, Table A1.

Research on non-linguistic pitch has revealed consistent associations between perceived pitch height and various sensory and perceptual attributes, too. Higher-pitched sounds are linked to small size, sharp shapes, lighter or brighter stimuli, and upward movement, while lower-pitched sounds are associated with larger size, rounded

shapes, darkness, heaviness, and downward movement (see reviews in Deroy & Spence, 2013; Spence, 2011).

The resemblance between form and meaning is not confined to the perception of sensory dimensions (e.g., shape and size), however, it also extends to the perception of basic affective meanings, i.e., emotions. Exploring the relationship between sound and emotion is worthwhile, as emotion helps individuals process emotionally relevant information, guide their actions, and adapt to changing circumstances (Kashima et al., 2020).

Building on the concept of iconicity, we define the relation between form and affective or emotional meaning as affective iconicity: A signal in any modality is iconic to the extent that language users perceive it through a sense of resemblance between pitch and affective meanings. This term describes emotional experiences that language users construct through speech forms (Aryani et al., 2019). A more detailed review of affective iconicity follows in the next section.

### 1.2.1 Affective iconicity of pitch

Few studies have looked into the relationship between lexical tones and affective or emotional meanings. Yao et al. (2013) analysed the lexical tones and emotional meanings of approximately 100 bisyllabic Chinese words and identified a direct link between tonal sequences and specific types of emotion. They observed a distribution bias, with falling tones more common in words denoting anger and joy, high-level tones in words denoting sadness, and rising and low-dipping tones in words denoting fear. Similarly, Yap et al. (2014) proposed that the direction of tonal pitch contours in Chinese words correlates with the emotional valence of their translation equivalents in English, with rising contours

perceived as the most positive and falling contours as the most negative.

While tones may relate to affective meaning, it remains unclear whether lexical tones, with their pitch variations, exhibit affective iconicity in tonal languages like Standard Chinese (SC). This question is important for two reasons. First, as previously reviewed, the relationship between pitch and emotional arousal and valence is well-established. High-arousing emotions are correlated with a higher pitch level, higher average pitch height, wider pitch range, and steeper pitch slope compared to low-arousing emotions (Bänziger & Scherer, 2005; Laukka et al., 2005; Scherer et al., 2003). Similarly, wider pitch ranges and higher average pitch height are associated with positive valence (Belyk & Brown, 2014; Kamiloğlu et al., 2020; Laukka et al., 2005).

Second, tonemes in SC are as crucial as phonemes in expressing lexical meanings. Previous research has shown that phonemes can be affectively iconic across languages, suggesting they convey emotional content through their acoustic properties (Schmidtke et al., 2014). This phenomenon, known as affective iconicity, underscores the essential role of emotion in communication (Adelman et al., 2018; Aryani et al., 2018, 2019; Auracher et al., 2011; Darwin, 1998; Davidson et al., 2009; Nielsen & Dingemanse, 2021; Whissell, 2003). Segmental components, such as consonants and vowels, have been shown to predict arousal (German in Aryani et al., 2018; German in Auracher et al., 2020; Japanese in Kambara & Umemura, 2021) and valence (German, Polish, Dutch, Spanish, and English in Adelman et al., 2018; Japanese and German in Körner & Rummer, 2023; Dutch and Chinese in Louwerse & Qu, 2017; Chinese and English in Yu et al., 2021). What remains unexplored is the affective iconicity of lexical tone.

To explore this question, we adopted a dimensional framework for describing emotions, as it captures the complexity of emotional experiences more effectively than the discrete emotion framework (Adelman et al., 2018; Aryani et al., 2018; Bänziger & Scherer, 2005; Kamiloğlu et al., 2020; Laukka et al., 2005). The discrete emotion framework conceptualises emotions as distinct categories, such as anger, happiness, sadness, fear, and surprise. Each category is thought to represent a universal, biologically hardwired response that is consistent across cultures (Ekman, 1992). While this model has been influential, it has been challenged by recent studies because of its unvalidated neural foundations (Posner et al., 2005).

In contrast, the dimensional approach describes emotions along continuous axes, most commonly arousal and valence (Russell, 1980). Arousal reflects the intensity or activation level of an emotion, ranging from calm or passive to highly excited or energetic. Valence, on the other hand, represents the emotional quality, ranging from negative (e.g., sadness, fear) to positive (e.g., joy, contentment) (Barrett & Russell, 1999). While categorical approaches capture the richness of emotional states, these categories are not rigidly discrete; instead, the dimensional framework describes their underlying structure and the relationships between emotions (Cowen & Keltner, 2017). For example, both anger and excitement are high-arousal emotions, but they differ in valence: anger is negative, whereas excitement is positive. Neuroscientific studies have shown that brain responses to emotions are continuously distributed rather than neatly categorised into discrete emotions (e.g., Lindquist et al., 2012). This view is further supported by studies such as Citron (2014) and Hamann (2012), which demonstrate that emotions involve distributed neural patterns rather than localised, discrete responses. The dimensional approach is

particularly well-suited for analysing linguistic features, such as tonemes, where emotional expression often operates on a spectrum rather than within discrete categories. This enables researchers to capture the dynamic interplay between acoustic properties and emotional perception.

In summary, we ask the question **R1**: 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 emotional arousal and valence ratings in non-tonal languages?

### 1.2.2 Linguistic pitch is predominant over phonemes in affective iconicity

Existing research suggests iconicity in SC phonemes and tonemes across various contexts, including onomatopoeias (tones in Thompson, 2018), size perception (tones and vowels /i, u, a/ in Chang et al., 2021), shape (tones and vowels /i, u, a/ in Chang et al., 2021; tones and vowels /i, u/ in Shang & Styles, 2017), power and gender associations (tones and frontal vowels in Shih et al., 2019), and character traits (tones and consonants /p, t, k/ in X. Wang, 2021). However, the interplay between phonemes and tonemes in iconicity remains unclear.

Fort et al. (2015) explored the respective contributions of consonants and vowels in the bouba–kiki effect in French, finding that consonants play a more significant role than vowels. Specifically, varying vowels had a marginally smaller (or negligible) influence on the shape–sound mapping of VCV (or CVCV) pseudowords compared to consonants. The authors proposed that the bouba–kiki effect is influenced by multiple factors, including acoustic, articulatory, and phonological properties of speech stimuli. Tone languages such as SC

offer a valuable context for exploring such interactions due to the complexity of their phonetic features across lexical tones, vowels, and consonants.

This study focuses on the interplay between phonemes and tonemes in affective iconicity. Segmental components, such as consonants and vowels, have been shown to predict arousal (e.g., German in Aryani et al., 2018; Japanese in Kambara & Umemura, 2021) and valence (e.g., German, Polish, Dutch, Spanish, and English in Adelman et al., 2018; Japanese and German in Körner & Rummer, 2023; Dutch and Chinese in Louwerse & Qu, 2017; Chinese and English in Yu et al., 2021). In these studies, the segments /i/, /u/, /t/, and /n/ are the most prevalent segments across languages (Styles & Gawne, 2017), which are also significantly relevant to affective iconicity investigations. For instance, research in Japanese and German demonstrated that /i/ is more frequently associated with positive expressions than /u/, suggesting a potential cross-linguistic consistency in vowel-valence iconicity (Körner & Rummer, 2023). Additionally, in poetic contexts, plosive sounds have been linked to pleasant moods and heightened activation, whereas nasal sounds are often associated with unpleasant moods and lower activation (e.g., in German, Chinese, Russian, and Ukrainian, see Auracher et al., 2011). However, findings by Louwerse and Qu (2017) challenge this trend, as Dutch and Chinese speakers rated nasal-initial words differently in terms of emotional valence—positively for Chinese speakers but negatively for Dutch speakers—underscoring the need for further research.

To investigate the interplay between phonemes and tonemes, we included four segmental phonemes (/i/, /u/, /t/, /n/), chosen for their distinct acoustic and articulatory features and frequent occurrence

across languages. It is plausible that tonemes exert a greater influence than phonemes due to their phonetic properties. On the one hand, lexical-prosodic features are considered facilitators in interpreting iconic meanings across languages (Dingemanse et al., 2016). On the other hand, studies on emotional expression consistently show that suprasegmental or prosodic features carry more weight than verbal content in languages such as German and SC (Filippi et al., 2017; Lin et al., 2020, 2021). While our goal is not to equate tonemes with emotional prosody and phonemes with verbal semantics, we propose that tonemes play a crucial role in iconicity due to their shared phonetic feature with emotional prosody, pitch, a highly salient perceptual dimension. An affective sensitivity to pitch in both lexical tones and emotional prosody may indicate shared or transferable mechanisms in emotional processing.

This study therefore addresses the following further questions: **R2**: How do SC phonemes exhibit iconicity compared to other languages? **R3**: How do phonemes and tonemes interact in affective iconicity in SC?

Exploring the affective roles of tonemes and their interaction with phonemes could bridge the gap between tonal and non-tonal languages in affective iconicity research. While phonemes have been extensively studied in non-tonal languages, little is known about how tonemes and phonemes contribute to emotional expression in tonal languages. Understanding this dynamic could reveal whether tonal languages prioritise pitch-based emotional communication, offering new insights into the universality and variability of affective iconicity.

### 1.2.3 Adaptive significance of linguistic pitch in affective iconicity

Although iconicity is prevalent across languages, its underlying mechanisms remain largely unexplored. Iconicity has long been considered a result of perceptual analogies between form and meaning, where linguistic features mirror the properties of the concepts they represent (Nielsen & Rendall, 2011; Sidhu & Pexman, 2018). For instance, Sidhu and Pexman (2018) outlined several potential mechanisms underlying iconicity, including statistical co-occurrence, shared perceptual and conceptual properties, neural factors related to brain structure or information processing, species-general associations such as the frequency code, and language-specific patterns. From a statistical co-occurrence perspective, phonemes may become linked to specific emotions due to the overlap between the musculature involved in articulation and the one used in emotional expression (e.g., Körner & Rummer, 2023). This suggests that iconicity may emerge from the cognitive capacity to perceive and exploit analogies between linguistic form and emotional meaning.

More recently, Adelman et al. (2018) proposed an emotion-centred perspective on iconicity, arguing that iconicity may have evolved due to its adaptive advantages for language users. Specifically, iconicity might facilitate emotional communication, improving the efficiency and effectiveness of language as a tool for conveying affective states. This perspective highlights the functional role of iconicity in emotional expression and suggests that its evolution may be driven by its utility in social and communicative contexts. Across five Indo-European languages, Adelman et al. (2018) explored whether the individual phonemes of words predict the emotional valence of those words. They

found that in English, Spanish, Dutch, German, and Polish, the first phoneme of a word predicted its valence better than subsequent phonemes. Interestingly, the faster the first phoneme is produced, the more negative the word is rated. Adelman et al. (2018) attributed this pattern to the phenomenon of automatic vigilance, which refers to the preferential attention to negative stimuli. Based on these observations, the authors suggested that iconicity might be explained as an adaptation, particularly relevant for understanding affective iconicity.

Nevertheless, the adaptation account of iconicity requires further exploration. For instance, investigations into this adaptation account have primarily focused on emotional valence, a dimension shaped significantly by social and cultural conventions (Citron, 2012). In contrast, emotional arousal, which is more closely tied to physiological responses and is considered a more ancient and innate dimension (Darwin, 1998), has received less attention in this context. Expanding the adaptation account to include arousal could provide a more comprehensive understanding of affective iconicity.

Additionally, the adaptation account has primarily been tested in Indo-European languages, where phonemes alone distinguish lexical meanings. In contrast, the majority of the world's languages, i.e., tonal languages, use both phonemes and lexical tones to convey meaning. Lexical tones, distinct from consonants and vowels, primarily signal lexical distinctions but share prosodic features, such as pitch, with emotional intonation. This raises an important question: can the adaptation account of iconicity observed in phonemes also be extended to explain affective iconicity of lexical tones? Given the shared prosodic foundation in lexical tone and emotional intonation, it is plausible that lexical tones might exhibit similar patterns of affective iconicity.

This study aims to extend the adaptation account by shifting the focus from phonemes to tonemes, from valence to both arousal and valence, and from non-tonal to tonal languages. Specifically, we address the research question (**R4**): Does the adaptation account underlying affective iconicity observed in segmental features also apply to suprasegmental features in tonal languages? By exploring this question, the study seeks to advance our understanding of the interplay between iconicity, emotional communication, and linguistic diversity.

## 1.3 Methodological approaches and stimuli used in affective iconicity research

In the field of iconicity, empirical approaches to measuring affective iconicity include both data-driven and behavioural methods. Data-driven approaches typically utilise large datasets of affective norms to identify phonological components, as demonstrated in German by Aryani et al. (2018). Behavioural approaches typically involve collecting experimental data from participants using rating tasks or guessing-meaning experiments (see review in McLean et al., 2023; Motamedi et al., 2019; Winter & Perlman, 2021). In rating tasks, participants are asked to evaluate how well a signal (e.g., orthographic words, images, or videos) resembles its referent, often using a Likert scale with 5, 7, or 9 points to capture varying degrees of similarity (Motamedi et al., 2019). While less commonly employed in the study of iconicity, slider scales with finer granularity (e.g., 100 points), commonly used in psycholinguistic research (Warriner et al., 2017), may offer valuable insights for future studies on iconicity. The scaled data (Likert/Slider) from these tasks allow for nuanced judgments, especially for real-word stimuli (e.g., Warriner et al., 2013).

In contrast, forced-choice tasks often present participants with an auditory (most likely artificial) or visual stimulus alongside two or more response options, asking them to select the option that best matches the given stimulus (e.g., Imai et al., 2008). Each of these methods provides unique insights into the study of affective iconicity, facilitating a deeper understanding of how signals relate to their referents.

Studies on iconicity employ both real words and pseudowords (or nonce words) to explore the relationship between linguistic forms and their meanings (Lockwood & Dingemanse, 2015; Motamedi et al., 2019). Real-word stimuli provide insights into how existing phonological or morphological patterns align with perceived meanings in natural languages (e.g., Adelman et al., 2018; Hinojosa et al., 2020; Winter et al., 2017, 2023). In contrast, pseudowords, i.e., artificially constructed non-lexical items, are frequently employed to control for prior semantic knowledge and isolate the effects of phonological features on perceived meaning. For instance, studies have used pseudowords to explore cross-linguistic similarities in sound-symbolic associations or to test how phonological properties influence affective or sensory judgments (e.g., D'Onofrio, 2014; Imai et al., 2008; Köhler, 1947; Monaghan et al., 2012). The combination of real-word and pseudoword stimuli provides complementary insights into the nature and mechanisms of iconicity.

## 1.4 Interim summary on affective iconicity

In Chapters 2, 3, and 4, we systematically investigate the affective iconicity of linguistic pitch in Standard Chinese (SC), focusing on three key aspects outlined in Section 1.2: the emotional associations of linguistic pitch (for lexical tone), the predominant role of tonemes over

phonemic components in affective communication, and the mechanisms underlying these associations. This comprehensive investigation aims to provide an integrated understanding of how affective iconicity functions in SC.

Chapter 2 addresses Research Question 1 (**R1**): Does pitch, a fundamental cue for emotional prosody, influence emotional responses in lexical tone languages like Standard Chinese (which use pitch to distinguish words) in a similar way to the influence of how segmental phonemes influence emotional responses, like arousal and valence, in non-tonal languages? To answer this question, three studies were conducted. Study 1 explored the predictive role of lexical tones on emotional arousal and valence ratings using two existing corpus datasets for written bisyllabic words. Emotional ratings in these datasets were collected using 7- and 9-point Likert scales. However, given the limitations of these corpora, such as uneven distributions of lexical tones, and the advantages of a more dynamic Slider scale (Warriner et al., 2013, 2017), Study 2 implemented an online Slider scale rating task. This task, designed for written bisyllabic words, featured improved stimulus balancing, an enhanced design, and advanced statistical modelling to address these limitations. Studies 1 and 2 offered insights into nuanced emotional judgments of real-word stimuli (e.g., Warriner et al., 2013). Study 3 shifted to the auditory domain to further investigate the effect of tonal pitch characteristics. Using spoken bisyllabic nonce words (devoid of semantic meaning), participants completed a two-alternative forced-choice task. This paradigm, commonly used in studies of unfamiliar stimuli, facilitates clear decision-making without overcomplicating the judgment process (e.g., Monaghan et al., 2012). Data across all three studies were

collected from native adult SC speakers, combining preexisting corpus data (Study 1) and newly collected experimental data (Studies 2 and 3).

We hypothesised that the studies would reveal a predictive relationship between lexical tones and emotional arousal and valence ratings, reflecting pitch-iconicity effects akin to those observed in emotional prosody. Specifically, we predicted that tonal pitch features—such as pitch level, average pitch height, pitch range, pitch slope, and pitch contour direction—would significantly influence emotional ratings. Higher overall pitch level, average pitch height, pitch range, and pitch slope were expected to correlate with higher emotional arousal (Bänziger & Scherer, 2005; Scherer et al., 2003; Thompson, 2018). Similarly, higher pitch level, wider pitch range, and upward pitch contour direction were predicted to align with positive valence (Belyk & Brown, 2014; Kamiloğlu et al., 2020; Yap et al., 2014). These predictions are grounded in prior findings on pitch, iconicity, and emotional prosody, which together suggest a robust interplay between tonal pitch features and affective judgments in a lexical tone language.

Chapter 3 addresses Research Questions 2 and 3. **R2**: How do phonemes /i, u, t, n/ predict affective iconicity in Standard Chinese? **R3:** How do phonemes /i, u, t, n/ and lexical tones (T1, T2, T3, and T4) interact in predicting affective iconicity in Standard Chinese? To address these questions, auditory contexts were designed to focus on the perceptual emotional experiences of tokens composed of phonemes and tonemes. The study of this chapter employs meaningless monosyllables within a two-alternative forced-choice task. Monosyllabic stimuli were selected over bisyllabic ones to isolate the effects of segmental and suprasegmental components, offering a more straightforward and more controlled examination of each speech unit's

contributions. This choice minimises the complexity introduced by tonal coarticulation or by interactions between tones and phonemes across syllables. Additionally, using monosyllables aligns with prior research on the iconicity of vowels and monosyllables (e.g., Chang et al., 2021; Shang & Styles, 2017; Tarte, 1982). The use of nonsense or nonce stimuli further ensures a clear distinction between the perceptual properties of the stimuli and the linguistic factors of interest, avoiding confounding effects of lexical semantics (Lockwood & Dingemanse, 2015). The 2AFC task was selected as the experimental paradigm to enable robust and reliable comparisons (Motamedi et al., 2019).

We hypothesised that both segmental and suprasegmental components contribute to affective iconicity in Standard Chinese. Specifically, we expected T4 would be associated with higher arousal and negative valence due to its fast pitch variation and downward contour. In contrast, T2 and T1 would be linked to positive valence due to their upward contour and stable high pitch level, respectively, and lower arousal due to their less dynamic contours. We anticipated replicating the correspondence of "/i/–positive" and "/u/–negative". We also aimed to explore the consistency of the proposed associations: "/t/–positive and high-arousing" and "/n/–negative and low-arousing". We intended to investigate how different vowels contribute to arousal ratings. For the interaction among phonemes and tonemes, we hypothesised that lexical tones would dominate phonemes in predicting affective iconicity, given that lexical tones rely on primary perceptual cues, namely pitch variations, which are highly salient in iconicity and emotional processing.

Chapter 4 addresses Research Question 4 (**R4**): Does the adaptation account underlying affective iconicity observed in

segmental features also apply to suprasegmental features in tonal languages? To address this question, we conducted corpus analyses using the same three corpora that were used in Chapter 2. Specifically, we analysed whether first-syllable tone in bisyllabic words predicted the emotional arousal and valence ratings of the words better than the second-syllable tone. Furthermore, how lexical tones at both syllables differ regarding their predictability of the arousal and valence ratings.

We hypothesised that lexical tone serves to signal both arousal and valence, thereby contributing to the emotional adaptive value of words in a tone language. Results from ERP studies show that higher-arousal emotions elicit more pronounced EPN and LPC components, compared to neutral stimuli. This suggests not only immediate and enhanced voluntary allocation of attention (EPN) but also more extensive and prolonged processing of emotions (LPC) (Fischler & Bradley, 2006; Herbert et al., 2008; Kissler et al., 2007, 2009; Schacht & Sommer, 2009). Such a time course of the neural processing of emotions may result not only from the initial but also the subsequent phonemes in a word that trigger higher emotional arousal. Hence, we anticipated the effects of tone on affective meanings throughout the word. Specifically, valence is more likely to be predicted by a word's initial lexical tone, whereas arousal by both tones of the entire word, due to the distinctive neural processing characteristics of emotional arousal and valence.

These three chapters together provide an overview of how phonemic pitch plays a role in affective iconicity, shaped by lexical tone's pitch properties, linguistic context, and emotional information processing. In the next section, we move toward the processing of linguistic pitch in another context.

## 1.5 Perceptual reorganisation

The ability to distinguish speech contrasts is fundamental to speech perception and language acquisition. Speech contrasts, which are differences in sounds that convey distinct word meanings, such as phonemes in non-tonal languages or tones in tonal languages, form the foundation for understanding and producing words (Gage & Baars, 2018). For example, in Japanese, the distinction between short and long vowels, such as *kado* ("corner") and *kaado* ("card"), is crucial for meaning differentiation (Hisagi et al., 2010). Similarly, in Dutch, the vowels /ɪ/ and /i/ distinguish words like *rit* ("ride") and *riet* ("reed") (L. Liu & Kager, 2016).

Adults are remarkably specialised in perceiving the speech contrasts of their native language(s), but this specialisation comes at a cost: a reduced ability to perceive contrasts absent from their linguistic environment. This raises a key question: when and how does this specialisation develop?

Research shows that early experiences begin shaping language processing even before birth (Gervain, 2015, 2018). For example, neonates exposed to English during pregnancy display stronger brain activity when hearing English (familiar) sentences compared to Tagalog (unfamiliar) sentences (May et al., 2011). This early sensitivity lays the groundwork for perceptual reorganisation, a process through which infants in their first year of life transition from being universal listeners, capable of distinguishing a wide variety of speech contrasts across languages, to becoming specialised in those relevant in their native language(s) (Werker & Polka, 1993; Werker & Tees, 1984).

By around six months, infants begin to show increased sensitivity to native contrasts and declining sensitivity to non-native contrasts, a

phenomenon known as perceptual narrowing or attunement (see reviews in Kuhl, 2004; Singh et al., 2022; Tsuji & Cristia, 2014). For instance, Japanese-learning infants at 6–8 months can still discriminate the English /r/–/l/ contrast, but this ability declines by the end of the first year due to its lack of relevance in Japanese (Kuhl et al., 2006; Tsushima et al., 1994). Conversely, English-learning infants improve in discriminating these sounds between 6–8 and 10–12 months (Kuhl et al., 2006).

Interestingly, perceptual narrowing is not a uniform process and varies depending on the specific speech contrast and linguistic exposure. For instance, Japanese-learning infants can discriminate Japanese vowel length contrasts (/mana/ vs. /ma:na/) at 4 and 7.5 months, but not at 9.5 months (Sato et al., 2010b). Similarly, Dutch-learning infants fail to distinguish the native /i/–/ɪ/ contrast at 5–6 months but succeed by 11–12 months. On the other hand, Mazuka et al. (2014) found that both 4.5-month-old Japanese-learning infants and 10-month-old Japanese-learning infants were able to discriminate syllables containing non-native German vowels (like /bu:k/-/by:k/ and /bi:k/-/be:k/), suggesting that vowel discrimination is influenced by their specific acoustic and/or articulatory properties. In addition, exposure to native vowel tokens that are acoustically similar to the German vowels may facilitate discrimination of the non-native vowels (Mazuka et al., 2014). Together, these findings suggest that the timing of perceptual reorganisation is contrast-dependent and shaped by language experience (L. Liu & Kager, 2016).

Some studies reveal a U-shaped trajectory in perceptual development, where sensitivity to certain contrasts temporarily diminishes before reemerging as infants develop more refined perceptual and cognitive mechanisms. For instance, de Klerk et al.

(2019) found that Dutch-learning infants at both 6 and 10 months could discriminate the native vowel contrast (/aː/-/eː/) and a non-native (English) vowel contrast (/ɛ/-/æ/), whereas 8-month-olds failed to discriminate the non-native contrast. The authors hypothesise that different age groups rely on distinct strategies, with younger infants leveraging early perceptual abilities and older infants using emerging phonetic categories, while 8-month-olds may be in transition between these mechanisms.

The diverse patterns observed across languages and segmental features suggest that perceptual sensitivity to speech contrasts is shaped by a combination of acoustic and/or articulatory properties, linguistic experience, and evolving perceptual strategies.

Beyond segmental features, many of the world's languages rely on suprasegmental features such as lexical tones, primarily cued by pitch variations, to distinguish meaning (Yip, 2002). Pitch is multifunctional, conveying emotional, grammatical, and lexical information, as demonstrated by Chen (2022) in the context of Standard Chinese. This property of pitch makes it a particularly intriguing domain for studying perceptual reorganisation. The next section will explore how these processes unfold for (tonal) pitch contrasts, offering insights into the diversity and complexity of speech perception development.

## 1.6 Perceptual reorganisation of pitch

Pitch plays a critical role in language acquisition, serving both universal and language-specific functions due to its varied roles across linguistic systems. For instance, in tonal languages like Standard Chinese (SC), pitch is acquired as a marker of lexical tone, emotion, and discourse functions, while in non-tonal languages like English and Dutch, pitch is not used to distinguish lexical meaning but is instead

used for sentence-level information and to mark stressed syllables. This dual nature underscores the universality of pitch sensitivity while highlighting its language-specific acquisition pathways.

Neonates exhibit early sensitivity to pitch. For example, French newborns can discriminate Japanese bisyllabic words differing in pitch contour (descending vs. ascending; such as candy /amé/ and rain /áme/) (Nazzi et al., 1998), and Hungarian infants show mismatch negativity (MMN) responses to pitch interval changes in sinusoidal tones, indicating pitch discrimination similar to adults (Stefanics et al., 2009). However, as infants approach their first birthday, their sensitivity to pitch contrasts becomes increasingly shaped by their native language; that is, it becomes perceptually reorganised.

In tonal language environments like Mandarin, pitch sensitivity remains robust and phonologically tuned for distinguishing lexical tones, as tone is crucial for making lexical distinctions. In contrast, in non-tonal environments like English, sensitivity to lexical pitch distinctions diminishes, while sensitivity to sentence-level intonation pitch cues is maintained. For example, Mattock and Burnham (2006) found that between 6 and 9 months, Chinese infants retained the ability to discriminate lexical tones, while English infants lost this ability for speech tones but retained it for nonspeech tones. Similarly, European Portuguese infants at both 5 and 9 months can discriminate intonational phrases consisting of a single prosodic word, distinguishing statements from yes-no questions (Frota et al., 2014). This divergence reflects how infants learning tonal languages encode pitch as phonologically contrastive, while those in non-tonal environments deprioritise pitch variation as lexically irrelevant.

However, mixed findings complicate the understanding of pitch reorganisation across linguistic contexts. Dutch infants who are not

learning Limburgian tones can discriminate these contrasts at both 6 and 12 months, just like Limburgian infants (Ramachers et al., 2018). Similarly, 12-month-old Dutch infants succeeded in discriminating SC tone contrasts (e.g., T2 vs. T4; T2 vs. T3) and their musical counterparts, despite failing to do so at 4 months (A. Chen et al., 2017; A. Chen & Kager, 2016). However, Liu and Kager (2014) observed a U-shaped trajectory in Dutch infants' discrimination of SC T1–T4 contrasts, which was discriminated at 5-6 months, not discriminated around 9 months, and discriminated again at 17–18 months.

Neuroimaging evidence links pitch processing to hemispheric lateralisation, with linguistic pitch primarily processed in the left hemisphere (Gandour et al., 2004; Zatorre et al., 1992) and emotional or intonational pitch in the right hemisphere (Gandour et al., 2004; Kotz et al., 2006; Schirmer & Kotz, 2006). Infants' hemispheric responses to pitch also align with developmental trends. Neonates show right-hemisphere responses to intonation (Arimitsu et al., 2011, 2018), and Japanese infants demonstrate a shift from bilateral to left-lateralised processing for lexical pitch-accent within word stimuli (but not for pure tone stimuli) between 4 and 10 months (Sato et al., 2010a). This transition suggests that while pitch discrimination persists, the underlying mechanisms adapt depending on linguistic experience.

The case of Dutch infants raises questions about how their neural processing of pitch-accent contrasts might differ from Japanese infants, given that pitch is not phonologically relevant in Dutch at the word level. Investigating neural responses in Dutch infants could clarify how acoustic properties, linguistic relevance, and neural plasticity interact in pitch reorganisation. By integrating behavioural and neuroimaging evidence, such investigations can elucidate whether behavioural discrimination correlates with hemispheric lateralisation patterns.

Comparing Dutch and Japanese infants would offer a unique opportunity to disentangle the influences of universal auditory principles and language-specific demands on pitch processing, especially during a period of rapid brain maturation (Leroy et al., 2011). Acoustic salience plays a critical role in shaping pitch perception. Certain contrasts, such as pitch accents, may be more easily discriminated due to their distinctive acoustic features, such as pitch contour shape and dynamic range. However, as infants gain linguistic experience, the relevance of these contrasts in their native language environment begins to exert a top-down influence on perception. For Dutch infants, who lack exposure to tone languages, the decline and re-emergence of pitch discrimination may be tied to reduced attention to non-native contrasts during the reorganisation process. Investigating these mechanisms further can clarify how acoustic properties interact with experience-driven factors to shape pitch perception.

We first explore (**R5**) whether Dutch-learning infants can distinguish lexical pitch contrasts, then explore (**R6**) whether Dutch- and Japanese-learning infants show different patterns of lateralisation while processing pitch contrasts. We designed a cross-linguistic study comparing Dutch- and Japanese-learning infants' neural responses when processing pitch contrasts, using functional near-infrared spectroscopy (fNIRS), building on the work by Sato et al. (2010a). The fNIRS technique is ideal for studying infants due to its non-invasive nature, high temporal and spatial resolution, and suitability for awake participants (Wilcox & Biondi, 2015) and has been extensively used in infant studies (Wilcox & Biondi, 2015; Yücel et al., 2021).

We hypothesised that Dutch infants could discriminate lexical pitch contrasts and that initial pitch processing is driven by general auditory mechanisms and acoustic properties. However, language-

specific experience would result in distinct processing patterns between Dutch and Japanese infants. Specifically, we anticipated that younger infants exposed to either Japanese or Dutch would show right-lateralised or bilateral responses to pitch contrasts in both non-speech and speech stimuli, reflecting general auditory processing based on pitch's perceptual properties. As they grow older, lateralisation patterns would diverge, particularly for pitch contrasts in speech stimuli, influenced by their linguistic experience with pitch functions. Japanese (older) infants would exhibit left-lateralised responses for lexical pitch contrasts due to their linguistic relevance, while Dutch (older) infants would process those same pitch contrasts as intonation, and would therefore show more bilateral processing.

## 1.7 Summary of the contents of this dissertation

This dissertation investigates the affective iconicity of linguistic lexical tonal pitch in Standard Chinese in (Chapters 2–4) and the perceptual reorganisation of linguistic pitch in Japanese pitch accent and Dutch lexical stress (Chapter 5). Chapter 2 investigates whether pitch, a key cue for emotional prosody, influences emotional responses in lexical tone languages like Standard Chinese, similar to how segmental phonemes shape emotional expressions, such as arousal and valence, in non-tonal languages (**R1**). Chapter 3 examines the role of phonemes /i, u, t, n/ in affective iconicity (**R2**) and their interaction with lexical tones (T1, T2, T3, and T4) in shaping these associations (**R3**). Chapter 4 extends the discussion by testing whether the adaptation account of affective iconicity observed in segmental features also applies to suprasegmental features in tonal languages (**R4**). Finally, Chapter 5 shifts focus to perceptual reorganisation, first assessing whether Dutch-learning infants can distinguish lexical

stress-related pitch contrasts (**R5**) and then comparing how Dutch- and Japanese-learning infants neurologically differ in their ability to discriminate these contrasts (**R6**).

# 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.

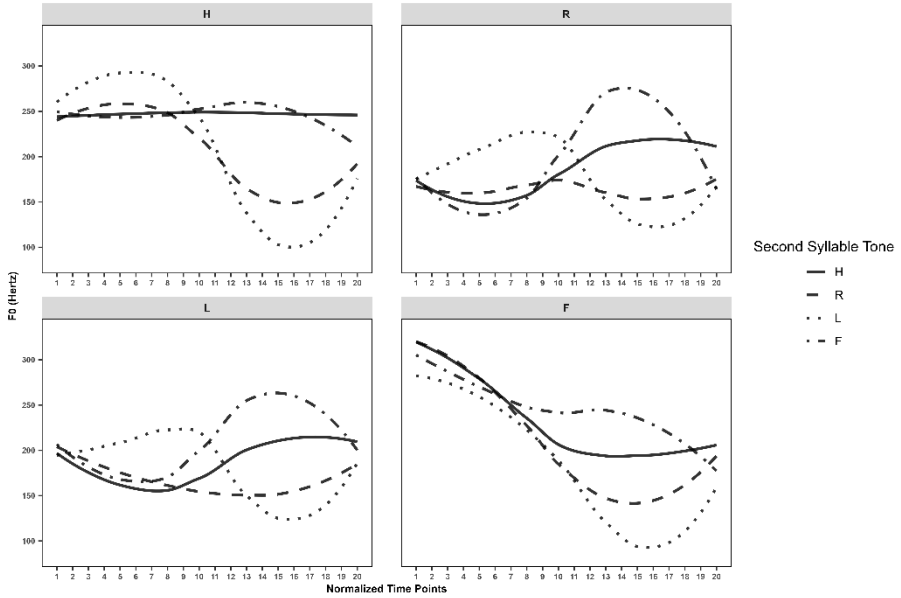
As mentioned earlier, our empirical domain is Standard Chinese, which includes four distinctive lexical tones: high-level (Tone 1, T1, or H), mid-rising (Tone 2, T2, or R), low-dipping (Tone 3, T3, or L) and high-falling (Tone 4, T4, or F). Our investigation centres on bisyllabic words primarily due 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 ($\Delta R^2$ = 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 ($\Delta R^2$ = 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.*

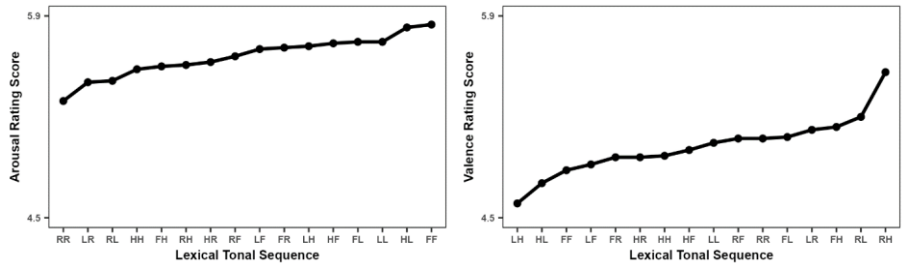
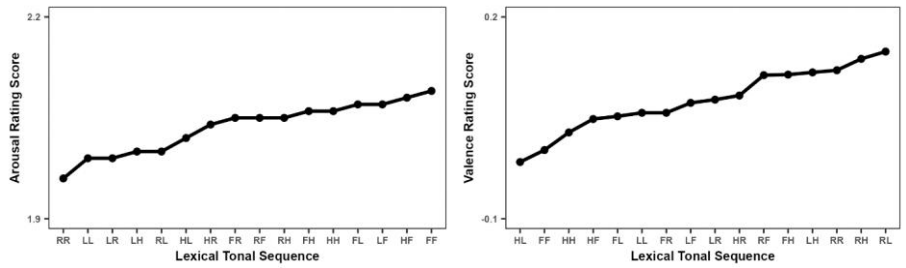


**Figure 2.3**

*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; $\Delta R^2$ = 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 ($\Delta R^2$ = 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 ($\Delta R^2$ = 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.

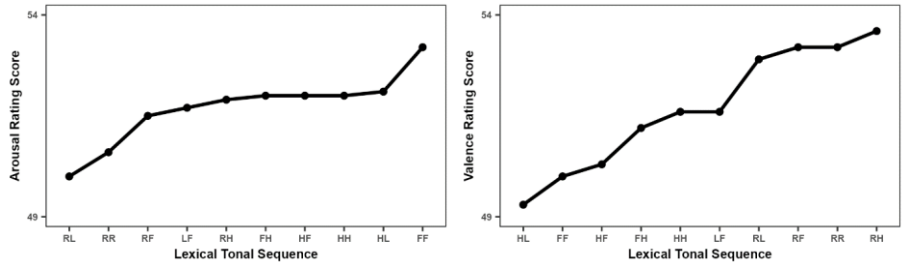
Similarly, in the GLMM analyses, comparisons between null and interest models revealed significant effects of lexical tonal sequence on arousal ratings in all datasets ($F$ = 31.10 ~ 34.35, *ps* < .001). Moreover, FF tonal sequence consistently yielded higher arousal ratings compared to RR and RL. However, 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.*

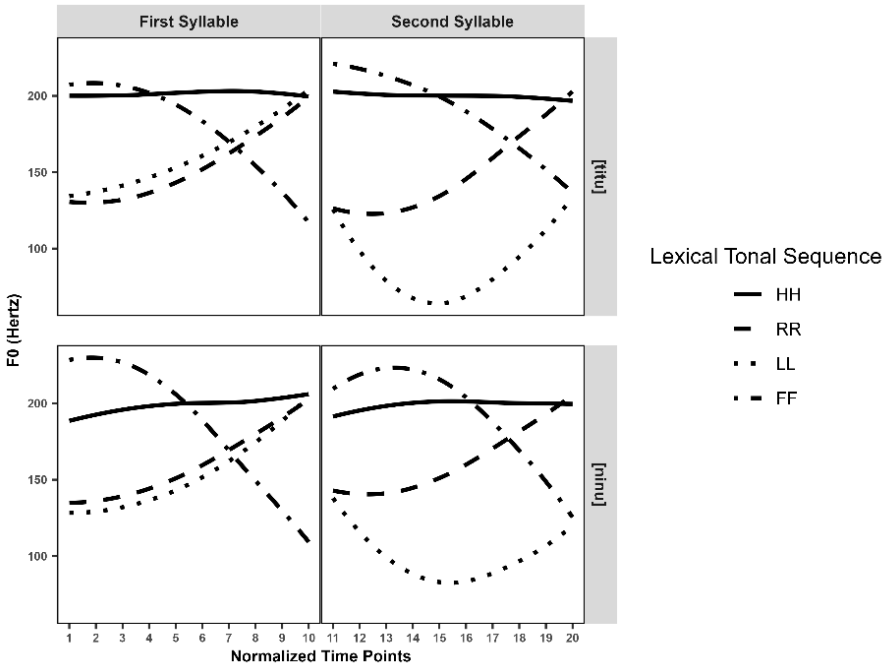


**Table 2.1**

*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.

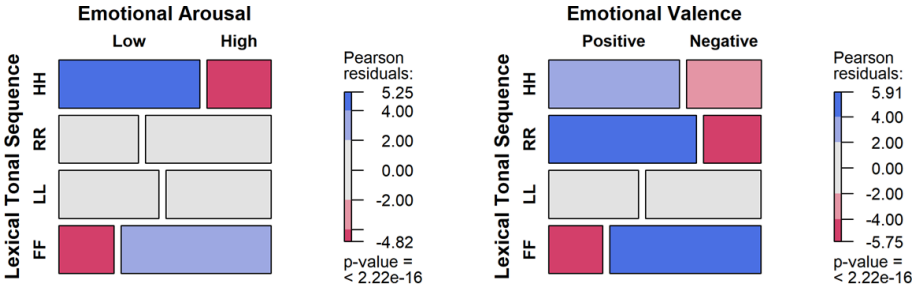
We utilised the Odds Ratio as the main measurement for discussing the outcomes. Odds represent the ratio of the probability of

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 ($\chi^2$ = 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 ($\chi^2$ = 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.

# Chapter 3
# Affective iconicity of tonemes and phonemes in monosyllabic units in Standard Chinese

A version of this chapter is under revision: Zheng, T., Levelt, C. C., & Chen, Y. (under revision). The affective iconicity of phonemes and tonemes in Standard Chinese.

**Abstract**: While both segmental and suprasegmental aspects of words have been recognised as potential factors influencing their iconic interpretations (Dingemanse et al., 2016; Svantesson, 2017), how these components collectively or independently drive the associations of form and affective meaning remains elusive. The current study addressed this issue in a lexical tonal language, Standard Chinese, where suprasegmental pitch information distinguishes word meanings. Specifically, we investigated how phonemes at both the segmental level (i.e., vowels and consonants) and suprasegmental level (i.e., lexical tones; also known as tonemes) may influence native Standard Chinese listeners' rating of auditory stimuli's emotional arousal and valence in two-alternative forced-choice tasks. The results indicated a consistent correlation between tonemes and the perceived arousal and valence ratings of the tone-carrying nonce words. At the segmental level, consonants were more consistently associated with arousal, while vowels correlated with valence. Regarding arousal ratings, nonce words with falling and rising tones tended to be rated with higher arousal than those with high and low-dipping tones, and those with an onset /t/ were rated higher in arousal than those with /n/. For valence ratings, nonce words with falling and low-dipping tones were more likely to receive negative ratings than those with high and rising tones. Stimuli containing /u/ were more inclined to be perceived negatively than those with /i/. When tonemes and phonemes predicted contrary ratings, tonemes were found more influential than segmental phonemes in biasing listeners' arousal and valence ratings. Though subtle and sporadic, these findings support the universal tendency of affective iconicity across segmental phonemes and suprasegmental tonemes, emphasising the predominant role of pitch in determining emotional arousal and valence ratings.

