


ARTICLE

Cross-talker lexical tone discrimination in infancy

Ye Feng¹ , René Kager², Regine Lai³ and Patrick C. M. Wong^{3,4}

¹School of Linguistics and Language Resources, Beijing Language and Culture University, Beijing, China, ²Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, Netherlands, ³Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong SAR, China and ⁴Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China

Corresponding author: Patrick C.M. Wong; Email: p.wong@cuhk.edu.hk

(Received 14 December 2023; revised 16 March 2025; accepted 27 March 2025)

Abstract

This study investigated how infants deal with cross-talker variability in the perception of native lexical tones, paying specific attention to developmental changes and the role of task demands. Using the habituation-based visual fixation procedures, we tested Cantonese-learning infants of different age groups on their ability to discriminate Cantonese Tone 1 (high level) and Tone 3 (mid level) produced by either multiple talkers or a single talker. Results demonstrated that the 12-month-old and 24-month-old groups showed reliable discrimination across talkers, whereas the 18-month-old group did not (Experiment 1), despite their ability to discriminate the same contrast when the talker was held constant (Experiment 2). In a task that included a novel object as a referent to the sound, the 18-month-olds discriminated the contrast across talkers from Tone 1 to Tone 3 (Experiment 3). These results revealed a U-shaped developmental path and perceptual asymmetry in native lexical tone discrimination across talkers.

Keywords: lexical tone discrimination; talker variability; infancy

摘要

本研究探讨了婴儿如何处理母语词汇声调感知过程中的话者间变异性的问题，重点关注该能力的发展性变化趋势及任务难度的作用。通过基于习惯化范式的视觉注视程序，我们测试了不同年龄阶段的粤语母语婴儿在辨别粤语高平调(T1)与中平调(T3)方面的表现，考察了跨说话人与单一说话人条件下的差异。结果表明，12个月和24个月的婴儿在跨说话人条件下均表现出可靠的声调辨别能力，而18个月的婴儿未能表现出这种能力(实验1)，尽管他们在单一说话人条件下可以成功辨别相同的声调差异(实验2)。然而，当任务采用一个新异物体与语音材料配对时，18个月的婴儿在跨说话人条件下成功辨别了从T1到T3的声调差异(实验3)。这些结果揭示了婴儿在跨说话人辨别母语词汇声调时呈现U型发展趋势和感知上的不对称性。

1. Introduction

The discrimination and categorization of phonetically contrastive speech sounds are fundamental prerequisites for early language development (Kuhl, 1983; Werker & Yeung, 2005). Remarkably, infants have been found to discriminate native speech sound contrasts within the first few months of life (e.g., Chen & Kager, 2016; Eimas *et al.*, 1971; Harrison, 2000; Kalashnikova *et al.*, 2023; Novitskiy *et al.*, 2022; Shi *et al.*, 2017; Swoboda *et al.*, 1976; Yeung *et al.*, 2013). However, the apparent ease with which young language learners develop sensitivity to distinguish native sound categories may belie a complex reality: there is no one-to-one correspondence between the acoustic manifestations and the perceived linguistic categories in the language input (Liberman *et al.*, 1967). In naturalistic language learning settings, the same speech sound may display varying acoustic properties coming from different talkers due to physiological differences, social status, and cultural background, among other factors (Nusbaum & Magnuson, 1997). Even the same talker may exhibit variability when producing multiple examples of a given category (Newman *et al.*, 2001; Wang *et al.*, 2021; Wang & Wong, 2024). For adult listeners, accommodating such talker variability in speech perception involves extra processing costs as demonstrated by longer reaction time and/or lower accuracy rate when presented with stimuli coming from multiple talkers, compared with when there is only one talker (see Luthra, 2024, for a comprehensive review). The present study examines how infants in the early stages of native language acquisition deal with talker variability when perceiving native contrasts.

1.1. *The role of talker variability in phonetic discrimination and novel word learning*

Previous investigations have indicated distinct effects of talker variability in tasks primarily involving phonetic discrimination, typically administered to younger infants, in comparison with tasks demanding sound-meaning mapping, typically tested in infants aged 14 months and older. In phonetic discrimination tasks, variations in the speech stimuli seem to impose a certain degree of perceptual challenge on infants. For example, Kuhl (1979) examined the ability of six-month-old infants to discriminate two spectrally dissimilar vowels, /a/ and /i/, across synthesized male, female, and child voices using conditioned head-turn procedures. Infants trained in one voice and tested in different voices were able to transfer correct categorizations across voices, showing a certain degree of tolerance in phonetic categorization. However, when extending these findings to a less distinct vowel contrast, /a/ and /ɔ/, Kuhl (1983) observed that six-month-old infants succeeded in discriminating the less salient vowel pair only when talker variability was gradually introduced via five progressive stages. Without this gradual introduction, infants could only accurately discriminate vowels when the pitch contour of the stimuli remained identical across voices, indicating limitations in handling changes in an additional dimension beyond formant frequencies and voice identities. Similarly, Jusczyk *et al.* (1992) found that two-month-old infants could detect the consonant change from /baɡ/ to /daɡ/ produced by multiple talkers in a high-amplitude-sucking paradigm, but when a two-minute delay was inserted between the habituation and test phases, infants only detected the change in the single-talker condition. These observations align with results of a word recognition study, where 10.5-month-old infants, but not their 7.5-month-old counterparts, recognized familiarized words across talker gender (Houston & Jusczyk, 2000; but see Singh, 2018, for successful word recognition across talker gender in eight-month-olds with modified task procedures), despite the fact that 7.5-month-olds

were previously reported able to capitalize on phonetic details in word recognition when talker variability was not involved (Jusczyk & Aslin, 1995).

However, in a more recent study, Quam et al. (2021) used habituation-based visual fixation procedures to compare directly the discrimination performance of 7.5-month-old infants in single- versus multiple-talker conditions and found different results. Their findings showed that infants successfully discriminated the native contrast (/b/–/p/) but failed to discriminate the non-native contrast (/n/–/ŋ/), regardless of the presence or absence of multiple talkers. This aligns with the results of a prior study by Chen and Kager (2016), which investigated lexical tone discrimination in the context of intra-talker variability – a less-explored area. In their study, the authors tested whether Dutch infants could discriminate Mandarin Tone 2 (T2, rising tone) and Tone 3 (T3, dipping tone) with varying tokens. It was found that 6- and 12-month-old non-tone-learning (NTL) infants successfully discriminated the non-native tonal contrast, while four-month-olds failed, in the both presence and absence of intra-talker variability. Together, these findings suggest a limited impact of talker variability on speech sound discrimination.

The study by Quam et al. (2021) was, in part, motivated by a growing body of literature highlighting the beneficial role of talker variability in early word learning. Specifically, in the classical switch task, 14-month-old infants commonly encounter challenges when attempting to associate two minimally contrastive words (e.g., “bih”–“dih”) with distinct meanings (e.g., Stager & Werker, 1997; Werker et al., 1998). However, successful associative word learning has been observed when the stimuli were produced by multiple talkers (Höhle et al., 2020; Rost & McMurray, 2009, 2010) or when there was increased acoustic variability (Galle et al., 2015). It has been suggested that exposure to multiple talkers provides infants with the opportunity to utilize variability along the noncontrastive dimension, such as indexical cues (Apfelbaum & McMurray, 2011; Rost & McMurray, 2010), or to consider the relational properties among various cues within the target word (Höhle et al., 2020). This enables them to identify the phonemically relevant changes within the contrast more effectively.

1.2. The potential factors and theoretical underpinnings

To reconcile the disparities in the literature, two critical factors must be taken into consideration: developmental level and task demands. Notably, early sensitivity to phonetic details in discrimination and the ability to employ this sensitivity in sound-meaning mappings follow distinct developmental trajectories (Werker, 2018; Werker & Yeung, 2005). While infants have been shown to possess the capacity to discriminate native phonetic contrasts within the first six months of life, their ability to associate these contrasts with different word meanings typically emerges later, often beyond 14 months of age (e.g., Byers-Heinlein et al., 2013; Curtin et al., 2009; Stager & Werker, 1997; Werker et al., 1998). Given that infants participating in cross-talker discrimination tasks are usually much younger than those involved in cross-talker word-learning experiments, it prompts an intriguing question: As young language learners mature, do they experience different effects from talker variability?