**Keywords**: Affective iconicity, Phoneme, Standard Chinese, Toneme, Emotional arousal, Emotional valence

## 3.1 Introduction

Iconicity, the correspondence between sound and meaning in speech, has been argued to be a fundamental aspect of all human languages (Dingemanse et al., 2015; Perniss et al., 2010; Winter et al., 2023). Among the various links between speech sounds and meaning, such as the association of /i/ with smallness (Blasi et al., 2016), the bouba–kiki effect linking rounded shapes to "bouba" and spiky shapes to "kiki" (Ćwiek et al., 2021), reduplication in ideophones (Dingemanse, 2015), vowel sounds in brand names (Shrum et al., 2012), and broader reviews on non-arbitrariness (Svantesson, 2017), the relationship between phonemes and affect has been particularly well-established (Aryani et al., 2018; Schmidtke et al., 2014; Whissell, 2003). This connection between sound and affective meaning is especially significant given the central role emotion plays in communication (Adelman et al., 2018; Aryani et al., 2019; Auracher et al., 2011; Darwin, 1998; Davidson et al., 2009; Nielsen & Dingemanse, 2021). A growing consensus in the literature suggests that phonemes and tonemes can convey emotions through iconicity, functioning as an early warning system in human language, akin to alarm calls in other species (Adelman et al., 2018; Zheng et al., 2025). It is important to note that the observed associations between phonemes/tonemes and affective meanings differ from systematicity, which refers to statistical regularities in sound-meaning mappings across languages. As outlined in the General Introduction, we define affective iconicity as the degree to which language users perceive a subjective connection between a linguistic signal's form (e.g., pitch, in this case) and its affective meaning.

The current study built upon the existing literature and aimed to investigate further possible interactions of phonemes and tonemes in conveying affective iconicity in a tonal language. We adopted the dimensional framework approach, which has been argued to better capture the complexity of emotional experiences than the discrete emotion perspective (Adelman et al., 2018; Aryani et al., 2018; Bänziger & Scherer, 2005; Kamiloğlu et al., 2020; Laukka et al., 2005). In this approach, rather than categorizing emotions into distinct entities (e.g., anger, fear) as in the discrete emotion approach, emotions are described using dimensions such as arousal (ranging from calm to exciting in terms of how activated an emotion is) and valence (ranging from sad to happy in terms of how positive or negative an emotion is) (Harmon-Jones et al., 2017).

There has been a productive line of research on how distinctive phonetic properties of speech help convey emotional arousal and valence of speech units. At the segmental level, consonant and vowel phonemes have been shown to predict both arousal (German in Aryani et al., 2018; Japanese in Kambara & Umemura, 2021) and valence values (German, Polish, Dutch, Spanish, and English in Adelman et al., 2018; Japanese and German in Körner & Rummer, 2023; Dutch and Chinese in Louwerse & Qu, 2017; Chinese and English in Yu et al., 2021). Within the suprasegmental domain, prosodic cues such as intonational pitch variation (both in terms of pitch level, average pitch height, and pitch contour) have been shown consistently to be associated with emotional arousal and valence (Bänziger & Scherer, 2005; Belyk & Brown, 2014; Kamiloğlu et al., 2020; Laukka et al., 2005; Scherer et al., 2003). Typically, higher global pitch level and wider pitch range are associated with higher arousal and more positive valence (cf. Scherer & Oshinsky, 1977; Stel et al., 2012).

Despite the significant insights into segmental and suprasegmental aspects of speech, the interplay between these aspects of speech in driving affective iconicity has yet to be fully elucidated. This study addresses this knowledge gap using data from Standard Chinese (SC). SC, with its use of suprasegmental pitch variation to distinguish meanings through lexical tones, provides an ideal and necessary testing ground for investigating the interaction of segmental and suprasegmental elements for affective iconicity. Lexical tones function just like segmental phonemes in tonal languages like Standard Chinese and are, therefore, also known as tonemes. Existing research suggests iconicity in SC phonemes and tonemes across various contexts such as onomatopoeias (tones in Thompson, 2018), the perception of size (tones and vowels /i, u, a/ in Chang et al., 2021) and shape (tones and vowels /i, u, a/ in Chang et al., 2021; tones and vowels /i, u/ in Shang & Styles, 2017, 2023), as well as the association with power and gender (tones and frontal vowels in Shih et al., 2019) and character traits (tones and consonants /p, t, k/ in X. Wang, 2021). Furthermore, there is also some, yet-to-be-replicated, evidence for the role of phonemic pitch in affective connotations in the general vocabulary (Yao et al., 2013; Yap et al., 2014; Zheng et al., 2024, 2025). In short, the existing literature suggests that the involvement of both phonemic segments and tonemic pitch variations for affective iconicity in Standard Chinese, but little is known about how exactly these two components interact. Given that across languages, segmental and suprasegmental components have been shown to exert varying degrees of influence on sound-meaning associations (Dingemanse et al., 2016; Fort et al., 2015), we set out to investigate whether, and if so, how consonants, vowels, and lexical tones differ and interact in conveying emotional arousal and valence.

In this study, we included all four distinctive lexical tones in Standard Chinese. When produced in isolation, the high-level tone (T1, H) maintains a high pitch level throughout the syllable. The rising tone (T2, R) starts at a mid-level pitch and gradually rises. The low-dipping tone (T3, L) starts at a mid-level pitch, dips, and then often rises again. The falling tone (T4, F) starts at a high-level pitch and steeply falls.

For segments, we included /i/, /u/, /t/, and /n/, chosen for their widespread occurrence across languages and their reported relevance in iconicity studies (Blasi et al., 2016; Styles & Gawne, 2017). Research in Japanese and German revealed that /i/ tends to be associated with positive expressions more frequently than /u/, suggesting a potential cross-linguistic consistency in vowel-valence iconicity (e.g., Körner & Rummer, 2023). The correspondence between vowels and emotional arousal is less clear, making our inclusion of /i/ and /u/ a valuable contribution.

As for consonants, the inclusion of /t/ and /n/ is due to the extensive study of their expression of emotions in poetic languages. In poems, plosive sounds are more often associated with pleasant moods and heightened activation, whereas nasal sounds tend to signify an unpleasant mood with lower activation (German, Chinese, Russian, Ukrainian in Auracher et al., 2011). Nevertheless, contradictory findings have been reported by Louwerse and Qu (2017), where Dutch and Chinese speakers rated nasal-initial words differently in terms of emotional valence, with positivity for Chinese speakers and negativity for Dutch speakers, highlighting the need for further investigation.

To explore the predictive role of tonemes and phonemes on emotional arousal and valence choices, we conducted an online experiment using two-alternative forced-choice tasks. For further procedural details, readers are referred to the Procedure section.

Grounded in previous studies, we hypothesised that both segmental and suprasegmental components should contribute to affective iconicity in Standard Chinese. Specifically, we expected T4 to be associated with higher arousal and negative valence due to its fast pitch variation and downward contour, while T2 and T1 associated with positive valence due to their upward contour and stable high pitch level and lower arousal due to their less dynamic contours. We also expected to replicate the "/i/–positive, /u/–negative" correspondence and explore the consistency of the "/t/–positive and high-arousing, /n/–negative and low-arousing" correspondence. Last but not least, we aimed to explore how different vowels contribute to the arousal ratings and examine the possible interactions of tonemes and phonemes on emotional ratings.

## 3.2 Method

### 3.2.1 Sample size estimation

We conducted a prior sample size estimation using G*Power 3.1 (Faul et al., 2009), setting the odds ratio at 2.5, with an α level of 0.05, a power of 0.80, and assuming no variance explained by other predictors with X parm π set as 0.5. This calculation resulted in a total required sample size of 128 (one-tailed, binomial).

### 3.2.2 Participants

A total of 179 native SC-speaking college students participated online in the study. All participants had no hearing, visual, speech, or alexithymia disorders and used Standard Chinese for more than 80% of their daily communication. For the final analysis of all the target stimuli, we only included those participants who accurately rated the

non-target validation items (see further details for the items in the Stimuli section). This resulted in a total of 121 participants for the arousal task (average age = 21.53 years; SD = 2.11 years; 72 females) and 135 participants for the valence task (average age = 21.52 years; SD = 2.13 years; 84 females).

### 3.2.3 Stimuli

#### *3.2.3.1 Auditory stimuli*

Two sets of auditory stimuli were used in this study. The first vowel set (V) comprised single vowels /i/ and /u/, each articulated with four SC tones (i.e., /i1/, /i2/, /i3/, /i4/, /u1/, /u2/, /u3/, /u4/). This set investigated the effects of lexical tone, vowel, and their interaction. The second consonant-vowel nonce word set (CVL) combined the consonants /t/ and /n/ with vowels /i/ and /u/ and the coda /l/ to form monosyllabic nonce words (/til/, /nil/, /tul/, /nul/). This set explored the effects of lexical tone, vowel, consonant, and their interactions. An additional (third) set included four disyllabic nonce words with consonant-vowel-consonant-vowel structure (CVCV), which was intended to conceptually replicate a previous study (Zheng et al., 2025) with nonce words so as to better control segmental and tonal compositions and exclude possible effects of lexical meanings. The use of nonce words in iconicity studies (McCormick et al., 2021) and auditory presentation (Cuskley et al., 2017) have often been adopted to bias listeners' processing towards sound properties rather than semantic meanings, thereby gaining insight into iconicity as a pre-semantic phenomenon (Sučević et al., 2015; Westbury, 2005).

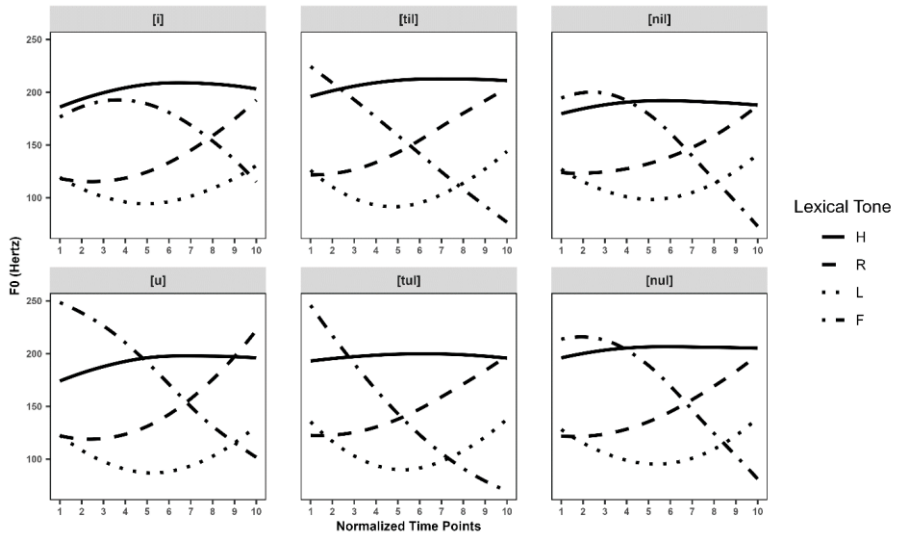
In addition to the target stimuli, we also included two real words *baozao* (/pɑʊ//tsɑʊ/ T4T4, meaning *irritable*) and *xingfen* (/ɕɪŋ//fən/

T1T4, meaning *excited*), as validation items. The arousal and valence ratings for *baozao* have been shown to be consistently negative and highly arousing, while those for *xingfen* have been shown to be consistently positive and highly arousing (Y. Wang et al., 2008; X. Xu et al., 2022). Therefore, they were chosen to evaluate participants' involvement in the online tasks, and their deviant inappropriate ratings were used as an objective criterion for participants' exclusion.

All stimuli were recorded in isolation with a Sennheiser MKH416T microphone (sample size 44.1 kHz, 16 bit) at Leiden University's Phonetics Lab by a male native speaker of Standard Chinese. He was born and raised in Beijing and was unaware of the study's purpose. He was asked to produce the stimuli as a statement without any emphasis. The stimuli were recorded three times, with the stimulus list randomised for each recording, and edited in Praat (Boersma & Weenink, 2024). Tokens judged to be most clearly articulated by the first author and another native speaker consultant were selected. Each stimulus was normalised to an average intensity of 70 dB SPL. We segmented the auditory stimuli and set the boundaries of the onset and offset for each syllable. Then, we defined the number of *f0* points (i.e., 10) and *f0* range (i.e., 45-400 Hz) to be extracted from the target rime. The pitch contours of the six tokens (two single vowels and four monosyllabic nonce words) are shown in Figure 3.1.

**Figure 3.1**

*Pitch contours of auditory stimuli.*




### *3.2.3.2 Visual stimuli*

Two pairs of emoji faces representing high vs. low arousal and positive vs. negative valence were selected from the EmojiGrid (Toet & van Erp, 2019). EmojiGrid was developed to facilitate more spontaneous and intuitive communication about affective experiences. Its assessments align well with other self-report tools like SAM and Likert scales and demonstrate the typical V-shaped relationship between arousal and valence. Figure 3.2 shows the emojis used in this study.

**Figure 3.2**

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


### 3.2.4 Procedure

The experiment was conducted online at the Gorilla platform (https://gorilla.sc/). After recruitment, each participant received a personalised link to the experiment, accessible only via a PC or laptop. A headphone test ensured that all participants completed the experiment in a quiet environment using headphones (Milne et al., 2021).

All instructions were given in written Chinese. Participants were first introduced to the target emotional dimensions (arousal and valence) and their corresponding emoji icons through a quiz on assigning emojis to emotional dimensions in specific scenarios. Feedback was provided, and participants could proceed only after correctly completing all assignments to ensure their correct understanding of the emotional dimensions and emojis used in the task.

For both valence and arousal ratings, a two-alternative forced-choice (2AFC) task was employed. Participants performed the valence and arousal rating tasks in a counterbalanced order. However, attrition during the participants' performance phase resulted in a slight difference in their final distribution. Specifically, across all

participants, 93 started with the valence task, while 86 started with the arousal task. Despite this minor imbalance, the counterbalancing process effectively minimised order effects across participants.

Each participant completed a series of five sessions. The sequence included a practice session comprising four trials using unrelated nonce words and single vowels to familiarise participants with the task. Subsequently, participants completed three experimental sessions (i.e., V, C/i/L, C/u/L), each consisting of eight trials. The order of these experimental sessions was counterbalanced across participants using a Latin-square design. The last session, including the CVCV stimuli, followed by the validation items, was presented at the end to minimise potential influences from tone-bearing units and ensure consistent presentation of tones. Notably, the three experimental sessions focused on stimuli with a single toneme as the independent variable, whereas the CVCV session involved stimuli with two tonemes, i.e., lexical tonal sequence patterns (see results in Zheng et al., 2025). The entire experiment lasted approximately 20 minutes. Additionally, a mandatory 10-second break was included after each session to minimise potential fatigue. Within each session, stimuli were presented in eight pseudo-randomised orders, ensuring alternation between onsets and a minimum distance between identical lexical tones. Each stimulus was played automatically once and could be manually replayed at most five times. Participants indicated their perceived arousal or valence of the auditory stimulus by selecting one of two emojis, which were horizontally presented with their positions counterbalanced across trials.

## 3.3 Data analysis

We employed factorial generalised linear mixed models (GLMMs) for the V dataset to examine the main effects and interactions of lexical tone and vowel on arousal and valence ratings of single vowels. Similarly, we employed factorial GLMMs for the CVL dataset to examine the main effects and interactions of lexical tone, vowel, and consonant on arousal and valence ratings of CVL monosyllables. Conducting separate analyses for the two datasets enabled us to replicate the vowel and lexical tone effects while providing deeper insight into the main effect of consonants and the interaction among lexical tone, vowel, and consonant. These analyses were conducted using the *lme4* (Bates et al., 2014) package in R (R Core Team, 2023). Logistic regression models with a binomial probability distribution and the logit link function were utilised to analyse binary outcomes. To account for potential participant heterogeneity, we initially included random intercepts and slopes by subject. However, after fitting the models, we encountered convergence issues and overfitting with the random slopes. Therefore, we opted for a simpler model with only by-subject intercepts. Overdispersion was assessed using residual ratios calculated as the sum of squared Pearson residuals divided by the residual degrees of freedom, evaluating whether the observed variability in the outcomes exceeds what would be expected under the assumed binomial distribution.

For pairwise multiple comparisons, we utilised the *emmeans* package (Lenth, 2023), incorporating Bonferroni correction to control the false positive rate (type I errors). Post-hoc power analysis was performed for each coefficient in all models using the *pwr* package to ensure sufficient power to detect the effects of each predictor

(Champely et al., 2020). This analysis helps to validate the reliability of our findings by confirming that the sample size was adequate to identify significant effects, thereby reducing the risk of type II errors.

We used the Odds Ratio (OR) as a measurement in our data analysis. The odds represent the ratio of the probability of an event occurring (e.g., high arousal) to the probability of it not occurring (e.g., being rated as low arousal). The OR is a statistical measure that compares the odds of an event occurring in one group to the odds of it occurring in another group. For example, an OR would indicate how much more likely high arousal is chosen in one category (e.g., /i/) compared to another category (e.g., /u/).

To visualise the results, mosaic plots were employed to illustrate how the frequency of arousal or valence choices varied with predictors (Friendly & Meyer, 2015). The tiles were coloured and shaded based on standardised residuals, indicating deviations from expected frequencies under the assumption of independence, where the expected probability is equal for all levels between categorical variables. Positive residuals denote observed higher frequencies than expected, while negative residuals indicate lower-than-expected frequencies. The shading provides information about relative frequencies within the same row. Importantly, standardised residuals in the mosaic plots measured departure from independence but do not directly correspond to the inferential statistics of fitted models. Instead, they offered additional insight into the strength and direction of the association between factors and emotional dimensions.

## 3.4 Results

Figures 3.3–3.12 present mosaic plots illustrating the frequency distribution of arousal or valence choices across various predictors,

including lexical tone, vowel, consonant, and their interactions. Residual ratios, ranging from 0.93 to 1.01, indicate no concerns regarding overdispersion across all logistic regression models.

### 3.4.1 V dataset

#### *3.4.1.1 Arousal*

We fitted four models: a null model and three alternative models. These alternatives included fixed effects by (a) lexical tone, (b) vowel, and (c) lexical tone, vowel, and their interaction. The random effect of the variable subject was negligible across all models (i.e., singular fit), indicating insignificant variability in arousal ratings attributed to individual differences, and was thus excluded from all the alternative models. All alternative models significantly predicted the emotional arousal ratings of stimuli. Specifically, model (a) yielded $\chi^2 = 164.14$, $p < .001$, with an $R^2$ of 7.13%; (b) $\chi^2 = 4.26$, $p < .05$, $R^2 = 0.32\%$; and (c) $\chi^2 = 171.83$, $p < .001$, $R^2 = 12.85\%$.

Pairwise comparisons indicated that T4 was rated significantly more arousing than T1 and T3 (OR = 3.20, 5.48), as was T2 compared to T1 and T3 (OR = 5.26, 9.09). However, no significant difference was found between /i/ and /u/ in their arousal ratings ($p = 0.12$). The significant interaction between lexical tone and vowel primarily reflected the main effect of toneme, suggesting the nonsignificant influence of single vowels on arousal ratings.

**Figure 3.3**

*Arousal frequency variations across lexical tones (left) and vowels (right) in the V dataset. Tones are denoted as H (T1: high-level), R (T2: rising), L (T3: low-dipping), and F (T4: falling).*

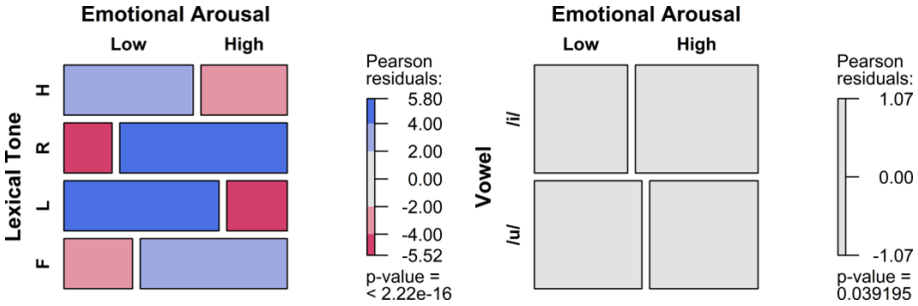
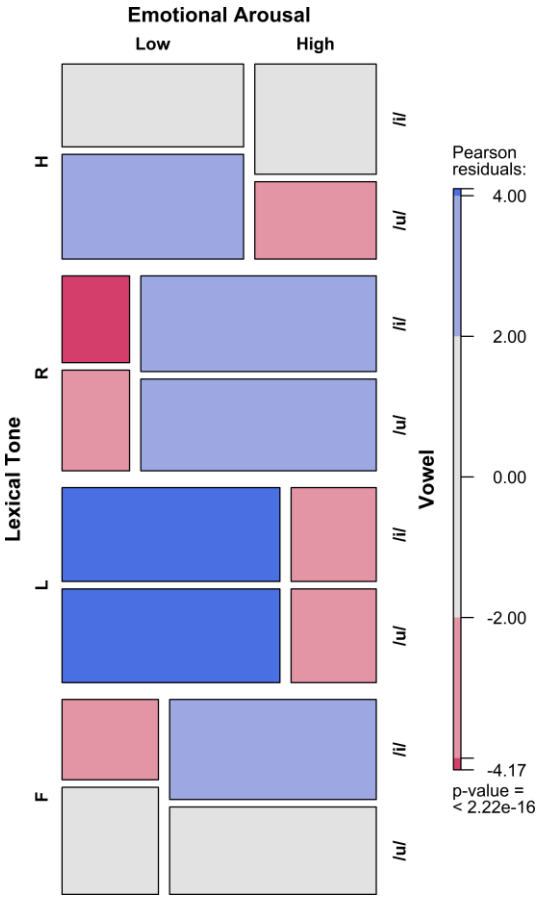


**Figure 3.4**

*Arousal frequency variations across the interaction of lexical tones and vowels in the V dataset.*




### *3.4.1.2 Valence*

We applied the same four models for valence ratings as for arousal ratings. In contrast to arousal ratings, the random effect of the variable subject was significant, indicating some variability in valence ratings attributed to individual differences. The variance was 0.04 (± 0.21) for model (a), 0.02 (± 0.16) for model (b), and 0.08 (± 0.27) for model (c). All alternative models significantly predicted the emotional valence

ratings of the utterances. Specifically, model (a) yielded $\chi^2 = 59.63$, $p < .001$, with an $R^2$ of 4.00%; for model (b) $\chi^2 = 26.53$, $p < .001$, $R^2 = 1.80\%$; and for model (c) $\chi^2 = 105.31$, $p < .001$, $R^2 = 7.07\%$.

Pairwise comparisons showed that T4, T1, and T3 were more likely to be rated negatively than T2 (OR = 2.76, 1.84, and 3.69, respectively). Moreover, T3 was twice as likely to be rated negatively as T1. Additionally, /u/ was rated more negatively than /i/ (OR = 1.89). The interaction effect between lexical tone and vowel revealed that T2 consistently predicted positive valence despite the vowels, while /i/ paired with T3 was rated as more negative (OR = 2.83) than /u/ paired with T2, suggesting a predominant effect of lexical tone (over vowel) on valence ratings.

**Figure 3.5**

*Valence frequency variations across lexical tones (left) and vowels (right) in the V dataset.*

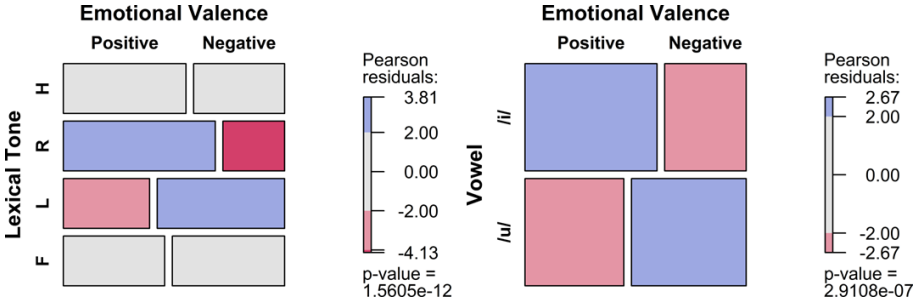
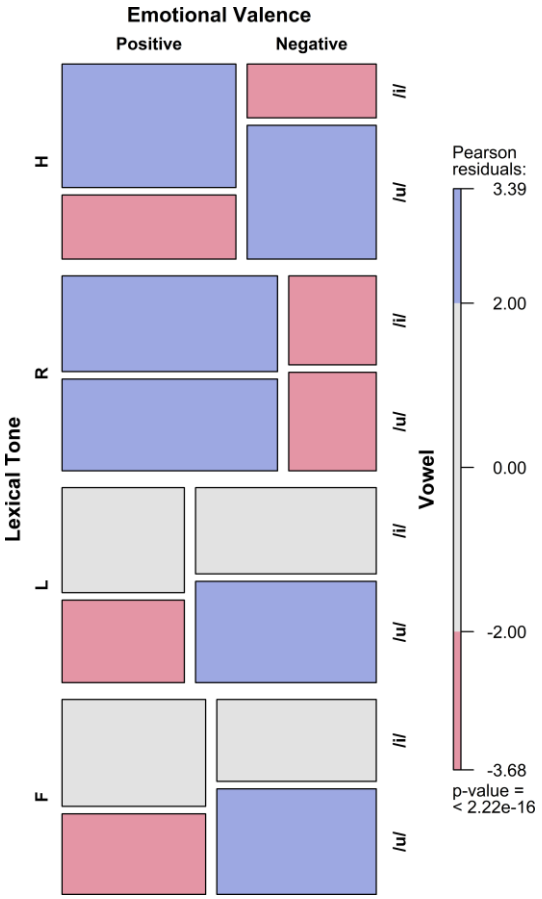


**Figure 3.6**

*Valence frequency variations across the interaction of lexical tones and vowels in the V dataset.*




## 3.4.2 CVL dataset

### *3.4.2.1 Arousal*

We fitted a null model and seven alternative models incorporating fixed effects of (a) lexical tone, (b) vowel, (c) lexical tone, vowel, and their interaction, (d) onset consonant, (e) lexical tone, onset consonant, and their interaction, (f) vowel, onset consonant, and their

interaction, and (g) lexical tone, vowel, onset consonant, and their interactions. The random effect of the subject was significant in models (a), (c), and (e) with a variance of 0.25 (± 0.50), as well as in models (b), (d), and (f) with a variance of 0.21 (± 0.46). Across all alternative models, fixed effects significantly predicted the emotional arousal rating. Specifically, model (a) yielded $\chi^2$ = 91.33, $p$ < .001, with an $R^2$ of 3.46%; (b) $\chi^2$ = 3.69, $p$ = 0.05, $R^2$ = 0.14%; (c) $\chi^2$ = 104.60, $p$ < .001, $R^2$ = 4.00%; (d) $\chi^2$ = 0.17, $p$ < .01, $R^2$ = 0.31%; (e) $\chi^2$ = 104.46, $p$ < .001, $R^2$ = 3.96%; (f) $\chi^2$ = 13.93, $p$ < .01, $R^2$ = 0.53%; and (g) $\chi^2$ = 94.64, $p$ < .001, $R^2$ = 3.59%.

Pairwise multiple comparisons demonstrated that T4 was rated to be more arousing than T1 and T3 (OR = 2.35, 3.10), as was T2 compared to T1 and T3 (OR = 1.85, 2.44). Additionally, /t/ predicted higher arousal than /n/ (OR = 1.31). These effects of lexical tone and consonant were consistent with their main effects in the interaction models. Furthermore, /n4/ and /n2/ predicted higher arousal than /t3/ (OR = 2.78 and 1.89) in model (e). In model (g), /nu4/ was more likely to be rated as higher arousal than /ti3/, /tu1/, and /tu3/ (OR = 3.23, 2.78 and 2.94, respectively). These interaction effects suggest a predominant effect of lexical tone on the arousal rating over onset consonant. Again, the vowel effect on arousal rating was not observed.

**Figure 3.7**

*Arousal frequency variations across lexical tones (left), vowels (middle), and onset consonants (right) in the CVL dataset.*

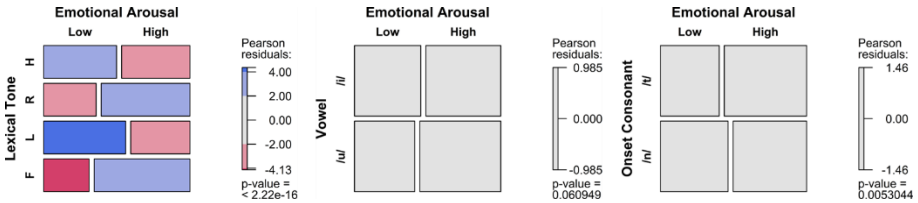


**Figure 3.8**

*Arousal frequency variations across the interaction of lexical tones and vowel (left), lexical tones and onset consonants (middle), and vowel and onset consonants (right) in the CVL dataset.*

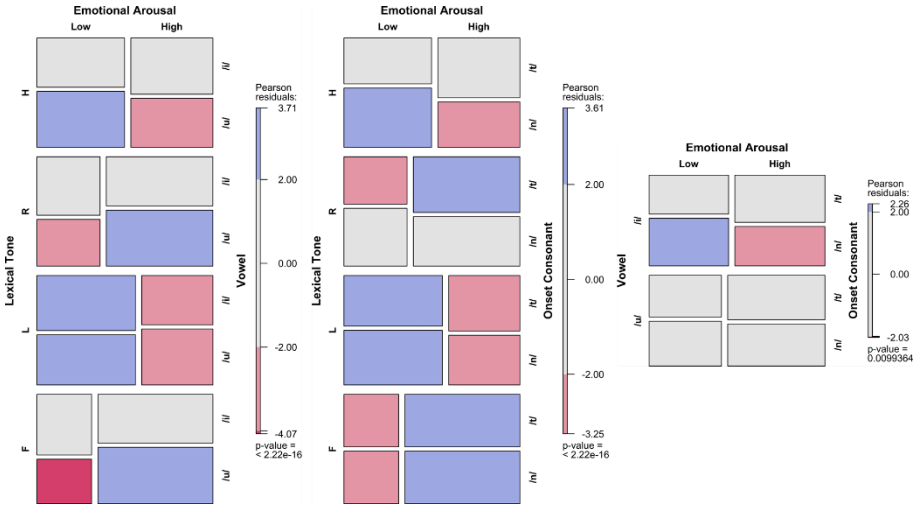
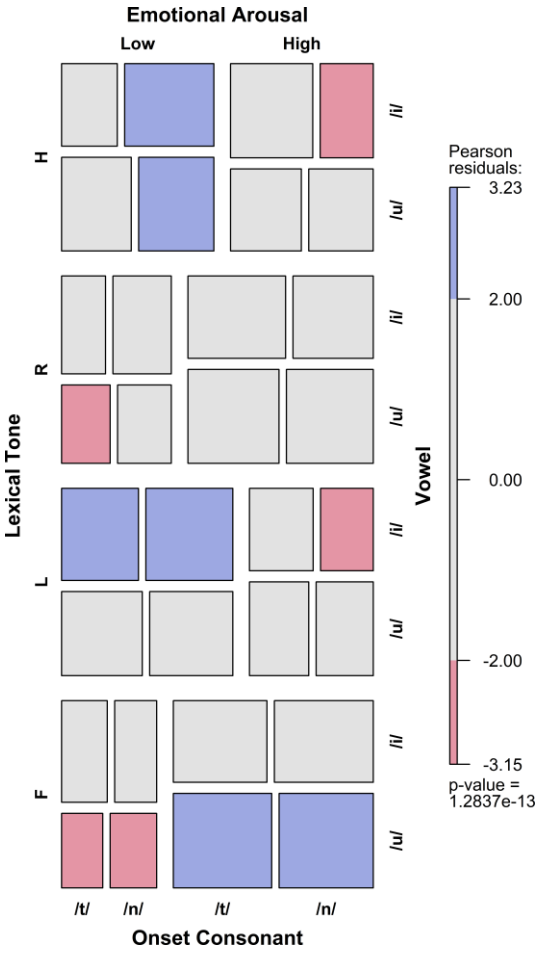


**Figure 3.9**

*Arousal frequency variations across the interaction of lexical tones, vowel, and onset consonants in the CVL dataset.*




### *3.4.2.2 Valence*

We fitted the same models for valence as for arousal. The random effect of the subject was significant except for model (g), with variances of 0.17 (± 0.41) for models (a), (c) and (e), and 0.10 (± 0.32) for models (b) and (d). For model (f), it was 0.11 (± 0.32). All alternative models significantly predicted the emotional valence ratings of the nonce

words, except for model (d). Specifically, model (a) yielded $\chi^2 = 235.78$, $p < .001$, with an $R^2$ of 7.93%; (b) $\chi^2 = 12.54$, $p < .001$, $R^2 = 0.42\%$; (c) $\chi^2 = 234.85$, $p < .001$, $R^2 = 7.90\%$; (d) $\chi^2 = 1.84$, $p = .176$, $R^2 = 0.06\%$; (e) $\chi^2 = 245.58$, $p < .001$, $R^2 = 8.26\%$; (f) $\chi^2 = 14.49$, $p < .01$, $R^2 = 0.49\%$; and (g) $\chi^2 = 253.71$, $p < .001$, $R^2 = 8.53\%$.

Pairwise multiple comparisons revealed that T4 was rated to be more negative than T1 and T2 (OR = 3.99, 4.08), as was T3 compared to T1 and T2 (OR = 4, 4.08). Vowel /u/ was rated to be more negative than /i/ (OR = 1.37). Regarding the interaction effect of lexical tone and vowel, /i4/ was rated to be more negative than /u1/ and /u2/ (OR = 2.75, 2.71), as was /i3/ compared to /u1/ and /u2/ (OR = 3.05, 3.00). These results suggest a predominant effect of lexical tone on valence ratings over vowel, reconfirmed in model (g). No significant effect of consonants on valence ratings was observed.

**Figure 3.10**

*Valence frequency variations across lexical tones (left), vowels (middle), and onset consonants (right) in the CVL dataset.*

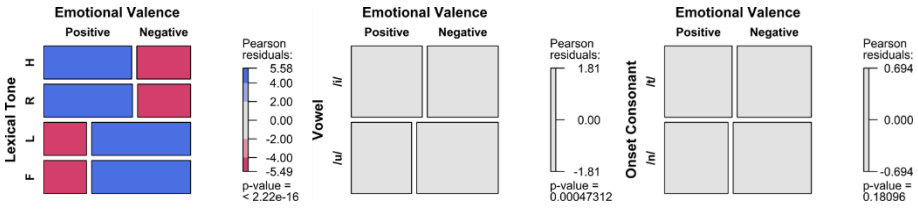


**Figure 3.11**

*Valence frequency variations across the interaction of lexical tones and vowels (left), lexical tones and onset consonants (middle), and vowels and onset consonants (right) in the CVL dataset.*

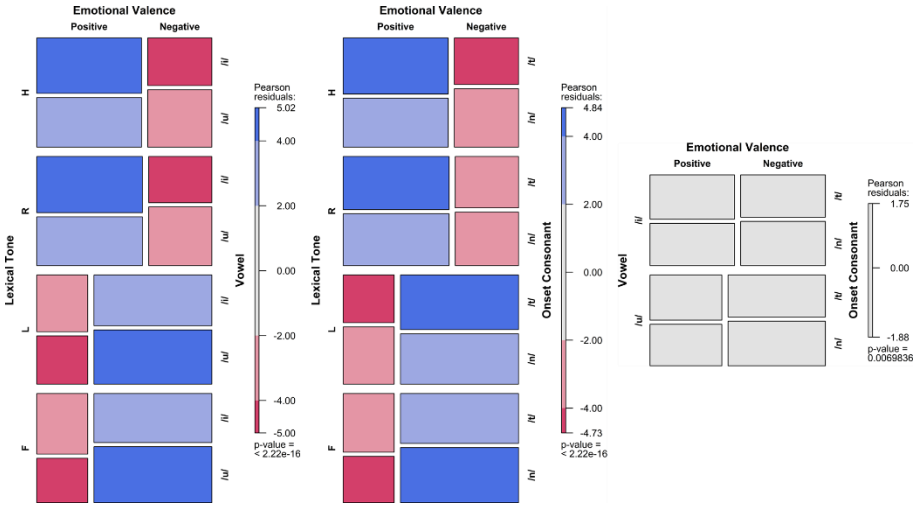
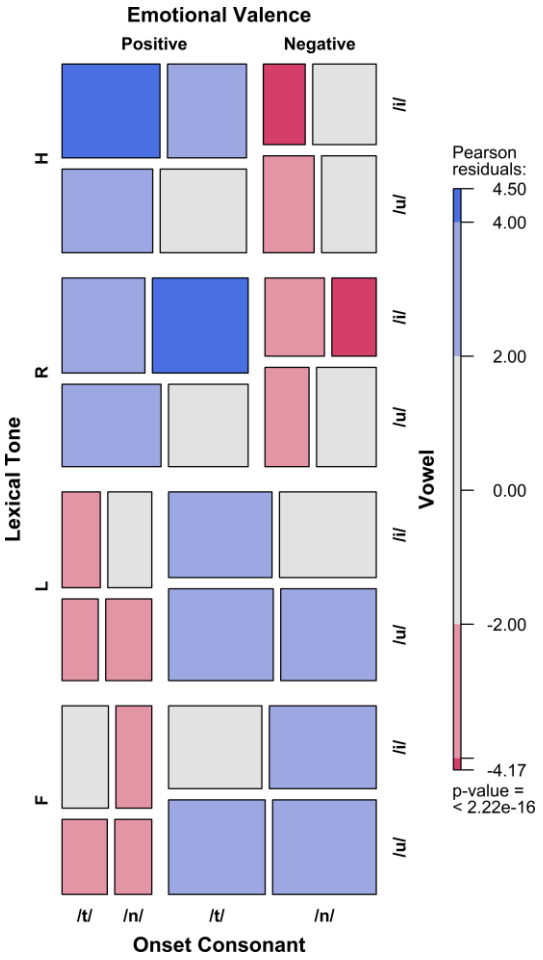


**Figure 3.12**

*Valence frequency variations across the interaction of lexical tones, vowels, and onset consonants in the CVL dataset.*




## 3.5 Discussion

The present study investigated how lexical tone, vowel, consonant, and their interactions contribute to the ratings of emotional arousal and valence by native Standard Chinese speakers. Results showed significant toneme–arousal, consonant–arousal, toneme–valence, and vowel–valence associations as predicted, suggesting that native

Standard Chinese speakers consider both phonemes and tonemes when interpreting the affective meanings of auditory stimuli. Our results confirmed that lexical tones shape emotional responses, demonstrating their inherent iconicity in nonce words, consistent with findings in real words (Zheng et al., 2025). Additionally, the study replicated findings on the correspondences between phonemes and emotions reported in the literature. While the absence of consonant-valence and vowel-arousal correspondences remains unclear, our results lend further support to the view that there are some systematic correspondences between sounds and meanings cross-linguistically (Svantesson, 2017).

One novel discovery of this study is the predominant impact of tonemes over consonants on arousal rating and over vowels on valence rating. Lexical-prosodic information or prosodic cues in vocalisations have been considered a facilitator in interpreting iconic meanings across languages (Dingemanse et al., 2016; Perlman, Clark, et al., 2015; Stel et al., 2012; Thompson, 2018). Our results highlight that, although lexical tones in a tonal language primarily serve to distinguish word meanings, speakers nevertheless prioritise tonemes in interpreting the overall affective meaning of a nonce word, which is particularly evident in cases when the tonemes and phonemes are not synergistic (e.g., /i4/ versus /u1/). This discovery expands the view of prosody as a facilitating element, highlighting its critical role in shaping affective responses.

The associations between speech elements and emotional responses can be attributed to the acoustic properties of phonemes and tonemes. For instance, the consonant-arousal association, where /t/ is linked to high arousal and /n/ to low arousal, can be explained by their spectral characteristics. The consonant /t/ involves a closure and

explosive release of airflow, producing a sudden burst and higher spectral components associated with increased arousal (see the consonant-arousal association in Aryani et al., 2018; see the consonant and size association in Winter & Perlman, 2021). In contrast, nasals like /n/ have lower acoustic frequencies, often connoting slower, rounder, and softer meanings. This pattern is evident in studies on the naming of birds (Berlin, 2006) and the tactile perception of textures (Sakamoto & Watanabe, 2018).

Similarly, the "/u/–negative, /i/–positive" correspondence (implicated in Aryani et al., 2013; see Körner & Rummer, 2022 for /i/–positive associations, and 2023 for both /i/–positive and /u/–negative associations) can be understood through both acoustic and articulatory perspectives. Research has demonstrated that vowel formants (F1, F2, and F3) play a crucial role in vowel-size and vowel-shape correspondences (Knoeferle et al., 2017), indicating that individuals rely on acoustic cues as the basis for iconic judgments. Specifically, front vowels like /i/, characterised by a higher second formant (F2) and an intrinsically higher fundamental frequency (*fo*), are consistently perceived as more positive (e.g., conveying a non-threatening attitude). In contrast, back vowels like /u/, which exhibit a lower second formant (F2) and an intrinsically lower *fo*, are associated with negative emotions (e.g., threat or danger) (see discussions in Ohala, 1994; Perlman & Cain, 2014). Articulatorily, /i/ involves muscle tension associated with smiling, whereas /u/ involves antagonistic muscle tension linked to negative emotions (Körner & Rummer, 2023). These factors are not mutually exclusive, contributing to its positive connotations.

In the case of vowel-only words, the abundance of homophones and interjections might have influenced participants' ratings. For

instance, the vowels /u1/ and /i2/ might be associated with the Chinese characters 污 and 咦, meaning dirty and surprise, respectively. However, similar vowel-valence associations were found in the CVL-nonce word, suggesting that these effects are not solely dependent on specific conventionalised lexical items or interjections. Additionally, the use of auditory presentation minimises the influence of orthography or typography (Cuskley et al., 2017).

Toneme-arousal and toneme–valence associations are likely influenced by pitch characteristics. Higher arousal ratings for T4- and T2-syllables, compared to T1- and T3-syllables, are possibly due to more dynamic pitch variations associated with the falling and rising tones. Although average pitch height might play a role (e.g., T4/T2 versus T3), its impact on arousal is marginal, as evident in the lack of difference between T4/T2 and T1. This supports the idea that pitch dynamics are more crucial in predicting arousal than pitch height (Zheng et al., 2025). See further discussions below. Previous studies show that T4 is generally more arousing than T2 in disyllabic words (Zheng et al., 2024, 2025). However, the similar arousal levels for T4 and T2 in the nonce word context suggest that pitch dynamics may overshadow semantic meanings, necessitating replication and further research to clarify these nuances.

Positivity ratings for T1 and T2 (versus T4 and T3) may, in part, relate to their higher pitch levels (e.g., T1 versus T4 and T3; T2 versus T3) and larger pitch ranges (e.g., T2 versus T3). Notably, while T4 has a greater range and higher pitch level than T2, it may be its falling contour that contributes to a negative valence. This reinforces the notion that rising contours (e.g., T2) are more likely to associate with positive affect and refer to positive objects, compared to falling contours (e.g., T4) (Frick, 1985; Ohala, 1994, p. 327; Yap et al., 2014;

Zheng et al., 2024). We'd like to stress the need for replication studies to understand how pitch contour trajectories, overall pitch level, average pitch height, and pitch range interact in predicting valence.

In brief, our current study confirms that the main speech dimensions, i.e., consonant, vowel, and tone, demonstrate iconicity in the emotional interpretations of SC monosyllabic nonce words. Our findings highlight the significant role of phonemic pitch in lexical tones' affective iconicity, despite its primary function of distinguishing lexical meaning in tonal languages. The concept of pitch iconicity is often discussed in the context of the "frequency code" (Gussenhoven, 2016; Hinton et al., 2006; Ohala, 1983, 1984, 1994), which suggests that higher (acoustic) frequency is associated with higher arousal and positive valence. The associations we found in this study, where dynamic pitch contours play a more dominant role, partially supported the frequency code hypothesis. Specifically, the intrinsic *fo* of phonemes is correlated with perceived emotional arousal (for consonants) or valence (for vowels). The *fo* difference between T2/T4 and T3 likely contributed to their arousal contrast, and differences in *fo* between T1/T2 and T3/T4 may explain their valence choice differences. However, it is important to note that the arousal and valence differences among lexical tones are more likely influenced by other pitch characteristics, such as pitch dynamics, pitch ranges, and pitch slope. Thus, we cautiously conclude that our results add another case where the frequency code may be argued to structure a sound-symbolic connection between pitch and its affective connotations, although the mapping between pitch characteristics and their possible interpretations is by no means straightforward (Morton, 1977; Perlman, 2024; Winter et al., 2021).

Further research is essential to explore the associations between vowels and arousal, and consonants and valence, particularly given their prevalence across languages. Future studies should also aim to clarify the stability and generalizability of the predominant effect of tonemes, particularly by expanding the range of phonemes and examining different linguistic contexts and tone systems. Such studies will help determine whether the associations we found are universally applicable or language-specific, and if the latter, in what specific ways and why. Additionally, further investigation into how the perceptual qualities of pitch and its underlying acoustic properties influence emotional interpretation could provide deeper insights into the mechanisms underpinning affective iconicity. Addressing these questions will not only enhance our theoretical understandings of sound-meaning associations in general but also have practical implications for fields like speech synthesis, language learning, and cross-cultural communication.

# Chapter 4
# Adaptive significance of tonemes for affective iconicity in Standard Chinese

A shorter version of this chapter is published as: Zheng, T., Levelt, C. C., & Chen, Y. (2024). *The adaptive value of Mandarin tones for affective iconicity*. Proc. Speech Prosody 2024, 562–566. https://doi.org/10.21437/SpeechProsody.2024-1

**Abstract**: Iconicity is prevalent across languages, yet its underlying mechanisms remain debated. While previous studies have explained iconicity through perceptual analogy (see a review in Sidhu & Pexman, 2018), recent research proposes an adaptation perspective. Specifically, Adelman et al. (2018) found that the first phoneme of a word predicted its valence better than subsequent phonemes in Indo-European languages. Additionally, word-initial phonemes that are produced more rapidly tend to convey negativity rather than positivity, especially in English and Spanish. This led Adelman et al. to suggest that iconicity may have evolved due to the adaptive advantage of negative emotions in survival and social interactions. However, this view is largely based on the affective iconicity of segmental phonemes in Indo-European languages and focuses only on valence. What has left open is whether emotional arousal, a more ancient and innate dimension of emotion (Darwin, 1998), may also play a role. Additionally, many languages of the world use not only segments but also suprasegmental lexical tones to distinguish word meanings. The primary perceptual cue for tone is pitch variation, which expresses emotions as well. Previous studies have demonstrated affective iconicity in Standard Chinese lexical tones (Zheng et al., 2025, under revision), but little is known about their adaptive significance. Using three Standard Chinese corpora with valence and arousal ratings for bisyllabic words, this study employed hierarchical linear regression models to examine whether lexical tones in the first and second syllables predict a word's emotional ratings differently. Results revealed that valence is predicted by first-syllable tone, with positive words more likely to carry a rising tone than a falling or high-level tone. This positional effect of tone for valence in a Sinitic tonal language conceptually replicates the adaptive significance of the first phoneme

for valence found in segments of an Indo-European language. Arousal, in contrast, is predicted by the tones of both syllables, with high-arousing words more likely to contain a falling tone than a rising or low-dipping tone. This distributed effect of tone on arousal suggests that the spread of crucial information, particularly through iconicity, may have evolved due to the amplified reactions and evaluations associated with high-arousing stimuli. This extends the adaptive significance of valence to encompass arousal, highlighting its role in shaping emotional and cognitive processes. These findings lend further support to the proposal that (affective) iconicity is an adaptation providing communication efficiency in speech.

**Keywords**: Affective iconicity, Standard Chinese, Emotional arousal, Emotional valence, Adaptation, First syllable tone, Second syllable tone

## 4.1 Introduction

Iconicity refers to the subjective and non-arbitrary association between a signal's form and meaning as perceived by language users, a phenomenon found across languages and linguistic components (Adelman et al., 2018; Aryani et al., 2018; Perniss et al., 2010; Winter et al., 2023; Yap et al., 2014; Zheng et al., 2025, under revision). Among its various dimensions, affective iconicity, where linguistic forms convey emotional meaning, has been widely studied across languages (e.g., German, English, Spanish, Polish, and Dutch in Adelman et al., 2018; German in Aryani et al., 2018; English in Whissell, 2003; Standard Chinese Yao et al., 2013; Standard Chinese and English in Yu et al., 2021).

Emotional expression is essential for survival and social interaction, guiding threat detection (e.g., recognising danger), adaptive responses (e.g., fight-or-flight responses), and social communication (Kashima et al., 2020). Psycholinguistic studies show that emotionally-valenced words (positive or negative) are processed more efficiently than neutral words, with faster reaction times and enhanced neural responses in lexical decision tasks (e.g., Kissler et al., 2007; Kousta et al., 2009). This processing advantage may reflect an adaptive mechanism where phonetic features help encode emotional meaning.

Supporting this, Adelman et al. (2018) demonstrated that across Indo-European languages (e.g., English, Spanish, Dutch, German, and Polish), phonemes in the general lexicon predict the emotional valence of words. Notably, in English and German, phonemes with shorter articulation times were found to be more likely to initiate negative words, while those with longer articulation times correlated better with

positive words. This form–affect association, termed emotional sound symbolism by Adelman et al. (2018), suggests that iconicity may have evolved to enhance communication efficiency, akin to alarm calls in non-human species. They argued that the phoneme-based structure reflects the adaptive significance of emotional sound symbolism, which enhances communication efficiency by prioritising negative signals. This efficiency further reinforces the emergence and development of iconicity in speech.

In this chapter, we use the term "affective iconicity" to refer to emotional sound symbolism (Adelman et al., 2018) with two important disclaimers. First, unlike traditional sound symbolism, which links specific phonetic features to particular meanings (e.g., the bouba/kiki sound-shape symbolism Ćwiek et al., 2021), we use "iconicity" to encompass broader forms, such as speech, sign, and word (Winter et al., 2023). Second, following the embodied view of language, we use "affect" to encompass behavioural and neural emotional responses across species and cultures (Aryani et al., 2019). Although this study does not directly examine neural responses, insights from neuroscience inform our understanding of affective iconicity's adaptive functions. See more details in the following paragraphs.

Existing theories or hypotheses on the emergence of iconicity or sound symbolism often attribute it to general cognitive faculties, such as analogy and generalisation (e.g., Monaghan et al., 2014; Sidhu & Pexman, 2018). The statistical co-occurrence hypothesis, for instance, suggests that sound-meaning correspondences emerge because certain features frequently co-occur (see an overview in Sidhu & Pexman, 2018). This is supported by findings that the articulatory muscles involved in phoneme production overlap with those used in emotional expression (e.g., Körner & Rummer, 2023). While cognitive analogies

explain why affective iconicity is possible, Adelman et al. (2018) argue that it fails to explain the positional effect of phonemes, especially how the temporal characteristics of phonemes influence the words' valence.

However, the adaptation account requires further refinement to clarify its scope and generality. First, while it provides a compelling explanation for valence-based sound symbolism, it treats emotional arousal (the degree of activation, another core dimension of emotion) as a control variable, rather than examining its independent role in shaping sound-affect associations (Adelman et al., 2018). Specifically, Adelman et al. examined how phonemes contribute to valence ratings after accounting for arousal. Yet, valence and arousal are both fundamental to emotional processing and may play distinct yet complementary roles in adaptation. Arousal, considered more ancient and evolutionarily fundamental than valence (Darwin, 1998), which reflects immediate and automatic physiological activation (Russell, 1980), whereas valence involves cognitive appraisal and contextual interpretation (Citron, 2012).

Despite these differences, valence and arousal are interrelated. They often form a V-shaped relationship, where both positive and negative emotions are typically high-arousing, while neutral emotions tend to be low-arousing (Kuppens et al., 2013). Their interaction plays a crucial role in evaluative processing (Robinson et al., 2004), suggesting that ignoring arousal's distinct role risks overlooking key aspects of affective iconicity. To fully understand the adaptive functions of affective iconicity, it is essential to consider both arousal and valence as independent yet interacting factors.

Second, prior research on affective iconicity primarily focused on segmental phonemes in Indo-European languages, while ignoring that suprasegmental features, such as lexical tone, play a similar role in

distinguishing word meanings, and they may also carry adaptive significance in emotional communication. Lexical tones in many languages rely on pitch variations, which are also a primary cue for emotional intonation, to distinguish lexical meaning, similar to phonemes. Research shows that pitch at the prosodic level correlates with emotional arousal and valence expression. Specifically, high-arousing emotions (e.g., fear, excitement) generally correlate with a higher pitch level and average pitch height, a wider pitch range, and a steeper pitch slope, in comparison to low-arousing emotions (Bänziger & Scherer, 2005; Laukka et al., 2005; Scherer et al., 2003). Positive emotions generally correlate with a higher pitch level and average pitch height and a wider pitch range, while negative emotions show the opposite (Belyk & Brown, 2014; Kamiloğlu et al., 2020; Laukka et al., 2005; c.f., Scherer & Oshinsky, 1977).

Studies on Standard Chinese (SC) suggest that lexical tones evoke similar emotional valence and arousal effects due to their phonetic similarities to emotion-induced pitch modulations at the sentence level, despite the conventional view that lexical tone is distinct from affective meaning (Zheng et al., 2025). For example, falling-falling tone sequence (i.e., T4T4) in both general lexicon and nonce words is associated with higher arousal than rising-rising tones (i.e., T2T2); and in nonce words, falling-falling tone sequence is associated with negative valence, while rising-rising tone sequence with positive valence. These findings suggest that pitch variations, despite their phonemic function at the lexical level in tone languages, remain iconic to convey affective meaning.

Zheng et al. (2025) argue that the tone–affect associations align with the frequency code hypothesis (Gussenhoven, 2016; Ohala, 1984), which posits that high or rising *fo* signals friendliness, excitement, or

non-threat, which can promote social bonding. On the other hand, low or falling *f0* signals authority, dominance, or threat, aiding conflict resolution or warning. These associations are also evolutionarily meaningful and may suggest an adaptive significance of tones similar to phonemes in Indo-European languages, lending support to the idea that iconicity in speech emerges and persists due to its evolutionary advantages. This study explored the positional effects of lexical tones in affective iconicity and further extended the current findings on valence to arousal.

Using three Standard Chinese corpora (CAWS, NORM, DCAWS; Y. Wang et al., 2008; X. Xu et al., 2022; Zheng et al., 2025), we analysed how the first- and second-syllable tones of bisyllabic words predict their emotional arousal and valence ratings. We included all four lexical tones (also known as tonemes): High-level tone (T1), Rising tone (T2), Low-dipping tone (T3), and Falling tone (T4) for each syllable. To minimise confounding effects, we incorporated factors such as word frequency (Brysbaert et al., 2018), onset consonant category (Louwerse & Qu, 2017), and part-of-speech (Perry et al., 2015) into our model analyses (as done in Zheng et al., 2025). Additionally, we examined the relationship between arousal and valence across each lexical tone and syllable position to further clarify their potential interactions in affective iconicity.

In this exploratory study, we hypothesised that lexical tone contributes to signalling both valence and arousal. Furthermore, we predicted that pitch features, including pitch level, average pitch height, pitch range and slope, and pitch contour, contribute to predicting arousal and valence ratings, in line with the effects of pitch in non-tonal languages to express emotion (Bänziger & Scherer, 2005; Frick, 1985;

Laukka et al., 2005) and prior research on tone–arousal and tone–valence iconicity (Zheng et al., 2025, under revision).

Importantly, we propose that the adaptive significance of tone–affect iconicity manifests differently for valence and arousal, given their distinct cognitive and neural processing characteristics. Specifically, we predicted a positional effect, such that the tone of the first syllable would have a stronger influence on valence, given that valence is extracted earlier than arousal in word and picture processing (Gianotti et al., 2008) and prioritised in phonemes at the beginning of words (Adelman et al., 2018). Moreover, lexical tones with higher overall pitch level, higher average pitch height, wider pitch ranges, and/or steeper slopes are more likely to predict negative valence. In contrast, lexical tones with lower overall pitch level, lower average pitch height, narrower pitch ranges, and/or more gradual slopes are more likely to predict positive valence. For example, typically exhibits sharper, more rapid pitch changes over a short temporal period, which may contribute to its association with negative affect. These patterns would reflect the adaptive significance of negative valence in survival and communication.

In contrast, we predict a distributed effect for arousal, as arousal-related information unfolds over a longer time course. ERP studies show that higher-arousal emotions elicit distinct neural responses, notably the Early Posterior Negativity (EPN), which reflects enhanced attention to emotional stimuli, and the Late Positive Component (LPC), which indicates more sustained and detailed emotional processing (Fischler & Bradley, 2006; Herbert et al., 2008; Kissler et al., 2007, 2009; Schacht & Sommer, 2009). This suggests that arousal-related effects of lexical tone may not be limited to the initial syllable but rather extend throughout the word (i.e., distributed effect). Arousal has been

shown to amplify reactions, leading to intensified evaluations and enhanced long-term memory for events. This suggests that the arousal dimension conveys information about the urgency or importance of stimuli (Storbeck & Clore, 2008). Given these effects, it is plausible that arousal, particularly high-arousing stimuli, contributes to the adaptive significance and emergence of iconicity, as its benefits in facilitating communication are well-documented. Furthermore, lexical tones with a higher average pitch height and more dynamic pitch variations, such as a wider range and steeper slopes, tend to be linked to high arousal, whereas tones with a lower average pitch height and less dynamic pitch variations tend to be associated with low arousal. For example, T4 is commonly linked to high arousal.

## 4.2 Method

### 4.2.1 Corpus datasets

We analysed three corpus datasets, wherein all bisyllabic words were annotated with their underlying lexical tones.

The Chinese Affective Word System (CAWS) dataset (Y. Wang et al., 2008) comprises arousal and valence ratings for 1,492 bisyllabic words. A total of 64 undergraduate students (aged 18–21 years) rated arousal and valence on a 9-point Likert scale. A retest was subsequently conducted with 30 undergraduates (aged 18–21 years). Arousal values ranged from low (1, e.g., peaceful) to high (9, e.g., alert), while valence ratings spanned from negative (1, e.g., unhappy) to positive (9, e.g., happy).

The Affective Norms for Chinese Words dataset (X. Xu et al., 2022), hereafter NORM, contains arousal and valence ratings of 9,573 bisyllabic words in Standard Chinese. Arousal ratings were collected

from 1,189 participants, and valence ratings from 1,232 participants (aged 18–62 years). Most participants (96.9%) received a college-level education or higher. The NORM dataset employed a 5-point Likert scale for arousal (0–4) and a 7-point scale for valence (-3 to +3).

The Developed Chinese Affective Word System (DCAWS) dataset (Zheng et al., 2025) includes arousal and valence ratings for 1,500 bisyllabic words. Arousal ratings were obtained from 51 participants (aged 18–27 years), while valence ratings were provided by 53 participants (aged 18–24 years). Ratings for both dimensions were recorded using a 1–100 slider scale with 1-point increments, ranging from low to high or negative to positive.

## 4.3 Data analysis

We first visually inspected the graphs displaying emotional arousal and valence ratings for all words to examine the relationship between the ratings of arousal and valence in each dataset.

We then explored how lexical tone in different positions (first vs. second syllable) predicts the emotional arousal and valence of bisyllabic words, separately in the three datasets. To observe how lexical tone contributes to account for the arousal and valence variances, we conducted a set of hierarchical linear regression analyses (HLM) in each dataset (R Core Team, 2023). The HLM helps to determine whether lexical tone explains a statistically significant amount of variance in the arousal and valence after accounting for all other variables, such as word frequency, onset consonant, and part-of-speech. The HLM consisted of three models: the first model includes factors such as word frequency, onset consonant, and part-of-speech, the second model added an individual tone (either on the 1st or 2nd syllable) beyond the factors in the first model, the third model includes

tones of both syllables along with the other factors in the first model. The dependent variables were the aggregated emotional arousal and valence rating scores. The HLM results highlighted the influence of individual tones on emotional ratings. Pairwise multiple comparisons, adjusted with Bonferroni correction using the *emmeans* package (Lenth, 2023), were employed to examine differences across individual tones.

Additionally, generalised linear mixed models (GLMMs) were conducted only for the DCAWS dataset, given that this is the only corpus with individual rating data points available. GLMMs for both arousal and valence consist of individual lexical tones as the fixed effect, and by-subject and by-item intercepts as the random effects to account for potential heterogeneity among raters and items.

## 4.4 Results

### 4.4.1 V-shaped relationship between arousal and valence

Figures 4.1–4.3 illustrate the relationship between emotional arousal and valence in CAWS, NORM, and DCAWS datasets, respectively. The numerical values on the axes correspond to Likert/Slider scale ratings. A distinctive V-shaped relationship emerged between arousal and valence in all datasets, confirming that positive and negative valence tend to associate with high arousal, while neutral valence aligns with low arousal. In each panel of the figure, “H” is used for high-level tone (T1), “R” for rising tone (T2), “L” for low-dipping tone (T3), and “F” for falling tone (T4).

**Figure 4.1**

*The V-shaped relationship between emotional arousal and valence in the CAWS dataset.*

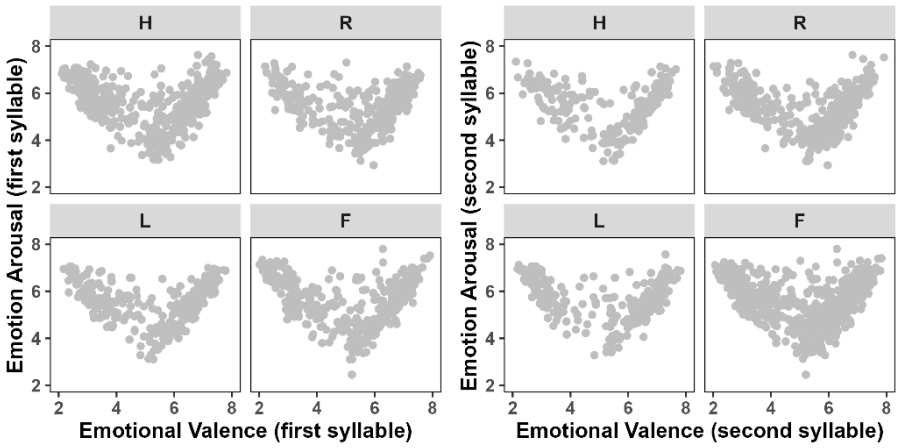



**Figure 4.2**

*The V-shaped relationship between emotional arousal and valence in the NORM dataset.*

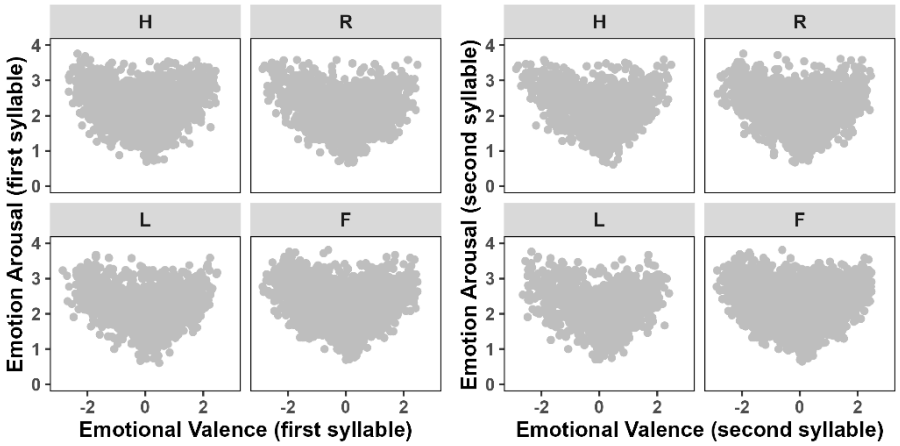
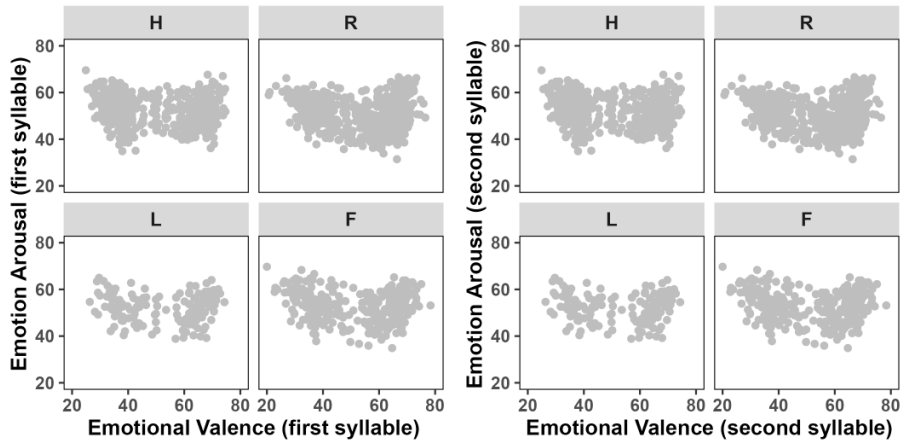


**Figure 4.3**

*The V-shaped relationship between emotional arousal and valence in the DCAWS dataset.*




### 4.4.2 Effects of individual lexical tone on arousal and valence ratings

Regression diagnostics of HLMs confirmed the fulfilment of linearity, homoscedasticity, and normality assumptions, with no identified collinearity issues. No violations of independence existed for the average scores of arousal and valence ratings across participants. GLMMs diagnostics also showed no violation of linear regression assumptions.

#### *4.4.2.1 Arousal ratings*

Regarding the arousal ratings, individual tones in both the first and second syllables significantly impacted the unique variance across the three datasets, showing a distributed effect. In the CAWS datasets, the model with first-syllable tones demonstrated a significant predictive effect on emotional arousal ratings ($F$ = 5.10, $p$ < 0.01). Notably, the first-syllable tone accounted for a significant portion of unique

variance in the arousal ratings ($\Delta R^2$ = 0.88%, $p$ < .001), with the falling tone ($M$ = 5.75) predicting higher arousal levels than the rising tone ($M$ = 5.50, $p$ < .001). Likewise, the model with second-syllable tones demonstrated a significant predictive effect on emotional arousal ratings ($F$ = 5.56, $p$ < 0.001). The results indicated a significant influence of the second-syllable tone on the unique variance in arousal ratings ($\Delta R^2$ = 0.96%, $p$ < .001). Further comparisons revealed the larger arousing effect of the falling tone ($M$ = 5.72) in comparison to the rising tone ($M$ = 5.50, $p$ < .01). The third model showed consistent effects of monosyllabic tones on emotional arousal ratings, with first and second syllable tones combined predicting emotional arousal ratings better than single-tone models.

The same HLM analyses were conducted for the NORM corpus and revealed significant predictive effects of both first and second-syllable tones on emotional arousal ratings ($F$ = 5.06, $p$ < 0.01; $F$ = 5.56, $p$ < 0.001). The first-syllable tone and second-syllable tone accounted for a considerable portion of unique variance in the arousal ratings ($\Delta R^2$ = 0.15%, $p$ < .001; $\Delta R^2$ = 0.22%, $p$ < .001). Pairwise comparisons revealed higher arousal levels for the falling tone ($M$ = 2.07) compared to the rising tone ($M$ = 2.02, $p$ < .01) and the low tone ($M$ = 2.02, $p$ < .05) at the first syllable. Similarly, at the second syllable, the falling tone ($M$ = 2.07) showed a larger arousing effect compared to the rising tone ($M$ = 2.01, $p$ < .001) and the low tone ($M$ = 2.03, $p$ < .05). The third model demonstrated consistent effects of monosyllabic tones on emotional arousal ratings, and the model with tones at both first and second syllable together were preferred in model comparisons.

HLM analyses for the DCAWS corpus revealed a significant predictive effect of both first- and second-syllable tones on emotional arousal ratings ($F$ = 4.42, $p$ < 0.01; $F$ = 3.33, $p$ < 0.05). Notably, both

first-syllable and second-syllable tones accounted for a significant portion of the unique variance in the arousal ratings ($\Delta R^2$ = 0.84%, $p$ < .001; $\Delta R^2$ = 0.64%, $p$ < .001). Pairwise comparisons revealed higher arousal levels for the falling tone ($M$ = 52.7) compared to the rising tone ($M$ = 51.0, $p$ < .01) at the first syllable. Although pairwise comparisons showed no significant differences among second-syllable tones, the falling tone had the highest arousal rating score ($M$ = 52.1), while the rising tone had the lowest ($M$ = 50.7) at the second syllable.

GLMM analysis for the DCAWS corpus revealed a similar predictive effect of both first- and second-syllable tones on emotional arousal ratings. Comparisons between the null model and the interest model provided compelling evidence that the first-syllable tone significantly predicted emotional arousal ratings ($F$ = 21.73, $p$ < .001). Pairwise multiple comparisons revealed that both the falling tone ($M$ = 53.0) and the high-level tone ($M$ = 52.5) predicted higher arousal ratings than the rising tone ($M$ = 51.1, $p$ < .001 and $p$ < 0.01, respectively). Similarly, the second syllable tone significantly predicted emotional arousal ratings ($F$ = 13.39, $p$ < .01). Pairwise multiple comparisons showed that the falling tone ($M$ = 52.5) yielded higher arousal ratings than both the rising tone ($M$ = 50.7, $p$ < .05) and the low tone ($M$ = 51.3, $p$ < .05). Again, the model with two syllable tones showed a better fitness than single-tone models.

#### *4.4.2.2 Valence ratings*

Regarding the valence ratings, only the lexical tone of the first syllable, demonstrating a positional effect, showed a noteworthy influence within the NORM ($F$ = 3.76, $p$ < 0.05, $\Delta R^2$ = 0.12%) and DCAWS ($F$ = 3.60, $p$ < 0.05, $\Delta R^2$ = 0.68%) datasets. Specifically, the rising tone ($M$ = 0.13) was perceived as more positive compared to both the high-level

tone ($M$ = 0.04, $p$ < .05) and the falling tone ($M$ = 0.04, $p$ < .05) in the NORM dataset. Similarly, in the DAWS dataset, the rising tone ($M$ = 53.2) was perceived as more positive compared to both the high-level tone ($M$ = 50.4, $p$ < .05) and falling tone ($M$ = 50.3, $p$ = 0.058).

Figures 4.4–4.6 illustrate further individual tones' different contributions to the emotional arousal and valence ratings in the three corpora, respectively. In these plots, lexical tones are organised in ascending order based on their average rating scores for each syllable (first syllable on the left) and emotional dimension in each panel (arousal rating scores on the top). The mean rating scores for each tone are connected to show the overall trends. Take the left-upper panel of Figure 4.4 as an example. For the first syllable, the rising tone is associated with the lowest mean arousal rating score (score 5.50), while the falling tone with the highest mean arousal rating score (score 5.75).

**Figure 4.4**

*Arousal and valence rating scores for each tone at the 1st and 2nd syllable in the CAWS dataset.*

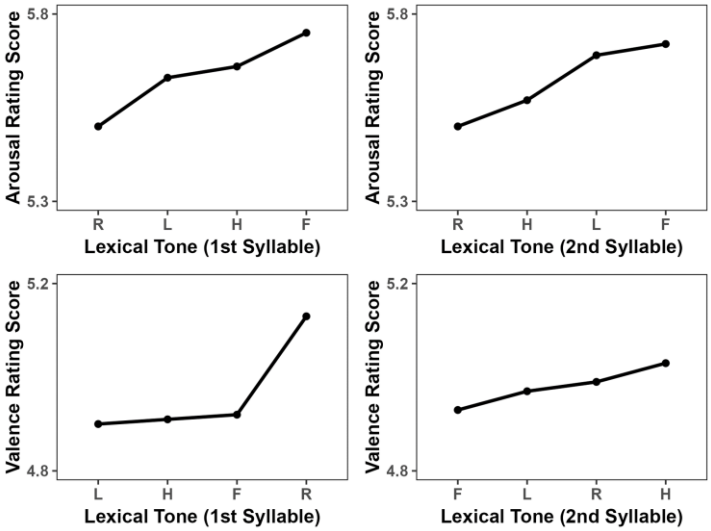



**Figure 4.5**

*Arousal and valence rating scores for each tone at the 1st and 2nd syllable in the NORM dataset.*

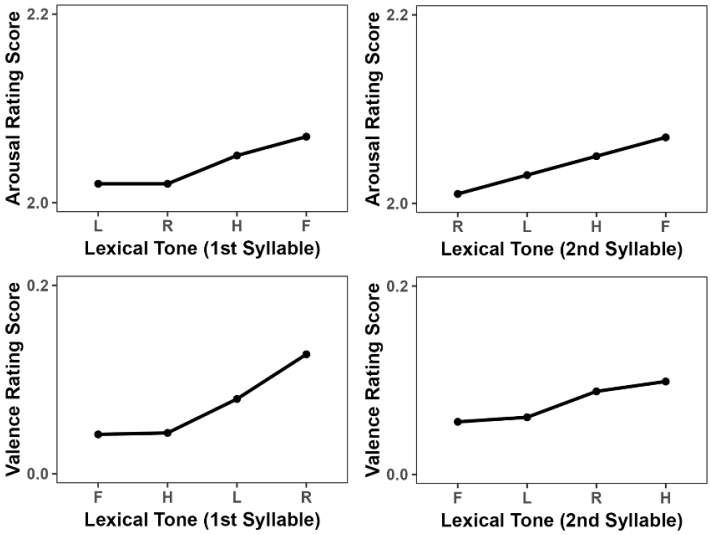
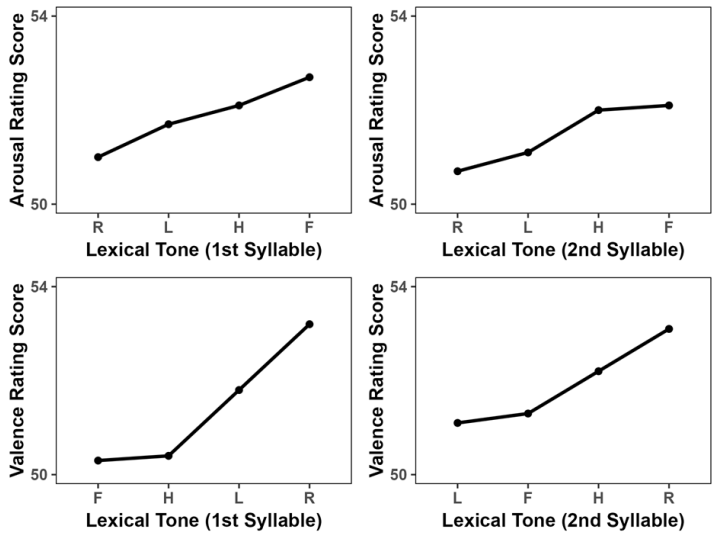


**Figure 4.6**

*Arousal and valence rating scores for each tone at the 1st and 2nd syllable in the DCAWS dataset.*




## 4.5 Discussion

This study explored whether lexical tones exhibit adaptive significance for both emotional arousal and valence in affective iconicity. Specifically, we investigated whether positional effects exist for valence and distributed effects for arousal when using lexical tones to predict emotional arousal and valence ratings of bisyllabic Standard Chinese words in three corpora (CAWS, NORM, DCAWS; Y. Wang et al., 2008; X. Xu et al., 2022; Zheng et al., 2025). Our results confirmed the commonly observed V-shaped relationship between arousal and valence (Citron et al., 2014; Kuppens et al., 2013), reinforcing their interdependent roles in emotional processing. This suggests that both arousal and valence should be taken into account in the investigation of affective iconicity to provide a comprehensive understanding of nuanced emotional expressions.

By incorporating key lexical factors (e.g., word frequency, part of speech, onset-consonant category), we found that individual tones in the first and second syllables of SC bisyllabic words significantly influenced arousal ratings, demonstrating a distributed effect of tone on arousal. Specifically, the falling tone (T4) correlated with high-arousing words, whereas the rising and low-dipping tones (T2 and T3) correlated with low-arousing words. In contrast, only first-syllable tones influenced valence ratings, indicating a positional effect; words beginning with a falling or high-level tone (T1 or T4) were more likely to be associated with negative valence. Although the magnitude of these tonal effects on arousal and valence was relatively small, it is comparable to the effects of phonemes on valence ratings found in prior research (Adelman et al., 2018), suggesting a subtle yet consistent role of sound in encoding emotional meaning across languages.

The positional effect of lexical tones on valence parallels findings from phoneme-based research, where the first phoneme of a word critically influences valence (Adelman et al., 2018). Adelman et al. (2018) proposed the concept of *negative priority*, suggesting that negative information is given preferential attention due to its survival relevance. This leads to faster pronunciation of the first phoneme in negative words, as they signal urgent or threatening information. However, in Standard Chinese, lexical tone *fo* contrasts are predominantly realised within the rhyme of a syllable, especially for syllables with a voiceless onset. So, the negative priority account attributed to temporal features of segmental affective iconicity may have limited applicability to lexical tones. Instead, the pitch characteristics of tones, particularly the falling contour (T4), may better explain the tendency for negative words to begin with falling tones. This aligns with the frequency code hypothesis (Ohala, 1994),

which posits that falling pitch contours signal threat or danger. By initiating a falling contour, speakers can efficiently convey urgency or negative valence, enhancing communication efficiency in contexts where rapid transmission of critical information is vital.

The contrastive effect of rising (T2) and falling (T4) contours also aligns with the conceptual metaphor of "Good is up, bad is down" (Ley-Flores et al., 2022; Yap et al., 2014). Worth noting is that the high-level tone (T1) is not as positively valenced as the rising tone (T2), which may be taken to emphasise the importance of rising contour in conveying positive emotional meaning (Laukka et al., 2005; cf. Scherer & Oshinsky, 1977). While these findings highlight the role of pitch direction in valence iconicity, further research is needed to disentangle the effects of average pitch height and pitch range, as well as the direction of pitch contour on valence ratings. This will provide a more nuanced understanding of how tonal (pitch) features contribute to affective iconicity.

For arousal, tones of both syllables influenced ratings, with the falling tone (T4) consistently associated with higher arousal. This association may be attributed to the pitch characteristics of this tone: the falling tone in SC is realised with a higher average pitch height, greater pitch range, and steeper slope compared to rising and low-dipping tones. These features likely contribute to the perception of urgency or intensity, aligning with the adaptive role of high-arousing stimuli in signalling importance or threat. The adaptive significance of high-arousing stimuli is well-documented, as such stimuli tend to capture attention and enhance memory, likely due to their role in signalling urgency or emotionally significant information (Storbeck & Clore, 2008). The association of T4 with high-arousing words may reflect an evolutionary mechanism where high-arousing tonal features

amplify the salience and memorability of urgent or emotionally significant messages.

Moreover, the link between arousal and the social transmission of information further supports the adaptive role of tonal effects. Berger (2011) demonstrated that participants experiencing high emotional arousal, whether positive or negative, were more likely to share neutral media content (e.g., an article or video) encountered afterwards. For our results, the correlation between falling tones (T4) and high-arousing words may facilitate the dissemination of emotionally charged or urgent information within a social group. This reinforces the communicative efficiency of tonal iconicity, as arousal plays a crucial role in attention and sustained processing. It is plausible that arousal information has been evolutionarily preserved in iconicity to optimise social communication, ensuring that emotionally significant messages are prioritised and retained.

Some may argue that arousal reflects the level of emotional activation, often correlating with stimulus intensity and increasing attentional engagement with more intense signals. Under this view, certain tonal patterns, such as high tones, may capture attention due to their acoustic salience, independent of emotional content (see insights from consonants in Baroni, 2014). However, when investigating the mechanisms underlying the relationship between tonal patterns and arousal, it is crucial to distinguish between arousal and attention, which are intertwined yet distinct concepts, with each engaging different neural substrates (Paus, 2000).

Arousal typically refers to the physiological and psychological changes influencing the overall state of alertness or readiness to respond to stimuli. In contrast, attention involves the selective focus on specific aspects of stimulation. Arousal can impact attentional

processes by increasing alertness and enhancing stimulus salience. Elevated arousal can lead to increased vigilance, manifested by the modulation of neural activity in sensory cortices, which, in turn, facilitates attention and influences subsequent cognitive performance (Asutay & Västfjäll, 2017). In our data, a falling tone seems to evoke higher arousal than a rising tone, which could lead to more engaged attention and conscious processing (but see (Lialiou et al., 2024) for an attention-orienting effect of a rising tone in German compared to a falling tone). Further research is necessary to fully understand the processing of pitch contour in relation to arousal and valence and its potential impact on cognitive processing.

It is important to note that the perceptual and cognitive analogy view of iconicity (e.g., Sidhu & Pexman, 2018) and the adaptive view (e.g., Adelman et al., 2018) are not mutually exclusive. Adaptive significance might have driven the co-occurrence of muscle movements involved in both the articulation of specific sounds and the expression of emotions, enhancing communicative efficiency and offering survival advantages.

In summary, this study extends the adaptation view of affective iconicity in three key ways: from segmental to suprasegmental features, from valence-only to both valence and arousal, as well as from Indo-European to Sinitic languages. Our findings emphasise the role of linguistic pitch in emotional communication, the distinct yet complementary contributions of emotional arousal and valence to emotional communication, and the potential cross-linguistic and adaptive or evolutionary significance of affective iconicity. However, the sources or acoustic correlates of the adaptive significance may differ between phonemes and tonemes. Further investigation is necessary to explore the positional influence of both segmental

phonemes and lexical tones on affective iconicity. It is plausible that there exist two distinct mechanisms for the positional effects of segmental and suprasegmental tonal iconicity, or a more general mechanism that extends and further refines the negative priority developed for segments (as defined in Adelman et al., 2018).

# Chapter 5
# Development of hemispheric lateralisation of lexical pitch processing: A cross-linguistic study in Dutch and Japanese infants

**Abstract**: Infants' ability to discriminate speech contrasts shifts from broadly perceiving contrasts across languages to focusing on those relevant to their native language. This phenomenon, known as perceptual reorganisation, is often linked to hemispheric lateralisation and is hypothesised to be driven by specific acoustic properties or linguistic relevance of auditory signals. While Dutch infants can discriminate pitch changes in speech, how they process these pitch changes remains unclear. Previous research found that 10-month-old Japanese infants exhibit left hemispheric lateralisation for pitch changes in speech stimuli but bilateral responses to pitch changes in non-speech stimuli, whereas 4-month-old Japanese infants show bilateral responses to both (Sato et al., 2010a). These findings suggest that lateralisation patterns reflect the functional role of pitch in Japanese. To further explore the relationship between perceptual reorganisation, hemispheric lateralisation, and linguistic experience, we conducted a cross-linguistic study comparing lateralisation patterns in response to pitch changes in speech and pure tone stimuli in Japanese (lexical pitch-accent language) and Dutch (stress language) infants at 4 and 10 months of age. At 4 months, Dutch infants exhibited stronger activation in the right hemisphere compared to the left hemisphere (i.e., rightward lateralisation) in the pure tone condition, while Japanese infants showed bilateral responses (i.e., no hemispheric lateralisation). For pseudowords, Dutch infants exhibited stronger activation in the right hemisphere compared to the left hemisphere as they responded to pure tone stimuli, whereas Japanese infants exhibited stronger activation in the left hemisphere compared to the right hemisphere (i.e., leftward lateralisation). By 10 months, both Dutch and Japanese infants demonstrated rightward lateralisation for pure tones. However, for pseudowords, Japanese

infants maintained leftward lateralisation, while Dutch infants shifted to bilateral responses. These findings suggest distinct developmental trajectories in perceptual reorganisation across language groups, shaped by linguistic experience and the functional role of auditory signals in the native language.