The second factor that warrants consideration is task demands. It is important to distinguish between tasks that assess speech sound discrimination or word recognition, which do not require linking phonemically specified patterns to novel object meanings, and tasks focused on associative novel word learning, which specifically involve this linkage. For example, in discrimination tasks utilizing habituation-based visual fixation

procedures (e.g., Stager & Werker, 1997, Experiment 4), infants are habituated to repeated auditory stimuli while presented with non-referential visual stimuli, often a black-and-white checkerboard. When their looking time decreases below a predetermined habituation criterion, a novel auditory stimulus is introduced. Infants typically demonstrate a rebound in attention upon detecting this change. In contrast, word-learning tasks often involve presenting two novel objects paired with repetitions of two speech sounds, commonly known as the switch task (e.g., Stager & Werker, 1997, Experiment 1). In this setup, infants are encouraged to associate the sounds with specific objects. After habituation to the word–object pairings, a switched pairing is presented, involving familiar visual and auditory stimuli in a novel combination. Infants are expected to show longer looking time only when they detect a change in the pairing of sound and meaning. Therefore, it is plausible that the nature of tasks may have a discernible influence on how infants perceive and respond to talker variability in speech stimuli. Additionally, even the same task may impose different cognitive loads on different participants. In the classical study by Stager and Werker (1997; Experiments 2 and 3), the authors found that the single word–object pairing task was taken differently by infants of 8 and 14 months of age. For the 14-month-olds, the single word–object association task involved word learning, which may lead to a reduced sensitivity to phonetic differences such as that between “bih” and “dih.” For infants of eight months, this may be a simple sound discrimination task, allowing them to easily distinguish between “bih” and “dih.” Therefore, the cognitive load required by the task plays a crucial role in early speech perception.

The notion that developmental level and task demands may be the key factors influencing how infants respond to talker variability finds support in the Processing Rich Information from Multidimensional Interactive Representations (PRIMIR) framework proposed by Werker and Curtin (2005). PRIMIR introduces three distinct representational planes (general perceptual, word form, and phoneme) for information storage. During speech processing, the information attended to is modulated by three attentional filters: perceptual biases, task demands, and developmental level. Initial biases are crucial for initiating speech perception and linking it to language acquisition, although their significance may diminish with development. In contrast, the importance of task demands and developmental level increases over time, jointly influencing the prioritization of information access.

Therefore, in the context of speech perception across multiple talkers, infants’ attention to the phonetic details and indexical information in the speech signals is likely modulated by their age and the task nature. First, infants need to reach a developmental stage where they possess at least some level of phonological knowledge of their native language to tell apart phonemically relevant and irrelevant information. Second, in tasks focused on acoustic discrimination, all information is accessed; thus, talker variability may hinder the discrimination of the target contrast. Conversely, in tasks involving post-processing decisions about meaningful words, attention will prioritize phonological/categorical information over phonetic and indexical details.

Another theoretical framework relevant to the present hypothesis is the perceptual attunement account put forth by Best *et al.* (2009). This framework highlights a crucial developmental stage occurring between 15 and 19 months, which aligns with the period of vocabulary growth, typically around 18 months. During this phase, young learners undergo an attentional shift from lower-level phonetic patterns to higher-order phonological regularities in their native language. This shift in perceptual attunement allows infants to perceive the underlying phonological structure amidst various surface

realizations (e.g., talker accents), leading to the development of phonological constancy (Mulak & Best, 2013).

1.3. The present study

Building on the cited empirical evidence and theoretical motivations, this study aimed to investigate how infants deal with cross-talker variability in the perception of native speech contrasts with a specific attention on developmental changes and task demands. While extensive research has explored infants' ability to handle talker variability in segmental contrasts, relatively less is known about their processing of such variability in the perception of supra-segmental contrasts. This is particularly pertinent in tone languages, such as Mandarin and Cantonese, where pitch changes alter lexical meanings. The varying pitch ranges across speakers inevitably introduce complexity into the talker normalization process for tone language speakers. Therefore, investigating lexical tone contrasts offers a unique opportunity to unravel the impact of talker variability on early native speech sound perception.

Lexical tones can differ in pitch height (e.g., high versus low) and/or contour shape (e.g., rising versus falling). For contour tones, the word-internal pitch contour offers more information about the speakers' pitch range, which may make normalization easier. Moreover, as contour tones vary in both pitch height and contour shape across speakers, they offer more cues, further facilitating the normalization process. In contrast, level tones differ only in relative pitch height and are more susceptible to the influence of talker variability. Previous research with 14-, 18-, and 24-month-old Cantonese-learning infants indicated that successful mapping of two level tones to different word meanings occurred only when speaker-matched precursor phrases were provided as a frame of reference (Feng et al., 2022). Even adult listeners rely on contextual cues in preceding or following sentences to identify level tones across speakers (e.g., Francis et al., 2006; Wong & Diehl, 2003).

Therefore, to mitigate ceiling effects and fully capture the impact of talker variability on early speech sound discrimination, the present study utilized Cantonese Tone 1 (T1, high-level tone) and Tone 3 (T3, mid-level tone) as the target speech contrast. We included Cantonese-learning infants aged 12, 18, and 24 months in the discrimination tasks. This age range covers infants who have not yet established consistent competence in sound-object mapping tasks (typically under 14 months) and those who exhibit more reliable sound-object mapping abilities (typically 18 months and older).

In the series of experiments that follow, we first investigated the developmental differences in cross-talker lexical tone discrimination among the three age groups in Experiment 1. In Experiments 2 and 3, we focused on the 18-month-old group, who were arguably at the stage of attentional shift, and explored the potential effect of task demands by manipulating the auditory and visual stimuli in the visual fixation procedures. Recruitment of participants primarily took place through social media platforms such as WhatsApp and Facebook. All infants included in the study came from Cantonese-speaking monolingual families in Hong Kong. Prior to the experiment, all caregivers provided written informed consent in accordance with the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee under the project name The Neural Basis of Language and Cognitive Development (CREC no.: CRE-2015.410).

2. Experiment 1

In Experiment 1, infants at 12, 18, and 24 months of age were assessed for their ability to discriminate the native Cantonese T1–T3 contrast in the presence of talker variability. The experiment adapted the visual fixation procedures utilized by Stager and Werker (1997, Experiment 4) such that the auditory stimuli were produced by six rather than only one single talker.

2.1. Methods

2.1.1. Participants

Sixty-six monolingual Cantonese-learning infants were included in this experiment: 24 12-month-olds (mean age = 369 days; range = 340–399 days; 14 girls), 24 18-month-olds (mean age = 544 days; range = 510–581 days; 9 girls), and 18 24-month-olds (mean age = 712 days; range = 671–753 days; 8 girls). An additional 13 infants participated but were excluded from the analysis due to language background ($n = 5$), equipment failure ($n = 1$), experimenter error ($n = 3$), fussiness ($n = 1$), failure to reach the habituation criterion ($n = 1$), and failure to complete the task ($n = 2$). There were no reported cases of prior perceptual or neurological disorders among the participants.

2.1.2. Stimuli

Auditory stimuli. The speech stimuli consisted of a minimal pair of Cantonese non-words that only differed in terms of their lexical tones, that is the CV syllable /pi/ in Cantonese T1 (high-level) and the same segments in T3 (mid-level). The non-words were chosen to ensure that they were acceptable lexical forms in Cantonese while remaining unfamiliar to the infants. All recordings were conducted in a sound-attenuated booth. The speech stimuli were recorded using Adobe Audition, with a microphone connected to a MacBook Pro via a sound card (Roland Quad-Capture). A sampling rate of 44100 Hz and a sampling precision of 16 bits were employed. Six female native speakers of Cantonese were instructed to read the non-words in a lively, child-directed manner. From each speaker, three tokens of each tone were recorded, following the methodology of Rost and McMurray (2009). The voice onset time (VOT) of the consonant was manipulated to 80 ms for all tokens in order to avoid unnecessary variations, which aligns with the average VOT values of 77 ms observed in Cantonese speakers for the same consonant (Lisker & Abramson, 1964). Each token was then normalized to 500 ms in length and 70 dB in volume. Acoustic analysis was performed in Praat (Boersma & Weenink, 2020) using the “prosodypro” script developed by Xu (2013). The average acoustic measurements of the vowels are presented in Table 1, and the pitch contours of the tokens are demonstrated in Figure 1(A).