**Keywords**: Lexical pitch, Dutch, Japanese, Infants, Hemispheric lateralisation, Perceptual reorganisation

## 5.1 Introduction

Infants' auditory systems are remarkably adaptable, with an inherent capacity to discern a wide range of speech contrasts, including segmental features at the phonemic level and suprasegmental features at the prosodic level. However, during their first year, the sensitivity to nonnative contrasts declines, while sensitivity to native contrasts improves. This process, known as perceptual narrowing or reorganisation, typically occurs in the second half of the first year and is thought to reflect increasing specialisation in infants' auditory processing of their L1. Initially, infants process speech sounds within a broad, general auditory framework, which gradually shifts toward more focused processing of native speech sounds (e.g., Kuhl, 2004; Werker, 2024; Werker & Tees, 1984). This developmental shift is crucial as infants fine-tune their ability to detect phonetic, prosodic, and lexical cues that carry linguistic meaning.

The human brain is specialised for language processing, which is widely regarded as a predominantly left-hemisphere function, particularly within the frontal and temporal regions (Broca, 1861; Wernicke, 1874). These regions are essential for the processing of phonological information and semantic tasks (e.g., Vigneau et al., 2006). Neuroimaging studies have shown evidence of hemispheric lateralisation for the perception of native speech in infants, which already resembles the patterns observed in adults. For example, both infants and adults consistently demonstrate left-lateralised brain activation in response to native speech stimuli, whereas non-native or non-speech stimuli typically elicit right-lateralised or bilateral responses (e.g., Dehaene-Lambertz et al., 2002; Molfese et al., 1975; Vannasing et al., 2016). Some studies have highlighted a

developmental trajectory in infants' responses to speech stimuli. For example, Minagawa-Kawai et al. (2007) found that Japanese infants aged 13 months exhibited clear left-hemisphere dominance for the processing of durational phonemic contrasts, whereas this lateralisation was not observed in 6- to 7-month-olds. Such developmental changes in the neural basis of lateralisation are believed to occur in parallel with perceptual reorganisation, as the brain adapts to the sounds of the native language (see reviews in Gervain, 2020; Werker, 2018).

Three main hypotheses attempt to explain the laterality of speech perception: a signal-driven, a domain-driven, and a learning bias hypothesis (see Minagawa-Kawai et al., 2011 for a full review). The signal-driven hypothesis proposes that left-hemispheric lateralisation emerges from this hemisphere being apt at processing rapid durational changes (or temporally variant sounds), such as those found in speech sounds, while the right hemisphere specialises in the fine spectral processing (or spectrally variant sounds) of slower pitch changes or emotional vocalisations (e.g., Zatorre & Belin, 2001). Yet, a review noted that relatively small changes in acoustic features, or their familiarity, may modify the degree to which the left vs. right auditory areas contribute to sound processing (Tervaniemi & Hugdahl, 2003). The domain-driven hypothesis suggests that left-hemispheric preference is shaped by linguistic relevance rather than by acoustic properties alone (Dehaene-Lambertz et al., 2006). Therefore, tone language speakers are expected to show a leftward lateralisation for the processing of tone. More specifically, it could be that the pathways from the primary auditory cortex radiate to higher-level processing areas in the left temporal cortex for acoustic features that are "language relevant" and to the right temporal cortex for more general properties

relevant to both speech and non-speech sounds (Aslin, 2012). Lastly, the learning bias hypothesis suggests that speech processing is primarily shaped by learning experiences (Minagawa-Kawai et al., 2011). Specifically, the formation of categorical phonetic units and the identification of words and linguistic rules rely on specialised learning mechanisms, which are believed to be particularly effective in the left hemisphere. This, in turn, leads to the left-lateralisation of speech processing.

While each hypothesis has extensive supporting evidence, there are nuances to consider for each of them. For example, regarding the acoustic hypothesis, the processing of spectral and temporal information is not exclusively contralateral (Bourke & Todd, 2021). Poeppel (2001) suggested that the left and right hemispheres are specialised for temporal information at different levels: the left hemisphere for temporal information at the syllable level and the right hemisphere for temporal information at the intonation level. Similarly, the domain-driven and the learning bias hypotheses intersect in explaining why language experience shapes lateralisation differently across acoustic features. For instance, phonemic pitch learned in a L1 tone language may not lead to left lateralisation when processing phonemic pitch in a L2 tone language, suggesting the interactive effects of pitch function, linguistic experience, and top-down cognitive processing, such as phonological knowledge and selective attention to pitch patterns (Y. Xu et al., 2006). Thus, further explorations are needed to test the underlying mechanisms of brain lateralisation, particularly in speech perception, where perceptual cues, such as pitch, serve multiple roles across languages.

Speech perception relies on mapping basic auditory information (acoustics) of the speech signal to conceptual-semantic representations

(linguistics). Among various perceptual cues, pitch is particularly significant because of its multifunctionality: it serves as a prosodic feature in all languages and also contributes to phonemic distinctions in all tone and pitch-accent languages (see a review in Kager, 2018). Universally, pitch conveys affective and grammatical meaning, and in over half of the world's languages, it serves as a lexical cue. Lexical tone relies on pitch variation to distinguish lexical meaning. For example, Standard Chinese and Japanese use pitch to differentiate lexical contrasts, whereas pitch in English and Dutch rarely affects lexical meaning (Ota et al., 2018; Yip, 2002). Investigating the lateralisation of pitch is crucial for understanding how the brain processes acoustic information across different linguistic contexts and offers insights into the lateralisation of speech processing as a whole.

The processing of pitch in adults reveals both signal-driven and experience-based lateralisation. Signal-driven lateralisation reveals that speech processing is predominantly a temporal process, lateralised to the left hemisphere, whilst the processing of music, being primarily a spectral process, is lateralised to the right (Zatorre & Belin, 2001; Zatorre & Gandour, 2008). Experience-based lateralisation, often intertwined with linguistic relevance, suggests that pitch processing is influenced by its linguistic function. Studies have shown that prosodic contrasts are primarily processed in the right hemisphere but are shaped by language experience. For example, lexical pitch is predominantly processed in the left hemisphere in Chinese speakers, reflecting its linguistic role, while it is processed in the right hemisphere in English speakers (Gandour et al., 2003, 2004). Although there is extensive evidence supporting right hemisphere lateralisation for non-speech pitch processing, some studies have reported left hemisphere lateralisation even for non-speech pitch

processing in tone language speakers (e.g., Gu et al., 2013), which has been interpreted as a result of the linguistic role of pitch in speech for these speakers. These findings suggest that adult hemispheric lateralisation of pitch is influenced not only by acoustic properties (bottom-up processing) but also by linguistic relevance/experience (top-down processing).

Pitch is the first and arguably most salient auditory feature humans are exposed to from birth (L. Liu et al., 2022), and even newborns can discriminate lexical pitch accents (Nazzi et al., 1998). As they develop, infants learn to associate pitch with linguistic (e.g., distinguishing word meaning), emotional (e.g., expressing emotions), and grammatical functions (e.g., signalling questions or statements) (Kager, 2018). Notably, their sensitivity to pitch in nonspeech remains stable throughout the first year. For example, English infants between 6 to 9 months consistently discriminate musical tones (Mattock & Burnham, 2006), and European Portuguese infants between 5 to 9 months can distinguish falling and falling-rising intonations uttered with bisyllabic pseudowords (Frota et al., 2016). However, linguistic experience influences infants' sensitivity to lexical tones. For example, sensitivity to nonnative lexical tones declines in Dutch infants around 9 months (L. Liu & Kager, 2014; Nazzi et al., 1998). Likewise, English infants lose sensitivity to Mandarin and Thai tones by 9 months (Mattock et al., 2008; Mattock & Burnham, 2006), and Cantonese infants of both 4- and 9-month-old show a preference for Cantonese tones over Mandarin tones (Yeung et al., 2013). These variations in pitch perception suggest that infants adapt to their linguistic environment, refining pitch sensitivity based on language-specific functions.

The discrimination of pitch in infants is also reflected in hemispheric lateralisation. Neonates show right temporal activity for

intonation (Arimitsu et al., 2011, 2018), while the function of pitch in a language can influence its lateralisation to the left hemisphere, as shown in Sato et al. (2010a)'s fNIRS study. They found that Japanese infants process pitch accent in the left hemisphere by 10 months (but not at 4 months), whereas pure tones are processed bilaterally at both ages. This leftward shift reflects the function of pitch in Japanese, which conveys lexical information, and aligns with broader evidence that segmental processing typically engages the left hemisphere, while intonation and emotional prosody engage the right hemisphere, and nonspeech or nonnative contrasts are processed bilaterally (Schirmer & Kotz, 2006; van Lancker, 1980). In combination with the behavioural evidence referred to above, it suggests that the linguistic function of pitch in the native language induces infants' lateralisation of pitch processing at least at 10 months.

In brief, pitch is a fundamental perceptual cue in speech, serving different linguistic functions across languages. While infants are born with a general sensitivity to pitch, their processing of pitch becomes increasingly shaped by the specific demands of their native language. This adaptation is reflected not only in their ability to discriminate pitch contrasts but also in the lateralisation of pitch processing in the brain. Evidence suggests that both the acoustic cues underlying pitch perception and its linguistic relevance contribute to this developmental trajectory. However, the extent to which early hemispheric lateralisation of pitch processing is driven by innate acoustic biases versus language-specific experience remains unclear.

Although infants exposed to non-tonal or stress languages often show a decline in their ability to discriminate pitch contrasts at the word level during their first year of life, findings across studies present a more complex picture of perceptual reorganisation. For example,

Dutch infants who are not learning Limburgian tones can still discriminate these contrasts at both 6 and 12 months, similar to Limburgian-learning infants (Ramachers et al., 2018). Likewise, 12-month-old Dutch infants successfully discriminated SC tone contrasts (e.g., T2 vs. T4; T2 vs. T3) and their musical counterparts, despite failing to do so at 4 months (A. Chen et al., 2017; A. Chen & Kager, 2016). Moreover, Liu and Kager (2014) observed a U-shaped trajectory in Dutch infants' discrimination of SC T1-T4 contrasts, with success at 5-6 months, a decline around 9 months, and re-emergence at 17–18 months. These discrepancies raise the question of how Dutch infants process pitch changes across development. Based on previous findings, both the linguistic function and the acoustic properties of lexical pitch may contribute to these patterns, highlighting the need for further investigation.

To further investigate these discrepancies, we designed a cross-linguistic study to examine the influences of universal acoustic principles and language-specific demands on early pitch processing. Specifically, we directly compared Japanese and Dutch infants, as these languages assign contrasting roles to pitch: in Japanese, pitch functions as a lexical cue that distinguishes word meanings, whereas in Dutch, it primarily serves as an affective or prosodic cue. By selecting these languages, we aimed to minimise confounding physiological factors and isolate the impact of the linguistic relevance of pitch on hemispheric lateralisation. This design allows us to determine whether early lateralisation patterns for pitch arise from innate acoustic processing biases or are shaped by the linguistic context in which infants are immersed.

We employed behavioural and neuroimaging methods across three experiments to examine if, and if so, where in the brain infants from

different language environments process pitch changes. To minimise the influence of lexical knowledge, we created pseudowords that fulfil the phonological rules of both Japanese (Okada, 1991) and Dutch (Gussenhoven, 1992). These pseudowords were designed to be perceived as novel words in each language, ensuring a controlled and comparable linguistic context (see details in the Stimuli section of Experiment 1). In Experiment 1, we examined 4- and 10-month-old Dutch infants' behavioural responses to pitch changes in a habituation task, to investigate whether they discriminate pitch changes in general. This was a necessary first step that had been taken in Sato et al. (2010a) too with Japanese 4- and 10-month-olds. Experiments 2 and 3 utilised functional Near Infrared Spectroscopy (fNIRS) (see reviews in Aslin & Mehler, 2005; Azhari et al., 2020; Nishiyori, 2016; Quaresima et al., 2012; Wilcox & Biondi, 2015) to observe 4- and 10-month-old Dutch and Japanese infants' brain responses to pitch changes in pseudoword and pure tone contexts. fNIRS is one of the best neuroimaging techniques to investigate infants' brain responses, given its good temporal and spectral resolution, the possibility of testing awake infants, and the fact that it is a non-invasive and safe procedure. (Wilcox & Biondi, 2015).

Referring to the theoretical accounts and empirical evidence in previous studies, and specifically Sato et al. (2010a), we hypothesised that infants' processing of pitch changes undergoes perceptual reorganisation, with neural responses in the temporal region (auditory cortex) reflecting the transition from general to specialised auditory processing across languages. Initially, we expected pitch processing to rely on general auditory mechanisms, dependent on acoustic properties, in both Dutch-learning and Japanese-learning infants. Over time, however, this general processing would diverge based on

language exposure: older infants exposed to a pitch-accented language (Japanese) and those exposed to a non-tonal language (Dutch) would develop distinct processing patterns, influenced by language-specific experience and the linguistic relevance of pitch. Specifically, we predicted that 4-month-old infants from both language groups would exhibit bilateral (or rightward) brain responses to pitch changes in both pseudowords and pure tone stimuli. By 10 months, we expected Japanese infants, who are exposed to a pitch-accented language, to show left temporal lateralisation for pitch changes in pseudowords, consistent with Sato et al. (2010a). In contrast, Dutch infants were not expected to show left lateralisation for either stimulus type. We also predicted that responses to pure tones would remain bilateral (or rightward) in both groups at 10 months.

### 5.1.1 Ethical approval

Experiments 1 and 2, which took place at Leiden University, were approved by the ethical committee of the Faculties of Humanities and Archaeology at Leiden University (2022/11). Experiment 3, which took place at The Institute of Physical and Chemical Research (RIKEN), was approved by the Institutional Review Board of RIKEN (Wako 3 30-2 (16)). Across experiments, parents were briefed on the intention of the study and subsequently gave their informed consent for their infants' participation prior to the experiments. All parents and infants were compensated for their voluntary participation.

## 5.2 Experiment 1. Habituation of lexical pitch in Dutch infants

The habituation paradigm is widely used in infant research to investigate speech perception and other aspects of language processing

(e.g., Kucharský et al., 2022; Oakes, 2010). In a typical habituation paradigm, infants are repeatedly exposed to a single type of stimulus during the habituation phase. Once they have become habituated, they proceed to the test phase, where their responses to both habituated and novel stimuli are measured. If infants show a longer response time to the novel stimulus compared to the habituated one, it suggests they can perceive the difference between them.

In Experiment 1, we used a modified habituation paradigm to examine whether Dutch infants can discriminate pitch changes in pseudowords. Previous studies have shown that Dutch infants aged 5–17 months can perceive pitch differences in words, even when such variations are not phonologically meaningful in Dutch (A. Chen et al., 2017; L. Liu & Kager, 2014; Ramachers et al., 2018). However, we aimed to investigate whether this discrimination ability persists when stimuli are presented in an alternating versus non-alternating manner, as in the subsequent fNIRS experiments. Our design involved a novel test stimulus featuring a sequence of pseudowords with alternating familiar and novel pitch patterns (see Section 5.2.1 for details), explicitly testing infants' ability to discriminate pitch changes based on alternation rather than general novelty. This approach aligns with the block design used in the subsequent fNIRS experiments (see Section 5.3.1), ensuring methodological consistency and strengthening the potential evidence to find evidence for hemispheric differences in pitch discrimination.

Based on the previous studies, we hypothesised that infants in both age groups would differentiate pitch changes within pseudowords. Specifically, we expected a decrease in looking time during the habituation phase. Once habituated, we predicted that infants would exhibit longer looking times for the trials with a novel pattern,

consisting of words with alternating HL and LH patterns, compared to the trials consisting of words maintaining the HL or LH pitch pattern they were habituated to.

### 5.2.1 Method

#### *5.2.1.1 Participants*

We tested 14 Dutch infants aged 4 months (8 girls; *M* = 124 days, range = 107–134 days) and 32 infants aged 10 months (19 girls; *M* = 309 days, range = 281–319 days) while awake. Twenty-one infants were excluded from further analysis due to language background (*n* = 10), excessive fussiness (*n* = 2), large motions (*n* = 2), no habituation after 28 trials (*n* = 2), and technical error (*n* = 5).

The final sample included 9 infants aged 4 months (5 girls; *M* = 125 days, range = 115–134 days) and 16 infants aged 10 months (11 girls; *M* = 308 days, range = 292–318 days).

#### *5.2.1.2 Stimuli*

The auditory stimuli consisted of 14 bisyllabic pseudoword pairs that varied only in their pitch pattern (High-low vs. Low-high), i.e., /ika/, /ipa/, /isa/, /ita/, /kapi/, /papi/, /pasa/, /pika/, /pipa/, /pisa/, /pita/, /sapa/, /sapi/, and /tapi/. These pseudowords were created based on shared phonotactic rules of Dutch and Japanese. All stimuli were recorded in isolation with a Sennheiser MKH416T microphone (sample size 44.1 kHz, 16 bit) at Leiden University's Phonetics Lab by a female native speaker of Japanese (from the Tokyo area). who was proficient in Dutch. The speech signals were digitised at a 44.1 kHz sampling rate with 16-bit resolution. She was asked to produce the stimuli without any emphasis. The stimuli were recorded three times, with the stimulus list randomised for each recording. Tokens judged to

be most clearly articulated by the authors were selected for further manipulations. Native Dutch and Japanese speakers then rated the wordlikeness of each token (i.e., whether it could plausibly be a Dutch word). While ratings varied, none of the tokens were deemed impossible. Additionally, consultations with phoneticians and phonologists confirmed that all tokens adhered to Dutch and Japanese phonotactic rules. Consequently, all tokens were included in the final stimulus set. For further details on stimulus formulation and evaluation, see Appendix B1.

#### *5.2.1.3 Procedure*

We configured the experimental procedure using Habit2 software (Oakes et al., 2019) on a Windows 10 desktop at the Babylab of Leiden University. The experiment took place in a dimly lit, sound-attenuated booth. Infants sat on their parents' laps, facing a stimulus monitor approximately 60 cm away. Parents wore headphones playing masking sounds to prevent any unintended influence on the infant's behaviour. Visual stimuli were displayed on a 43-inch monitor (110 cm width) and reflected onto rear-facing mirrors positioned on the booth's back wall, making them visible to the experimenter who sat outside the booth. Auditory stimuli were presented through the monitor's speaker at a sound pressure level of 60 dB. The experimenter recorded the infant's looking time while observing a live video stream from a camera on a monitor.

The experiment employed a modified habituation paradigm (see a review in Kucharský et al., 2022), consisting of two phases: habituation and test. In the habituation phase, infants were repeatedly presented with the same stimulus until they decreasedly responded to the stimulus (see below). Each trial was preceded by a quiet visual

attention-getter (e.g., Looming circle; The ManyBabies Consortium, 2020) to engage the infant's focus. The experimenter initiated each trial by pressing a key once the infant looked at the attention-getter on the screen.

In each trial, a dynamic, shiny, and colourful checkerboard was displayed on a screen alongside an auditory stimulus. Four wordlists with a single tone pitch pattern (either HL or LH throughout) were randomly presented together with the checkerboard during the habituation phase. The experimenter pressed a key once the infant fixated on the screen and released it when the infant looked away. The trial ended either when the infant looked away for more than 2 s or when the wordlist finished playing. The habituation phase ended when the average looking time over the last four trials was less than 65% of the average of the first four trials, with a maximum of 28 habituation trials.

Following habituation, the procedure automatically proceeded to the test phase, which included two trials: a "Same" trial, presenting one of the wordlists from the habituation phase, and a "Change" trial, presenting a wordlist with alternating HL and LH tone patterns. The order of "Same" and "Change" trials was counterbalanced across infants.

### 5.2.2 Data analysis

To thoroughly examine the habituation process, we conducted inferential statistical analyses using the online recordings of looking times during both the habituation and test phases. Although offline recordings were available as a backup, they were not used in the current analyses. In the habituation phase, we averaged looking times from the first four and last four trials, creating comparison pairs

(Initial vs. Final). A decrease in looking time from the Initial to Final trials indicates successful habituation. In the test phase, we compared looking times between the "Same" and "Change" trials.

We analysed habituation and test phases separately using a Generalised Linear Mixed Model (GLMM) via the lme4 package (Bates et al., 2014) to account for individual variability. The GLMM included fixed effects of trials, groups, and their interactions, with by-subject random intercepts to address heterogeneity across subjects. Pairwise multiple comparisons across all conditions were conducted using the *emmeans* package (Lenth, 2023) with Bonferroni corrections to control for type I error.

To further elucidate results, we conducted paired-sample t-tests to assess looking time differences across trials within each group in both phases. Given the limited sample size, we supplemented these analyses with Bayesian methods using JASP (Coon et al., 2024; Visser et al., 2024) to strengthen our inferences. The Bayes factors (BF+0) were computed directly within JASP by comparing the likelihood of the observed data under the null and alternative hypotheses, using default prior distributions specified by the software. The Bayes factor quantifies the relative evidence for the null hypothesis versus the alternative hypothesis. This approach provides a robust complement to our frequentist analyses by offering an additional layer of evidence regarding the strength of the effects. Moreover, a Bayesian sequential analysis was conducted to further illustrate evidence accumulation with each infant added (Visser et al., 2024). Unlike the paired-sample t-tests, which provide a static assessment of differences with BF+0, this approach updates the BF+ iteratively, offering a dynamic assessment of the strength of evidence as the sample grows.

### 5.2.3 Results

#### *5.2.3.1 Generalised linear mixed-effects modelling (GLMM)*

#### *Habituation phase*

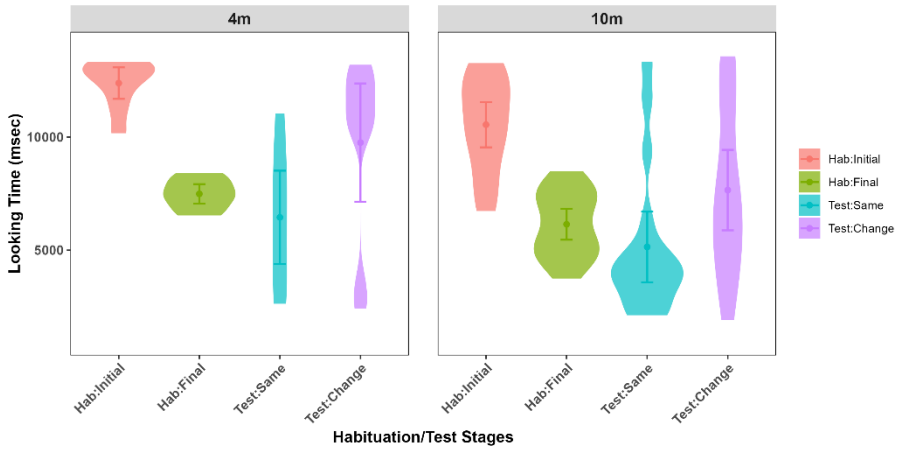
Model diagnostics indicated no violations of linear regression assumptions. Comparing the null and interest models revealed a significant decrease in looking times by the end of the habituation phase ($F$ = 87.93, $p$ < .001). Pairwise comparisons showed that both four-month-olds and ten-month-olds exhibited shorter looking times in Final trials than in Initial trials, with reductions of 4893 ms ($p$ < 0.001) and 4384 ms ($p$ < 0.001), respectively.

#### *Test phase*

Model diagnostics confirmed no violations of linear regression assumptions. Comparing the null and interest models revealed that looking times were significantly longer for the Change stimulus than for the Same stimulus ($F$ = 10.37, $p$ < 0.01). T-tests further demonstrated that both four-month-olds and ten-month-olds responded with longer looking times in Change trials than in Same trials, showing increases of 2561 ms ($p$ < 0.05) and 2561 ms ($p$ < 0.05), respectively. Figure 5.1 displays violin plots of looking times at each stage in both phases.

**Figure 5.1**

*Looking times in habituation (initial vs. final) and test (same vs. change) phases.*




### *5.2.3.2 Bayesian analysis*

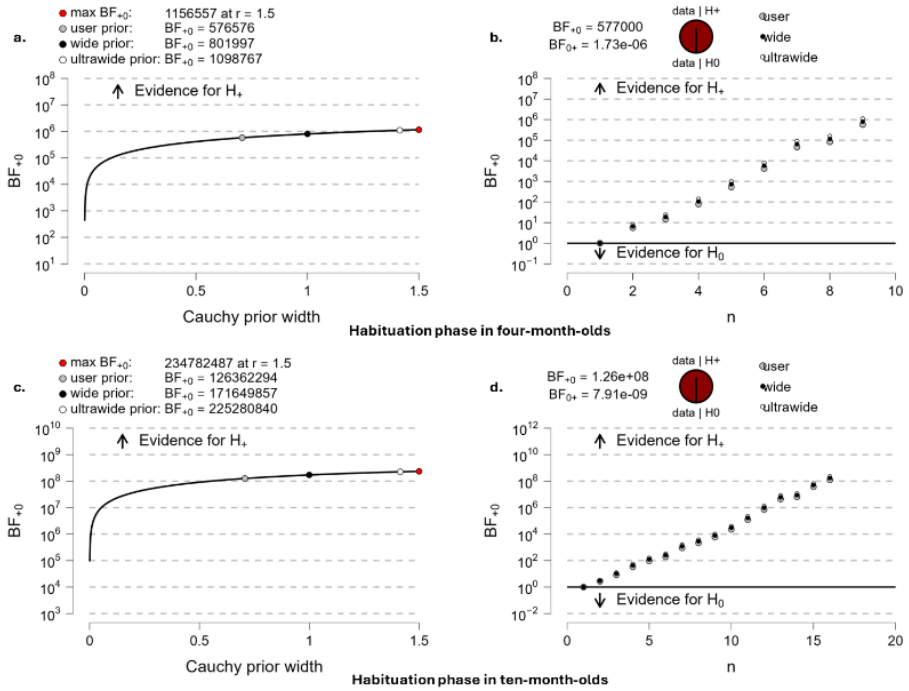
#### *Habituation phase*

The Bayesian Repeated Measures ANOVA revealed the best model as one including both Trial and Group factors (*BF* = 3.99). Both age groups habituated successfully, exhibiting shorter looking times in Final trials compared to Initial trials. The Bayes Factor Robustness Check showed that the t-tests of both 4- and 10-month-olds were robust across prior width adjustments (default prior width = 0.707; see Figure 2a and 2c). In Bayesian analyses, the prior width ($r$) determines the range of plausible effect sizes under the alternative hypothesis, with wider priors reflecting less certainty and allowing for larger effect sizes. A wider prior means more conservative assumptions about the size of the effect. The maximum $BF_{+0}$ was obtained with a prior width $r$ = 1.5 for both age groups, indicating strong evidence for the alternative hypothesis under this specification. Importantly, the evidence for the

alternative hypothesis remained stable across a variety of prior distributions (e.g., user prior, wide prior, and ultrawide prior), suggesting that the results were not highly sensitive to changes in prior specifications. This provides robust support for the alternative hypothesis in both age groups. Sequential plots further illustrated progressively stronger evidence supporting the alternative hypothesis (Initial > Final; Figure 5.2b and 5.2d). The BFs were substantially greater than 10, with the trend line showing accumulating evidence in favour of the alternative hypothesis as each datapoint was added for both age groups.

**Figure 5.2**

*Bayes factor robustness check and sequential plots for habituation phase by age group.*

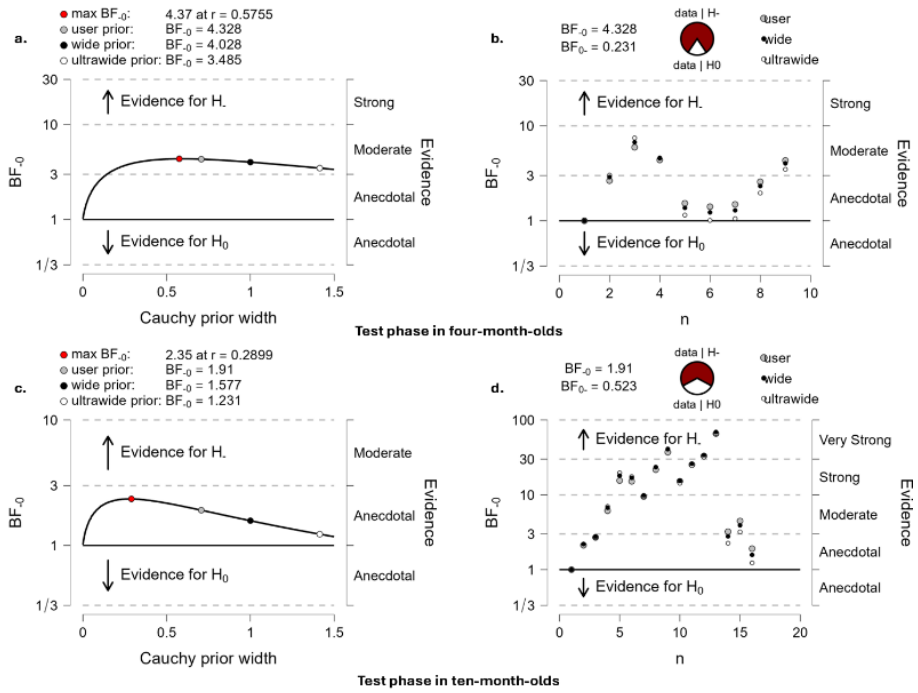


***Test phase***

The Bayesian Repeated Measures ANOVA revealed the optimal model as one including only the Trial factor (*BF* = 2.91), indicating that, on average, infants looked longer at the Change trial than the Same trial. The Bayes Factor Robustness Check showed that the t-tests of both 4- and 10-month-olds were fairly robust across prior width adjustments (default prior width = 0.707; see Figure 3a and 3c). Specifically, the maximum $BF_{+0}$ was obtained when the prior width *r* was automatically set to 0.6 for four-month-olds and 0.3 for ten-month-olds, reflecting narrower ranges of expected effect sizes for these groups. Evidence for the alternative hypothesis remained stable for four-month-olds but was more variable for ten-month-olds, suggesting that the analysis was relatively robust for the younger group but less so for the older group. This variability may reflect differences in the strength or consistency of the effect between the two age groups. Sequential plots provided moderately strong evidence supporting the alternative hypothesis (Change > Same; Figure 5.3b and 5.3d). For four-month-olds, the BF ranged between 3 and 10, indicating moderate evidence for the alternative hypothesis, with the trend line suggesting unstable accumulation of evidence as more data points were added. In contrast, for ten-month-olds, the BF remained relatively small, with the trend line initially showing stable accumulation followed by a sudden drop in the last few data points. These patterns suggest potential individual variability in the ability to discriminate pitch changes in Dutch infants at both ages.

**Figure 5.3**

*Bayes factor robustness check and sequential plots for test phase by age group.*




## 5.2.4 Discussion

In this habituation experiment, both age groups successfully habituated and discriminated the "Change" trial from the "Same" trial. However, the evidence for discrimination was weaker than typically observed, likely due to variability among infants. This variability may stem from the complexity of the "Change" stimuli (the alternating stimuli) or developmental factors influencing infants' responses. While some infants may have focused on the similarities between the alternating and the non-alternating stimuli, i.e., recognising the pattern they were habituated to in the alternating stimuli, others may have focused on the difference, namely the alternating versus non-

alternating characteristics of the test trials. A meta-analytic study by Cristia (2018) found that the overall effects of behavioural studies using the alternating/non-alternating paradigm were not robust.

Nonetheless, both our findings and those of Chen et al. (2017) and Liu & Kager (2014) and Ramachers et al. (2018) suggest that Dutch infants are in principle able to detect pitch changes. One possible explanation for the variability in our data is the developmental stage of 10-month-old Dutch infants, who are in a critical period of tuning into their native language. Since pitch changes do not induce lexical contrasts in Dutch, these infants may show inconsistent sensitivity to such changes. Supporting this interpretation, prior research has shown that while Dutch infants can discriminate Chinese tone contrasts (e.g., high-level vs. falling tones) at 5-6 months, they experience a decline in tone sensitivity around 9 months, but regain this sensitivity by 15 months (L. Liu & Kager, 2014). In addition, 6- to 12-month-old Dutch infants can distinguish Limburgian tones (Ramachers et al., 2018). Furthermore, 12-month-old Dutch infants can discriminate T2–T3 lexical tones and musical tones, whereas 4-month-olds do not (A. Chen et al., 2017). According to Chen et al. (2017), these findings suggest that Dutch infants rely on general acoustic pitch discrimination rather than representing pitch patterns in words as phonologically meaningful categories. Thus, while Dutch infants are capable of distinguishing pitch changes, they may not interpret pitch as linguistically significant. To explore this hypothesis further, we conducted an fNIRS experiment (Experiment 2) to measure brain responses of Dutch infants while listening to pitch changes.

## 5.3 Experiment 2. Hemispheric lateralisation of lexical pitch in Dutch infants (NIRS study)

In Experiment 2, we used fNIRS to investigate brain responses to both language-related pitch changes (within bisyllabic pseudowords) and non-linguistic pitch changes (within pure tones). If Dutch infants perceive both types of pitch changes similarly, this would support the hypothesis that they process pitch changes as general prosodic auditory information rather than as linguistically meaningful cues.

### 5.3.1 Method

#### *5.3.1.1 Participants*

We tested 28 Dutch infants aged 4–5 months (12 girls; *M* = 140 days, range = 117–158 days) and 31 infants aged 10 months (15 girls; *M* = 305 days, range = 288–320 days) while awake. All infants were healthy, full-term, and either monolingual or bilingual. Bilingual infants could only participate when their second language was a non-tonal Germanic language with a stress pattern similar to Dutch, like English or German. Twelve infants were excluded from further analysis due to crying (*n* = 1), data file corruption (*n* = 1), language background (*n* = 2), fussiness (*n* = 2), excessive motion artifacts (*n* = 3), and insufficient data, defined as having either fewer than three valid trials or fewer than two valid channels per ROI (*n* = 3).

The final sample included 21 infants aged 4–5 months (11 girls; *M* = 140 days, range = 117–158 days) and 26 infants aged 10 months (12 girls; *M* = 305 days, range = 288–320 days).

#### *5.3.1.2 Stimuli*

The same stimuli as in Experiment 1 were used. However, because the stimuli would be used in both the fNIRS study with Dutch infants and the fNIRS study with Japanese infants, four pseudowords were excluded due to their similarity to real Japanese words, leaving ten pseudowords: /ipa/, /kapi/, /pasa/, /pika/, /pipa/, /pisa/, /pita/, /sapa/, /sapi/, and /tapi/. The pseudowords were presented with two pitch patterns: High-Low (HL) or Low-High (LH).

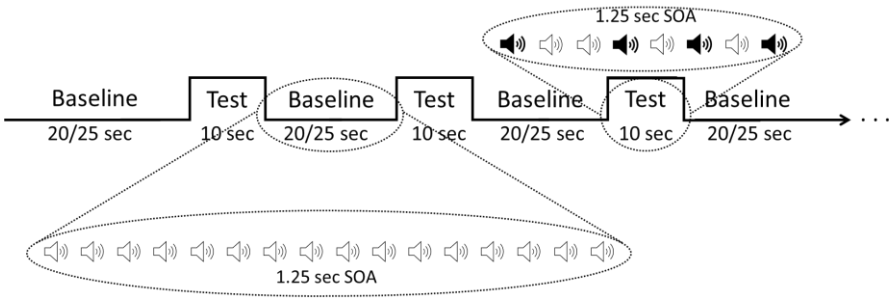
Ten pure tone pairs were generated by extracting the fundamental frequencies from the pseudoword tokens. First, the mean fundamental frequencies of each syllable in a word were measured. Based on these mean values, we created complex pure tones using sinewave shapes, which included the fundamental frequency along with its second and third overtones. The tones from each syllable were then combined at the nearest zero crossing point to form continuous pure tone sequences, either HL or LH.

#### *5.3.1.3 Procedure*

Auditory stimuli were played at approximately 60 dB SPL through a loudspeaker positioned below the presentation monitor. Infants sat on their parents' laps throughout the procedure. To reduce head movement, a silent, non-synchronised movie played continuously on the monitor, and an experimenter engaged the infant noiselessly with toys. Experimenters and parents wore headsets playing unrelated music to prevent them from hearing the stimuli and possibly influencing the infants' behaviour.

**Figure 5.4**

*Procedure for Experiments 2 and 3.*




*Note*. Baseline blocks played auditory stimuli with a consistent pitch pattern, while the test blocks played stimuli with alternating pitch patterns in a pseudorandomised order. The duration of a test block was 10 seconds, whereas baseline blocks lasted for either 20 or 25 seconds, with the duration randomly determined by the experiment program (PsychoPy).

Each infant was tested in two conditions: pseudoword (WD) and pure tone (PT). Both conditions consisted of alternating baseline and test blocks. In the baseline block, infants either heard a sequence of pseudowords (WD condition) or pure tones (PT condition) presented with either high-low (HL) or low-high (LH) pitch patterns, with a 1.25-second stimulus onset asynchrony (SOA). There were eight baseline sequences in total: two HL and two LH sequences of 20 seconds each, and two HL and two LH sequences of 25 seconds each. The test blocks, lasting 10 seconds each, contained a mix of HL and LH words or pure tones, with four 10-second sequences created for each HL-LH and LH-HL alternation. A PT test block always immediately followed a PT baseline block, while a WD test block always immediately followed a WD baseline block. Research indicates that the hemodynamic response typically takes more than 10 seconds to return to baseline (Taga et al.,

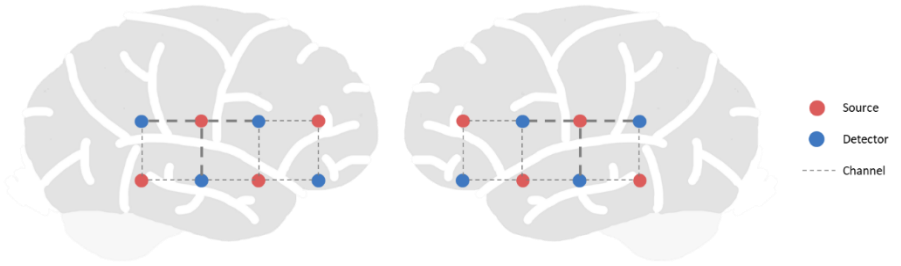
2003), therefore, the baseline block was intentionally longer than the test block to ensure that the hemodynamic response elicited during the experimental condition had returned to baseline levels (Lloyd-Fox et al., 2010) before the next test block. All sequences in both the baseline and test blocks were presented in a pseudorandom order. The presentation of HL-LH and LH-HL alternations, as well as the conditions, were counterbalanced among participants. Each condition included at least five pairs of baseline and test blocks.

#### *5.3.1.4 NIRS recording: NIRScout*

Hemodynamic responses to lexical pitch pattern changes in Dutch infants were recorded with a multichannel NIRS system (NIRScout 8×16, NIRx Medical Technologies, Berlin, Germany). The system used LEDs at 760 nm and 850 nm wavelengths, with a 15.625 Hz sampling rate. Eight sources and eight detectors were positioned to cover the bilateral frontal, temporal, and parietal cortices, creating ten channels per hemisphere. Optodes were placed in a stretchy Easycap with a 2cm source-detector separation. Head circumference was measured beforehand to select the appropriate cap size. The arrays were custom-built, with the lower row's midpoint aligned over T3 and T4 according to the International 10-20 system. See Figure 5.5 for an illustration of the NIRScout system's optode array configuration.

**Figure 5.5**

*Optode array configurations in Experiment 2 (NIRScout).*




*Note.* Coloured circles represent the light incident probes: red for source optodes and blue for detector optodes. Dashed lines indicate channels, with bold dashed lines highlighting the channels of interest for the temporal lobe.

### *5.3.1.5 NIRS data preprocessing*

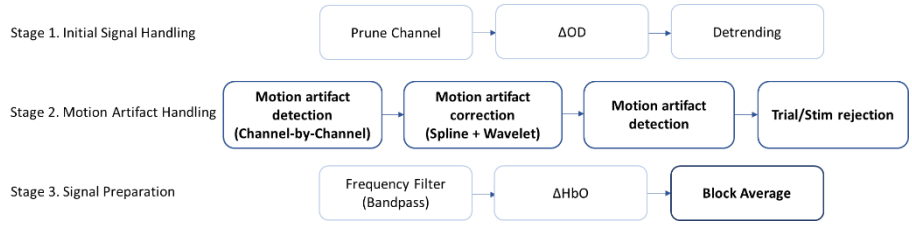
The preprocessing of fNIRS data in infants remains a topic of ongoing discussion, with no clear consensus or standardised procedures for handling noise, motion artefacts, hemodynamic response variability and so on. This lack of agreement is partly due to the relatively recent adoption of fNIRS as a research tool and the still incomplete understanding of the infant hemodynamic response, including its development and variability (Gemignani & Gervain, 2021). Differences in processing pipelines, step-by-step methodological choices, and parameter settings can substantially affect the resulting data. Recent studies have explored these issues by comparing various processing approaches, using both simulated and real data, to move toward a more unified framework for analysing infant fNIRS data (e.g., Di Lorenzo et al., 2019; Gemignani & Gervain, 2021; Gervain et al., 2023; Pinti et al., 2019; Yücel et al., 2021). In general, a typical preprocessing procedure includes steps such as a signal quality check of the raw light intensity data, optical density conversion, physiological noise reduction, motion artefact detection and correction, frequency filtering, baseline

correction, haemoglobin signal conversion and extraction. Among these steps, motion detection and correction are critical given their influence on the signal (e.g., Di Lorenzo et al., 2019). Though previous studies have provided "best" algorithms and parameters regarding motion detection and correction, we deemed it worthwhile to apply different parameters to get more insight into the robustness of haemoglobin response patterns.

We preprocessed the NIRS data using Homer3 (Huppert et al., 2009), with the processing pipeline outlined in Figure 5.6. First, channels with extreme light intensities were excluded. The light intensity data were then converted to optical density (OD) and detrended to eliminate slow drifts. Motion artefacts were identified on a channel-by-channel basis and corrected using Spline and Wavelet algorithms (Di Lorenzo et al., 2019), which facilitated the retention of valid trials. Subsequently, a reassessment of motion artefacts was conducted to exclude any residual invalid trials. Further, a band-pass filter (0.05–0.2 Hz) with a 3$^{rd}$-order low-pass and 5$^{th}$-order high-pass was applied to remove high-frequency noise, such as physiological noise from hair, skin, or bone structures and cardiac activity. The denoised signals were converted to concentration changes using the modified Beer-Lambert law (with the partial pathlength factor set to 1.0, as per Homer3 default). Oxygenated and deoxygenated haemoglobin concentrations were calculated, followed by baseline correction by subtracting the mean signal of a 5-second pre-stimulus window from the entire hemodynamic response function (e.g., Pinti et al., 2015; Zhang et al., 2022). Finally, datasets were included only if they contained at least three valid trials and two valid channels per hemisphere for each condition.