Table 1. Average acoustic measurements and standard deviations of the vowels of the stimuli in Experiments 1 and 3

Lexical tone	F0 (Hz)		F1 (Hz)		F2 (Hz)	
	M	SD	M	SD	M	SD
T1 (/pi1/)	260	27	335	57	2860	215
T3 (/pi3/)	217	17	355	42	2831	201

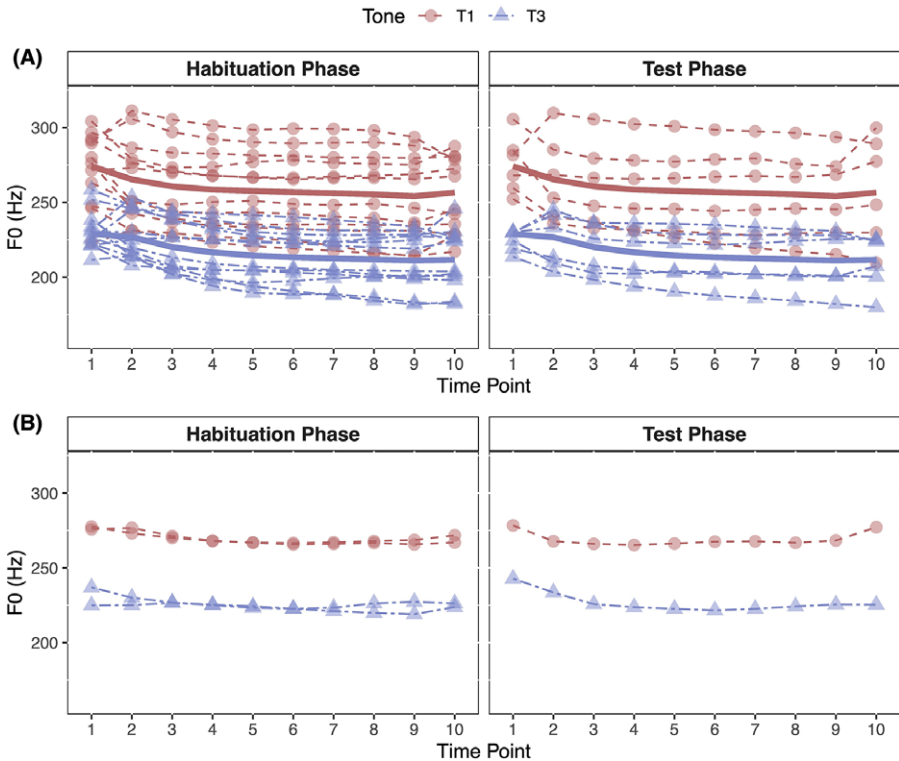


Figure 1. Pitch contours of the tokens used in Experiments 1 and 3 (panel A) and in Experiment 2 (panel B), separated by experimental phase. The dashed lines, marked with circles, depict the F0 contours of T1 tokens, while the dash-dotted lines with triangles illustrate the F0 contours of T3 tokens. The circles and triangles on these lines denote the 10 equidistant time points sampled along each tone contour. In (A), the thick solid lines indicate the mean F0 for all tokens of each tone (with T1 above T3). See the online article for the colour version of this figure.

In the habituation phase, stimuli strings were created by concatenating repetitions of two tokens for each lexical tone into 18-second strings at 1-second intervals, while the remaining tokens were used to form the stimuli string in the test phase. Additionally, a Cantonese non-word, /nu/, carrying T2 (high-rising) was used as both pre- and post-test stimuli. These tokens were recorded by a male native Cantonese speaker and were normalized to the same amplitude (70 dB) and length (500 ms).

Visual stimuli. During all experimental trials, a static black-and-white checkerboard was presented against a white background as the visual stimulus. This choice ensured that infants were unlikely to form associations between the visual stimulus and any specific speech stimulus. An identical checkerboard, set in motion alongside music, was employed as an attention-getter between trials. Figure 2 illustrates the presentation of stimuli during the habituation and test phases.

2.1.3. Apparatus and procedure

Following Quam et al. (2021), we utilized an adapted version of the visual fixation procedures. The experiment was conducted in a testing booth covered with white curtains

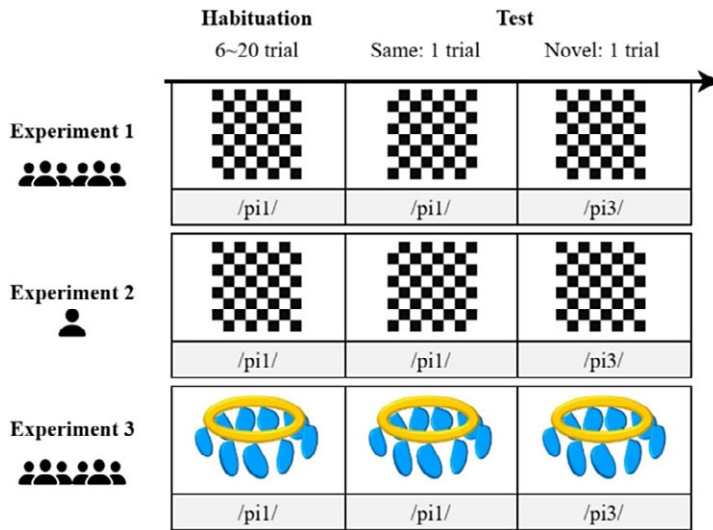


Figure 2. A demonstration of visual and auditory stimuli used during the habituation and test phases in Experiments 1 to 3. See the online article for the colour version of this figure.

to eliminate potential visual distractions for our young participants, leaving only the screen as the focal point. This screen was positioned at an approximate distance of 70 cm from the infants. Hidden behind this screen was a Bose SoundLink Color II speaker through which auditory stimuli were delivered. The experiment was administered by an experimenter located in an adjacent control room using Habit2 (Oakes et al., 2019) on a Macintosh computer. The infants' gaze behaviour was captured via a concealed camera mounted above the screen and transmitted in real time to the experimenter's computer. During the experiment, each infant was seated on the lap of their caregiver. The caregiver, wearing sound-proof headphones (Bose QC II) and listening to masking music, was instructed not to engage with the infant throughout the experiment unless necessary.

Before each trial, there was an attention-getter to make sure that the infant was directed towards the screen. Each trial was structured such that the static checkerboard picture was presented for 18 seconds, and the non-word was delivered 12 times within that time frame. Importantly, the duration of each trial was infant-controlled. The infant's looking behaviour was coded online via Habit2 by an experimenter who was blind to the trial type presented. A trial ended either when the infant averted their gaze for a continuous period of 2 seconds or when the looking time reached the predetermined maximum trial length of 18 seconds. If the infant's look-away time fell short of 2 seconds, the trial persisted. In cases where the infant's looking time within a trial was less than 1 second, the trial was repeated. In this way, infants' looking time to the visual stimuli during each trial was recorded and calculated as an indicator of their attention to the auditory stimuli.

The experiment commenced with a single-trial pretest, consisting of repeated presentations of a non-word /nu/ in Cantonese T2 (high rising tone) produced by a male native speaker of Cantonese. Following the pretest, the habituation phase ensued. In this phase, infants were exposed to repeated presentations of tokens of the non-word /pi/, each carrying either Cantonese T1 or T3 (counterbalanced across participants). These tokens

were produced by six female native speakers of Cantonese. This phase continued until infants reached the preset habituation criterion: a 50% decrease in the total looking time during three consecutive trials compared to the total looking time in the longest three habituation trials. Consequently, infants underwent a minimum of six trials and a maximum of 20 trials during the habituation phase.