**Figure 5.6**

*fNIRS data preprocessing pipeline for Experiment 2 (Dutch).*




*Note*. Dark blue circles and bold font indicate critical preprocessing steps where specific parameters (see items a, b, and c) were applied.

We followed the motion-related analysis guidelines by Di Lorenzo et al. (2019) to evaluate eight preprocessing pipelines, varying in three key aspects: a. Motion artefact detection; b. Motion artefact correction, and c. Trial rejection and block averaging. These pipelines were applied to four datasets with varying levels of data quality manipulation (e.g., removal of noisy channels, trials, and time periods), yielding 32 possible combinations.

The details on preprocessing pipelines are as follows:

a) Parameters for motion artefact detection(s): We set thresholds of 1 for both tMotion and tMask, 13 for STDEVthresh, and 0.4 for AMPthresh. However, visual inspection of detection efficiency revealed that some artefacts were not identified. To address this, we applied a second set of parameters with a lower amplitude threshold (0.1) to capture additional artefacts.

b) Parameters for motion artefact correction: The Spline correction coefficient was set to 0.99, as recommended, and the interquartile range (*iqr*) for wavelet correction was set to 0.5 and 0.8. These *iqr* values were chosen based on Di Lorenzo et al. (2019), who found them most effective for

recovering the hemodynamic response function (HRF) and maximally trial retention.

c) Parameters for trial rejection and block average: Hemodynamic responses occurring within 20 and 25 seconds after stimulus onset were considered valid. The HRF typically returns to baseline within 15 seconds (e.g., Abboub et al., 2016; Gemignani & Gervain, 2021). Since the test block of sound sequences lasted 10 seconds, responses to these stimuli would be expected within 25 seconds after onset. We analysed response windows of [0, 20] and [0, 25], with the former preserving more trials and the latter allowing for a longer recovery period.

A summary of the eight pipelines and their parameter configurations is provided in Table 5.1.

**Table 5.1**

*Summary of preprocessing pipelines and parameters.*

| Pipeline | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Motion detection (Channel-by-Channel) | 0.4 | 0.4 | 0.1 | 0.1 | 0.4 | 0.4 | 0.1 | 0.1 |
| Motion correction (Wavelet *iqr*) | 0.5 | 0.5 | 0.5 | 0.5 | 0.8 | 0.8 | 0.8 | 0.8 |
| Motion detection | 0.4 | 0.4 | 0.1 | 0.1 | 0.4 | 0.4 | 0.1 | 0.1 |
| Stim/Trial rejection & Block average | 20 | 25 | 20 | 25 | 20 | 25 | 20 | 25 |

While Homer3 is capable of dealing with most artefacts, it can be helpful to remove noisy channels and trials beforehand to improve removal precision. To compare the impact of different preprocessing

strategies, we created four datasets reflecting varying levels of data quality manipulation in order to evaluate their respective outcomes: (1) the original dataset where noisy or malfunctioning channels, trials, and time periods were retained, (2) a dataset with noisy or malfunctioning channels removed, (3) a dataset with noisy/malfunctioning channels and noisy trials removed, and (4) a dataset with noisy/malfunctioning channels, noisy trials, and noisy time periods removed. These datasets allowed us to assess how different levels of data cleaning influenced the outcomes.

To select the optimal pipeline and dataset, we applied the following criteria: first, we selected datasets with 2 good channels per hemisphere and at least 3 remaining epochs (i.e., predefined time windows of fNIRS signal, including pre- and post-stimulus periods) after preprocessing. In some cases, no epochs remained after preprocessing, which, after visual inspection, was possibly due to overly strict motion detection criteria. Pipelines such as 3, 4, 7, and 8 utilised these overly strict motion detection criteria and were therefore not considered to provide insights into the pitch processing. Though we tried both *iqr* values of 0.5 and 0.8, we followed Di Lorenzo et al. (2019)'s recommendation to use 0.5 for analysing data collected with stimuli of s or longer in order to maximally benefit from the motion correction, which should result in higher trial retention and better estimation of hemodynamic responses in our case. Therefore, we focus on pipelines 1 and 2, which differ only in the duration of the time window of trial rejection and block average (e.g., the epoch duration).

The results of all four datasets generated by the two pipelines were largely consistent. We report findings from Dataset 3 (noisy channels and trials removed) processed with Pipeline 1 because it strikes a balance between artefact removal and data retention. Removing noisy

channels and trials improves preprocessing accuracy, while still preserving a sufficient number of valid epochs for analysis. Given the factors that we outlined, we think that the choice for this pipeline provides us with a reliable representation of the observed effects. A summary of all preprocessing results is in Appendix B2.

### 5.3.2 Data analysis

We first conducted Kolmogorov-Smirnov tests (V. W. Berger & Zhou, 2014) to examine the normality of the oxygenated haemoglobin (Oxy-Hb) data. Subsequently, we compared Oxy-Hb responses between baseline and test blocks to validate the responses to lexical pitch pattern changes during the test blocks using Wilcoxon signed-rank tests (Scheff, 2016). The grand-average method, a commonly used approach for analysing fNIRS signals, was employed to identify group differences in the brain's hemodynamic response across test conditions (J. Y. Chan et al., 2023). This method involves averaging haemoglobin changes across the time window and region of interest (ROI), followed by statistical testing to identify significant differences. Block-averaged fNIRS responses were calculated by subtracting the baseline average (i.e., the 5 s preceding stimulus onset) from each block. For each participant, the responses within each ROI were calculated as the grand mean of block-averaged fNIRS responses across channels in that ROI. Specifically, Oxy-Hb signals were averaged over the test window ([0 15] s) while preserving the direction of change. The 5-second period following the test block was included to capture the full response to the test stimuli, as Oxy-Hb peaks typically occur around 10 seconds, consistent with Sato et al. (2010a). This time window was selected based on previous fNIRS studies on infant hemodynamic responses, which have shown that the peak response typically occurs within this

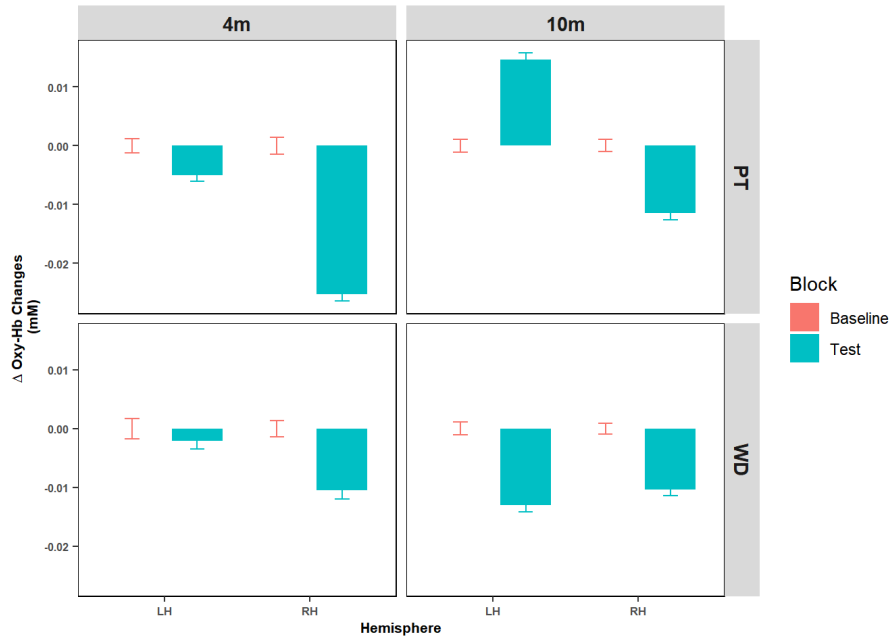
range (Sato et al., 2010a). Additionally, visual inspection of the grand-averaged time course confirmed that the key response patterns were indeed captured within this window. The resulting mean Oxy-Hb values were then used to quantify response strength, with more negative values reflecting greater decreases in Oxy-Hb and more positive values reflecting greater increases.

Following the approach by Sato et al. (2010a), we conducted multiple analyses to examine the effects of condition, hemisphere, and age on mean Oxy-Hb responses. First, we performed two separate two-way repeated measures ANOVAs for 4-month-olds and 10-month-olds, with condition (pseudoword vs. pure tone) and hemisphere (left vs. right) as within-subject factors. Additionally, we applied Generalised Linear Mixed-effects Models (GLMMs) to account for both fixed effects (condition and hemisphere) and random effects (subjects) in each age group. To assess interactions across age groups, we conducted a three-way ANOVA with condition, hemisphere, and age as factors. Similarly, we ran a full GLMM, incorporating condition, hemisphere, and age as fixed effects, with random effects of subjects. In all GLMMs, the random effects of subjects include random intercepts and slopes estimated for condition within subjects, accounting for individual variation in responses to the two conditions.

Figure 5.7 presents the averaged baseline and test block Oxy-Hb values for each combination of condition, hemisphere, and age. Figure 5.8 shows the averaged Oxy-Hb responses, with standard error bars, for each condition and hemisphere within each age group. Figure 5.9 illustrates the time course of Oxy-Hb changes from the ROIs.

**Figure 5.7**

*Averaged Oxy-Hb changes during baseline and test blocks across conditions, hemispheres, and age groups in Dutch infants.*




*Note.* Only error bars for Baseline conditions were plotted; the bars did not show up due to the fact that the mean values of Oxy-Hb Changes were too close to 0.

**Figure 5.8**

*Averaged Oxy-Hb changes in test blocks (0–15 seconds) across conditions, hemispheres, and age groups in Dutch infants.*

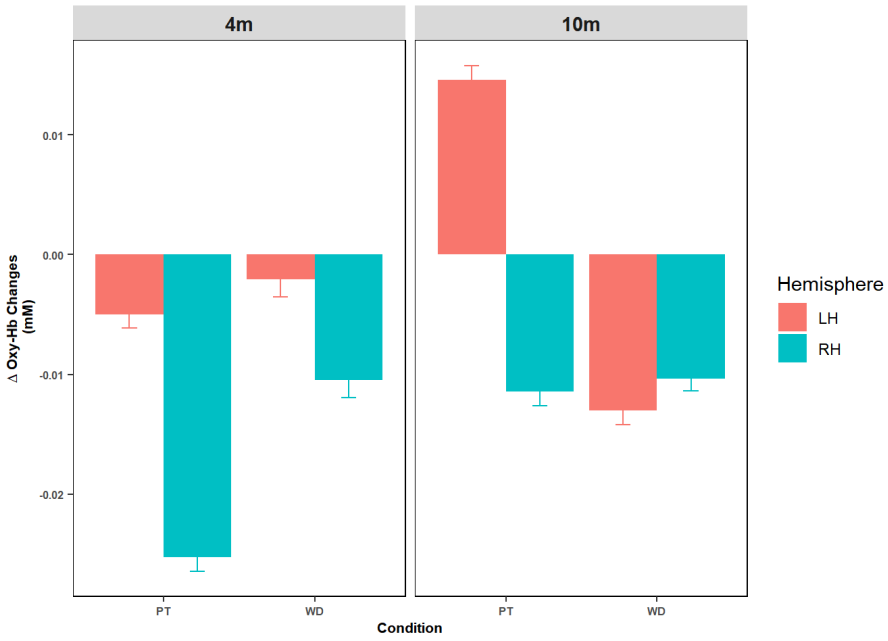
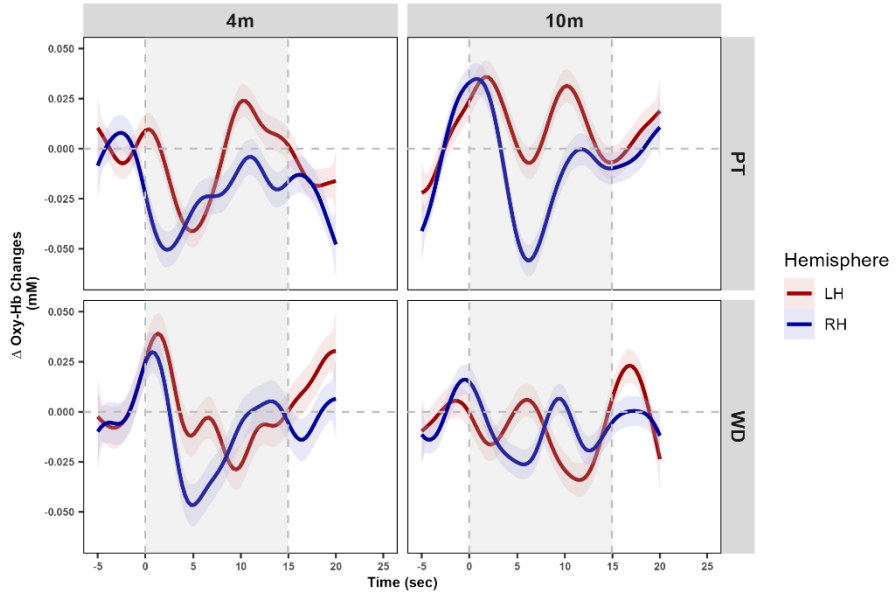


**Figure 5.9**

*Grand average time courses of Oxy-Hb changes from ROIs in the left (red) and right (blue) hemispheres across conditions and age groups in Dutch infants. Colour shaded areas represent standard errors.*




### 5.3.3 Results

The Kolmogorov-Smirnov test revealed that Oxy-Hb changes in both age groups, across conditions and hemispheres, were not normally distributed ($D$ = 0.41-0.42, $ps$ < 0.001). As a result, nonparametric tests were employed to analyse the data. Wilcoxon signed-rank tests were conducted to compare Oxy-Hb activations within the 0–15 s time window after stimulus onset against baseline values (calculated as the 5 s preceding stimulus onset) for each combination of condition, hemisphere, and age group. The results indicated significant differences between baseline and test blocks for certain combinations (4m: PT, LH, $W$ = 954820, $p$ = 0.1169; PT, RH, $W$ = 1103340, $p$ <

0.001; WD, LH, $W = 917151$, $p = 0.8823$; WD, RH, $W = 927003$, $p = 0.7636$; 10m: PT, LH, $W = 1386709$, $p < 0.001$; PT, RH, $W = 1636534$, $p < 0.05$; WD, LH, $W = 1662045$, $p < 0.01$; WD, RH, $W = 1651091$, $p < 0.01$).

As shown in Figure 5.9, we primarily observed inverted responses, where Oxy-Hb decreased after stimulus onset (timepoint = 0), for both types of stimuli in Dutch infants. The mean Oxy-Hb values within the test window, depicted in Figure 5.8, reflect the strength of these responses: more negative values indicate stronger negative responses. Notably, for the pure tone condition in 10-month-olds, the Oxy-Hb value in the left hemisphere was above zero, whereas all other values were below zero. Thus, in the 10-month-old Dutch infants, the right hemisphere exhibited stronger negative responses than the left in the pure tone condition. We now turn to statistical details from the ANOVA and GLMM analyses, comparing hemispheric responses across both conditions and age groups.

In 4-month-olds, a two-way ANOVA revealed stronger negative responses under the pure tone condition compared to the word condition ($F = 44.95$, $p < 0.001$), and stronger negative responses in the right hemisphere compared to the left hemisphere across both conditions ($F = 117.81$, $p < 0.001$). Furthermore, a three-way ANOVA indicated significantly stronger negative responses in the right hemisphere under the pure tone condition compared to the word condition ($p < 0.001$). These findings were confirmed by the GLMM analysis for 4-month-olds and the full GLMM analysis, which consistently showed a hemispheric difference under both conditions ($Bs = 0.00842–0.02025$, $SE = 0.00159–0.00165$, $Z = 5.106–12.709$, $ps < 0.001$).

In 10-month-olds, stronger negative responses were observed for pseudoword stimuli compared to pure tone stimuli ($F$ = 137.6, $p$ < 0.001), and they were stronger in the right compared to the left hemisphere ($F$ = 106.3, $p$ < 0.001). In the pure tone condition, stronger negative responses were found in the right than the left hemisphere ($p$ < 0.001), a result confirmed by the GLMM analysis for 10-month-olds ($B$ = 0.02601, $SE$ = 0.00136, $Z$ = 19.147, $p$ < 0.001). However, no significant hemispheric differences were observed under the word condition. The GLMM analysis for 10-month-olds and full GLMM analysis further confirmed the hemispheric difference under the pure tone condition ($Bs$ = 0.0260, $SE$ = 0.00136/0.00140, $Z$ = 18.616/19.147, $ps$ < 0.001) and the absence of a significant hemispheric effect in the word condition ($B$ = -0.00266, $SE$ = 0.00140, $Z$ = -1.904, $p$ = 1.0).

### 5.3.4 Discussion

In Experiment 2, we examined Dutch infants' brain responses in the temporal lobes while they passively listened to pitch changes under two conditions: pseudoword (speech) and pure tone (non-speech). The results showed more right-hemispheric activity for both speech and non-speech stimuli in 4-month-olds, whereas 10-month-olds exhibited more right-hemispheric activity only for non-speech stimuli. Notably, no lateralisation to the left hemisphere was observed in response to pitch changes in speech stimuli for the 10-month-olds.

The stronger activity in the right hemisphere observed in 4-month-olds across both conditions aligns with the acoustic hypothesis, which states that the right hemisphere is specialised in processing spectral information (Belin et al., 1998; Zatorre & Belin, 2001). The continued predominantly right-hemispheric activity in 10-month-olds when processing pure tones indicates that this acoustic processing of pitch

remains stable over time. This finding also confirms the role of the right temporal region in processing suprasegmental features of speech (Arimitsu et al., 2011; Cutler et al., 1997; Homae et al., 2006). The lack of significant left lateralisation in 10-month-olds when processing pitch in speech stimuli is likely due to the limited linguistic significance of pitch in Dutch, where it does not function to indicate lexical contrast. Instead, Dutch infants may rely on the acoustic differences between the tonal pitch patterns rather than representing them as phonologically meaningful categories (A. Chen et al., 2017). As we hypothesised in Experiment 1, while Dutch infants can distinguish pitch changes, they do not appear to interpret pitch as linguistically significant. The shift to bilateral activation in speech processing in 10-month-olds possibly reflects the increasing role of the left hemisphere when processing speech as infants approach their first year (Binder et al., 2000).

Experiment 3 will examine brain responses in Japanese infants, for whom pitch does serve as a lexical contrast.

## 5.4 Experiment 3. Hemispheric lateralisation of lexical pitch in Japanese infants (NIRS study)

Unlike Dutch, Japanese uses pitch contrasts as a distinguishing feature in its lexicon. By comparing the brain responses of Japanese infants with those of Dutch infants, we aim to determine whether pitch changes in pure tones and words elicit distinct patterns of hemispheric activation, reflecting the different roles of pitch across languages.

We expect to replicate the results of Sato et al. (2010a), where Japanese 10-months-old infants showed left lateralised brain responses to linguistic pitch changes, while activation related to changes in pure tones remained predominantly in the right hemisphere,

suggesting that their brains had attuned to pitch in speech being a meaningful linguistic cue (Sato et al., 2010a).

### 5.4.1 Method

#### *5.4.1.1 Participants*

We tested 27 Japanese infants aged 4 to 5 months (M = 134 days, range = 70–152 days) and 31 infants aged 10 months (M = 302 days, range = 290–325 days). All infants were healthy, full-term, and from Japanese-speaking families in the Tokyo area. Twenty-eight infants were excluded due to issues such as crying and probe refusal ($n$ = 1), sleeping ($n$ = 1), excessive motion artifacts ($n$ = 1), age ($n$ = 1), insufficient trials or channels ($n$ = 6), lack of gain or signal ($n$ = 5), undershoot/overshoot signal ($n$ = 6), and excessive noises ($n$ = 7).

The final sample included 16 infants aged 4–5 months (7 girls; M = 136 days, range = 125–152 days) and 14 infants aged 10 months (7 girls; M = 300 days, range = 290–311 days).

#### *5.4.1.2 Stimuli*

Same as in Experiment 2.

#### *5.4.1.3 Procedure*

Same as in Experiment 2.

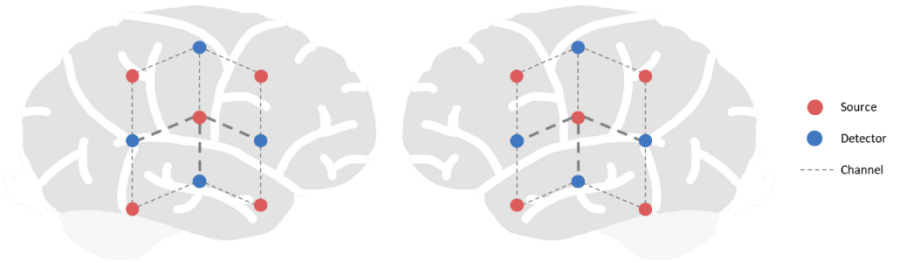
#### *5.4.1.4 NIRS recording: ETG-4000*

Hemodynamic responses to lexical pitch pattern changes in Japanese infants were recorded using a multichannel NIRS system (ETG-4000, Hitachi Medical Co., Tokyo, Japan) that employed continuous near-infrared lasers at two wavelengths (695 nm and 830 nm), with a sampling rate of 10 Hz. The fNIRS probes consisted of nine optical

fibres arranged in a 3x3 array, held in place by a soft silicone holder. On each hemisphere, five sources and four detectors were placed 3 centimetres apart, creating twelve channels per hemisphere. The middle probes on the lowest row were positioned near the T3 and T4 locations, according to the International 10-20 system, covering the bilateral temporal areas. A DIY cap with an adjustable string was used to fix the probes in place. See Figure 5.10 for an illustration of the ETG-4000 system's optode array configuration.

**Figure 5.10**

*Optode array configurations in Experiment 3 (ETG-4000).*




*Note.* Coloured circles represent the light incident probes: red for source optodes and blue for detector optodes. Dashed lines indicate channels, with bold dashed lines highlighting the channels of interest for the temporal lobe.

### *5.4.1.5 NIRS data preprocessing*

We preprocessed the NIRS data using Homer3 (Huppert et al., 2009), following the same processing streams as for the Dutch data. In contrast to the Dutch data, we observed negative intensity values in the Japanese data. These values potentially arise from factors such as noise fluctuations, ambient light interference, or minor inconsistencies in device calibration. They may also result from small numerical rounding errors during data transfer between devices or software (e.g., from HITACHI to Homer3, or from .csv to .snirf formats). The difference

between the Dutch and Japanese datasets may be due to variations in the sensitivity of the light sources and detectors used in the two systems (NIRScout for the Dutch data and ETG-4000 for the Japanese data). In addition, the data format exported from NIRScout is .nirs, which can be directly read by Homer3, potentially reducing the likelihood of rounding errors during data conversion. Since the negative values in the Japanese data appeared primarily at the beginning of the time series, we attributed them to rounding errors and did not consider them indicative of poor data quality. [1]

To correct the negative values in the Japanese data, we applied a positive offset to all data points, ensuring all values remained positive without altering haemoglobin signal shapes. This correction was necessary for subsequent frequency filtering.

After this adjustment, the data proceeded through the preprocessing pipeline illustrated in Figure 5.6, using all parameters detailed in Section 5.3.1.5 and Table 5.1.

### 5.4.2 Data analysis

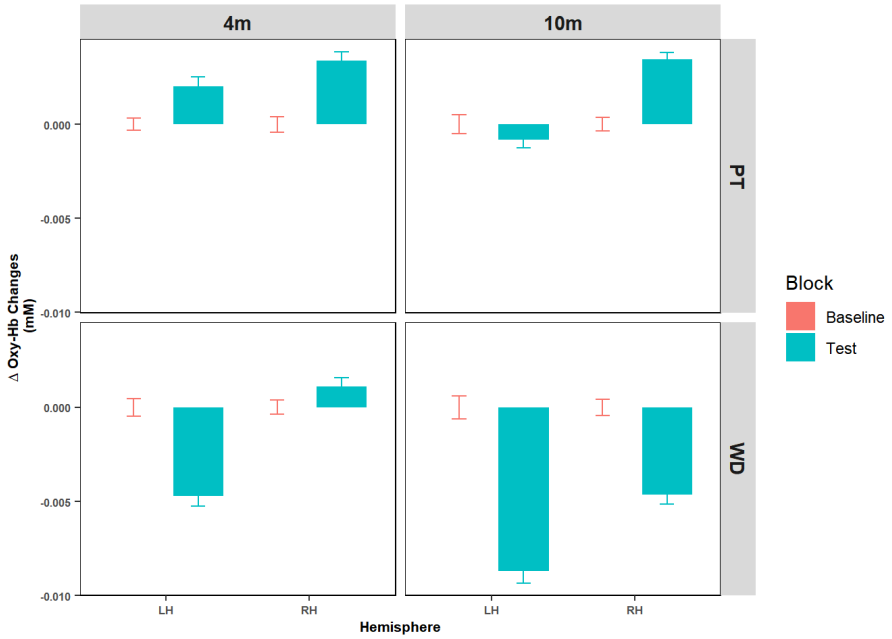
The analyses were the same as in Experiment 2. Figure 5.11 displays the averaged baseline and test block values for each combination of condition, hemisphere, and age group. Figure 5.12 shows the averaged Oxy-Hb responses with standard error bars for each condition and

[1] It is worth noting that negative intensity values are a known issue in the fNIRS community (see reports on the OPENFNIRS forum via this link https://openfnirs.org/community/?wpfs=PreprocessIntensity_Negative). While the exact causes, to the best of our knowledge, are not explicitly documented in other studies and do require further exploration, the presence of negative values has not raised concerns about the overall quality of the data.

hemisphere across age groups. Figure 5.13 illustrates the time course of the Oxy-Hb changes from the ROIs.

**Figure 5.11**

*Averaged Oxy-Hb changes during baseline and test blocks across conditions, hemispheres, and age groups in Japanese infants.*




*Note.* Only error bars for Baseline conditions were plotted; the bars did not show up due to the fact that the mean values of Oxy-Hb Changes were too close to 0.

**Figure 5.12**

*Averaged Oxy-Hb changes in test blocks (0–15 seconds) across conditions, hemispheres, and age groups in Japanese infants.*

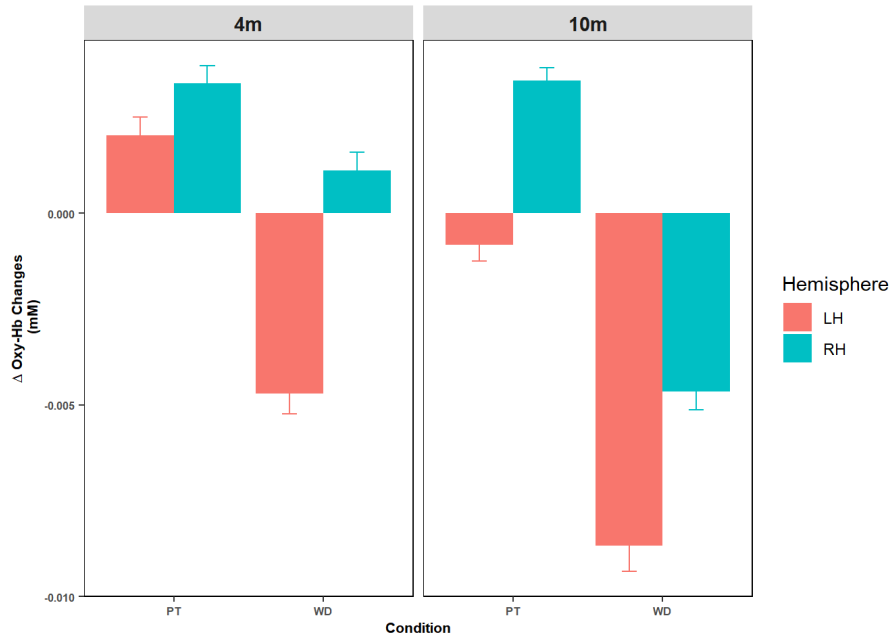
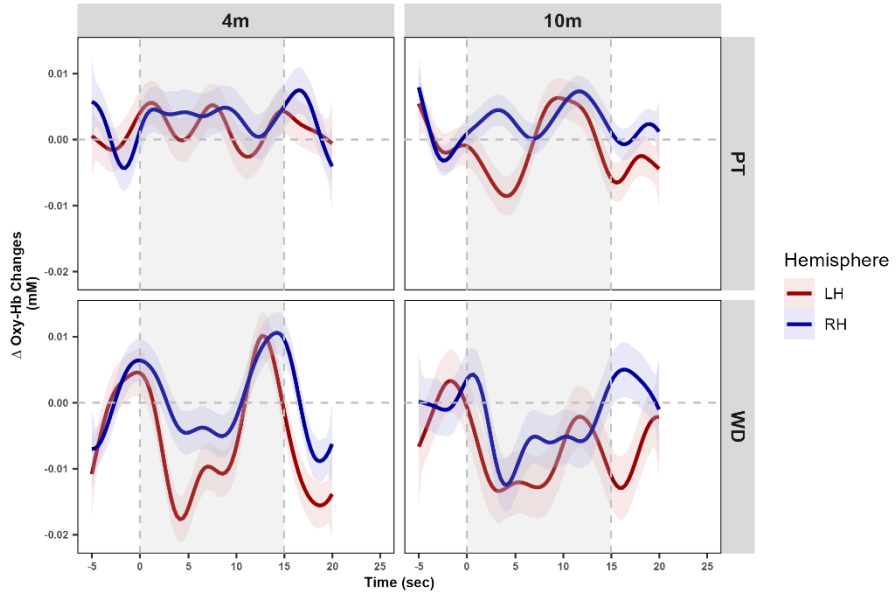


**Figure 5.13**

*Grand average time courses of Oxy-Hb changes from ROIs in the left (red) and right (blue) hemispheres across conditions and age groups in Japanese infants. Colour shaded areas represent standard errors.*




## 5.4.3 Results

The Kolmogorov-Smirnov test revealed that Oxy-Hb changes, across both age groups and conditions, in the left and right hemispheres were not normally distributed (*Ds* = 0.47, *ps* < 0.001). Consequently, Wilcoxon signed-rank tests identified significant differences between baseline (5 seconds before stimulus onset) and test blocks (0–15 seconds post-stimulus) for each combination of condition, hemisphere, and age group (4m: PT, LH, W = 224821, $p < 0.05$; PT, RH, W = 208876, $p < 0.001$; WD, LH, W = 263049, $p < 0.01$; WD, RH, W = 227098, $p = 0.07161$; 10m: PT, LH, W = 185624, $p = 0.9216$; PT, RH,

W = 153912, $p < 0.001$; WD, LH, W = 205309, $p < 0.01$; WD, RH, W = 192861, $p = 0.2314$).

As shown in Figure 5.13, Oxy-Hb responses in the pure tone condition typically followed a canonical pattern, whereas responses in the word condition were inverted in Japanese infants. The mean Oxy-Hb values within the test window, depicted in Figure 5.12, reflected response strength, where more negative values correspond to stronger negative responses and more positive values to stronger positive responses. Notably, for 10-month-olds in the pure tone condition, the Oxy-Hb value in the left hemisphere was below zero, whereas for 4-month-olds in the word condition, the Oxy-Hb value in the right hemisphere was above zero. This indicates that in 10-month-old Japanese infants, the right hemisphere exhibited stronger positive responses than the left hemisphere in the pure tone condition, while in 4-month-old Japanese infants, the left hemisphere exhibited stronger negative responses than the right hemisphere in the word condition. The following paragraphs provide statistical details from the ANOVA and GLMM analyses, comparing hemispheric responses across both conditions and age groups.

In the 4-month-olds, a two-way ANOVA revealed a significant interaction between condition and hemisphere, indicating stronger negative Oxy-Hb responses in the left hemisphere compared to the right hemisphere in the word condition ($p < 0.01$). This was further confirmed by the GLMM analysis ($B = -0.005816$, $SE = 0.000559$, $Z = -10.403$, $p < 0.001$).

For 10-month-olds, neither the two-way ANOVA nor the GLMM analysis showed a significant interaction between condition and hemisphere. However, both the three-way ANOVA and the full GLMM analysis revealed a significant condition-by-hemisphere-by-age

interaction, suggesting the following patterns: stronger positive responses in the right hemisphere than in the left hemisphere in the pure tone condition in 10-month-olds ($B$ = -0.00429, $SE$ = 0.000576, $Z$ = -7.454, $p$ < 0.001), stronger negative responses in the left hemisphere than the right hemisphere in the word condition in 10-month-olds ($B$ = -0.00403, $SE$ = 0.000576, $Z$ = -7.005, $p$ < 0.001), and stronger negative responses in the left hemisphere than the right hemisphere in the word condition in 4-month-olds ($B$ = -0.00582, $SE$ = 0.000539, $Z$ = -10.794, $p$ < 0.001).

Additionally, the three-way ANOVA showed that in the word condition, stronger negative responses were observed in both hemispheres in the 10-month-olds than in the 4-month-olds ($p$ < 0.001).

### 5.4.4 Discussion

In Experiment 3, we examined Japanese infants' brain responses in the temporal lobes to pitch changes in pseudowords (speech) and pure tones (non-speech). The results showed that Japanese infants exhibited bilateral responses or right-hemispheric activation for pitch changes in non-speech stimuli, a pattern similar to that observed in Dutch infants. In contrast, both 4- and 10-month-old Japanese infants demonstrated left-hemispheric lateralisation for speech stimuli.

The right-hemispheric activation for non-speech stimuli in 10-month-olds supports the notion that pitch processing is initially driven by lower-level acoustic properties (Belin et al., 1998; Zatorre & Belin, 2001). However, unlike Sato et al. (2010a), who reported no lateralisation for pure tone stimuli at either age group, we found rightward lateralisation for pure tones in 10-month-olds. This discrepancy may stem from differences in stimulus complexity: our

pure tone stimuli included second and third overtones, which could have enhanced right-hemisphere engagement. This aligns with the view that subtle acoustic variations influence hemispheric contributions to sound processing (Tervaniemi & Hugdahl, 2003).

The bilateral processing of pure tones in 4-month-old Japanese infants may reflect early cognitive demands in adapting to the linguistic features of their native language (Lovčević & Tsuji, 2024). Gu et al. (2013) proposed that experience with linguistic pitch could carry over to pure tone perception, potentially increasing left-hemisphere activation. This interaction may explain why younger infants did not show a clear right-hemispheric preference for non-speech stimuli.

Regarding speech stimuli, our findings show a more complex pattern than those reported by Sato et al. (2010a). While Sato et al. found left-lateralisation for pitch changes in pseudowords only in 10-month-olds, we observed this pattern in both 4- and 10-month-olds. This suggests that linguistic stimuli engage left-lateralised processing mechanisms even at an earlier age, and prior to the establishment of the first lexical representations. Similar early left-lateralisation has been reported by Telkemeyer et al. (2011) in 3-month-olds in response to slow acoustic modulations of speech, and by Ren et al. (2024), who observed left-lateralised processing in 2–4.5-month-old Chinese-learning infants when processing word sequences with tonal contrasts (T2 vs. T4). The early emergence of left-lateralised activation when processing pitch at the word level may reflect a developmental readiness to categorise pitch contrasts in a linguistically relevant way, facilitating later lexical meaning.

With increasing linguistic experience, left-hemispheric activation appears to strengthen. Our results show that 10-month-olds exhibit stronger left-hemispheric activation than 4-month-olds, suggesting

growing specialisation for processing pitch contrasts relevant to Japanese. This may reflect the increasing importance of pitch as a lexical cue as infants' linguistic exposure expands.

While these developmental interpretations are compelling, methodological differences between our study and Sato et al. (2010a) could also account for the discrepancies. Differences in fNIRS preprocessing pipelines, such as sensitivity to subtle Oxy-Hb changes or handling of systematic noise, may have influenced the detection of left-lateralised activity. Future research should systematically evaluate how preprocessing choices impact findings on hemispheric lateralisation.

## 5.5 General discussion

Across three experiments, we explored the developmental trajectory of pitch processing in Dutch and Japanese infants, uncovering both cross-linguistic similarities and differences in hemispheric activation and lateralisation, shaped by the acoustic properties and linguistic relevance of pitch. In Experiment 1, Dutch infants at 4 and 10 months distinguished pitch changes in speech stimuli, though individual differences were observed in the consistency of their discrimination (see Results in 5.2.3.2). Experiments 2 and 3 revealed that at 4 months, Japanese infants exhibited bilateral activity, while Dutch infants displayed predominantly the right-hemispheric activation in response to pitch changes in pure tones. By 10 months, both groups demonstrated right-hemispheric activation to pitch changes in pure tones. However, their responses to pitch changes in speech stimuli diverged: Dutch infants shifted from right-hemispheric to bilateral activation between 4 to 10 months; in contrast, Japanese infants

exhibited consistent left-hemispheric activation at both ages, reflecting the linguistic significance of pitch in their native language.

The leftward lateralisation in Japanese 10-month-olds aligns with the linguistic role of pitch in Japanese, confirming previous findings in adults and infants responding to real-word stimuli (Dehaene-Lambertz et al., 2006; Sato et al., 2007, 2010a). However, differing from Sato et al. (2010a), who reported bilateral responses in 4-month-old Japanese infants, we observed left-hemispheric activation at this early age. This suggests that sensitivity to the linguistic role of pitch may emerge earlier than previously thought. Additionally, we observed stronger left lateralisation at 10 months than at 4 months, indicating developmental reinforcement of pitch discrimination within speech. Similar early left-lateralisation has been reported in German-learning 3-month-olds for slow acoustic modulations of speech (Telkemeyer et al., 2011) and in 2-4.5-month-old Chinese-learning infants processing tonal contrasts (T2 vs. T4) (Ren et al., 2024). This early left-lateralised activation suggests a developmental readiness to categorise pitch contrasts in a linguistically relevant way. However, further research is needed to determine the precise timeline of this specialisation.

Although Sato et al. (2010a) observed bilateral responses to pure tone stimuli in Japanese infants at both ages, we found a shift from bilateral to right-hemispheric activation between 4 and 10 months. A key difference may lie in stimulus complexity: our study used complex pure tones incorporating second and third overtones, increasing spectral variation while maintaining temporal characteristics. This enhanced complexity likely contributed to rightward lateralisation, consistent with the right hemisphere's sensitivity to acoustic complexity (Tervaniemi & Hugdahl, 2003). However, the reason for the bilateral response in 4-month-olds remains unclear. One possible

explanation is the carryover hypothesis (Gu et al., 2013), which suggests that linguistic pitch processing may influence non-speech stimuli, potentially enhancing left-hemisphere activation. At 4 months, Japanese infants may not yet fully differentiate linguistic and non-linguistic pitch processing, leading to a more bilateral response. By 10 months, however, pitch processing in speech and non-speech contexts may have become more distinct, reducing the influence of linguistic pitch processing on pure tone perception and leading to right-hemispheric specialisation. Further investigation is needed to illustrate the underlying mechanisms.

For Dutch infants, both age groups exhibited rightward lateralisation in response to pure tones, consistent with the right hemisphere's specialisation for spectral variation and slower pitch changes (Arimitsu et al., 2011; Belin et al., 1998; Cutler et al., 1997; Homae et al., 2006; Zatorre & Belin, 2001). In speech stimuli, Dutch infants shifted from right-hemispheric to bilateral activation from 4 to 10 months. The right-hemispheric dominance at 4 months likely reflects an early-stage reliance on acoustic properties for pitch processing. By 10 months, the shift to bilateral activation may indicate an emerging sensitivity to prosodic structure, aligning with adult-like bilateral pitch processing of linguistic prosody in Dutch (e.g., Witteman et al., 2011, 2014). This pattern supports the idea that infants' neural specialisation for pitch develops in response to their native language environment.

The comparison between Dutch and Japanese infants suggests that both language groups successfully discriminate pitch in both pure tones and pseudowords. However, pitch in non-linguistic contexts is predominantly processed in the right temporal cortex (Aslin, 2012), while linguistic pitch processing develops based on the language-

specific relevance of acoustic features. In Japanese, where pitch serves as a lexical cue, acoustic information is routed from the primary auditory cortex to higher-level processing areas in the left temporal cortex. In contrast, Dutch infants did not show explicit hemispheric lateralisation for speech stimuli, indicating a tendency to process word prosody as intonation. This functional specialisation emerges through language-specific experience, shaping distinct developmental trajectories in Japanese and Dutch infants.

Moreover, perceptual narrowing in speech perception is guided by implicit learning mechanisms, where language-specific acoustic features become prioritised (Bosseler et al., 2013). During this process, both bottom-up lateralisation (driven by acoustic features) and top-down modulation (shaped by prior knowledge and/or attention) interact to drive hemispheric specialisation (Hugdahl, 2000). Supporting this, Homae et al. (2011) and Ren et al. (2024) found that the frontal lobe plays a crucial role in speech processing and category perception. Specifically, greater frontal region involvement is observed for tone categorisation as infants aged (Ren et al., 2024), and the relationship between the frontal and temporal regions becomes stronger after exposure to speech stimuli (Homae et al., 2011). Although our observation is limited to temporal activity, the comparison across groups also provided valuable insights on how bottom-up (acoustic features) and top-down modulation (language experience or prior knowledge) interact to drive hemispheric specialisation. Future studies should incorporate both frontal and temporal lobe measures to better understand their interaction in hemispheric specialisation and perceptual reorganisation.