The subsequent test phase consisted of two trials: a same trial and a novel trial. In the same trial, infants were exposed to the same tone (albeit different tokens) as that presented during habituation. Conversely, the novel trial introduced a change in lexical tone. Specifically, the tone changed from either T1 to T3 or T3 to T1, contingent upon the tone heard during the habituation phase. The order of the two trials was counterbalanced. It is noteworthy that in this paradigm, successful discrimination is indicated by an increase in looking time to the novel trial compared with that to the same trial (Chen & Kager, 2016). The experiment ended with a single-trial post-test, which was the same as the pretest. The inclusion of this post-test was aimed at ensuring the sustained engagement and attentiveness of the young participants throughout the task. This was vital in averting potential discrimination failures during the test phase caused by fatigue or a complete loss of attention, thereby reducing the risk of a type II error.

Statistical analysis. Linear mixed-effects (LME) models were used for all major analyses included in this study. Analyses were conducted using the lme4 package (Bates et al., 2014) in R (version 3.6.1) (R Core Team, 2016). P-values were computed using the lmerTest package (Kuznetsova et al., 2017), and pairwise comparisons were conducted using Tukey's honestly significant difference (HSD) test with the emmeans package (Lenth, 2023) where appropriate. The dependent variable for all LME models in this study was infants' looking time (in milliseconds), and the independent variables for Experiment 1 were Trial Type (same, novel), Age Group (12 months, 18 months, 24 months), and Habituation Condition (whether infants were habituated to T1 or T3). Models were fit using the maximum-likelihood estimation. The initial full model included Trial Type, Age Group, Habituation Condition, and all possible interactions as fixed effects, with a random intercept specified for Subject. The full model was then compared against reduced models with fixed effects removed one by one to assess whether the fixed effect in question was significant or not. Model fit was evaluated using likelihood ratio tests (LRTs), Akaike's information criterion (AIC), and Bayesian information criterion (BIC). The final model was selected based on parsimony, statistical significance of terms, and minimization of AIC/BIC values, prioritizing theoretical interpretability and simplicity.

2.2. Results

The accumulated amount of habituation time before testing did not show significant differences across age groups [12-month-olds: 75472.54 milliseconds; 18-month-olds: 75965.71 milliseconds; 24-month-olds: 83914.89 milliseconds; $F(2, 327) = 1.187$, $p = 0.306$]. As previously employed in related studies (e.g., Byers-Heinlein et al., 2013; Singh et al., 2016), a preliminary analysis was performed, comparing infants' looking time in the last habituation trial to that in the post-test trial to ensure infants had fully re-engaged in the task. An LME model with Trial Type as fixed effect and Subject as random effect confirmed that infants recovered to the post-test trial from habituation [$\Delta\chi^2(1) = 64.79$, $p < 0.001$]. Their looking time in the post-test trial significantly exceeded that in the last habituation trial ($\beta = 6630$, $SE = 665$, $t = 9.964$, $p < 0.001$).

The critical analyses aimed to assess infants' looking time in the same versus novel trials during the test phase. Mean looking time separated by age group is displayed in Figure 3. Model comparisons revealed that the model including Trial Type, Age Group, and the interaction of Trial Type \times Age Group as fixed effects with random intercepts specified for Subject fit best for the data [$\Delta\chi^2(1) = 10.92, p = 0.004$]. Neither the Trial Type \times Habituation Condition interaction [$\Delta\chi^2(1) = 1.06, p = 0.303$] nor the three-way interaction [$\Delta\chi^2(2) = 0.75, p = 0.686$] significantly affected model fit when removed.

The final model revealed a significant main effect of Trial Type [$\Delta\chi^2(1) = 15.78, p < 0.001$], with looking times in novel trials exceeding those in same trials by an estimated 1857 ms ($SE = 510$). Critically, the Trial Type \times Age Group interaction was significant [$\Delta\chi^2(2) = 10.66, p = 0.005$], indicating that developmental differences modulated infants' discrimination patterns (see Table S1 in the Supplementary material for the summary of the final model). The interaction was examined using the emmeans package (Lenth, 2023). Interestingly, as shown in Figure 3, both 12-month-old and 24-month-old groups demonstrated a significant effect of Trial Type (12 months: $\beta = 1857, SE = 522, t = 3.557, p < 0.001$; 24 months: $\beta = 2217, SE = 603, t = 3.678, p < 0.001$), that is, both groups looked noticeably longer to the novel trial than to the same trial. The effect sizes (Cohen's *d*) for novel versus same trial were 0.78 for the 12-month-old group and 0.81 for the 24-month-old group. In contrast, the 18-month-old infants did not show differences in the looking time to the same versus novel trials during the test phase ($\beta = 143, SE = 522, t = 0.274, p = 0.785$).

These results suggested that Cantonese-learning infants were able to discriminate Cantonese T1 and T3 across speakers at the age of 12 months, while failed to notice the

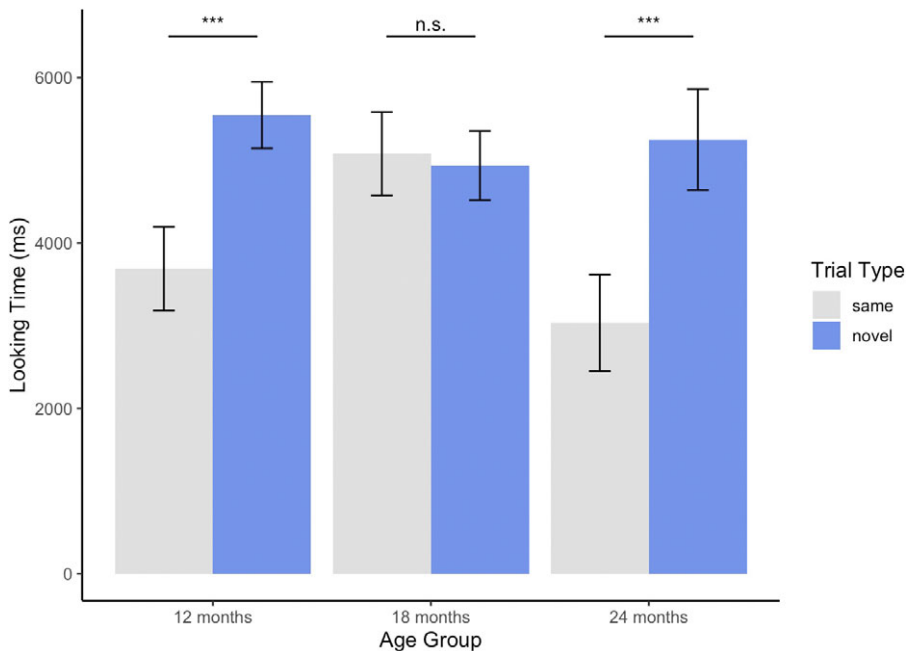


Figure 3. Fixation times to the visual stimulus for the same and novel trials in the test phase in Experiment 1 divided by age group (error bars: SEM). For non-significant results: “n.s.” or “not significant.” *** $p < 0.001$.

tonal contrast across speakers at the age of 18 months. When reaching the end of the second year, infants again successfully adapted to talker variability in the T1–T3 discrimination task.

Given that previous studies have highlighted the role of vocabulary size in early speech perception (e.g., Werker et al., 2002) and that 18 months is commonly associated with a vocabulary spurt (Nazzi & Bertoncini, 2003), it may provide insight to examine whether the vocabulary sizes of the infants in this experiment influenced their discrimination performance. Vocabulary size was measured using the Chinese Communicative Development Inventory – Cantonese version (CCDI-C) (Tardif & Fletcher, 2008), with scores obtained from parental reports on the Words and Sentence checklist. The raw CCDI-C scores for both vocabulary production and sentence complexity of the 18-month-olds were normalized against gender-specific norms provided in Tardif and Fletcher (2008) and obtained z-scores for boys and girls, respectively. Discrimination performance was represented by a discrimination value (DV) for each participant to normalize individual differences in looking time, following the method outlined by Dar et al. (2018). This calculation involved dividing the looking time in the novel test trial by the sum of the looking times in both the same and novel test trials. Specifically, a DV greater than 0.5 indicates a longer looking time to the novel trial, suggesting discrimination, while a DV equal to or below 0.5 suggests a failure to detect the tonal contrast across talkers.