Interestingly, our study revealed predominantly inverted haemoglobin responses. Canonical haemoglobin responses,

characterised by an increase in regional blood oxygenation, were observed only in response to pure tone pitch changes in Japanese infants. In contrast, inverted responses, marked by decreased oxygenation, were observed for speech stimuli in Japanese infants and both speech and non-speech stimuli in Dutch infants at both ages. As reviewed by Issard and Gervain (2018), inverted responses are common in infant fNIRS studies and may reflect stimulus complexity, cognitive demands, or developmental factors. Given that our stimuli and experimental design were identical across groups, we hypothesise that these differences arise from variations in cognitive processing development specific to each language environment. Japanese and Dutch infants may engage different neural mechanisms for pitch processing, shaped by their linguistic environments. In Japanese, where pitch is crucial for distinguishing meaning, infants may develop an early augmented sensitivity to pitch variations, leading to a canonical hemodynamic response for pure tones. However, the inverted response to speech stimuli suggests that linguistic pitch processing imposes additional cognitive demands, possibly due to ongoing refinement of speech-specific pitch perception and processing. In contrast, since pitch is not a lexical cue in Dutch, infants may rely on general auditory mechanisms when processing pitch differences, which might require more cognitive effort, resulting in inverted responses regardless of stimulus type. Alternatively, the inverted responses could have resulted from the influence of our preprocessing pipelines and algorithms on the signal extractions. Further research is needed to uncover the mechanisms underlying these haemoglobin response patterns.

In summary, the dual acoustic-functional properties of lexical pitch accent provide valuable insights into the perceptual

reorganisation and the concomitant hemispheric lateralisation during infancy. By comparing Dutch and Japanese infants, we explored whether the location of pitch processing in the brain is determined by the function of pitch in distinguishing lexical meaning. The observed similarities and differences highlight the role of language experience in shaping auditory processing pathways in infants. That is, while Japanese infants seem to develop a linguistically-oriented sensitivity to pitch changes, Dutch infants do not show such a specialisation. Finally, data preprocessing played a crucial role in the observed results. Due to the fact that the Japanese and the Dutch infants were tested in two different labs, different fNIRS systems and probes were used in Experiments 2 and 3. To minimise differences, we employed consistent experimental designs and procedures across the two labs. During data processing, we carefully adjusted parameters for motion detection and correction to ensure reliable signal extraction. By comparing results across multiple pipelines, we identified the most suitable approach for both Dutch and Japanese datasets. However, other preprocessing factors, such as filtering techniques and motion correction algorithms, may also influence extracted signals. Future studies should systematically explore these methodological variations to enhance data extraction and interpretation and to contribute to standardising fNIRS data analyses in infant research.

# Chapter 6
# General discussion

This dissertation investigates the multifaceted roles of phonemic pitch in two distinct contexts: (1) the affective iconicity of linguistic pitch in Standard Chinese (SC) tones and (2) the developmental hemispheric lateralisation of linguistic pitch during perceptual reorganisation in Dutch- and Japanese-learning infants. In the first context, the study explores: (**R1**) whether the linguistic pitch of SC lexical tones biases emotional responses (arousal and valence) in adult speakers, (**R2**) how segmental phonemes influences emotional responses, (**R3**) whether pitch dominates segmental phonemic cues in impacting emotional responses, and (**R4**) the potential mechanisms through which lexical pitch affects emotional responses. In the second context, the study examines cross-linguistic differences in infants' brain responses to lexical pitch, comparing learners of a pitch accent language (Japanese) and a lexical stress (Dutch) language. Specifically, it addresses research questions (**R5**) whether Dutch-learning infants can distinguish lexical pitch contrasts, and (**R6**) how Dutch- and Japanese-learning infants differ in their ability to discriminate lexical pitch contrasts.

By addressing these research questions, this work provides insights into the relationships between linguistic pitch, emotional responses, language experience, brain development, and hemispheric lateralisation. The following two sections summarise each chapter, highlighting key findings, contributions, limitations, and directions for future research.

## 6.1 Research questions and findings

**Chapter 2** explored the potential pitch iconic effects of lexical tones on affective interpretations in bisyllabic words in Standard Chinese (**R1**). Our findings provide compelling evidence that affective iconicity in Standard Chinese (SC) lexical tones systematically influences both

the general lexicon (as analysed across three corpora) and spoken nonce words. Consistent with our hypotheses, the FF tonal sequence, characterised by extensive pitch variation in range and slope, was consistently associated with higher arousal compared to RR across both written words (Studies 1 and 2) and spoken nonce words (Study 3). In spoken nonce words, HH was more likely to be rated as low arousal compared to the other three tonal sequences: RR, RL (LL), and FF. Additionally, a significant influence of tonal sequences on valence was observed exclusively in nonce words. Specifically, FF and RL tonal sequences were more likely to be rated as having negative valence, whereas RR and HH sequences were associated with positive valence.

The pitch iconic effects of lexical tones on affective interpretations in bisyllabic words in SC mirror the intonational pitch variations in emotional expressions, suggesting a universal tendency for pitch cues to shape emotional communication. Specifically, 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).

**Chapter 3** examined how lexical tones, vowels, consonants, and their interactions contribute to affective iconicity. We first replicated widely observed phonemes-emotional meaning correspondences (**R2**). For vowels, /u/ was associated with negative valence and /i/ with positive valence across both vowel-only and CV(L) monosyllabic syllables, consistent with previous studies. Regarding consonants, /t/ was more likely to be rated as high arousal, while /n/ was associated with low arousal. At the monosyllabic level, lexical tones also showed

systematic effects: T2 and T4 were associated with higher arousal compared to T1 and T3, while T1 and T2 were more likely to be associated with positive valence compared to T3 and T4. In line with our hypothesis, lexical tones emerged as stronger predictors of emotional arousal and valence compared to phonemes. Specifically, when lexical tones and phonemes predicted opposing directions for valence and arousal, lexical tones were more decisive. The results revealed significant associations between tonemes and both arousal and valence, consonants and arousal, and vowels and valence, supporting the idea that native Standard Chinese speakers integrate both phonemes and tonemes when interpreting the affective meanings of auditory stimuli. However, no significant consonant-valence or vowel-arousal correspondences were observed in this study.

The predominant effect of lexical pitch over phonemes in affective interpretations lends further support to the notion that lexical-prosodic features facilitate the interpretation of iconic meanings (Dingemanse et al., 2016). This finding aligns with previous research demonstrating the dominance of emotional prosody over verbal content in emotion recognition (Filippi et al., 2017; Lin et al., 2020, 2021). The shared phonetic feature between lexical tones and emotional prosody, pitch, a highly salient perceptual dimension, suggests that these systems may rely on shared or transferable mechanisms in emotional processing. Specifically, the prominence of pitch in both lexical tones and emotional prosody highlights its central role in conveying affective meaning, whether through linguistic or paralinguistic channels.

**Chapter 4** investigated the adaptive significance of lexical tones in Standard Chinese in affective iconicity. Using the three corpus datasets from Chapter 2, we found that individual tones significantly

influenced the arousal and valence ratings of bisyllabic Standard Chinese words. Both the lexical tones of the initial and second syllables predicted the arousal ratings of words. Specifically, high-arousing words were characterised by a falling tone (T4) in both syllables, while low-arousing words were more likely to be characterised by a rising (T2) tone or a low-dipping tone (T3). In contrast, the valence ratings of the bisyllabic words were predominantly influenced by the lexical tone of the first syllable. Negative words were more likely to initiate with a falling tone (T4), while positive words tended to start with a rising tone (T2). These findings align with our hypotheses and suggest a potential mechanism of affective iconicity due to adaptive significance (Adelman et al., 2018). Specifically, lexical tones with falling contours are associated with negative, high-arousing words, signalling the urgency and importance of the stimulus or information, thereby facilitating communication efficiency. Such an interpretation of the lexical tone effects on emotional expression lends further support to the adaptive significance proposal by Adelman et al. (2018), which is based on findings from segmental phonemes.

**Chapter 5** explored the developmental trajectory of pitch processing in Dutch and Japanese infants, revealing cross-linguistic similarities and differences in hemispheric lateralisation influenced by the acoustic properties and linguistic relevance of pitch. Across three experiments, we observed that Dutch infants at both 4 and 10 months could distinguish pitch changes within speech stimuli, though with some individual differences in the consistency of their discrimination (Experiment 1). In Experiments 2 and 3, Dutch infants showed right-lateralised responses to both pure tones and pseudowords at 4 months. By 10 months, their responses to pure tones remained right-lateralised, but for pseudowords, they shifted to a bilateral pattern. Japanese

infants exhibited bilateral responses to pure tones at 4 months, but by 10 months, their responses became right-lateralised. For pseudowords, they exhibited left-lateralised responses at both ages, with stronger activation at 10 months than at 4 months. These findings align with our hypotheses, highlighting the influence of both acoustic and linguistic factors in early pitch processing. However, the left-lateralised response to pseudowords in 4-month-old Japanese infants may indicate an early developmental sensitivity to linguistically relevant pitch contrasts. This difference between Dutch and Japanese infants, particularly in the older group, reflects how pitch processing is shaped by its linguistic function, emphasising the influence of language experience.

## 6.2 Contributions, limitations, and future directions

This dissertation provides new insights into the affective and developmental implications of linguistic (phonemic) pitch, revealing its multifaceted roles across acoustic, cognitive, and linguistic domains. The findings emphasise that pitch processing is not solely determined by acoustic properties but also shaped by cognitive processing, linguistic experiences, and relevance.

### 6.2.1 Phonemic pitch in affective iconicity

Our exploration of affective iconicity highlights how iconicity adapts to specific linguistic systems, particularly Standard Chinese (SC), and its universal role in human communication.

#### *Lexical tones and emotional responses*

We demonstrated that lexical tones significantly shape emotional responses, including arousal and valence. Lexical tones at both the monosyllabic and bisyllabic levels, regardless of lexical meaning,

predict emotional arousal. Valence predictions, however, were observed primarily in bisyllabic nonce words. There are a few possibilities explaining this difference.

One is that iconicity itself is pre-semantic. 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). These findings suggest that iconicity operates at the acoustic level, independent of semantic content, with pitch features such as overall level, average height, range, contour direction, and slope driving emotional associations. For instance, the falling contour, large pitch range, and steep slope of falling tone (T4) may explain its association with high arousal and negative valence, while the nearly static pitch and high level of high-level tone (T1) align with its low-arousal, positive valence tendency. Similarly, the rising contour and moderate pitch range of the rising tone (T2) appear linked to positive valence. Notably, bisyllabic falling-falling (T4T4) tonal sequences evoked higher arousal than rising-rising (T2T2) tonal sequences, despite there being no significant difference in the monosyllabic context. This discrepancy suggests that pitch dynamics, syllable structure, and tonal articulation might interact, warranting further exploration with these considerations. In addition, our stimuli in the corpora and nonce words are limited; future research may consider larger datasets and diverse phonemes to thoroughly investigate tone–arousal and tone–valence association.

Another possibility is that while both associations may be pre-semantic, arousal is more pronounced than valence for iconicity at the

level of lexical tone, given their different effect sizes in Chapter 2. This aligns with findings by Aryani et al. (2018), which demonstrated that phonemic segments more strongly influence arousal than valence. As two key emotional dimensions, arousal is a more ancient and universal dimension tied to physiological responses, whereas valence is a more abstract dimension influenced by cultural and semantic contexts (Darwin, 1998; Russell, 2003). The differences between the two dimensions might lead to varying degrees of iconic associations with tone. For example, tone–valence association is easily shadowed by lexical meanings and conventions. Further replication and exploration are necessary to test these possibilities. Cross-linguistic studies could further investigate whether naïve speakers with no knowledge of SC would perceive affective iconicity in tonal sequences. Additionally, experiments with homophonic SC bisyllabic words, where tones vary but lexical meanings remain constant, might clarify whether lexical tone alone influences valence ratings.

### *Affective iconicity and the adaptation hypothesis*

Our findings revealed distinct patterns of tone–arousal and tone–valence associations and supported the adaptation account of iconicity. We observed positional effects of lexical tones on emotional arousal and valence ratings: the first-syllable tone of bisyllabic words predicted valence, while both syllables' tones predicted arousal. This positional effect of lexical tones on emotional valence rating replicates findings from Indo-European languages, where the initial phoneme often predicts valence, and extends the adaptation hypothesis to include tonemes.

These results align with the distinctions for arousal and valence (Citron, 2012; Darwin, 1998) and their neural processing mechanisms

(Fischler & Bradley, 2006; Herbert et al., 2008; Kissler et al., 2007, 2009; Schacht & Sommer, 2009). Moreover, Aryani et al. (2019) revealed a potential neural mechanism underlying affective iconicity in an fMRI study. They found that affective iconic words, compared to their non-iconic counterparts, elicited additional brain activity in the left amygdala, a region associated with the multimodal representation of emotions. This effect was modulated by sound processing in the left superior temporal gyrus and language processing in the left inferior frontal gyrus.

Future research, particularly neuroimaging studies, could further investigate the neural pathways involved in pitch iconicity and their interactions with emotional and linguistic processing regions. Examining the mechanisms underlying pitch iconicity may offer new insights into the neurocognitive foundations of tonal and affective processing. Specifically, tones may engage distinct neural pathways that integrate both linguistic and emotional processing, potentially involving interactions between auditory and affective brain regions. Additionally, exploring whether speakers of tonal and non-tonal languages process pitch iconicity differently could shed light on the brain's plasticity in adapting to diverse linguistic environments. Such findings would have broader implications for understanding how language experience shapes neural responses to emotionally salient speech cues.

### *Lexical tones dominate over phonemes in emotional predictions*

Lexical tones demonstrated a stronger influence than phonemes in predicting emotional responses. While phoneme-based associations such as "/i/–positive," "/u/–negative," "/t/–high arousal," and "/n/–

low arousal" were replicated, tonal predictions often overrode phonemic ones, particularly in cases of conflict (e.g., /i4/ versus /u1/ were rated as negative versus positive). Lexical-prosodic information has been considered a facilitator in interpreting iconic meanings across languages (Dingemanse et al., 2016; Perlman, Clark, et al., 2015; Stel et al., 2012; Thompson, 2018). Our findings further underscore the prioritisation of suprasegmental features over segmental ones in affective interpretation, emphasising the critical role of prosody in tonal languages.

This dominance of lexical tones suggests that tonemes may function as a dual-purpose tool for conveying both lexical and emotional meaning. Investigating the relative contributions of segmental versus suprasegmental features in tonal and non-tonal languages could provide deeper insights into how linguistic systems harness pitch for communication.

As discussed in the *Lexical tones and emotional responses* section, we propose that iconicity operates at the acoustic level, possibly independent of semantic content. This suggests that pitch, beyond its primary role in distinguishing lexical meaning in tonal languages, retains a residual capacity to convey emotion. Even with this residual capacity, pitch can exert a stronger influence than phonemes in emotional expressions, highlighting its salience in conveying affective meanings. However, further studies replicating this phenomenon across diverse language users and dialects are needed to provide more robust evidence.

### *Pitch iconicity and the frequency code hypothesis*

Some of our findings on pitch iconicity in SC lexical tones provide further support for the frequency code hypothesis, which posits

universal principles of pitch-iconicity based on acoustic properties (Gussenhoven, 2016; Hinton et al., 2006; Ohala, 1984). According to this hypothesis, higher and/or rising *f0* is associated with smallness, submission, and politeness, whereas lower and/or falling *f0* conveys largeness, dominance, and aggression.

Particularly, we assume that the *f0* difference between T2/T4 and T3 (high versus low-dipping) likely contributed to their arousal contrast, and differences in *f0* between T1/T2 and T3/T4 (high/rising versus low-dipping/falling) may explain their valence choice differences. However, it is important to note that these arousal and valence differences among lexical tones are more likely influenced by the tones' pitch characteristics, such as pitch contour directions, pitch ranges, and pitch slope.

It is important to bear in mind that while the frequency code may contribute to the sound-symbolic link between pitch and affective connotations, the relationship is not entirely straightforward, as shown in studies such as Morton (1977), Perlman (2024), and Winter et al. (2021). Future research with more nuanced experimental designs is needed to clarify how specific pitch features function in affective iconicity and how these functions relate to the frequency code hypothesis.

### *Broader implications and future directions*

The investigation of tonemes as affective carriers enriches our understanding of iconicity by extending it from segmental to suprasegmental features. Cross-linguistic studies are crucial to determining whether the affective iconicity of tonal sequences, rooted in pitch variations, is universal. Additionally, research on the interaction of tone and intonation could shed light on how tonal

languages integrate prosodic and lexical functions to convey meaning efficiently. Affective iconicity may also play a role in reducing cognitive load by allowing listeners to infer emotional meaning directly from acoustic cues. This effect could be particularly relevant in tonal languages, where pitch is a dominant linguistic feature. Furthermore, iconicity has been shown to facilitate word learning, which may be especially beneficial for children and second-language learners (e.g., Asano et al., 2015; Imai et al., 2008, 2015; Lockwood et al., 2016; Nielsen & Dingemanse, 2021). By advancing the study of pitch iconicity, this research not only deepens our understanding of tonal languages but also contributes to broader theories of language acquisition, sound symbolism, and emotional communication. Future work should explore how pitch-based iconicity interacts with other linguistic and cognitive processes, shedding further light on the fundamental links between speech, emotion, and meaning.

### 6.2.2 Phonemic pitch in perceptual reorganisation

Investigating perceptual reorganisation and its associated hemispheric lateralisation in infants is crucial for understanding the developmental trajectory of language acquisition and the neural mechanisms underlying this process. Between 6 and 12 months, infants transition from being universal listeners, capable of discriminating most phonemic contrasts, to becoming specialists attuned to their native language(s). This perceptual reorganisation process forms the foundation for acquiring phonology. Understanding how and why infants lose sensitivity to nonnative contrasts provides critical insights into the timing and mechanisms of language learning and acquisition.

### *Hemispheric lateralisation and perceptual reorganisation of pitch*

Across the lifespan, pitch remains an indispensable feature of auditory perception, serving multiple functions in linguistic processing, emotional and social communication, and musical perception. The early shift from broad pitch sensitivity to language-specific specialisation reflects the brain's ability to adapt to environmental input, a hallmark of perceptual reorganisation (Gervain, 2020; Werker, 2018).

Our findings provide further evidence on the role of pitch in hemispheric lateralisation by comparing Japanese and Dutch infants. While the acoustic properties of pitch remain constant across languages, its linguistic functions differ, serving as a phonemic contrast in Japanese but not in Dutch. This cross-linguistic comparison sheds light on how perceptual properties interact with language experience and linguistic relevance or function to shape phonemic pitch processing. Perceptual reorganisation reflects experience-dependent plasticity, wherein the brain selectively enhances sensitivity to relevant phonetic cues while deprioritising those that are not functionally meaningful in the infant's linguistic environment (Minagawa-Kawai et al., 2011).

At 10 months, Japanese and Dutch infants demonstrated diverging neural responses to pitch-based speech stimuli, supporting the hypothesis that pitch processing pathways are shaped by the linguistic relevance of acoustic features. Specifically, pitch cues deemed "language-relevant" appear to be routed from the primary auditory cortex to higher-level processing areas in the left temporal cortex, whereas general acoustic properties are predominantly processed in

the right temporal cortex (Aslin, 2012). This functional specialisation underscores the role of language experience in shaping developmental trajectories. Our findings suggest that Japanese infants develop a phonologically oriented sensitivity to pitch, reflecting the importance of pitch-accent distinctions in their native language, while Dutch infants show no such specialisation, treating pitch as a general auditory feature.

We found an early emergence of left-lateralised processing for linguistic pitch in 4-month-old Japanese infants, aligning with previous studies (e.g., Ren et al., 2024; Telkemeyer et al., 2011). This early lateralisation may reflect a developmental readiness to acquire pitch contrasts in a linguistically relevant way. Moreover, by 10 months, left-hemisphere activation became stronger, suggesting increasing specialisation for processing pitch contrasts relevant to Japanese. This developmental shift likely results from both accumulating linguistic experience and the critical role of pitch in Japanese. These findings suggest a hierarchical developmental trajectory between phonological and lexical processing. One possibility is that phonological pitch processing emerges early, providing a foundation for later lexical development. Alternatively, phonological pitch processing may gradually strengthen over time, eventually integrating with lexical-level processing. Further systematic investigations across different language backgrounds are needed to clarify these developmental mechanisms.

The hemispheric response to pitch in non-speech stimuli primarily reflected processing based on acoustic properties (Aslin, 2012), with slower pitch variations predominantly processed in the right hemisphere. However, Dutch infants exhibited a shift from right-lateralised (at 4 months) to bilateral responses (at 10 months) for

speech stimuli. The right-hemispheric dominance at 4 months likely reflects an early-stage reliance on acoustic properties for pitch processing. By 10 months, the shift to bilateral activation may indicate an emerging sensitivity to prosodic structure, aligning with adult-like bilateral pitch processing to linguistic prosody in Dutch (e.g., Witteman et al., 2011, 2014). This further supports the idea that infants' neural specialisation for pitch develops in response to their native language environment.

Moreover, previous studies have shown increased involvement of the frontal region in tone categorisation as infants mature (Ren et al., 2024), and stronger connectivity between the frontal and temporal regions following exposure to speech stimuli (Homae et al., 2011). These findings suggest that other top-down modulations, shaped by prior knowledge, attention, and/or executive functions, also play a role in pitch processing. Future research should incorporate measures of both frontal and temporal lobe activity to better understand their interaction in hemispheric specialisation and perceptual reorganisation.

In summary, the dual acoustic-functional properties of lexical pitch accent provide valuable insight into the developmental trajectory of perceptual reorganisation and the corresponding hemispheric lateralisation of auditory processing. By comparing Dutch and Japanese infants, we explored whether early pitch processing is influenced by language-specific phonological categories or remains a general auditory skill at this developmental stage. The observed similarities and differences highlight the role of linguistic, experiential, and cognitive contexts in shaping early auditory processing pathways. Specifically, Japanese infants may develop a phonologically-oriented sensitivity to pitch, while infants in Dutch show no such specialisation.

### *Methodological considerations and future directions*

From a methodological perspective, fNIRS remains one of the most effective non-invasive neuroimaging techniques for studying infant brain responses due to its good temporal and spectral resolution, suitability for testing awake infants (Wilcox & Biondi, 2015). The adoption of fNIRS has facilitated the investigation of pitch processing variations across language groups and auditory contexts. Although there is no consensus on fNIRS preprocessing practices, we employed well-justified pipelines and further refined them by applying different parameters critical for haemoglobin signal extraction. However, key preprocessing steps, such as motion correction and baseline correction, require further refinement to improve consistency and comparability across studies. Standardising these techniques will enhance the reliability of fNIRS research on infant speech perception and neural specialisation.

Current studies in Dutch and Japanese infants suggest that multiple factors influence pitch processing, leading to variations in hemispheric lateralisation. Future research could explore more salient pitch contrasts, such as the distinction between T1 and T4 in Standard Chinese, to better understand how infants from different language backgrounds process pitch variations during perceptual reorganisation. Additionally, investigating how pitch saliency and relevance vary across language contexts, shaped by both pitch function and cognitive processing, could provide deeper insights into early pitch perception mechanisms.

## 6.3 Conclusion

This dissertation highlights two interrelated roles of pitch processing: affective iconicity in Standard Chinese (SC) speakers and perceptual reorganisation in Japanese (JP) and Dutch (NL) infants. Despite their differences, both cases illustrate how pitch perception emerges from the interplay between acoustic properties and linguistic experience.

In the affective iconicity context, SC speakers systematically associate pitch with emotional valence and arousal, demonstrating that pitch perception extends beyond linguistic functions to iconic sound-meaning mappings. Meanwhile, in the developmental context, JP and NL infants process phonemic pitch differently, with JP infants showing greater left-hemisphere involvement, which reflects the linguistic relevance of pitch in their native languages.

These findings underscore pitch as a dynamic cue shaped by experience and communicative function, whether as a lexical contrast (in JP and NL) or an emotional signal (in SC). By bridging insights from affective iconicity and hemispheric lateralisation, this work deepens our understanding of how the brain adapts to the auditory environment, refining pitch processing based on both linguistic experience and cognitive demands.

# References

Abboub, N., Nazzi, T., & Gervain, J. (2016). Prosodic grouping at birth. *Brain and Language*, *162*, 46–59. https://doi.org/10.1016/j.bandl.2016.08.002

Adelman, J. S., Estes, Z., & Cossu, M. (2018). Emotional sound symbolism: Languages rapidly signal valence via phonemes. *Cognition*, *175*, 122–130. https://doi.org/10.1016/j.cognition.2018.02.007

Akita, K. (2021). A typology of depiction marking: The prosody of Japanese ideophones and beyond. *Studies in Language*, *45*(4), 865–886. https://doi.org/10.1075/sl.17029.aki

Anikin, A., & Johansson, N. (2019). Implicit associations between individual properties of color and sound. *Attention, Perception, & Psychophysics*, *81*(3), 764–777. https://doi.org/10.3758/s13414-018-01639-7

Arimitsu, T., Minagawa, Y., Yagihashi, T., O. Uchida, M., Matsuzaki, A., Ikeda, K., & Takahashi, T. (2018). The cerebral hemodynamic response to phonetic changes of speech in preterm and term infants: The impact of postmenstrual age. *NeuroImage: Clinical*, *19*, 599–606. https://doi.org/10.1016/j.nicl.2018.05.005

Arimitsu, T., Uchida-Ota, M., Yagihashi, T., Kojima, S., Watanabe, S., Hokuto, I., Ikeda, K., Takahashi, T., & Minagawa-Kawai, Y. (2011). Functional hemispheric specialization in processing phonemic and prosodic auditory changes in neonates. *Frontiers in Psychology*, *2*, Article 202. https://doi.org/10.3389/fpsyg.2011.00202

Aryani, A., Conrad, M., & Jacobs, A. M. (2013). Extracting salient sublexical units from written texts: "Emophon," a corpus-based approach to phonological iconicity. *Frontiers in Psychology*, *4*, Article 654. https://doi.org/10.3389/fpsyg.2013.00654

Aryani, A., Conrad, M., Schmidtke, D., & Jacobs, A. (2018). Why "piss" is ruder than "pee"? The role of sound in affective meaning making. *PLoS ONE*, *13*(6), e0198430. https://doi.org/10.1371/journal.pone.0198430

Aryani, A., Hsu, C.-T., & Jacobs, A. M. (2019). Affective iconic words benefit from additional sound-meaning integration in the left amygdala. *Human Brain Mapping*, *40*(18), 5289–5300. https://doi.org/10.1002/hbm.24772

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex*, *63*, 196–205. https://doi.org/10.1016/j.cortex.2014.08.025

Aslin, R. N. (2012). Questioning the questions that have been asked about the infant brain using NIRS. *Cognitive Neuropsychology*, *29*(1–2), 7–33. https://doi.org/10.1080/02643294.2012.654773

Aslin, R. N., & Mehler, J. (2005). Near-infrared spectroscopy for functional studies of brain activity in human infants: Promise, prospects, and challenges. *Journal of Biomedical Optics*, *10*(1), 011009. https://doi.org/10.1117/1.1854672

Asutay, E., & Västfjäll, D. (2017). Exposure to arousal-inducing sounds facilitates visual search. *Scientific Reports*, *7*(1), 10363. https://doi.org/10.1038/s41598-017-09975-8

Auracher, J., Albers, S., Zhai, Y., Gareeva, G., & Stavniychuk, T. (2011). P is for happiness, N is for sadness: Universals in sound iconicity to detect emotions in poetry. *Discourse Processes*, *48*(1), 1–25. https://doi.org/10.1080/01638531003674894

Auracher, J., Menninghaus, W., & Scharinger, M. (2020). Sound predicts meaning: Cross-modal associations between formant frequency and emotional tone in stanzas. *Cognitive Science*, *44*(10), e12906. https://doi.org/10.1111/cogs.12906

Azhari, A., Truzzi, A., Neoh, M. J.-Y., Balagtas, J. P. M., Tan, H. H., Goh, P. P., Ang, X. A., Setoh, P., Rigo, P., Bornstein, M. H., & Esposito, G. (2020). A decade of infant neuroimaging research: What have we learned and where are we going? *Infant Behavior and Development*, *58*, 101389. https://doi.org/10.1016/j.infbeh.2019.101389

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591. https://doi.org/10.1006/jmla.2000.2756

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46*(3–4), 252–267. https://doi.org/10.1016/j.specom.2005.02.016

Baroni, A. (2014). On the importance of being noticed: The role of acoustic salience in phonotactics (and casual speech). *Language Sciences*, *46*, 18–36. https://doi.org/10.1016/j.langsci.2014.06.004

Barrett, L. F., & Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science*, *8*(1), 10–14. https://doi.org/10.1111/1467-8721.00003

Barton, D. N., & Halberstadt, J. (2018). A social Bouba/Kiki effect: A bias for people whose names match their faces. *Psychonomic Bulletin & Review*, *25*(3), 1013–1020. https://doi.org/10.3758/s13423-017-1304-x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv:1406.5823 [Stat]*. http://arxiv.org/abs/1406.5823

Belin, P., Zilbovicius, M., Crozier, S., Thivard, L., Fontaine, A. A., Masure, M.-C., & Samson, Y. (1998). Lateralization of speech and auditory temporal processing. *Journal of Cognitive Neuroscience*, *10*(4), 536–540. https://doi.org/10.1162/089892998562834

Belyk, M., & Brown, S. (2014). The acoustic correlates of valence depend on emotion family. *Journal of Voice*, *28*(4), 523.e9-523.e18. https://doi.org/10.1016/j.jvoice.2013.12.007

Berger, J. (2011). Arousal increases social transmission of information. *Psychological Science*, *22*(7), 891–893. https://doi.org/10.1177/0956797611413294

Berger, V. W., & Zhou, Y. (2014). Kolmogorov–Smirnov test: Overview. In *Wiley StatsRef: Statistics Reference Online*.

John Wiley & Sons. https://doi.org/10.1002/9781118445112.stat06558

Berlin, B. (2006). The first congress of ethnozoological nomenclature. *Journal of the Royal Anthropological Institute*, *12*(s1), S23–S44. https://doi.org/10.1111/j.1467-9655.2006.00271.x

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*(5), 512–528. https://doi.org/10.1093/cercor/10.5.512

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818–10823. https://doi.org/10.1073/pnas.1605782113

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.4.12) [Computer software]. http://www.praat.org/

Bosseler, A. N., Taulu, S., Pihko, E., Mäkelä, J. P., Imada, T., Ahonen, A., & Kuhl, P. K. (2013). Theta brain rhythms index perceptual narrowing in infant speech perception. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00690

Bourke, J. D., & Todd, J. (2021). Acoustics *versus* linguistics? Context is Part and Parcel to lateralized processing of the parts and parcels of speech. *Laterality*, *26*(6), 725–765. https://doi.org/10.1080/1357650X.2021.1898415

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different

shape–taste matches to Westerners. *Cognition*, *126*(2), 165–172. https://doi.org/10.1016/j.cognition.2012.09.007

Broca, M. P. (1861). Remarques sur le siége de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin de la Société Anatomique*, *6*, 330–357.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*(6), e10729. https://doi.org/10.1371/journal.pone.0010729

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & Rosario, H. D. (2020). *pwr: Basic functions for power analysis* (Version 1.3.0) [Computer software]. https://cran.r-project.org/web/packages/pwr/index.html

Chan, J. Y., Hssayeni, M. D., Wilcox, T., & Ghoraani, B. (2023). Exploring the feasibility of tensor decomposition for analysis of fNIRS signals: A comparative study with grand averaging method. *Frontiers in Neuroscience*, *17*, 1180293. https://doi.org/10.3389/fnins.2023.1180293

Chan, M. K. M. (1996). Some thoughts on the typology of sound symbolism and the Chinese language. In C. Cheng, J. Packard, J. Yoon, & Y. You (Eds.), *Proceedings of the Eighth North American Conference on Chinese Linguistics* (Vol. 2, pp. 1–15). Los Angeles: GSIL Publications.

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), 1–15. https://doi.org/10.1163/22134808-bja10068

Chen, A., & Kager, R. (2016). Discrimination of lexical tones in the first year of life. *Infant and Child Development*, *25*(5), 426–439. https://doi.org/10.1002/icd.1944

Chen, A., Stevens, C. J., & Kager, R. (2017). Pitch perception in the first year of life, a comparison of lexical tones and musical pitch. *Frontiers in Psychology*, *8*, Article 297. https://doi.org/10.3389/fpsyg.2017.00297

Chen, Y. (2022). Mind the subtle f0 modifications: The interaction of tone and intonation in Sinitic varieties. *Stellenbosch Papers in Linguistics Plus*, *62*(02), 113–136. https://doi.org/10.5842/62-2-904

Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, *36*(4), 724–746. https://doi.org/10.1016/j.wocn.2008.06.003

Chen, Y.-C., Huang, P.-C., Woods, A., & Spence, C. (2016). When "bouba" equals "kiki": Cultural commonalities and cultural differences in sound-shape correspondences. *Scientific Reports*, *6*, 26681. https://doi.org/10.1038/srep26681

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception*, *41*(3), 339–353. https://doi.org/10.1068/p7161

Citron, F. M. M. (2012). Neural correlates of written emotion word processing: A review of recent electrophysiological and hemodynamic neuroimaging studies. *Brain and Language*, *122*(3), 211–226. https://doi.org/10.1016/j.bandl.2011.12.007

Citron, F. M. M., Gray, M. A., Critchley, H. D., Weekes, B. S., & Ferstl, E. C. (2014). Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, *56*(100), 79–89. https://doi.org/10.1016/j.neuropsychologia.2014.01.002

Coon, J., Silva, P. N., Etz, A., & Sarnecka, B. W. (2024). Bayesian tools of the trade for developmental psychologists: A quick-start guide using JASP. *Journal of Cognition and Development*, *26*(1), 1–49. https://doi.org/10.1080/15248372.2024.2386032

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Crisinel, A.-S., Cosser, S., King, S., Jones, R., Petrie, J., & Spence, C. (2012). A bittersweet symphony: Systematically modulating the taste of food by changing the sonic properties of the soundtrack playing in the background. *Food Quality and Preference*, *24*(1), 201–204. https://doi.org/10.1016/j.foodqual.2011.08.009

Crisinel, A.-S., & Spence, C. (2009). Implicit association between basic tastes and pitch. *Neuroscience Letters*, *464*(1), 39–42. https://doi.org/10.1016/j.neulet.2009.08.016

Crisinel, A.-S., & Spence, C. (2010a). A sweet sound? Food names reveal implicit associations between taste and pitch. *Perception*, *39*(3), 417–425. https://doi.org/10.1068/p6574

Crisinel, A.-S., & Spence, C. (2010b). As bitter as a trombone: Synesthetic correspondences in nonsynesthetes between tastes/flavors and musical notes. *Attention, Perception &*

*Psychophysics*, *72*(7), 1994–2002. https://doi.org/10.3758/APP.72.7.1994

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition*, *170*, 312–327. https://doi.org/10.1016/j.cognition.2017.09.016

Cruttenden, A. (1997). *Intonation* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139166973

Cuskley, C., Simner, J., & Kirby, S. (2017). Phonological and orthographic influences in the bouba–kiki effect. *Psychological Research*, *81*(1), 119–130. https://doi.org/10.1007/s00426-015-0709-2

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201. https://doi.org/10.1177/002383099704000203

Cutler, A., & van Donselaar, W. (2001). *Voornaam* is not (really) a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech*, *44*(2), 171–195. https://doi.org/10.1177/00238309010440020301

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., … Winter, B. (2021). The *bouba/kiki* effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1841), 1–13. https://doi.org/10.1098/rstb.2020.0390

Darwin, C. (1998). *The expression of the emotions in man and animals* (3rd ed, With an introduction, afterword, and commentaries by Paul Ekman). Oxford University Press.

Davidson, R. J., Sherer, K. R., & Goldsmith, H. H. (Eds.). (2009). *Handbook of affective sciences*. Oxford University Press.

De Carolis, L., Marsico, E., Arnaud, V., & Coupé, C. (2018). Assessing sound symbolism: Investigating phonetic forms, visual shapes and letter fonts in an implicit bouba-kiki experimental paradigm. *PLoS ONE*, *13*(12), e0208874. https://doi.org/10.1371/journal.pone.0208874

De Klerk, M., De Bree, E., Kerkhoff, A., & Wijnen, F. (2019). Lost and found: Decline and remergence of non-native vowel discrimination in the first year of life. *Language Learning and Development*, *15*(1), 14–31. https://doi.org/10.1080/15475441.2018.1497490

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, *298*(5600), 2013–2015. https://doi.org/10.1126/science.1077066

Dehaene-Lambertz, G., Hertz-Pannier, L., & Dubois, J. (2006). Nature and nurture in language acquisition: Anatomical and functional brain-imaging studies in infants. *Trends in Neurosciences*, *29*(7), 367–373. https://doi.org/10.1016/j.tins.2006.05.011

Deroy, O., & Spence, C. (2013). Why we are not all synesthetes (not even weakly so). *Psychonomic Bulletin & Review*, *20*(4), 643–664. https://doi.org/10.3758/s13423-013-0387-2

Di Lorenzo, R., Pirazzoli, L., Blasi, A., Bulgarelli, C., Hakuno, Y., Minagawa, Y., & Brigadoi, S. (2019). Recommendations for

motion correction of infant fNIRS data applicable to multiple data sets and acquisition systems. *NeuroImage*, *200*, 511–527. https://doi.org/10.1016/j.neuroimage.2019.06.056

Dingemanse, M. (2015). Ideophones and reduplication: Depiction, description, and the interpretation of repeated talk in discourse. *Studies in Language*, *39*(4), 946–970. https://doi.org/10.1075/sl.39.4.05din

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615. https://doi.org/10.1016/j.tics.2015.07.013

Dingemanse, M., Schuerman, W., Reinisch, E., Tufvesson, S., & Mitterer, H. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language*, *92*(2), e117–e133. https://doi.org/10.1353/lan.2016.0034

D'Onofrio, A. (2014). Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm. *Language and Speech*, *57*(3), 367–393. https://doi.org/10.1177/0023830913507694

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekström, A. G., Nirme, J., & Gärdenfors, P. (2022). Motion iconicity in prosody. *Frontiers in Communication*, *7*, 994162. https://doi.org/10.3389/fcomm.2022.994162

Evans, K. K., & Treisman, A. (2011). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), Article 6. https://doi.org/10.1167/10.1.6

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fernández-Prieto, I., Navarra, J., & Pons, F. (2015). How big is this sound? Crossmodal association between pitch and size in infants. *Infant Behavior and Development*, *38*, 77–81. https://doi.org/10.1016/j.infbeh.2014.12.008

Filippi, P., Ocklenburg, S., Bowling, D. L., Heege, L., Güntürkün, O., Newen, A., & de Boer, B. (2017). More than words (and faces): Evidence for a Stroop effect of prosody in emotion word processing. *Cognition and Emotion*, *31*(5), 879–891. https://doi.org/10.1080/02699931.2016.1177489

Fischler, I., & Bradley, M. (2006). Event-related potential studies of language and emotion: Words, phrases, and task effects. *Progress in Brain Research*, *156*, 185–203. https://doi.org/10.1016/S0079-6123(06)56009-1

Fort, M., Martin, A., & Peperkamp, S. (2015). Consonants are more important than vowels in the bouba-kiki effect. *Language and Speech*, *58*(2), 247–266. https://doi.org/10.1177/0023830914534951

Fort, M., & Schwartz, J.-L. (2022). Resolving the bouba-kiki effect enigma by rooting iconic sound symbolism in physical properties of round and spiky objects. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-23623-w

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, *97*(3), 412–429. https://doi.org/10.1037/0033-2909.97.3.412

Friendly, M., & Meyer, D. (2015). *Discrete data analysis with R: Visualization and modeling techniques for categorical and count data*. CRC Press.

Frota, S., Butler, J., Lu, S., & Vigário, M. (2016). Infants' perception of native and non-native pitch contrasts. *Speech Prosody 2016*, 692–696. https://doi.org/10.21437/SpeechProsody.2016-142

Frota, S., Butler, J., & Vigário, M. (2014). Infants' perception of intonation: Is it a statement or a question? *Infancy*, *19*(2), 194–213. https://doi.org/10.1111/infa.12037

Gage, N. M., & Baars, B. J. (2018). Language and thought. In *Fundamentals of cognitive neuroscience: A beginner's guide (2nd ed.)* (pp. 185–214). Academic Press. https://doi.org/10.1016/B978-0-12-803813-0.00006-4

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*(7), 1191–1203. https://doi.org/10.3758/BF03193720

Gandour, J., Dzemidzic, M., Wong, D., Lowe, M., Tong, Y., Hsieh, L., Satthamnuwong, N., & Lurito, J. (2003). Temporal integration of speech prosody is shaped by language experience: An fMRI study. *Brain and Language*, *84*(3), 318–336. https://doi.org/10.1016/S0093-934X(02)00505-9

Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., Li, X., & Lowe, M. (2004). Hemispheric roles in the perception of speech prosody. *NeuroImage*, *23*(1), 344–357. https://doi.org/10.1016/j.neuroimage.2004.06.004

Gemignani, J., & Gervain, J. (2021). Comparing different pre-processing routines for infant fNIRS data. *Developmental*

*Cognitive Neuroscience*, *48*, 100943. https://doi.org/10.1016/j.dcn.2021.100943

Gervain, J. (2015). Plasticity in early language acquisition: The effects of prenatal and early childhood experience. *Current Opinion in Neurobiology*, *35*, 13–20. https://doi.org/10.1016/j.conb.2015.05.004

Gervain, J. (2018). The role of prenatal experience in language development. *Current Opinion in Behavioral Sciences*, *21*, 62–67. https://doi.org/10.1016/j.cobeha.2018.02.004

Gervain, J. (2020). Typical language development. In A. Gallagher, C. Bulteau, D. Cohen, & J. L. Michaud (Eds.), *Handbook of Clinical Neurology* (Vol. 173, pp. 171–183). Elsevier. https://doi.org/10.1016/B978-0-444-64150-2.00016-2

Gervain, J., Minagawa, Y., Emberson, L., & Lloyd-Fox, S. (2023). Using functional near-infrared spectroscopy to study the early developing brain: Future directions and new challenges. *Neurophotonics*, *10*(2), 023519. https://doi.org/10.1117/1.NPh.10.2.023519

Getz, L. M., & Kubovy, M. (2018). Questioning the automaticity of audiovisual correspondences. *Cognition*, *175*, 101–108. https://doi.org/10.1016/j.cognition.2018.02.015

Gianotti, L. R. R., Faber, P. L., Schuler, M., Pascual-Marqui, R. D., Kochi, K., & Lehmann, D. (2008). First valence, then arousal: The temporal dynamics of brain electric activity evoked by emotional stimuli. *Brain Topography*, *20*(3), 143–156. https://doi.org/10.1007/s10548-007-0041-2

Godoy, M. C., & Ananias, T. (2022). The bouba-kiki effect in a production task. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual*

*Conference of the Cognitive Science Society* (Vol. 44, pp. 763–770).