A correlation analysis was conducted between the z-scores and the DV. Interestingly, there was a marginally significant negative correlation between the vocabulary production of the 18-month-olds and their discrimination value in Experiment 1 [$t(22) = -1.933$, $p = 0.066$, $r = -0.38$] (Figure 4). Due to the limited number of participants in our experiment, we were unable to achieve statistically robust results. Nevertheless, the data suggested a trend indicating that sensitivity to cross-talker speech sound discrimination may be influenced by vocabulary size. Specifically, children who produced more

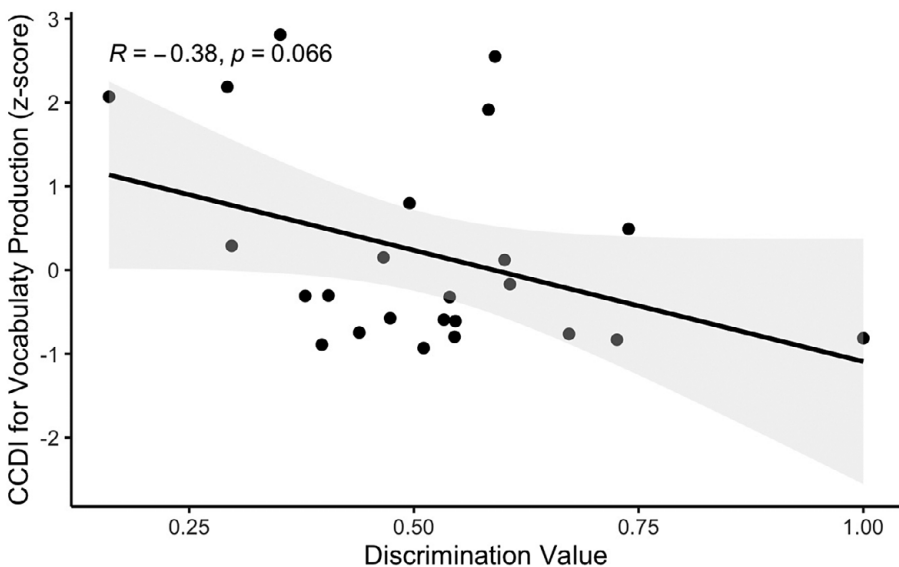


Figure 4. Correlation Between Vocabulary Production Scores and Discrimination Values for the 18-month-olds in Experiment 1.

words at 18 months showed a decreased likelihood of discriminating the T1–T3 contrast across multiple speakers.

3. Experiment 2

The findings from Experiment 1 unveiled an interesting U-shaped developmental trajectory in cross-talker lexical tone discrimination during the second year of life. In Experiment 2, our attention was turned to 18-month-olds with the objective of assessing their ability to discriminate the same tonal contrast under conditions devoid of talker variability.

3.1. Methods

Participants. Eighteen 18-month-old monolingual Cantonese-learning infants were included in this experiment (mean age = 555 days; range = 525–578 days; 6 girls). An additional two infants participated but were excluded from the analysis due to caregiver interference ($n = 1$) and failure to complete the task ($n = 1$). There were no reported cases of prior perceptual or neurological disorders among the participants.

Stimuli. The auditory stimuli in Experiment 2 consisted of the same monosyllabic non-word, /pi/, carrying either Cantonese T1 or T3 as employed in Experiment 1. The key distinction lay in the number of talkers. In Experiment 2, the stimuli were delivered by a single female native Cantonese speaker instead of the six speakers used in Experiment 1. For each tone (T1 and T3), three tokens were selected, resulting in a total of six tokens. Acoustic measurements are presented in Table 2, and Figure 1(B) displays the pitch contours employed in Experiment 2. During the habituation phase, stimuli strings were assembled by repeating two tokens for each lexical tone, creating 18-second sequences with 1-second intervals. The remaining token was utilized to form the stimuli strings for the test phase. Visual stimuli remained consistent with those used in Experiment 1, as depicted in Figure 2.

Apparatus and procedure. The apparatus and procedure were identical to those in Experiment 1. A visual representation of the stimulus presentation process during the habituation and test phases in this experiment is provided in Figure 2.

4. Results

Again, infants’ looking time in the last habituation trial and post-test trial were compared to ensure recovery of attention. An LME model with Trial Type as fixed effect and Subject as random effect confirmed that infants had recovered to the post-test [$\Delta\chi^2(1) = 18.25$,

Table 2. Average acoustic measurements and standard deviations of the vowels of the stimuli in Experiment 2

Lexical tone	F0 (Hz)		F1 (Hz)		F2 (Hz)	
	M	SD	M	SD	M	SD
T1 (/pi1/)	267	1	296	29	2866	81
T3 (/pi3/)	221	1	321	27	2813	116

$p < 0.001$], as their looking time in the post-test trial was significantly longer than that in the last habituation trial ($\beta = 6547$, $SE = 1230$, $t = 5.322$, $p < 0.001$).

Infant's looking time in the same versus novel trials during the test phase was compared as an indicator of discrimination. The mean looking times are presented in Figure 5. Model comparisons revealed that the model including only Trial Type as a fixed effect with random intercepts specified for Subject fit best for the data [$\Delta\chi^2(1) = 13.24$, $p < 0.001$] (see Table S2 in the Supplementary material for the summary of the final model). In the novel trial, infants exhibited significantly longer looking times than in the same trial ($\beta = 1941$, $SE = 439$, $t = 4.421$, $p < 0.001$), with a difference of $1940.9 \text{ ms} \pm 439.0$ (standard errors). The effect sizes (Cohen's d) for novel versus same trial were $d = 0.79$. There was no significant interaction observed for Trial Type \times Habituation Condition [$\Delta\chi^2(2) = 1.10$, $p = 0.578$].

Results from Experiment 2 indicate that 18-month-old infants successfully discriminated the T1–T3 contrast in the absence of inter-talker variability, thus ruling out the possibility that 18-month-olds' failure to discriminate the T1–T3 contrast across talkers in Experiment 1 was due to a general insensitivity to the contrast at this age.

5. Experiment 3

In Experiment 3, we continued our focus on the 18-month-old group and set out to explore the potential effect of task demands. Around 18 months of age, children typically

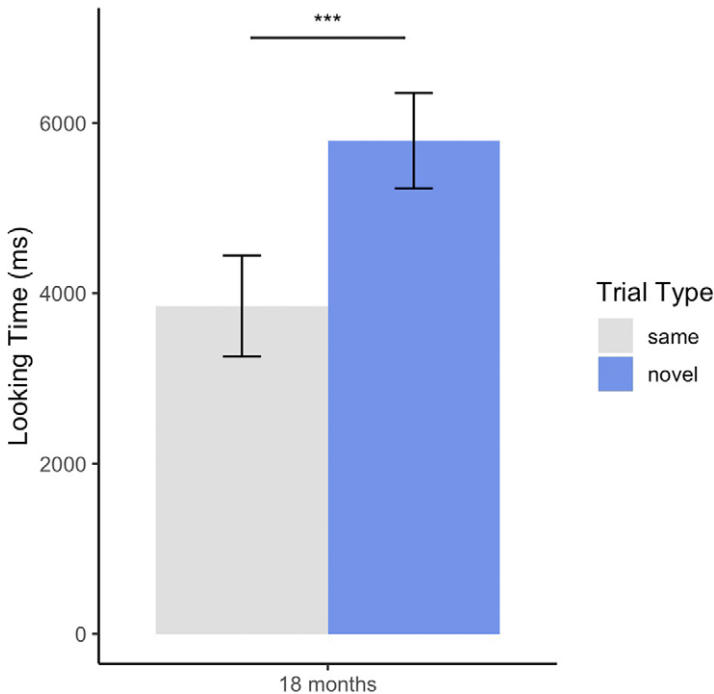


Figure 5. Fixation times to the visual stimulus for the same and novel trials in the test phase in Experiment 2 (error bars: SEM). For non-significant results: “n.s.” or “not significant.” *** $p < 0.001$.

experience a vocabulary spurt (Nazzi & Bertoncini, 2003). This period is also marked by the emergence of their ability to achieve phonological constancy across various regional accents (Best *et al.*, 2009; Mulak *et al.*, 2013; Potter & Saffran, 2017; White & Aslin, 2011). Therefore, it is possible that infants in this developmental stage are undergoing a shift in attention or learning mechanisms, which may influence how they process talker variability in different tasks. In Experiment 1, the speech stimuli from different speakers overlapped in pitch height across the two tonal categories, while the visual stimuli, consisting of a meaningless pattern, did not encourage any meaning association of the sound. This might have led the infants to allocate their focus on fine phonetic details and perceive the variations as occurring within a single category across talkers.

Experiment 3 was designed to investigate this possibility. In this experiment, instead of the checkerboard pattern, a picture of a novel object was presented synchronously with the lexical tones (Stager & Werker, 1997, Experiment 2). This single word–object pairing task would encourage the infants to process the speech stimuli as a label for the object, which was referred to as a simplified version of the word-learning task (*i.e.*, two word–object pairings) in some of the previous studies (*e.g.*, Singh *et al.*, 2016), and is theoretically more cognitively demanding than a pure sound discrimination task. This design allows for a better understanding of whether variations in task demands affect 18-month-olds' discrimination performance in the context of talker variability.

5.1. Methods

Participants. Twenty-four 18-month-old infants were included in Experiment 3 (mean age = 535 days; range = 513–569 days; 12 girls). An additional six infants participated but were excluded from the analysis due to language background ($n = 2$), fussiness ($n = 1$), failure to reach the habituation criterion ($n = 2$), and failure to complete the task ($n = 1$). Inclusion criterion was the same as the previous two experiments.

Stimuli. The auditory stimuli were identical to those in Experiment 1. The visual stimuli in the habituation and test phases were a novel object used in Singh *et al.* (2016) bouncing in the centre of the screen. A visual representation of the stimulus presentation process during the habituation and test phases in this experiment can be found in Figure 2.

Apparatus and procedure. The apparatus and procedure were identical to those in Experiments 1 and 2.

5.2. Results

An LME model with Trial Type as fixed effect and Subject as random effect confirmed that infants recovered to the post-test trial from habituation [$\Delta\chi^2(1) = 54.68$, $p < 0.001$]. Infants looked significantly longer to the post-test trial ($\beta = 9412$, $SE = 905$, $t = 10.399$, $p < 0.001$).

Model comparisons for infants' looking performance in the test phase revealed that the best fit model included Trial Type, Habituation Condition, and an interaction of Trial Type \times Habituation Condition as fixed effects with random intercepts specified for Subject (see Table S3 in the Supplementary material for the summary of the final model). Looking times to the same and novel trials in the test phase were plotted in Figure 6, and results were split by Habituation Condition in Figure 7 for an inspection of the interaction. Specifically, a main effect of Trial Type was observed [$\Delta\chi^2(1) = 9.51$, $p = 0.002$],

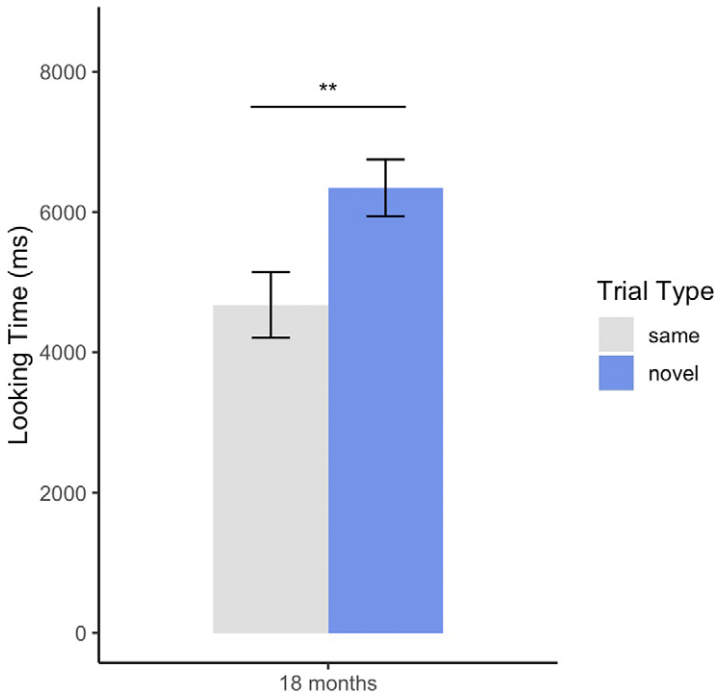


Figure 6. Mean fixation times to the visual stimulus for the same and novel trials in the test phase in Experiment 3 (error bars: SEM). ** $p < 0.01$.

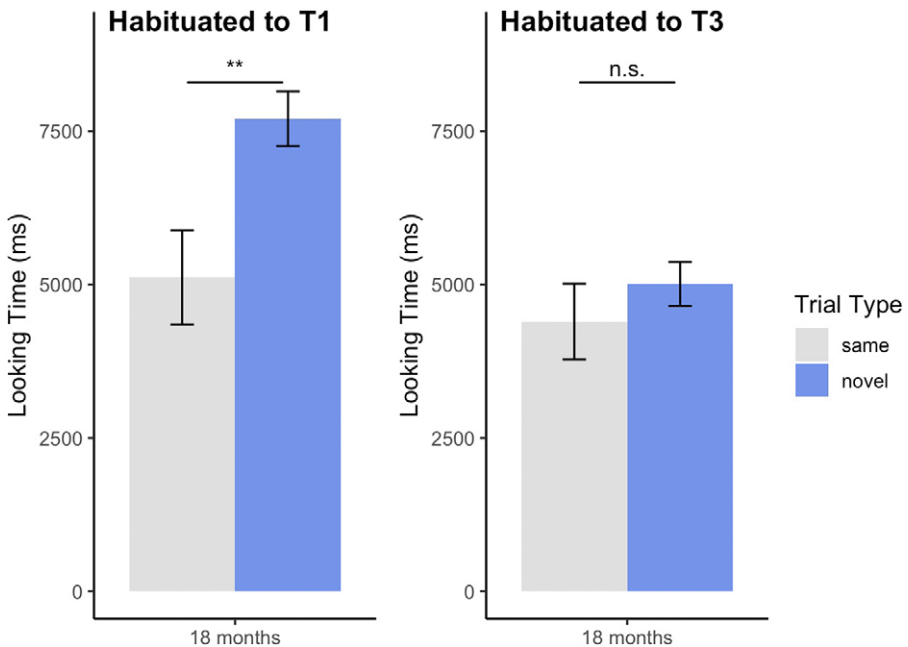


Figure 7. Mean fixation times to the visual stimulus for the same and novel trials in Experiment 3 split by Habituation Condition (error bars: SEM). ** $p < 0.01$.

Table 3. Results from cross-experiment comparisons on SES, CCDI-C, and habituation measures of the 18-month-olds

Exp.	SES (SD)	CCDI-C (SD)	No. of trials to habituation (SD)	Total habituation time (SD)
1 (N = 24)	50.46 (10.86)	106.67 (130.34)	8.88 (3.34)	75965.71 (41482.72)
2 (N = 18)	52.28 (6.32)	128.51 (76.52)	9.61 (3.33)	84728.50 (49016.41)
3 (N = 24)	47.19 (9.31)	104.75 (97.65)	9.71 (3.17)	107208.54 (44974.09)
F (df = 2)	1.667	0.304	0.435	2.952
Pr(>F)	0.197	0.739	0.649	0.059

with looking time in the novel trial longer than that in the same trial ($\beta = 2588$, $SE = 692$, $t = 3.743$, $p = 0.001$, Cohen's $d = 0.68$). There was also a significant interaction between Trial Type and Habituation Condition [$\Delta\chi^2(1) = 4.09$, $p = 0.043$]. Post hoc analyses showed that only those who were habituated to T1 successfully noticed the tonal contrast across speakers ($\beta = 2588$, $SE = 692$, $t = 3.743$, $p = 0.001$), while those who were exposed to T3 during habituation phase did not exhibit significantly longer looking time to the novel trial in test ($\beta = 610$, $SE = 692$, $t = 0.882$, $p = 0.387$).