González-Alvarez, J., & Sos-Peña, R. (2022). Perceiving body height from connected speech: Higher fundamental frequency is associated with the speaker's height. *Perceptual and Motor Skills*, *129*(5), 1349–1361. https://doi.org/10.1177/00315125221110392

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Gu, F., Zhang, C., Hu, A., & Zhao, G. (2013). Left hemisphere lateralization for lexical and acoustic pitch processing in Cantonese speakers as revealed by mismatch negativity. *NeuroImage*, *83*, 637–645. https://doi.org/10.1016/j.neuroimage.2013.02.080

Gussenhoven, C. (1992). Dutch. *Journal of the International Phonetic Association*, *22*(1–2), 45–47. https://doi.org/10.1017/S002510030000459X

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616983

Gussenhoven, C. (2016). Foundations of intonational meaning: Anatomical and physiological factors. *Topics in Cognitive Science*, *8*(2), 425–434. https://doi.org/10.1111/tops.12197

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: Controversies and consensus. *Trends in Cognitive Sciences*, *16*(9), 458–466. https://doi.org/10.1016/j.tics.2012.07.006

Hamilton-Fletcher, G., Pisanski, K., Reby, D., Stefańczyk, M., Ward, J., & Sorokowska, A. (2018). The role of visual experience in the emergence of cross-modal correspondences. *Cognition*, *175*, 114–121. https://doi.org/10.1016/j.cognition.2018.02.023

Harmon-Jones, E., Harmon-Jones, C., & Summerell, E. (2017). On the importance of both dimensional and discrete models of emotion. *Behavioral Sciences*, *7*(4), Article 4. https://doi.org/10.3390/bs7040066

Haudricourt, A.-G. (1954). De l'origine des tons en vietnamien. *Journal Asiatique*, *242*, 69–82.

Herbert, C., Junghofer, M., & Kissler, J. (2008). Event related potentials to emotional adjectives during reading. *Psychophysiology*, *45*(3), 487–498. https://doi.org/10.1111/j.1469-8986.2007.00638.x

Hinojosa, J. A., Haro, J., Magallares, S., Duñabeitia, J. A., & Ferré, P. (2020). Iconicity ratings for 10,995 Spanish words and their relationship with psycholinguistic variables. *Behavior Research Methods*, *53*(3), 1262–1275. https://doi.org/10.3758/s13428-020-01496-z

Hinton, L., Nichols, J., & Ohala, J. J. (Eds.). (2006). *Sound symbolism*. Cambridge University Press.

Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. (2010). Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and electrophysiological measures. *Brain Research*, *1360*, 89–105. https://doi.org/10.1016/j.brainres.2010.08.092

Hockett, C. F. (1958). *A course in modern linguistics*. Macmillan.

Homae, F., Watanabe, H., Nakano, T., Asakawa, K., & Taga, G. (2006). The right hemisphere of sleeping infant perceives

sentential prosody. *Neuroscience Research*, *54*(4), 276–280. https://doi.org/10.1016/j.neures.2005.12.006

Homae, F., Watanabe, H., Nakano, T., & Taga, G. (2011). Large-scale brain networks underlying language acquisition in early infancy. *Frontiers in Psychology*, *2*, Article 93. https://doi.org/10.3389/fpsyg.2011.00093

Huang, Y.-H., Pratoomraj, S., & Johnson, R. C. (1969). Universal magnitude symbolism. *Journal of Verbal Learning and Verbal Behavior*, *8*(1), 155–156. https://doi.org/10.1016/S0022-5371(69)80028-9

Hugdahl, K. (2000). Lateralization of cognitive processes in the brain. *Acta Psychologica*, *105*(2), 211–235. https://doi.org/10.1016/S0001-6918(00)00062-7

Huppert, T. J., Diamond, S. G., Franceschini, M. A., & Boas, D. A. (2009). HomER: A review of time-series analysis methods for near-infrared spectroscopy of the brain. *Applied Optics*, *48*(10), D280–D298. https://doi.org/10.1364/AO.48.00D280

Hyman, L. M. (2014). What is phonological typology? *UC Berkeley PhonLab Annual Report*, *10*(10), 101–118.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65. https://doi.org/10.1016/j.cognition.2008.07.015

Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015). Sound symbolism facilitates word learning in 14-month-olds. *PLoS ONE*, *10*(2), e0116494. https://doi.org/10.1371/journal.pone.0116494

Issard, C., & Gervain, J. (2018). Variability of the hemodynamic response in infants: Influence of experimental design and

stimulus complexity. *Developmental Cognitive Neuroscience*, *33*, 182–193. https://doi.org/10.1016/j.dcn.2018.01.009

Kager, R. (2018). How do infants disaggregate referential and affective pitch? *Frontiers in Psychology*, *9*, 1–3. https://doi.org/10.3389/fpsyg.2018.02093

Kambara, T., & Umemura, T. (2021). The relationships between initial consonants in Japanese sound symbolic words and familiarity, multi-sensory imageability, emotional valence, and arousal. *Journal of Psycholinguistic Research*, *50*(4), 831–842. https://doi.org/10.1007/s10936-020-09749-w

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, *27*(2), 237–265. https://doi.org/10.3758/s13423-019-01701-x

Kashima, Y., Coman, A., Pauketat, J. V. T., & Yzerbyt, V. (2020). Emotion in cultural dynamics. *Emotion Review*, *12*(2), 48–64. https://doi.org/10.1177/1754073919875215

Kissler, J., Herbert, C., Peyk, P., & Junghofer, M. (2007). Buzzwords: Early cortical responses to emotional words during reading. *Psychological Science*, *18*(6), 475–480. https://doi.org/10.1111/j.1467-9280.2007.01924.x

Kissler, J., Herbert, C., Winkler, I., & Junghofer, M. (2009). Emotion and attention in visual word processing—An ERP study. *Biological Psychology*, *80*(1), 75–83. https://doi.org/10.1016/j.biopsycho.2008.03.004

Knoeferle, K., Li, J., Maggioni, E., & Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-05965-y

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

Köhler, W. (1947). *Gestalt psychology (2nd ed.)*. New York: Liveright.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koriat, A., & Levy, I. (1979). Figural symbolism in Chinese ideographs. *Journal of Psycholinguistic Research*, *8*(4), 353–365. https://doi.org/10.1007/BF01067139

Körner, A., & Rummer, R. (2022). Articulation contributes to valence sound symbolism. *Journal of Experimental Psychology: General*, *151*(5), 1107–1114. https://doi.org/10.1037/xge0001124

Körner, A., & Rummer, R. (2023). Valence sound symbolism across language families: A comparison between Japanese and German. *Language and Cognition*, *15*(2), 337–354. https://doi.org/10.1017/langcog.2022.39

Kotz, S. A., Meyer, M., & Paulmann, S. (2006). Lateralization of emotional prosody in the brain: An overview and synopsis on the impact of study design. *Progress in Brain Research*, *156*, 285–294. https://doi.org/10.1016/S0079-6123(06)56015-7

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473–481. https://doi.org/10.1016/j.cognition.2009.06.007

Kubozono, H. (2012). Varieties of pitch accent systems in Japanese. *Lingua*, *122*(13), 1395–1414. https://doi.org/10.1016/j.lingua.2012.08.001

Kucharský, Š., Zaharieva, M., Raijmakers, M., & Visser, I. (2022). Habituation, part II. Rethinking the habituation paradigm. *Infant and Child Development*, *33*(1), e2383. https://doi.org/10.1002/icd.2383

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, *139*(4), 917–940. https://doi.org/10.1037/a0030811

Lapolla, R. J. (1995). An experimental investigation into phonetic symbolism as it relates to Mandarin Chinese. In J. Nichols, J. J. Ohala, & L. Hinton (Eds.), *Sound symbolism* (pp. 130–147). Cambridge University Press. https://doi.org/10.1017/CBO9780511751806.010

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, *19*(5), 633–653. https://doi.org/10.1080/02699930441000445

Lenth, R. V. (2023). emmeans: Estimated marginal means, aka least-squares means. *R Package Version 1.8.7*. https://CRAN.R-project.org/package=emmeans

Leroy, F., Glasel, H., Dubois, J., Hertz-Pannier, L., Thirion, B., Mangin, J.-F., & Dehaene-Lambertz, G. (2011). Early

maturation of the linguistic dorsal pathway in human infants. *The Journal of Neuroscience*, *31*(4), 1500–1506. https://doi.org/10.1523/JNEUROSCI.4141-10.2011

Lester, D. (1974). Symbolism in the Chinese language. *International Journal of Symbology*, *5*(1), 18–21.

Ley-Flores, J., Alshami, E., Singh, A., Bevilacqua, F., Bianchi-Berthouze, N., Deroy, O., & Tajadura-Jiménez, A. (2022). Effects of pitch and musical sounds on body-representations when moving with sound. *Scientific Reports*, *12*, Article 2676. https://doi.org/10.1038/s41598-022-06210-x

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

Li, F. (2017). Analyses on arbitrariness of Chinese characters from the perspective of morphology. *ProtoSociology*, *34*, 181–196. https://doi.org/10.5840/protosociology20173411

Li, V. P. H. (1986). Philology and power: Ezra pound and the regulation of language. *Boundary 2*, *15*(1/2), 187–210. https://doi.org/10.2307/303430

Lialiou, M., Grice, M., Röhr, C. T., & Schumacher, P. B. (2024). Auditory processing of intonational rises and falls in German: Rises are special in attention orienting. *Journal of Cognitive Neuroscience*, *36*(6), 1099–1122. https://doi.org/10.1162/jocn_a_02129

Lin, Y., Ding, H., & Zhang, Y. (2020). Prosody dominates over semantics in emotion word processing: Evidence from cross-channel and cross-modal Stroop effects. *Journal of Speech, Language, and Hearing Research*, *63*(3), 896–912. https://doi.org/10.1044/2020_JSLHR-19-00258

Lin, Y., Ding, H., & Zhang, Y. (2021). Unisensory and multisensory Stroop effects modulate gender differences in verbal and nonverbal emotion perception. *Journal of Speech, Language, and Hearing Research*, *64*(11), 4439–4457. https://doi.org/10.1044/2021_JSLHR-20-00338

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *The Behavioral and Brain Sciences*, *35*(3), 121–143. https://doi.org/10.1017/S0140525X11000446

Liu, L., Götz, A., Lorette, P., & Tyler, M. D. (2022). How tone, intonation and emotion shape the development of infants' fundamental frequency perception. *Frontiers in Psychology*, *13*, Article 906848. https://doi.org/10.3389/fpsyg.2022.906848

Liu, L., & Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition*, *133*(2), 385–394. https://doi.org/10.1016/j.cognition.2014.06.004

Liu, L., & Kager, R. (2016). Perception of a native vowel contrast by Dutch monolingual and bilingual infants: A bilingual perceptual lead. *International Journal of Bilingualism*, *20*(3), 335–345. https://doi.org/10.1177/1367006914566082

Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods*, *44*, 1042–1051. https://doi.org/10.3758/s13428-012-0203-3

Lloyd-Fox, S., Blasi, A., & Elwell, C. E. (2010). Illuminating the developing brain: The past, present and future of functional near infrared spectroscopy. *Neuroscience & Biobehavioral*

*Reviews*, *34*(3), 269–284. https://doi.org/10.1016/j.neubiorev.2009.07.008

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, *6*, Article 1246. https://doi.org/10.3389/fpsyg.2015.01246

Lockwood, G., Hagoort, P., & Dingemanse, M. (2016). How iconicity helps people learn new words: Neural correlates and individual differences in sound-symbolic bootstrapping. *Collabra*, *2*(1), 7. https://doi.org/10.1525/collabra.42

Louwerse, M., & Qu, Z. (2017). Estimating valence from the sound of a word: Computational, experimental, and cross-linguistic evidence. *Psychonomic Bulletin & Review*, *24*(3), 849–855. https://doi.org/10.3758/s13423-016-1142-2

Lovčević, I., & Tsuji, S. (2024). The developmental pattern of native and non-native speech perception during the 1st year of life in Japanese infants. *Infant Behavior and Development*, *76*, 101977. https://doi.org/10.1016/j.infbeh.2024.101977

Lowe, M. L., & Haws, K. L. (2017). Sounds big: The effects of acoustic pitch on product perceptions. *Journal of Marketing Research*, *54*(2), 331–346. https://doi.org/10.1509/jmr.14.0300

Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (*Pan troglodytes*) and humans. *Proceedings of the National Academy of Sciences*, *108*(51), 20661–20665. https://doi.org/10.1073/pnas.1112605108

Luk, G., & Bialystok, E. (2005). How iconic are Chinese characters? *Bilingualism: Language and Cognition*, *8*(1), 79–83. https://doi.org/10.1017/S1366728904002081

Maddieson, I. (2023). Tone is not predominant; Tone is not primordial. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1901–1905). International Phonetic Association.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, *87*(1/2), 173–188. https://doi.org/10.2307/1422011

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(3), 384–394. https://doi.org/10.1037/0096-1523.13.3.384

Marks, L. E., Hammeal, R. J., Bornstein, M. H., & Smith, L. B. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, *52*(1). https://doi.org/10.2307/1166084

Massaro, D. W. (2015). Speech perception. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 235–242). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.52021-2

Matsui, M. (2020). Effects of pitch of short pure tones on onomatopoeic expressions: A cross-linguistic study of Japanese and Chinese. *Acoustical Science and Technology*, *41*(2), 472–480. https://doi.org/10.1250/ast.41.472

Mattock, K., & Burnham, D. (2006). Chinese and English Infants' tone perception: Evidence for perceptual reorganization. *Infancy*, *10*(3), 241–265. https://doi.org/10.1207/s15327078in1003_3

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, *106*(3), 1367–1381. https://doi.org/10.1016/j.cognition.2007.07.002

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. https://doi.org/10.1111/j.1467-7687.2006.00495.x

May, L., Byers-Heinlein, K., Gervain, J., & Werker, J. F. (2011). Language and the newborn brain: Does prenatal language experience shape the neonate neural response to speech? *Frontiers in Psychology*, *2*, Article 222. https://doi.org/10.3389/fpsyg.2011.00222

Mazuka, R., Hasegawa, M., & Tsuji, S. (2014). Development of non-native vowel discrimination: Improvement without exposure. *Developmental Psychobiology*, *56*(2), 192–209. https://doi.org/10.1002/dev.21193

McCormick, K., Lacey, S., Stilla, R., Nygaard, L. C., & Sathian, K. (2021). Neural basis of the sound-symbolic crossmodal correspondence between auditory pseudowords and visual shapes. *Multisensory Research*, *35*(1), 29–78. https://doi.org/10.1163/22134808-bja10060

McLean, B., Dunn, M., & Dingemanse, M. (2023). Two measures are better than one: Combining iconicity ratings and guessing experiments for a more nuanced picture of iconicity in the lexicon. *Language and Cognition*, *15*(4), 716–739. https://doi.org/10.1017/langcog.2023.9

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human*

*Perception and Performance*, *15*(1), 69–79. https://doi.org/10.1037/0096-1523.15.1.69

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, *116*(4), 323–336.

Michaud, A., & Sands, B. (2020). *Tonogenesis*. HAL open archive. https://halshs.archives-ouvertes.fr/halshs-02519305

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Minagawa-Kawai, Y., Cristià, A., & Dupoux, E. (2011). Cerebral lateralization and early speech acquisition: A developmental scenario. *Developmental Cognitive Neuroscience*, *1*(3), 217–232. https://doi.org/10.1016/j.dcn.2011.03.005

Minagawa-Kawai, Y., Mori, K., Naoi, N., & Kojima, S. (2007). Neural attunement processes in infants during the acquisition of a language-specific phonemic contrast. *The Journal of Neuroscience*, *27*(2), 315–321. https://doi.org/10.1523/JNEUROSCI.1984-06.2007

Mitterer, H., & Cutler, A. (2006). Speech perception. In K. Brown (Ed.), *Encyclopedia of Language & Linguistics (2nd ed.)* (pp. 770–782). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00029-8

Mok, P. P. K., Li, G., Li, J. J., Ng, H. T. Y., & Cheung, H. (2019). Cross-modal association between vowels and colours: A cross-linguistic perspective. *The Journal of the Acoustical Society of America*, *145*(4), 2265–2276. https://doi.org/10.1121/1.5096632

Molfese, D. L., Freeman, R. B., & Palermo, D. S. (1975). The ontogeny of brain lateralization for speech and nonspeech stimuli. *Brain and Language*, *2*, 356–368. https://doi.org/10.1016/S0093-934X(75)80076-9

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1152–1164. https://doi.org/10.1037/a0027747

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 1–12. https://doi.org/10.1098/rstb.2013.0299

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 133–136. https://doi.org/10.3758/CABN.4.2.133

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855–869. https://doi.org/10.1086/283219

Motamedi, Y., Little, H., Nielsen, A., & Sulik, J. (2019). The iconicity toolbox: Empirical approaches to measuring iconicity. *Language and Cognition*, *11*(2), 188–207. https://doi.org/10.1017/langcog.2019.14

Nazzi, T., Floccia, C., & Bertoncini, J. (1998). Discrimination of pitch contours by neonates. *Infant Behavior and Development*, *21*(4), 779–784. https://doi.org/10.1016/S0163-6383(98)90044-3

Needle, J. M., Pierrehumbert, J. B., & Hay, J. B. (2022). Phonotactic and morphological effects in the acceptability of pseudowords. In A. D. Sims, A. Ussishkin, J. Parker, & S. Wray (Eds.), *Morphological Diversity and Linguistic Cognition* (1st ed., pp. 79–112). Cambridge University Press. https://doi.org/10.1017/9781108807951.005

Nielsen, A., & Dingemanse, M. (2021). Iconicity in word learning and beyond: A critical review. *Language and Speech*, *64*(1), 52–72. https://doi.org/10.1177/0023830920914339

Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *65*(2), 115–124. https://doi.org/10.1037/a0022268

Nielsen, A., & Rendall, D. (2018). Comparative perspectives on communication in human and non-Human primates: Grounding meaning in broadly conserved processes of voice production, perception, affect, and cognition. In S. Frühholz & P. Belin (Eds.), *The Oxford handbook of voice perception* (pp. 278–308). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.13

Nikroshkina, S. V. (2019). Linguistic and psychological aspects of universal sound symbolism studies. *Proceedings of the Internation Conference on "Humanities and Social Sciences: Novations, Problems, Prospects" (HSSNPP 2019)*, *333*, 262–266. https://doi.org/10.2991/hssnpp-19.2019.49

Nishiyori, R. (2016). fNIRS: An emergent method to document functional cortical activity during infant movements.

*Frontiers in Psychology*, *7*, Article 533. https://doi.org/10.3389/fpsyg.2016.00533

Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, *28*, 225–252. http://www.jstor.org/stable/223394

Oakes, L. M. (2010). Using habituation of looking time to assess mental processes in infancy. *Journal of Cognition and Development*, *11*(3), 255–268. https://doi.org/10.1080/15248371003699977

Oakes, L. M., Sperka, D., DeBolt, M. C., & Cantrell, L. M. (2019). Habit2: A stand-alone software solution for presenting stimuli and recording infant looking times in order to study infant development. *Behavior Research Methods*, *51*(5). https://doi.org/10.3758/s13428-019-01244-y

Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica*, *40*(1), 1–18. https://doi.org/10.1159/000261678

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of $f_0$ of voice. *Phonetica*, *41*, 1–16. https://doi.org/10.1159/000261706. PMID: 6204347.

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (1st ed.) (pp. 325–347). Cambridge University Press. https://doi.org/10.1017/CBO9780511751806.022

Ohala, J. J. (1997). Sound symbolism. *Proceedings of the 4th Seoul International Conference on Linguistics (SICOL)*, 98–103.

Okada, H. (1991). Japanese. *Journal of the International Phonetic Association*, *21*(2), 94–96. https://doi.org/10.1017/S002510030000445X

Ota, M., Yamane, N., & Mazuka, R. (2018). The effects of lexical pitch accent on infant word recognition in Japanese. *Frontiers in Psychology*, *8*, Article 2354. https://doi.org/10.3389/fpsyg.2017.02354

Pakosz, M. (1983). Attitudinal judgments in intonation: Some evidence for a theory. *Journal of Psycholinguistic Research*, *12*, 311–326. https://doi.org/10.1007/BF01067673

Pan, Y., & Schmitt, B. (1996). Language and brand attitudes: Impact of script and sound matching in Chinese and English. *Journal of Consumer Psychology*, *5*(3), 263–277. https://doi.org/10.1207/s15327663jcp0503_03

Parise, C., & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters*, *442*(3), 257–261. https://doi.org/10.1016/j.neulet.2008.07.010

Parise, C., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research*, *220*(3), 319–333. https://doi.org/10.1007/s00221-012-3140-6

Paus, T. (2000). Functional anatomy of arousal and attention systems in the human brain. *Progress in Brain Research*, *126*, 65–77. https://doi.org/10.1016/S0079-6123(00)26007-X

Peiffer-Smadja, N., & Cohen, L. (2019). The cerebral bases of the bouba-kiki effect. *NeuroImage*, *186*, 679–689. https://doi.org/10.1016/j.neuroimage.2018.11.033

Peng, D., & Wang, C. (1997). Hanzi jiagong de jiben danyuan: Laizi bihuashu xiaoying he bujianshu xiaoying de zhengju [Basic processing unit of Chinese character recognition: Evidence

from stroke number effect and radical number effect]. *Acta Psychologica Sinica*, *29*(1), 9–17.

Perlman, M. (2017). Debunking two myths against vocal origins of language. *Interaction Studies*, *18*(3), 376–401. https://doi.org/10.1075/is.18.3.05per

Perlman, M. (2024). Iconic prosody is deeply connected to iconic gesture, and it may occur just as frequently. In O. Fischer, K. Akita, & P. Perniss (Eds.), *Oxford handbook of iconicity in language*. Oxford University Press.

Perlman, M., & Cain, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture*, *14*(3), 320–350. https://doi.org/10.1075/gest.14.3.03per

Perlman, M., Clark, N., & Johansson Falck, M. (2015). Iconic prosody in story reading. *Cognitive Science*, *39*(6), 1348–1368. https://doi.org/10.1111/cogs.12190

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science*, *2*(8), 150152. https://doi.org/10.1098/rsos.150152

Perniss, P., Thompson, R., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, *1*, Article 227. https://doi.org/10.3389/fpsyg.2010.00227

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS ONE*, *10*(9), 1–17. https://doi.org/10.1371/journal.pone.0137147

Pike, K. L. (1948). *Tone languages: A technique for determining the number and type of pitch contrasts in a language, with*

*studies in tonemic substitution and fusion*. University of Michigan Press.

Pinti, P., Cardone, D., & Merla, A. (2015). Simultaneous fNIRS and thermal infrared imaging during cognitive task reveal autonomic correlates of prefrontal cortex activity. *Scientific Reports*, *5*, 17471. https://doi.org/10.1038/srep17471

Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., & Tachtsidis, I. (2019). Current status and issues regarding pre-processing of fNIRS neuroimaging data: An investigation of diverse signal filtering methods within a general linear model framework. *Frontiers in Human Neuroscience*, *12*, Article 505. https://doi.org/10.3389/fnhum.2018.00505

Pittayaporn, P., & Kirby, J. (2017). Laryngeal contrasts in the Tai dialect of Cao Bằng. *Journal of the International Phonetic Association*, *47*(1), 65–85. https://doi.org/10.1017/S0025100316000293

Poeppel, D. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive Science*, *25*(5), 679–693. https://doi.org/10.1207/s15516709cog2505_3

Pollet, T. V., Stulp, G., Henzi, S. P., & Barrett, L. (2015). Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. *American Journal of Primatology*, *77*(7), 727–740. https://doi.org/10.1002/ajp.22405

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology.

*Development and Psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

Quaresima, V., Bisconti, S., & Ferrari, M. (2012). A brief review on the use of functional near-infrared spectroscopy (fNIRS) for language imaging studies in human newborns and adults. *Brain and Language*, *121*(2), 79–89. https://doi.org/10.1016/j.bandl.2011.03.009

R Core Team. (2023). *R: A language and environment for statistical computing*. (Version 4.3.2) [Computer software]. https://www.R-project.org/

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3–34.

Ramachers, S., Brouwer, S., & Fikkert, P. (2018). No perceptual reorganization for Limburgian tones? A cross-linguistic investigation with 6- to 12-month-old infants. *Journal of Child Language*, *45*(2), 290–318. https://doi.org/10.1017/S0305000917000228

Ren, J., Cai, L., Jia, G., & Niu, H. (2024). Cortical specialization associated with native speech category acquisition in early infancy. *Cerebral Cortex*, *34*(4), bhae124. https://doi.org/10.1093/cercor/bhae124

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.3.6) [Computer software]. https://CRAN.R-project.org/package=psych

Robinson, M. D., Storbeck, J., Meier, B. P., & Kirkeby, B. S. (2004). Watch out! That could be dangerous: Valence-arousal interactions in evaluative processing. *Personality and Social*

*Psychology Bulletin*, *30*(11), 1472–1484. https://doi.org/10.1177/0146167204266647

Rojczyk, A. (2011). Sound symbolism in vowels: Vowel quality, duration and pitch in sound-to-size correspondence. *Poznań Studies in Contemporary Linguistics*, *47*(3), 602–615. https://doi.org/10.2478/psicl-2011-0030

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Sakamoto, M., & Watanabe, J. (2018). Bouba/kiki in touch: Associations between tactile perceptual qualities and Japanese phonemes. *Frontiers in Psychology*, *9*, Article 295. https://doi.org/10.3389/fpsyg.2018.00295

Sato, Y., Sogabe, Y., & Mazuka, R. (2007). Brain responses in the processing of lexical pitch-accent by Japanese speakers. *NeuroReport*, *18*(18), 2001–2004. https://doi.org/10.1097/WNR.0b013e3282f262de

Sato, Y., Sogabe, Y., & Mazuka, R. (2010a). Development of hemispheric specialization for lexical pitch-accent in Japanese infants. *Journal of Cognitive Neuroscience*, *22*(11), 2503–2513. https://doi.org/10.1162/jocn.2009.21377

Sato, Y., Sogabe, Y., & Mazuka, R. (2010b). Discrimination of phonemic vowel length by Japanese infants. *Developmental Psychology*, *46*(1), 106–119. https://doi.org/10.1037/a0016718

Schacht, A., & Sommer, W. (2009). Time course and task dependence of emotion effects in word processing. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 28–43. https://doi.org/10.3758/CABN.9.1.28

Scheff, S. W. (2016). Nonparametric statistics. In *Fundamental statistical principles for the neurobiologist* (pp. 157–182). Academic Press. https://doi.org/10.1016/B978-0-12-804753-8.00008-7

Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In C. E. Izard (Ed.), *Emotions in personality and psychopathology* (pp. 493–529). Springer. https://doi.org/10.1007/978-1-4613-2892-6_18

Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. In H. Wagner & A. Manstead (Eds.), *Handbook of social psychophysiology* (pp. 165–197). John Wiley & Sons.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 433–456). Oxford University Press.

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, *1*(4), 331–346. https://doi.org/10.1007/BF00992539

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*(1), 24–30. https://doi.org/10.1016/j.tics.2005.11.009

Schmidtke, D., Conrad, M., & Jacobs, A. M. (2014). Phonological iconicity. *Frontiers in Psychology*, *5*, Article 80. https://doi.org/10.3389/fpsyg.2014.00080

Shafer, V. L., Yu, Y. H., & Garrido-Nag, K. (2012). Neural mismatch indices of vowel discrimination in monolingually and bilingually exposed infants: Does attention matter? *Neuroscience Letters*, *526*(1), 10–14. https://doi.org/10.1016/j.neulet.2012.07.064

Shang, N., & Styles, S. J. (2017). Is a high tone pointy? Speakers of different languages match Mandarin Chinese tones to visual shapes differently. *Frontiers in Psychology*, *8*, Article 2139. https://doi.org/10.3389/fpsyg.2017.02139

Shang, N., & Styles, S. J. (2023). Implicit association test (IAT) studies investigating pitch-shape audiovisual cross-modal associations across language groups. *Cognitive Science*, *47*(1), e13221. https://doi.org/10.1111/cogs.13221

Shih, S. S., Ackerman, J., Hermalin, N., Inkelas, S., Jang, H., Kavitskaya, D., Kawahara, S., Oh, M., Starr, R. L., & Yu, A. (2019). Cross-linguistic and language-specific sound symbolism: Pokémonastics. *Manuscript*. https://ling.auf.net/lingbuzz/004725

Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, *55*(2), 167–177. https://doi.org/10.1016/j.jml.2006.03.002

Shrum, L. J., Lowrey, T. M., Luna, D., Lerman, D. B., & Liu, M. (2012). Sound symbolism effects across languages: Implications for global brand names. *International Journal of*

*Research in Marketing*, *29*(3), 275–279. https://doi.org/10.1016/j.ijresmar.2012.03.002

Sidhu, D. M. (2025). Sound symbolism in the lexicon: A review of iconic-systematicity. *Language and Linguistics Compass*, *19*(1), e70006. https://doi.org/10.1111/lnc3.70006

Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*, *25*(5), 1619–1643. https://doi.org/10.3758/s13423-017-1361-1

Sidhu, D. M., Vigliocco, G., & Pexman, P. M. (2022). Higher order factors of sound symbolism. *Journal of Memory and Language*, *125*, 104323. https://doi.org/10.1016/j.jml.2022.104323

Singh, L., Rajendra, S. J., & Mazuka, R. (2022). Diversity and representation in studies of infant perceptual narrowing. *Child Development Perspectives*, *16*(4), 191–199. https://doi.org/10.1111/cdep.12468

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America*, *100*(4), 2471–2485. https://doi.org/10.1121/1.417955

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995. https://doi.org/10.3758/s13414-010-0073-7

State Language Commission. (1988). *Xiandai hanyu changyong zibiao [List of frequently used characters in modern Chinese]*. Yuwen Publishing House.

State Language Commission. (2013). *Tongyong guifan zibiao [List of commonly used characters in modern Chinese]*. The Commercial Press.

Stefanics, G., Háden, G. P., Sziller, I., Balázs, L., Beke, A., & Winkler, I. (2009). Newborn infants process pitch intervals. *Clinical Neurophysiology*, *120*(2), 304–308. https://doi.org/10.1016/j.clinph.2008.11.020

Stel, M., Dijk, E. van, Smith, P. K., Dijk, W. W. van, & Djalal, F. M. (2012). Lowering the pitch of your voice makes you feel more powerful and think more abstractly. *Social Psychological and Personality Science*, *3*(4), 497–502. https://doi.org/10.1177/1948550611427610

Storbeck, J., & Clore, G. L. (2008). Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and Personality Psychology Compass*, *2*(5), 1824–1843. https://doi.org/10.1111/j.1751-9004.2008.00138.x

Styles, S. J., & Gawne, L. (2017). When does Maluma/Takete fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *I-Perception*, *8*(4), 1–17. https://doi.org/10.1177/2041669517724807

Sučević, J., Savić, A. M., Popović, M. B., Styles, S. J., & Ković, V. (2015). Balloons and bavoons versus spikes and shikes: ERPs reveal shared neural processes for shape–sound-meaning congruence in words, and shape–sound congruence in pseudowords. *Brain and Language*, *145–146*, 11–22. https://doi.org/10.1016/j.bandl.2015.03.011

Sun, X., Li, X., Ji, L., Han, F., Wang, H., Liu, Y., Chen, Y., Lou, Z., & Li, Z. (2018). An extended research of crossmodal correspondence between color and sound in psychology and cognitive ergonomics. *PeerJ*, *6*, e4443. https://doi.org/10.7717/peerj.4443

Svantesson, J.-O. (2017). Sound symbolism: The role of word sound in meaning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(5), e1441. https://doi.org/10.1002/wcs.1441

Swoboda, P. J., Morse, P. A., & Leavitt, L. A. (1976). Continuous vowel discrimination in normal and at risk infants. *Child Development*, *47*(2), 459–465. https://doi.org/10.2307/1128802

Taga, G., Asakawa, K., Maki, A., Konishi, Y., & Koizumi, H. (2003). Brain imaging in awake infants by near-infrared optical topography. *Proceedings of the National Academy of Sciences*, *100*(19), 10722–10727. https://doi.org/10.1073/pnas.1932552100

Tarte, R. D. (1982). The relationship between monosyllables and pure tones: An investigation of phonetic symbolism. *Journal of Verbal Learning and Verbal Behavior*, *21*(3), 352–360. https://doi.org/10.1016/S0022-5371(82)90670-3

Telkemeyer, S., Rossi, S., Nierhaus, T., Steinbrink, J., Obrig, H., & Wartenburger, I. (2011). Acoustic processing of temporally modulated sounds in infants: Evidence from a combined near-infrared spectroscopy and EEG Study. *Frontiers in Psychology*, *2*, Article 62. https://doi.org/10.3389/fpsyg.2011.00062

Tervaniemi, M., & Hugdahl, K. (2003). Lateralization of auditory-cortex functions. *Brain Research Reviews*, *43*(3), 231–246. https://doi.org/10.1016/j.brainresrev.2003.08.004

The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in*

*Psychological Science*, *3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Thompson, A. L. (2018). Are tones in the expressive lexicon iconic? Evidence from three Chinese languages. *PloS ONE*, *13*(12), e0204270. https://doi.org/10.1371/journal.pone.0204270

Thompson, A. L., & Do, Y. (2019). Defining iconicity: An articulation-based methodology for explaining the phonological structure of ideophones. *Glossa: A Journal of General Linguistics*, *4*(1), Article 72. https://doi.org/10.5334/gjgl.872

Toet, A., & van Erp, J. B. F. (2019). The EmojiGrid as a tool to assess experienced and perceived emotions. *Psych*, *1*(1), Article 1. https://doi.org/10.3390/psych1010036

Tse, C.-S., Yap, M. J., & Chan, Y.-L. (2024). Neighborhood in Chinese lexicon: A megastudy analysis of lexical decision and naming of two-character Chinese words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(9), 1489–1515. https://doi.org/10.1037/xlm0001357

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191. https://doi.org/10.1002/dev.21179

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., Menyuk, P., & Best, C. (1994). Discrimination of English /r-l/ and /w-y/ by Japanese infants at 6-12 months: Language-specific developmental changes in speech perception abilities. *3rd International Conference on Spoken Language Processing*, 1695–1698. https://doi.org/10.21437/ICSLP.1994-438

Vainio, L., Kilpeläinen, M., Wikström, A., & Vainio, M. (2023). Front is high and back is low: Sound-space iconicity in Finnish.

*Language and Speech*, *67*(4), 1001–1019. https://doi.org/10.1177/00238309231214176

Vainio, L., Wikström, A., Repetto, C., & Vainio, M. (2023). Sound-symbolic association between speech sound and spatial meaning in relation to the concepts of up/down and above/below. *Language and Cognition*, *15*(4), 884–903. https://doi.org/10.1017/langcog.2023.31

van Lancker, D. (1980). Cerebral lateralization of pitch cues in the linguistic signal. *Paper in Linguistics*, *13*(2), 201–277. https://doi.org/10.1080/08351818009370498

Vannasing, P., Florea, O., González-Frankenberger, B., Tremblay, J., Paquette, N., Safi, D., Wallois, F., Lepore, F., Béland, R., Lassonde, M., & Gallagher, A. (2016). Distinct hemispheric specializations for native and non-native languages in one-day-old newborns identified by fNIRS. *Neuropsychologia*, *84*, 63–69. https://doi.org/10.1016/j.neuropsychologia.2016.01.038

Vigneau, M., Beaucousin, V., Hervé, P. Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B., & Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *NeuroImage*, *30*(4), 1414–1432. https://doi.org/10.1016/j.neuroimage.2005.11.002

Visser, I., Kucharský, Š., Levelt, C., Stefan, A. M., Wagenmakers, E.-J., & Oakes, L. (2024). Bayesian sample size planning for developmental studies. *Infant and Child Development*, *33*(1), e2412. https://doi.org/10.1002/icd.2412

Walker, P. (2012). Cross-sensory correspondences and cross talk between dimensions of connotative meaning: Visual

angularity is hard, high-pitched, and bright. *Attention, Perception, & Psychophysics*, *74*(8), 1792–1809. https://doi.org/10.3758/s13414-012-0341-9

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21–25. https://doi.org/10.1177/0956797609354734

Walker, P., & Smith, S. (1985). Stroop interference based on the multimodal correlates of haptic size and auditory pitch. *Perception*, *14*(6), 729–736. https://doi.org/10.1068/p140729

Wang, X. (2021). Sound symbolism in Chinese children's literature. *Cognitive Linguistics*, *33*(1), 95–120. https://doi.org/10.1515/cog-2021-0019

Wang, Y., Zhou, L., & Luo, Y. (2008). The pilot establishment and evaluation of Chinese Affective Words System. *Chinese Mental Health Journal*, *22*(8), 608–612.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Warriner, A. B., Shore, D. I., Schmidt, L. A., Imbault, C. L., & Kuperman, V. (2017). Sliding into happiness: A new tool for measuring affective responses to words. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *71*(1), 71–88. https://doi.org/10.1037/cep0000112

Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, *39*(4), 703–728. https://doi.org/10.1017/S0142716418000152

Werker, J. F. (2024). Phonetic perceptual reorganization across the first year of life: Looking back. *Infant Behavior and Development*, *75*, 101935. https://doi.org/10.1016/j.infbeh.2024.101935

Werker, J. F., & Polka, L. (1993). Developmental changes in speech perception: New challenges and new directions. *Journal of Phonetics*, *21*(1–2), 83–101. https://doi.org/10.1016/S0095-4470(19)31322-1

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3

Wernicke, C. (1874). *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert.

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language*, *93*(1), 10–19. https://doi.org/10.1016/j.bandl.2004.07.006

Whalen, D., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, *49*(1), 25–47. https://doi.org/10.1159/000261901

Whissell, C. (2003). The emotional symbolism of two English e-sounds: /I/ as in “cheap” is pleasant and /I/ as in “chip” active. *Perceptual and Motor Skills*, *96*(1), 149–165. https://doi.org/10.2466/pms.2003.96.1.149

Wilcox, T., & Biondi, M. (2015). fNIRS in the developmental sciences. *WIREs Cognitive Science*, *6*(3), 263–283. https://doi.org/10.1002/wcs.1343

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2023). Iconicity ratings for 14,000+ English words. *Behavior Research Methods*, *56*, 1640–1655. https://doi.org/10.3758/s13428-023-02112-6

Winter, B., Oh, G. E., Hübscher, I., Idemaru, K., Brown, L., Prieto, P., & Grawunder, S. (2021). Rethinking the frequency code: A meta-analytic review of the role of acoustic body size in communicative phenomena. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1840), 20200400. https://doi.org/10.1098/rstb.2020.0400

Winter, B., & Perlman, M. (2021). Iconicity ratings really do measure iconicity, and they open a new window onto the nature of language. *Linguistics Vanguard*, *7*(1), 20200135. https://doi.org/10.1515/lingvan-2020-0135

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? Iconicity in English sensory words. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *18*(3), 443–464. https://doi.org/10.1075/is.18.3.07win

Witteman, J., Goerlich-Dobre, K. S., Martens, S., Aleman, A., Van Heuven, V. J., & Schiller, N. O. (2014). The nature of hemispheric specialization for prosody perception. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(3), 1104–1114. https://doi.org/10.3758/s13415-014-0255-1

Witteman, J., van IJzendoorn, M. H., van de Velde, D., van Heuven, V. J. J. P., & Schiller, N. O. (2011). The nature of hemispheric

specialization for linguistic and emotional prosodic perception: A meta-analysis of the lesion literature. *Neuropsychologia*, *49*(13), 3722–3738. https://doi.org/10.1016/j.neuropsychologia.2011.09.028

Wong, K. W. Y., & Kang, Y. (2019). Sound symbolism of gender in Cantonese first names. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 2129–2133).

Xiao, W., & Treiman, R. (2012). Iconicity of simple Chinese characters. *Behavior Research Methods*, *44*(4), 954–960. https://doi.org/10.3758/s13428-012-0191-3

Xu, D. (2012). Reduplication in languages: A case study of languages of China. In *Plurality and Classifiers across languages in China* (pp. 43–62). https://hal.science/hal-01386244

Xu, X., Li, J., & Chen, H. (2022). Valence and arousal ratings for 11,310 simplified Chinese words. *Behavior Research Methods*, *54*(1), 26–41. https://doi.org/10.3758/s13428-021-01607-4

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*, 61–83.

Xu, Y., Gandour, J., Talavage, T., Wong, D., Dzemidzic, M., Tong, Y., Li, X., & Lowe, M. (2006). Activation of the left planum temporale in pitch processing is shaped by language experience. *Human Brain Mapping*, *27*(2), 173–183. https://doi.org/10.1002/hbm.20176

Yao, Y., Lin, J., & Huang, C.-R. (2013). *Lexicalized emotion? Tonal patterns of emotion words in Mandarin Chinese*. The 25th North American Conference on Chinese Linguistics.

Yap, D., Casasanto, L. S., & Casasanto, D. (2014). Metaphoric iconicity in signed and spoken languages. *Proceedings of the*

*Annual Meeting of the Cognitive Science Society*, *36*, 1808–1813.

Yeung, H. H., Chen, K. H., & Werker, J. F. (2013). When does native language input affect phonetic perception? The precocious case of lexical tone. *Journal of Memory and Language*, *68*(2), 123–139. https://doi.org/10.1016/j.jml.2012.09.004

Yip, M. (2002). *Tone*. Cambridge University Press.

Yu, C. S.-P., McBeath, M. K., Glenberg, A. M., & Benjamin, A. S. (2021). The gleam-glum effect: /I:/ versus /λ/ phonemes generically carry emotional valence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(7), 1173–1185. https://doi.org/10.1037/xlm0001017

Yuan, J., & Chen, Y. (2014). 3rd tone sandhi in Standard Chinese: A corpus approach. *Journal of Chinese Linguistics*, *42*(1), 218–236.

Yücel, M. A., Lühmann, A. v, Scholkmann, F., Gervain, J., Dan, I., Ayaz, H., Boas, D., Cooper, R. J., Culver, J., Elwell, C. E., Eggebrecht, A., Franceschini, M. A., Grova, C., Homae, F., Lesage, F., Obrig, H., Tachtsidis, I., Tak, S., Tong, Y., … Wolf, M. (2021). Best practices for fNIRS publications. *Neurophotonics*, *8*(1), 012101. https://doi.org/10.1117/1.NPh.8.1.012101

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, *11*(10), 946–953. https://doi.org/10.1093/cercor/11.10.946

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, *256*(5058), 846–849. https://doi.org/10.1126/science.256.5058.846

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1087–1104. https://doi.org/10.1098/rstb.2007.2161

Zhang, F., Gervain, J., & Roeyers, H. (2022). Developmental changes in the brain response to speech during the first year of life: A near-infrared spectroscopy study of Dutch-learning infants. *Infant Behavior and Development*, *67*, 101724. https://doi.org/10.1016/j.infbeh.2022.101724

Zheng, T., Levelt, C. C., & Chen, Y. (2024). The adaptive value of Mandarin tones for affective iconicity. *Proc. Speech Prosody 2024*, 562–566. https://doi.org/10.21437/SpeechProsody.2024-114

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

Zheng, T., Levelt, C. C., & Chen, Y. (under revision). *The affective iconicity of phonemes and tonemes in Standard Chinese*.

Zhou, X., & Marslen-Wilson, W. (1995). Morphological structure in the Chinese mental lexicon. *Language and Cognitive Processes*, *10*(6), 545–600. https://doi.org/10.1080/01690969508407114

# Appendices

## Appendix A. Affective iconicity in Standard Chinese: Related studies overview and supplementary results

### Table A1

*An overview of studies on pitch iconicity.*

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| Lester (1974) | American English | Mandarin tones (presented by visual symbols ¯ˊˇˋ) | No significant recognition of Mandarin tones presented visually (52% correct) |
| Marks (1974) | English | Pitch height | Higher-pitched sounds were associated with brighter lights. |
| Tarte (1982) | English | Pitch height | Low tones were perceived as large, heavy, slow, dull, low, and masculine, whereas high tones were perceived as small, light, fast, sharp, high, and feminine. |
| Walker & Smith (1985) | English | Pitch height | Participants responded more slowly when pitch was incongruent with the multimodal features of test words. |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| Marks (1987) | English | Pitch height | Response times were faster when auditory and visual stimuli were congruent (e.g., high-pitched beeps with fast-flashing lights). |
| Marks et al. (1987) | English | Pitch height | Children and adults matched high pitch with brightness and low pitch with dimness. Pitch-size associations emerged around age 11. |
| Melara & O'Brien (1987) | English | Pitch height | Participants classified dot location and tone height faster in congruent conditions (e.g., a high tone with a dot at a high location). |
| Melara (1989) | English | Pitch height | Faster and more accurate responses when pitch and colour were congruent (e.g., high-pitched tones with white dots). |
| Lapolla (1995) | Mandarin Chinese | Mandarin tones | High-level tones were linked to "coarse" and "wide," while falling tones were associated with "largeness." |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| Lapolla (1995) | Cantonese | Cantonese tones | Mandarin speakers linked high-level tones with "smallness" and falling tones with "largeness." |
| M. K. M. Chan (1996) | Chinese | Overall pitch or pitch register | Across Standard Cantonese, Xiamen, and Wu dialects, higher pitch (Yin register) was linked to lightness, while lower pitch (Yang register) was linked to heaviness. |
| Ohala (1984, 1994, 1997) | Ewe, Yoruba, Cantonese | Tones in West African languages | High tones were associated with smallness and low tones with largeness. |
| Mondloch & Maurer (2004) | (Canada) English | Pitch height | Children reliably matched higher-pitched sounds with smaller, lighter objects. |
| Gallace & Spence (2006) | English | Pitch height | Participants reacted faster when a low-frequency sound was paired with a larger disk. |
| Shintel et al. (2006) | American English | Pitch height | Speakers described an upward-moving dot with a higher fundamental frequency. |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| Parise & Spence (2008) | English | Pitch height | High-pitched sounds were associated with smaller sizes, and low-pitched sounds with larger sizes. |
| Crisinel & Spence (2009) | British English | Pitch height | Faster responses to congruent pitch-taste pairings (e.g., high-pitched notes with sweet tastes). |
| Crisinel & Spence (2010a) | English | Pitch height | Sweet and sour tastes were associated with high-pitched sounds; bitter and salty tastes lacked a clear association with low pitch. |
| Crisinel & Spence (2010b) | English | Pitch height | Non-synesthetes showed systematic associations between tastes and musical notes (e.g., bitterness with lower-pitched notes, sweetness with higher-pitched notes). |
| Evans & Treisman (2011) | English | Pitch height | Pitch spontaneously mapped onto visual position, size, and spatial frequency but not contrast. |
| Perniss et al. (2010) | NA | NA | Pitch functions as a form of iconicity in sound symbolism and signed |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| | | | languages (e.g., prosodic marking). |
| Walker et al. (2010) | English | Pitch height | Infants looked longer at animations where pitch changed congruently with motion (e.g., rising pitch for rising objects). |
| Ludwig et al. (2011) | NA | Pitch height | Chimpanzees and humans performed better in congruent pitch-colour associations (e.g., high pitch with white, low pitch with black). |
| Rojczyk (2011) | Polish | Pitch height | Lowered pitch did not significantly influence size ratings, except for the vowel /u/. |
| Spence (2011) | NA | Pitch height | High pitch corresponded with smaller objects, higher elevation, brighter colours, angular shapes, higher spatial frequency, and upward motion. |
| Chiou & Rich (2012) | English | Pitch height | Matching auditory pitch facilitated visual elevation judgments, suggesting |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| | | | pitch-location mapping at an attentional level. |
| Chang et al. (2021) | Mandarin Chinese | Mandarin tones | Tone was more often matched with rounded shapes and large size, and tone with angular shapes and small size. |
| Crisinel et al. (2012) | English | Pitch height | Taste perception was influenced by background music (e.g., toffee tasted more bitter with low-pitched sounds). |
| Parise & Spence (2012) | British English | Pitch height | High pitch was linked to smaller, sharper shapes, while low pitch was linked to larger, more rounded shapes. |
| Stel et al. (2012) | NA | Pitch height | Lowering one's voice pitch increased feelings of power, but only when self-produced. |
| Walker (2012) | English | Pitch height | High-pitched words were associated with angular shapes and low-pitched words with curved shapes. |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| Deroy & Spence (2013) | NA | Pitch height | Higher pitch was consistently matched to brightness, small size, high elevation, angularity, and upward movement. |
| Yao et al. (2013) | Mandarin Chinese | Mandarin tones | Words expressing sadness had higher tonal levels, while joy and anger had steeper tonal contours. |
| Perlman & Cain (2014) | English | Pitch average: fundamental frequency in Hz; Pitch range: the absolute value of the difference between the maximum and minimum *fo*; Pitch change: the ordered difference between the maximum and minimum *fo*. | Smooth textures are vocalised with higher pitch than rough ones. Downward space corresponds to falling and lower pitch, while upward space aligns with higher pitch. Sharp shapes elicit higher pitch than dull shapes. Positive appraisals have a wider pitch range and higher pitch than negative ones. Male voices are lower with less pitch decrease than female voices. Only hearing children consistently used pitch for magnitude, unexpectedly associating larger items with higher |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| | | | pitch, contrary to previous English-speaking findings. |
| Fernández-Prieto et al. (2015) | English | Pitch height | Crossmodal pitch-size correspondence effects emerged in 6-month-old infants but not in younger ones. |
| Perlman, Dale, et al. (2015) | English | Pitch height | Readers used lower pitch when narrating "big" stories compared to "small" ones. |
| Perlman, Dale, et al. (2015) | English | Pitch height<br>Pitch range | Smooth textures are vocalised with higher pitch. Downward space corresponds to falling and lower pitch, while upward space aligns with higher pitch. Positive appraisals have a wider pitch range and higher pitch than negative ones. |
| Lowe & Haws (2017) | English | Pitch height | Lower pitch in voice or music led consumers to infer larger product sizes. |
| Perlman (2017) | English | Pitch height<br>Pitch contour<br>Pitch range | Iconic vocalisations in vocal charades followed these pitch patterns: smooth-higher, downward- |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| | | | falling/lower, sharp-higher, good appraisal-larger range/higher, male-lower/less decrease, cut action-higher. |
| Shang & Styles (2017) | Mandarin Chinese | Mandarin tones | Native speakers: T1-curvy, T4-pointy.<br>English speakers: T1-pointy, T3-curvy.<br>Bilinguals: bivalent pattern. |
| Svantesson (2017) | NA | High vs. low tone | Kammu, Yoruba, and Ewe onomatopoeic ideophones followed the frequency code (higher pitch = smaller size). |
| Getz & Kubovy (2018) | English | Pitch height | AVCs rely on both bottom-up and top-down processing. Top-down influence was strongest for size, weakest for height, while bottom-up effects were strongest for height, weakest for brightness. |
| Hamilton-Fletcher et al. (2018) | NA | Pitch height | Blind individuals showed reduced pitch-shape correspondence but maintained pitch-size and pitch-weight associations. |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| Nielsen & Rendall, (2018) | NA | Pitch height | Pitch-affect connections, often labelled as synesthetic, were also observed in non-synesthetes. |
| Shih et al. (2019) | Japanese, English, Mandarin, Cantonese, Korean, and Russian | Mandarin tones, Cantonese tones | In Mandarin, tone was linked to male names and powerful Pokémon attributes. In Cantonese, tone was negatively correlated with height and power. |
| Sidhu & Pexman (2018) | NA | Pitch height | High-pitched sounds were perceived as brighter, sharper, and faster. |
| Sun et al. (2018) | NA | Pitch height | High pitch was associated with red and yellow, while low pitch was linked to blue and orange. |
| Thompson (2018) | Mandarin, Cantonese, and Taiwanese | Mandarin tones, Taiwanese tones, and Cantonese tones | Across Mandarin, Cantonese, and Taiwanese, sound symbolic strata were skewed toward specific tonal categories. |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| Anikin & Johansson (2019) | English | Pitch height[2] | High pitch showed weak associations with blue, light grey, high saturation, and high luminance. |
| Wong & Kang (2019) | Cantonese | Cantonese tones | Rising tones showed a significant preference for female names. |
| Matsui (2020) | Japanese Chinese | Pith height | Vowels /u/ and /o/ were responded for low frequencies, /i/ for high frequencies across speakers of different languages. There are common relationships between the pitch of pure tones and onomatopoeic expressions. |
| Akita (2021) | Japanese | Japanese pitch accent | Low-pitched quotatives after exclamatory quotations and ideophonic adverbs in Japanese serve as backgrounding depiction markers, suppressing description while |

[2] In this article, the authors pointed out that pitch is usually considered a metathetic dimension, in the sense that higher pitch is not "larger" or "greater" than low pitch, but qualitatively different.

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| | | | highlighting depiction. Ideophone pitch contours, including HLL and HHL patterns, carry partial semantic motivation. |
| X. Wang (2021) | Chinese | Mandarin tones | Nature: rising tones (T35) were more common in human names than in animal or monster names, though post hoc analysis showed no significant difference.<br>Gender: rising tones (T35) were more frequent in male names than female names.<br>Personality traits: high-level tones (T55) were prevalent in positive character names, while negative characters often featured rising tones.<br>Size: no significant tonal distinctions were found. |
| Winter et al. (2021) | Japanese, English, Catalan, Spanish, Dutch, Korean, | Pitch height | A meta-analysis of speech production experiments across multiple languages (Korean, Japanese, Chinese, Catalan, Austrian German, German, Russian) found |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| | German, Austrian, Russian, Chinese | | that speakers lower their pitch when addressing an imagined superior compared to a friend or peer. |
| Ekström et al. (2022) | NA | Pitch height | Motion-prosody congruent pairings, particularly those with a declining *fo*, were more readily selected than incongruent ones, except for Turkish-speaking participants. |
| González-Alvarez & Sos-Peña (2022) | Spanish | Pitch height | Listeners more accurately perceived the speaker's body size when *fo* was raised. |
| Shang & Styles (2023) | Mandarin | Pitch height | All language groups exhibited basic pitch-height congruence (high-pointy, low-curvy) for the non-linguistic stimuli. |
| Vainio, Kilpeläinen, et al. (2023) | Finnish | Pitch height | The results reveal a novel sound-space symbolism phenomenon, where spatial concepts of forward/front and backwards/back are iconically linked to high- |

| Citation | Language | Pitch parameters | Main finding |
|---|---|---|---|
| | | | and low-pitched speech sounds. |
| Vainio, Wikström, et al. (2023) | Finnish | Pitch height | This study replicated the pitch-elevation effect, showing an increase in vocalisation pitch when responding to an up-directed stimulus. |

**Table A2**

*Summary of inferential statistics for LOOCV for arousal in the DCAWS dataset.*

| Sub set | GLMM F (HLM, $\Delta R^2$) coefficients | Significance of multiple comparisons: *p-values in GLMM (HLM)* | | | | |
|---|---|---|---|---|---|---|
| | | **FF-RL** | **FF-RR** | **HF-RL** | **HL-RL** | **FH-RL** |
| 1 | 36.7 (0.016) | 0 (0) | 0.006 (0.014) | 0.006 | 0.007 | 0.014 |
| 2 | 31.1 (0.015) | 0 (0.001) | 0.008 (0.037) | 0.058 | 0.018 | *n.s.* |
| 3 | 33.1 (0.015) | 0 (0) | 0.007 (0.031) | 0.015 | 0.022 | *n.s.* |
| 4 | 29.1 (0.014) | 0 (0.002) | 0.012 (0.051) | 0.048 | 0.048 | *n.s.* |
| 5 | 31.2 (0.015) | 0 (0.003) | 0.005 (0.028) | 0.041 | 0.060 | *n.s.* |
| 6 | 32.4 (0.015) | 0 (0) | 0.005 (0.027) | 0.052 | 0.050 | 0.046 |
| 7 | 33.9 (0.016) | 0 (0) | 0.004 (0.018) | 0.016 | 0.026 | 0.038 |
| 8 | 34.0 (0.017) | 0 (0) | 0.003 (0.014) | 0.042 | 0.013 | *n.s.* |
| 9 | 34.4 (0.016) | 0 (0) | 0.004 (0.019) | 0.016 | 0.015 | 0.038 |
| 10 | 31.5 (0.016) | 0 (0) | 0.004 (0.014) | 0.031 | *n.s.* | *n.s.* |

The HLM indicated that the ratings of emotional valence were not significantly influenced by the lexical tonal sequence. Additionally, the impact of the lexical tonal sequence on emotional valence seems comparatively minor when contrasted with its influence on emotional arousal. Table A3 indicates the detailed R-squared coefficients.

**Table A3**

*$\Delta R^2$ of lexical tonal sequences in explaining emotional arousal and valence across all corpora.*

| Corpus | Emotional arousal | Emotional valence |
|---|---|---|
| **CAWS** | 2.22%*** | 0.88% |
| **NORM** | 0.44%*** | 0.22% |
| **DCAWS** | 1.60%** | 0.84% |

*Note*. The differential R-squared coefficients *($\Delta R^2$)* were obtained by subtracting the $R^2$ of the two-block HLM models in each corpus analysis.

## Appendix B. Pseudowords formulation and their wordlikeness ratings in Dutch and Japanese

A set of pseudowords was created based on the shared phonotactic rules of Dutch and Japanese. First, phonemes that exist in both languages were identified. These phonemes were then composed into pseudowords with vowel (V) or consonant-vowel (CV) syllabic structures. The pseudowords were subsequently manipulated and recorded by a bilingual speaker fluent in both Japanese and Dutch. Finally, native speakers of Japanese and Dutch evaluated the pseudowords for their wordlikeness in each language. More details are shown as follows.

### Phonemes

Six co-existing phonemes, i.e., /p/, /t/, /k/, /s/, /i/, and /a/, were selected (Gussenhoven, 1992 for Dutch; Okada, 1991 for Japanese). Specifically, /p/ is a voiceless bilabial plosive consonant, /t/ a voiceless alveolar plosive, /k/ a voiceless velar plosive, and /s/ a voiceless alveolar fricative. Furthermore, the vowel /i/ is a close (or, high), front, and unrounded vowel, while /a/ open (or, low), front and unrounded vowel.

### Words composition

All pseudowords were created according to the phonotactic rules of both languages, based on previous studies of acceptability judgments for pseudowords (Bailey & Hahn, 2001; Needle et al., 2022). The chosen phonotactic framework followed a (C)VCV disyllabic structure. Six distinct syllabic constituents (i.e., /i/, /pi/, /pa/, /ta/, /ka/, and /sa/) were concatenated to make ten disyllabic pseudowords, including

/ipa/, /kapi/, /pasa/, /pika/, /pipa/, /pisa/, /pita/, /sapa/, /sapi/, and /tapi/.

### Words recording

Two pitch patterns, High-Low (HL) and Low-High (LH), were used for all pseudowords during recording. All stimuli were recorded in isolation with a Sennheiser MKH416T microphone (sample size 44.1 kHz, 16 bit) at Leiden University's Phonetics Lab by a female native speaker of Japanese (from the Tokyo area) who is proficient in Dutch. The speech signals were digitised at a 44.1 kHz sampling rate with 16-bit resolution. She was asked to produce the stimuli as a statement without any emphasis. The stimuli were recorded three times, with the stimulus list randomised for each recording. Tokens judged to be most clearly articulated by the first author were selected for further manipulations.

### Words manipulation

To optimise the stimuli for use in both languages, the pseudowords were further resynthesised. All steps were performed using Praat (Boersma & Weenink, 2024).

#### *Segmentation*

Since all target words contained a second syllable starting with an obstruent onset, segmentation was relatively straightforward. The acoustic waveforms, accompanied by corresponding spectrograms and auditory verification, provided clear cues of spectral shifts in a zoomed-in display to identify reliable syllable boundary locations.

The first syllables had either a stop onset (/p/, /k/, /t/), a fricative onset (/s/), or a vowel onset (/i/), which is often preceded by a phonetic

glottal closure. We marked the onset of the first syllable at the first positive-going zero crossing in the waveform for the vowel /i/, right before the noise burst for the release of a stop closure (/p/, /k/, /t/ and /ʔ/), or at the start of the medium amplitude noise for /s/.

The end of the first syllable was marked at the last negative-going zero crossing for both the vowel and the nasal coda before the closure for the following stop onset.

We took the end of the first syllable as the start of the following syllable, while the offset of the second syllable was marked as the last negative-going zero crossing for both the vowel and nasal coda.

### *Annotation*

After segmenting the sounds, we cut them into monosyllables using TextGrid files and moved the start and end of each monosyllable to the nearest zero crossings. According to phoneticians and phonologists' consultations, some of the syllables were manipulated in length or replaced with better-recorded ones.

### *Duration modification*

To enhance the experimental parameters, we further manipulated the duration lengths of the concatenated words, given that duration proved the most reliable correlate of stress in Dutch and can be exploited for recognising spoken words (Cutler & van Donselaar, 2001; Sluijter & van Heuven, 1996). Based on studies of vowel discrimination, the durations of V syllables were calibrated to 220 ms, the CV syllables to 250 ms, and the pause interval between syllables, if any, was adjusted to 100 ms (De Klerk et al., 2019; Shafer et al., 2012; Swoboda et al., 1976). Note that /s/ was adjusted to 150 ms to avoid harsh sound

effects. All pseudowords were evaluated as natural as real words in two languages by phoneticians and phonologists.

### Ratings on the pseudowords

After manipulating the lexical items, we conducted an online rating task using Qualtrics (Provo, UT, USA. Copyright © 2020 Qualtrics. https://www.qualtrics.com) to evaluate their perceived qualities. Native speakers of Japanese and Dutch assessed all pseudowords to determine the extent to which they resonated with the auditory characteristics of actual Dutch or Japanese vocabulary. Participants rated each word on a 5-point scale, where 5 indicated that a word sounded very much like a possible Dutch/Japanese word, and 1 indicated that a word did not sound like a possible Dutch/Japanese word at all.

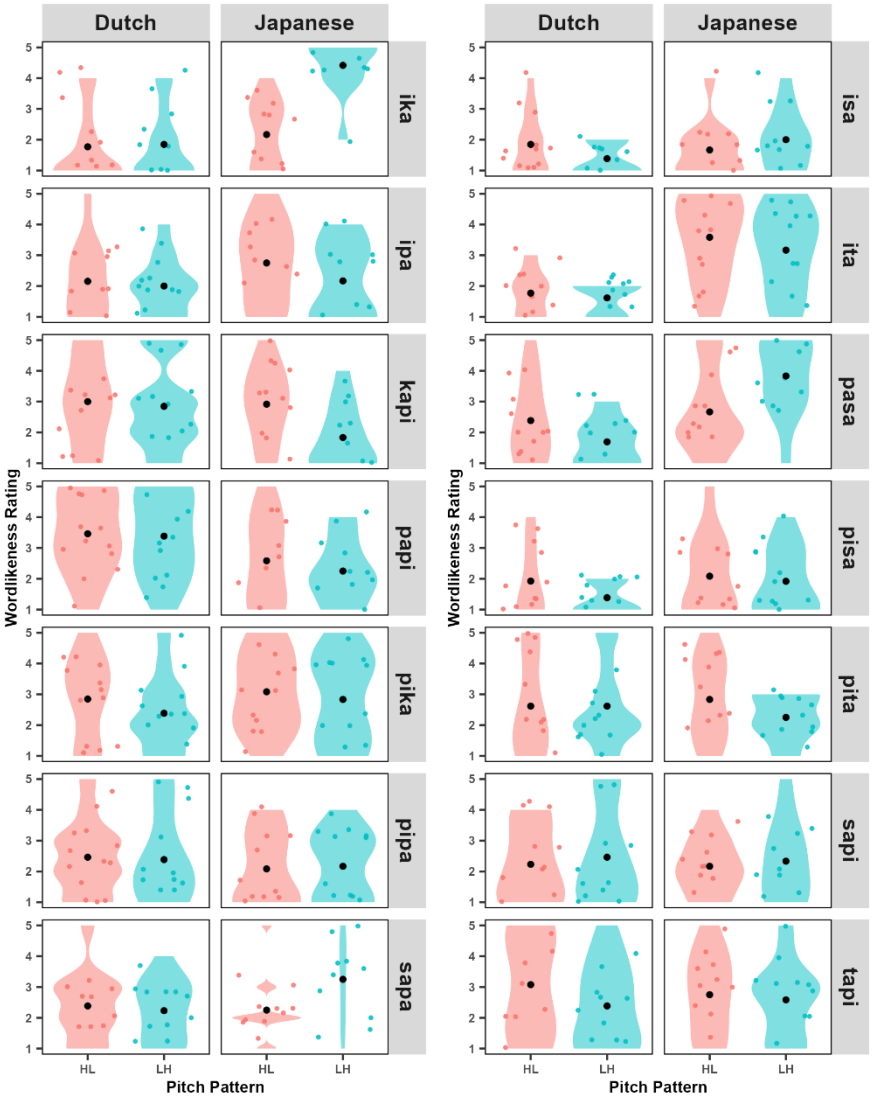
**Table B1**

*The demographics information of Dutch and Japanese raters.*

| Rater | Male | Female | Age range | Mean age |
|---|---|---|---|---|
| Dutch | 3 | 10 | 20–73 | 55.70 |
| Japanese | 5 | 7 | 19–48 | 35.50 |

**Figure B1**

*The likeness rating on all words by native language speakers in Japanese and Dutch.*



## Appendix C. Outcomes from all fNIRS data preprocessing pipelines in Experiments 2 and 3 of Chapter 5

This table summarises the outcomes of all preprocessing pipelines and datasets, including dataset inclusion counts and hemispheric response patterns for pitch processing from 4 months (4m) to 10 months (10m).

- Pipeline-Dataset: Represents the pipeline and dataset numbers (e.g., 1-2 indicates Pipeline 1 applied to Dataset 2).
- Inclusions (4m/10m): Indicates the number of datasets included in statistical analyses (e.g., 21/27 means 21 datasets for 4m and 27 for 10m).
- Hemispheric Response (4m–10m): Shows dominant hemisphere patterns for PT and Word conditions. For example, RH-BH under "Word" indicates a shift from right-hemispheric (RH) dominance at 4m to bilateral (BH) response at 10m. "n.s." (non-significant): Indicates that the interaction effect of Condition × Hemisphere × Age was not statistically significant, meaning no clear hemispheric dominance was observed.

**Table C1**

*Summary of outcomes across pipelines and datasets.*

| Pipeline -Dataset | NL | | | JP | | |
|---|---|---|---|---|---|---|
| | Inclusions 4m/ 10m | Hemispheric response 4m-10m | | Inclusions 4m/ 10m | Hemispheric response 4m-10m | |
| | | PT | Word | | PT | Word |
| 1-1 | 21/27 | RH-RH | RH-BH | 16/14 | BH-RH | LH-LH |
| 1-2 | 21/27 | RH-RH | RH-BH | 16/14 | BH-RH | LH-LH |

| Pipeline -Dataset | NL | | | JP | | |
|---|---|---|---|---|---|---|
| | Inclusions 4m/ 10m | Hemispheric response 4m-10m | | Inclusions 4m/ 10m | Hemispheric response 4m-10m | |
| | | PT | Word | | PT | Word |
| 1-3 | 20/26 | RH-RH | RH-BH | 16/14 | BH-RH | LH-LH |
| 1-4 | 18/26 | RH-RH | RH-BH | 16/14 | RH-n.s. | LH-n.s. |
| 2-1 | 21/27 | RH-RH | RH-BH | 16/14 | BH-RH | LH-LH |
| 2-2 | 21/27 | RH-RH | RH-BH | 16/14 | BH-RH | LH-LH |
| 2-3 | 20/26 | RH-RH | RH-BH | 16/14 | BH-RH | LH-LH |
| 2-4 | 17/24 | RH-RH | RH-LH | 16/14 | n.s.-RH | n.s.-LH |
| 3-1 | 21/27 | RH-BH | LH-LH | 8/2 | LH-LH | LH-BH |
| 3-2 | 21/27 | RH-BH | LH-LH | 7/2 | LH-LH | LH-BH |
| 3-3 | 21/27 | RH-RH | LH-RH | 7/2 | LH-LH | LH-RH |
| 3-4 | 17/26 | RH-RH | LH-RH | 7/2 | LH-LH | LH-RH |
| 4-1 | 21/27 | RH-BH | LH-LH | 7/2 | LH-LH | LH-BH |
| 4-2 | 21/27 | RH-BH | LH-LH | 7/2 | LH-LH | LH-BH |
| 4-3 | 21/27 | RH-RH | LH-RH | 7/2 | LH-LH | LH-RH |
| 4-4 | 16/26 | RH-RH | LH-BH | 7/2 | BH-LH | RH-RH |
| 5-1 | 20/26 | RH-RH | RH-LH | 16/9 | BH-RH | LH-BH |
| 5-2 | 20/26 | RH-RH | RH-LH | 16/9 | BH-RH | LH-BH |
| 5-3 | 20/26 | RH-RH | RH-LH | 16/9 | RH-RH | LH-BH |
| 5-4 | 16/24 | RH-RH | RH-LH | 16/9 | RH-RH | LH-RH |
| 6-1 | 20/26 | RH-RH | RH-LH | 16/9 | BH-RH | LH-BH |
| 6-2 | 20/26 | RH-RH | RH-LH | 16/9 | BH-RH | LH-BH |
| 6-3 | 19/24 | RH-RH | RH-LH | 16/9 | BH-RH | LH-BH |
| 6-4 | 16/24 | RH-RH | RH-LH | 16/9 | BH-RH | LH-LH |
| 7-1 | 21/27 | RH-BH | BH-LH | 8/2 | LH-LH | BH-RH |
| 7-2 | 21/27 | RH-BH | BH-LH | 8/2 | LH-LH | BH-RH |
| 7-3 | 21/26 | RH-BH | BH-LH | 8/2 | LH-LH | BH-RH |
| 7-4 | 17/24 | RH-RH | BH-BH | 8/2 | LH-BH | BH-RH |

| Pipeline -Dataset | NL | | | JP | | |
|---|---|---|---|---|---|---|
| | Inclu-sions 4m/ 10m | Hemispheric response 4m-10m | | Inclu-sions 4m/ 10m | Hemispheric response 4m-10m | |
| | | PT | Word | | PT | Word |
| 8-1 | 21/27 | RH-BH | BH-LH | 8/2 | LH-LH | BH-RH |
| 8-2 | 21/27 | RH-BH | BH-LH | 8/2 | LH-LH | BH-RH |
| 8-3 | 21/26 | RH-BH | BH-RH | 8/2 | BH-LH | BH-RH |
| 8-4 | 16/24 | RH-RH | BH-LH | 6/2 | BH-RH | BH-RH |

# English summary

Among the various perceptual cues in speech, pitch plays a universal and multifaceted role. It distinguishes grammatical structures (e.g., questions vs. statements), conveys emotion, and—crucially in tone languages—differentiates lexical meanings. Beyond these established functions, pitch can also convey meaning iconically, through its perceptual features (such as height and contour direction) and shared socio-linguistic conventions. Moreover, infants' acquisition of pitch depends not only on its acoustic properties but also on its linguistic function. Yet, two fundamental questions remain open: (1) Do lexical tones in tone languages like Standard Chinese exhibit affective iconicity—systematically shaping interpretations of emotional arousal and valence due to their pitch characteristics? and (2) How does pitch processing develop in infants acquiring a pitch-accent language versus a stress language, particularly regarding hemispheric lateralisation during perceptual reorganisation?

This dissertation addresses these questions across six chapters.

Chapter 1 lays the conceptual foundation, reviewing relevant literature on iconicity in language, affective iconicity, and the neural development of pitch perception. It also introduces the methodological approaches—including behavioural rating tasks and neuroimaging (fNIRS)—and outlines the cross-linguistic comparison between Dutch and Japanese infants.

Chapters 2 through 4 focus on affective iconicity. Using behavioural tasks such as Likert scales, Slider scales, and forced-choice paradigms, Chapter 2 reveals that tonal features such as pitch height, range, slope, and contour direction influence emotional arousal ratings in both real and nonce bisyllabic words. Valence effects were more

limited, emerging only in auditory nonce words. These results suggest that arousal is more closely linked to perceptual properties of pitch, while valence is shaped by higher-level lexical and cognitive influences.

Chapter 3 extends the investigation to monosyllabic tokens. While rising and falling tones were rated as more arousing than level tones, these distinctions were less pronounced than in bisyllabic sequences. Notably, lexical tones were more predictive of emotional responses than consonants or vowels—especially when segmental and suprasegmental cues conflicted—highlighting the dominant role of pitch in emotional interpretation.

Chapter 4 explores potential mechanisms behind these effects. Tones in the first syllable were associated with emotional valence, mirroring patterns observed in Indo-European languages. Rising tones in the first syllable more frequently appeared in words rated as positive, while falling tones were more common in words rated as negative. Arousal was influenced by tones in both syllables, with falling tones occurring more often in high-arousal words than rising or low-dipping tones. These findings suggest that affective iconicity in lexical tone may serve adaptive communicative functions, aligning with evolutionary accounts of iconicity.

Chapter 5 uses functional near-infrared spectroscopy (fNIRS) to examine Dutch and Japanese infants' neural responses to pitch changes in speech and non-speech stimuli at 4 and 10 months. At 4 months, both groups showed rightward lateralisation for non-speech pitch. By 10 months, Japanese infants displayed leftward lateralisation for speech, while Dutch infants showed more bilateral responses. These findings suggest that pitch processing becomes specialised for language in infants acquiring pitch-relevant languages, while remaining general in non-tonal language learners.

Chapter 6 summarises the main findings, discusses their implications for affective iconicity and developmental lateralisation in pitch processing, and outlines limitations and directions for future research.

Taken together, this dissertation demonstrates that lexical tones in Standard Chinese not only distinguish word meanings but also have an emotional effect due to their pitch characteristics. These tones contribute to affective meaning through mechanisms of pitch-based iconicity, which may serve adaptive communicative roles. In addition, it constitutes evidence that neural processing of pitch develops differently in infants depending on the linguistic relevance of pitch in their environment. This work integrates insights from affective iconicity and hemispheric lateralisation during perceptual reorganisation, highlighting how the brain processes pitch across linguistic, experiential, and cognitive contexts.

# Samenvatting in het Nederlands

Binnen de verschillende perceptuele kenmerken van spraak speelt toonhoogte een universele en veelzijdige rol. Door middel van toonhoogte kunnen grammaticale structuren onderscheiden worden (bijv. vragen versus mededelingen), kunnen emoties worden uitgedrukt en — essentieel in toontalen — kunnen lexicale betekenissen worden onderscheiden. Naast deze bekende functies kan toonhoogte ook betekenis overbrengen op iconische wijze, via perceptuele kenmerken als hoogte en contourrichting, en gedeelde sociolinguïstische conventies. Daarnaast hangt de verwerving van toonhoogte in spraak bij baby's niet alleen af van akoestische eigenschappen, maar ook van de linguïstische functie ervan. Toch blijven twee fundamentele vragen onbeantwoord: (1) Vertonen lexicale tonen in toontalen zoals het standaard Chinees ook affectieve iconiciteit, waardoor ze systematisch vorm geven aan interpretaties van emotionele opwinding en valentie door hun toonhoogte kenmerken? en (2) Hoe ontwikkelt de verwerking van toonhoogte in taal zich bij baby's die een taal met toonpatroon verwerven versus baby's die een taal met woordklemtoon leren, in het bijzonder met betrekking tot hemisferische lateralisatie tijdens perceptuele reorganisatie?

Dit proefschrift behandelt deze vragen in zes hoofdstukken.

Hoofdstuk 1 legt de conceptuele basis en bespreekt relevante literatuur over iconiciteit, affectieve iconiciteit en de neurale ontwikkeling van toonhoogteperceptie. Daarnaast worden methodologische benaderingen zoals gedragsbeoordelingstaken en neuro-imaging (fNIRS) geïntroduceerd, evenals de cross-linguïstische vergelijking tussen Nederlandse en Japanse baby's.

Hoofdstukken 2 tot en met 4 richten zich op affectieve iconiciteit. Door gebruik te maken van gedragstaken zoals Likert-schalen, Slider-schalen en gedwongen-keuze paradigma's, laat hoofdstuk 2 zien dat tonale kenmerken zoals toonhoogte, bereik, helling en contourrichting van invloed zijn op de beoordeling van emotionele opwinding die wordt opgeroepen in zowel bestaande als fictieve bisyllabische woorden. Effecten op valentie waren beperkter en traden alleen op bij fictieve woorden. Dit suggereert dat emotionele opwinding nauwer verbonden is met perceptuele eigenschappen van toonhoogte, terwijl valentie wordt gevormd door lexicale en cognitieve invloeden op een hoger niveau.

Hoofdstuk 3 breidt het onderzoek uit naar monosyllabische woorden. Hoewel stijgende en dalende tonen als opwindender werden beoordeeld dan vlakke tonen, waren deze verschillen hier minder uitgesproken dan bij bisyllabische sequenties. Opvallend was dat lexicale tonen emotionele reacties beter voorspelden dan medeklinkers of klinkers, vooral wanneer segmentele en suprasegmentele signalen tegenstrijdig waren, hetgeen de dominante rol van toonhoogte bij emotionele interpretatie benadrukt.

Hoofdstuk 4 verkent mogelijke mechanismen die deze effecten veroorzaken. Tonen in de eerste lettergreep bleken gekoppeld aan emotionele valentie, vergelijkbaar met patronen die waargenomen worden in Indo-Europese talen. Stijgende tonen in de eerste lettergreep kwamen vaker voor in woorden die als positief werden beoordeeld , terwijl dalende tonen vaker voorkwamen in woorden die als negatief werden beoordeeld. Opwinding werd beïnvloed door tonen in beide lettergrepen, waarbij dalende tonen vaker voorkwamen in woorden met een hoge graad van opwinding dan stijgende tonen of contourtonen. Deze bevindingen suggereren dat lexicale toon

adaptieve communicatieve functies kan hebben, affectieve iconiciteit, wat aansluit bij evolutionaire theorieën over iconiciteit.

Hoofdstuk 5 gebruikt functionele near-infrared spectroscopie (fNIRS) om de neurale respons op toonhoogteveranderingen in zowel talige als niet-talige stimuli te onderzoeken bij Nederlandse en Japanse baby's van 4 en 10 maanden oud. Bij 4 maanden oude baby's vertoonden beide groepen rechts-gelateraliseerde activatie bij niet-talige stimuli. Bij 10 maanden oude baby's vertoonden Japanse baby's links-gelateraliseerde activatie voor spraakstimuli, terwijl Nederlandse baby's meer bilaterale activatie lieten zien. Deze resultaten suggereren dat toonhoogte linguistisch verwerkt wordt bij baby's die een taal verwerven waarin toonhoogte lexicaal relevant is, terwijl deze specialisatie niet plaatsvindt bij baby's die een niet-tonale taal verwerven.

Hoofdstuk 6 vat de belangrijkste bevindingen samen, bespreekt implicaties voor affectieve iconiciteit en lateralisatie bij toonhoogteverwerking tijdens de taalontwikkeling, en benoemt beperkingen en toekomstige onderzoeksrichtingen.

Samenvattend toont dit proefschrift aan dat lexicale tonen in het standaard Chinees naast hun lexicale functie ook emotionele betekenissen uitdrukken door middel van hun toonhoogtekenmerken. Toon draagt bij aan de affectieve betekenis, via toonhoogte-gebaseerde mechanismen van iconiciteit, die mogelijk adaptieve communicatieve functies hebben. Het proefschrift laat ook zien dat de neurale verwerking van toonhoogte zich verschillend bij baby's ontwikkelt, afhankelijk van de taalspecifieke relevantie van toonhoogte in hun moedertaal. Dit werk integreert inzichten uit affectieve iconiciteit en hemisferische lateralisatie tijdens de ontwikkeling, en laat zien hoe het

brein toonhoogte verwerkt in linguïstische, ervarings- en cognitieve contexten

# 中文摘要

在言语感知的诸多线索中，音高具有普遍且多元的功能。一方面，音高能够区分语法功能（如疑问与陈述）、表达情绪，在声调语言中尤其关键的是区分词汇意义。除这些已知功能之外，音高还可以象似地传达含义，并受其感知特征（如高度和走向）以及社会语言习俗等影响。另一方面，婴儿对音高的习得不仅依赖于声学特性，也与其在语言中的功能息息相关。然而，两个基本问题仍待解决：（1）像汉语这样的声调语言的词汇声调，是否具有情感象似性，即其音高特征是否系统地塑造情绪唤醒和效价的解释？（2）在习得日语（音高重音）与荷兰语（重音）的婴儿中，音高加工如何发展，尤其在感知重组期间的大脑半球侧化表现如何？

本论文通过六个章节探讨上述问题。

第 1 章回顾语言中的象似性、情感象似性及音高感知神经发展的文献，介绍了包括行为评定任务与神经成像在内的研究方法，并概述荷兰和日本婴儿的跨语言对比。

第 2 至 4 章关注情感象似性。通过里克特量表、滑块量表和二选一范式的行为任务，第 2 章揭示声调的高度、范围、斜率和走向等特征显著影响了双音节词（包括真实词和虚构词）的情绪唤醒评分。效价方面的影响较有限，仅出现在听觉呈现的虚构词中。这表明唤醒度更紧密地与音高的感知属性相关，而效价则可能更多受到词汇和认知层面因素的影响。

第 3 章进一步考察了单音节词，尽管升调与降调比平调的唤醒度更高，但其区分程度不如双音节词明显。值得注意的是，当音段与超音段线索发生冲突时，声调比辅音或元音更能预测情绪反应，这凸显了声调在情绪解释中的主导地位。

第 4 章探讨了上述效应背后的机制。双音节词的第一音节声调与情绪效价相关，即第一音节升调更常见于表达积极情绪的词汇，而降调更常见于表达消极情绪的词汇，这与印欧语言中发现的音调音高与情绪的模式类似。情绪唤醒度受两个音节声调共同影响，高唤醒词汇中降调的比例更高。这提示词汇声调的情感象似性可能具有适应性沟通功能，与象似性的进化论观点一致。

第 5 章利用近红外技术考察荷兰和日本婴儿（4 及 10 个月大）对言语和非言语音高变化的神经反应。4 个月时两组婴儿对非言语音高表现为右半球偏侧化；10 个月时日本婴儿对言语表现为左半球偏侧化，而荷兰婴儿则未表现出偏侧化。这表明音高处理在习得音高相关语言的婴儿中逐渐语言特化，而在非声调语言婴儿中未出现此类特化。

第 6 章总结主要发现，讨论其对情感象似性及音高处理发展性侧化的启示，提出研究局限与未来方向。

综上所述，本论文证实汉语词汇声调不仅区分词义，还因其音高特性具备情感含义。这种象似机制可能具有适应性的沟通功能。同时，婴儿大脑对音高的神经加工也因语言环境中音高的不同功能而发展出不同模式。该研究整合了情感象似性和感知重组期间半球侧化的发现，展示了大脑如何在语言、经验与认知背景下调整对音高的加工方式。

# Curriculum Vitae

Tingting Zheng (郑婷婷) was born on 10 September 1992 in Shanxi, China. She completed both her bachelor's and master's degrees at Shaanxi Normal University, specializing in research on emotion and decision-making research. In 2018, she started her PhD at the Leiden University Centre for Linguistics, and this dissertation presents the findings of her doctoral research.