Infants' looking performance did not correlate with their vocabulary sizes ($p > 0.1$) in this experiment. The significant main effect of Trial Type indicated that the discrimination values (DVs) were less dispersed than those in Experiment 1 and exhibited a tendency to approach the 0.5 threshold, which likely accounted for the lack of a correlation effect.

In addition, the 18-month-old infants in Experiments 1, 2, and 3 were compared with respect to their family socioeconomic status (SES), vocabulary size, and habituation measures. To determine SES, participants' scores were computed by coding parental educational levels and occupational prestige using the Hollingshead Index (Hollingshead, 1975). Habituation measures included the number of trials participants experienced to reach habituation and their total habituation time. As illustrated in Table 3, there were no significant differences in these factors among the 18-month-olds across experiments. However, Experiment 3 did exhibit slightly longer habituation times, which was likely attributable to the use of more engaging visual stimuli.

The results from Experiment 3 highlight the crucial influence of task demands on infants' ability to handle cross-talker variability in native lexical tones. Furthermore, an interesting observation emerged in the interaction between Trial Type and Habituation Condition, revealing a perceptual asymmetry. Infants habituated to T1 extended their discrimination abilities across talkers, while those habituated to T3 did not. This finding suggests that the effectiveness of cross-talker discrimination is contingent on the specific tonal properties to which infants are habituated.

6. General discussion

This study aims to investigate how infants manage the challenges posed by cross-talker variability in the perception of native lexical tones, with a specific focus on developmental changes and the influence of task demands. Utilizing habituation-based visual fixation procedures, we tested Cantonese-learning infants on their abilities to distinguish between

Cantonese T1–T3 contrast, presented by either multiple talkers (Experiments 1 and 3) or a single talker (Experiment 2).

Results from Experiment 1 indicated that infants could effectively accommodate talker differences and discriminate the tonal categories at 12 months of age. Intriguingly, this sensitivity appeared to diminish at 18 months, only to be regained by the end of the second year, revealing a U-shaped developmental trajectory. Experiment 2 eliminated inter-talker variability from the procedures and confirmed that 18-month-olds could discriminate the T1–T3 contrast when the talker remained constant, which is in line with previous studies on lexical tone discrimination (e.g., Harrison, 2000; Shi et al., 2017; Yeung et al., 2013). This ruled out the possibility of a general insensitivity to the tonal contrast at this age and confirmed the role of developmental level observed in cross-talker lexical tone perception in Experiment 1. Experiment 3 introduced a manipulation of task demands by concurrently presenting a novel object with the auditory stimuli, making it a more cognitively demanding single word–object pairing task. Notably, 18-month-olds who were habituated to T1 during habituation succeeded in cross-talker discrimination in this task, highlighting the influence of task-specific processing strategies. However, it appeared that the subgroup habituated to T3 did not show reliable discrimination. To simply put, infants reliably noticed the change from T1 to T3 across talkers, but not in the reversed order, revealing a perceptual asymmetry.

The U-shaped trajectory. The present U-shaped developmental trajectory in cross-talker lexical tone discrimination coincides, to some extent, with the U-shaped trajectory of tonal reorganization (e.g., Götz et al., 2018; Liu & Kager, 2014), although the turning point in our study emerges later. The key differences lie in the language background of the infants, that is tone learning (TL) versus NTL and whether talker variability is involved in the discrimination.

The U-shaped trajectory of tonal reorganization is built upon the fact that NTL infants gradually lose the sensitivity to tonal contrasts between 6 and 9 months as they are tuned to their native languages which do not use lexical tones to convey word meanings (Mattock et al., 2008; Mattock & Burnham, 2006; Yeung et al., 2013). However, NTL infants continue to grapple with intonation in their native languages, which share the same acoustic cues, that is pitch, with lexical tones. Evidence has shown that discrimination between Mandarin T1 and T4 with shrunken F0 contours is evident in infants at 5–6 months and 17–18 months but not at 8–9, 11–12, or 14–15 months (Liu & Kager, 2014). Similarly, German-learning infants demonstrate the ability to discriminate Cantonese T2 and T3 at 6 and 18 months but not at 9 months (Götz et al., 2018). It appears that NTL infants lose sensitivity to lexical tones after 6 months but regain it around 18 months, which may result from the acquisition of intonation or cognitive maturation.

In contrast, this study assessed TL infants on their capacity to accommodate talker variability in discriminating native tonal contrasts. The finding of the U-shaped developmental trajectory in this study provides support for the PRIMIR framework, suggesting that developmental level and task demands jointly influence the prioritization of information access. In the context of this study, the two turning points in perception between 12 and 18 months (decline) and between 18 and 24 months (recovery) indicated that the three age groups may have approached the task differently.

The success observed in 12-month-olds was not unexpected, given the previously reported proficiency in cross-talker discrimination of native vowels at 6 months (Kuhl, 1979, 1983) and of native consonants at 2 months (Jusczyk et al., 1992) and 7.5 months (Quam et al., 2021). Moreover, 6- and 12-month-olds have demonstrated the ability to discriminate lexical tones in the presence of intra-talker variability (Chen & Kager, 2016).

It is crucial to note that discrimination tasks at these early ages, even across multiple talkers or tokens, may not necessarily imply the emergence of abstract phonological categories. Tasks utilizing a checkerboard as visual stimulus may not encourage sound-meaning mapping, as there is no inherent association implied between the auditory and visual stimuli. Consequently, discrimination at these early stages could be grounded purely in acoustic differences. As infants mature, their sensitivity to native speech sound categories increases, and they become more readily available for tasks that require more abstract representations. Both factors influence their performance in speech perception tasks.

The most classic example of this is the study by Stager and Werker (1997), where eight-month-olds succeeded in a single word-object pairing task, while 14-month-olds failed. The authors posited that for infants of 14 months, the single word-object association task involved word learning, which may lead to a reduced sensitivity to phonetic differences such as that between “bih” and “dih.” For infants of eight months, however, this may be a simple sound-discrimination task, allowing them to easily distinguish between “bih” and “dih.” These findings indicated that infants who approach the task as a pure discrimination task will focus more closely on phonetic details in the stimuli, whereas those who perceive it as a word-learning task may pay less attention to these details. This aligns with the concept of “acquired equivalence” (Miller & Dollard, 1941), which posits that cues associated with the same event become less discriminable. In the context of Stager and Werker’s task, the object association may have reinforced the acquired equivalence and encouraged the infants to focus less on phonetic distinctions.

Therefore, although our findings may initially seem to diverge from those reported by Stager and Werker (1997), a closer examination reveals that the differences are not as pronounced as they might appear. It is noteworthy that our study involved complex within-category variations not considered in their study or most of the previous studies for that matter. In our experiments, tonal contrasts (Cantonese T1–T3, both level tones) were produced by multiple speakers with varying pitch ranges, resulting in both within-category and cross-category variations based on pitch height. This complexity means that greater attention to phonetic details could make the task more challenging for children who are on the cusp of word learning. The 18-month-olds may have focused more on within-category variations, which may hinder their ability to effectively discriminate between categories. Conversely, as indicated in the findings of our Experiment 3, tasks that require word-object associations appear to facilitate easier discrimination of these tonal categories. In such cases, the task reinforces acquired equivalence and encourages children to discount within-category variations, allowing them to attend to between-category cues and abstract representations, which is consistent with our findings.

The difference between the two older groups aligns with research indicating that phonological development extends well into childhood, particularly regarding Cantonese lexical tone acquisition, which concludes late in terms of both perception (Ciocca & Lui, 2003) and production (Mok *et al.*, 2020; Wong & Leung, 2018). Infants around 24 months may have developed better lexical tone representations and are more capable of talker adaptations in their language environment than the 18-month-olds and thus less influenced by the within-category variations in the discrimination task.

The joint influence of developmental level and task demands on cross-talker lexical tone perceptual is also demonstrated in the potential effect of vocabulary sizes. The marginally significant negative correlation between the vocabulary production scores of the 18-month-olds and their discrimination value in our Experiment 1 suggested a trend wherein children who produced more words at 18 months were less likely to discriminate

the T1–T3 contrast across multiple speakers. While previous studies that did not involve talker variability (e.g., Werker et al., 2002) have shown that children with larger vocabulary sizes may tend to pay more attention to phonetic details, leading to more success in discrimination and word-learning tasks, this study specifically investigates the impact of talker variability. Thus, the increased attention children devote to phonetic details may have made the discrimination task more taxing due to extensive within-category variations.

Another, but not necessarily alternative, explanation is that by the age of 18 months, infants may begin to realize the presence of “noise” in their language input, which may lead to a temporary over-generalization of the stimuli or an over-adaptation to the variations. Converging evidence from cross-accent studies has shown that 19-month-old infants, but not those younger (e.g., 15 months), demonstrated the ability to adapt to a new accent for familiarized words (Best et al., 2009; Mulak et al., 2013; Potter & Saffran, 2017; van Heugten & Johnson, 2014; White & Aslin, 2011). This suggests that by the age of one and a half years, infants start to recognize phonetic variations introduced by talker differences in speech signals. Given that the tonal contrast used in the current study partly overlaps across speakers with varying pitch ranges, it is possible that the 18-month-olds may have over-adapted to the speaker variations in the stimuli such that the tokens from a different category were mistaken as exemplars from the same category. It was not until the age of 24 months, or when cognitive demands of the task shifted their attention away from phonetic details (as in our Experiment 3), that young learners were able to discern the distribution of variability within each tonal category.

The perceptual asymmetry. Previous studies have reported the directional effects of presenting lexical tone stimuli on discriminating tonal contrasts (Tsao, 2008; Yeung et al., 2013). The current study stands as one of the first to report perceptual asymmetry in native lexical tone discrimination across talkers in infancy.

Acoustic salience has been discussed in previous studies to interpret the asymmetry. Specifically, Tsao (2008) tested Mandarin-learning 12-month-olds on their discrimination of Mandarin lexical tones using conditioned head-turn procedures. They found that Mandarin T1 (level tone) to T3 (dipping tone) is easier for infants to discriminate than the reverse direction. The authors attributed this to general auditory perception, proposing that syllables with a level F0 contour may be more difficult to detect from the ones with varied F0 contours compared with the reverse. Contrastingly, using head-turn preference procedures, Yeung et al. (2013) found that Cantonese-learning infants at both 4 months and 9 months showed discrimination when familiarized with Cantonese T2 (high-rising tone), but not when familiarized with Cantonese T3 (mid-level tone). This suggests that the asymmetry observed may not be solely attributed to the acoustic properties of pitch contours. Moreover, in the present study, both tones used are level tones.

An alternative account proposed by Yeung et al. (2013) suggests that the Cantonese high-rising tone is situated more towards the periphery of the tone space compared with the mid-level tone, which may offer more information about the pitch range, thereby aiding in discrimination. This explanation is applicable to the asymmetry observed in the present study. While Cantonese T1 (high-level tone) used in the current experiments may not convey information about the talkers’ entire pitch range, it is indeed the tone with the highest F0 among all tones and is positioned more towards the periphery of the talkers’ tone space compared to the mid-level tone. Additionally, since the present study involved talker variability in the discrimination task, listeners had to normalize among talkers, in which case the high-level tone would serve as a better anchor than the mid-level tone. Interestingly, asymmetries in the perception of categories are common in vowel perception, where they also seem to point to the primacy of peripheral vowels. It has been

proposed that a vowel with extreme articulatory–acoustic properties (peripheral in the vowel space) within a contrast serves as a reference or perceptual anchor and plays an important role in early language development (Polka & Bohn, 2003, 2011). Although the differences in experimental paradigms and research focuses limit direct comparison, our findings align with this peripherality hypothesis, suggesting that infants may rely on the lexical tone that is more towards the periphery of the talkers' tone space as an anchor point to resolve talker variability in lexical tone discrimination.

Another plausible explanation for the perceptual asymmetry may be attributed to the unbalanced representational strength of lexical tones during the early stages of language acquisition. Previous research has indicated a higher prevalence of T1 words compared to T3 words in Cantonese (Leung *et al.*, 2004). Based on the acquisition patterns of Cantonese-speaking children, T1 is acquired before the other tones and T3 may be the last among the three level tones (Tse, 1978; Wong & Leung, 2018). Therefore, with respect to our data, it is plausible that the 18-month-olds may already possess a more stable representation of T1 and are more familiar with the distribution pattern of this tone in their linguistic environment. This could enhance their sensitivity to detecting a shift from the well-established T1 category to the less familiar T3 category during habituation. In contrast, habituation to T3 – a tone with weaker representational grounding – may elicit weaker dishabituation to T1, as the latter is already a highly familiar category.

In addition, it is noteworthy that the tokens of T1 exhibit more variation than the tokens of T3, as evident in the pitch range (see Figure 1). The observed pitch range difference across talkers between these two lexical tones may contribute to the interpretation of perceptual asymmetry. Notably, subtle differences in a narrow pitch range have been proposed as one of the underlying causes why tonal pairs like the Cantonese T3 (mid level) and T6 (low level) are later acquired in children (Ciocca & Lui, 2003) and are merging in adults (Mok *et al.*, 2013).

The present study is not without limitations. First, as shown in Table 3, the 18-month-olds in our Experiment 3 were more attentive to the experimental procedure than those in Experiment 1. Admittedly, the visual stimulus in Experiment 3 was more engaging than that in the first two experiments, potentially contributing to the longer time taken by infants in Experiment 3 to reach the habituation criterion. Longer habituation time may have facilitated talker normalization, allowing exposure to a greater number of tokens from the same lexical tone category and potentially aiding in detecting tonal differences. Future research is needed to further explore whether increased habituation time contributes to the observed effects, potentially by manipulating the visual stimuli or adjusting the habituation criterion. Second, the present study only employed six speakers of the same gender in the multiple-talker experiments. For the 18-month-old group, facilitation might occur with larger talker variations, as reported in the word-learning studies that utilized 18 male and female talkers (Rost & McMurray, 2009, 2010). Third, future studies could employ neurophysiological measures (e.g., EEG) or more sensitive behavioural paradigms to disentangle whether 18-month-olds' difficulty in discriminating tones across talkers stems from perceptual insensitivity to talker-invariant tonal features or increased cognitive load in processing variability. Additionally, the potential influence of vocabulary size on cross-talker lexical tone discrimination among 18-month-old children revealed in our study warrants further investigation with larger sample sizes or longitudinal designs. Lastly, how the current findings might extend to segmental contrasts or tonal contrasts involving pitch contour changes remains a topic for further investigation.

In summary, this study is among the first to unveil a U-shaped developmental trajectory in native lexical tone discrimination across talkers during the second year of

life, indicating a joint influence of developmental stage and task demands. Additionally, a perceptual asymmetry was found, providing additional evidence that infants may go through certain developmental transitions around 18 months of age in terms of how they perceive and adapt to talker variability in lexical tone discrimination. These findings contribute to discussions on the early development of phonological constancy.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/S0305000925000212>.

Data availability statement. Our numeric data are available at *Open Science Framework* (<https://osf.io/5nhuy/>).

Acknowledgments. This work was supported by grants from the Humanities and Social Sciences Youth Foundation, the Ministry of Education of the People's Republic of China (23YJC740011), the University Grants Committee (HKSAR) (RGC34000118), the Innovation and Technology Fund (HKSAR) (ITS/067/18), Dr. Stanley Ho Medical Development Foundation, and the Global Parent Child Resource Centre Limited. Part of this study was completed as part of the first author's doctoral dissertation. We thank the Chinese University of Hong Kong (CUHK) – Utrecht University (UU) Joint Centre for Language, Mind and Brain, and CUHK – NTU – WSU Joint Laboratory for Infant Research. We are also grateful to all our infant participants and their caregivers for their invaluable contributions to the study.

References

- Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, 35(6), 1105–1138. <https://doi.org/10.1111/j.1551-6709.2011.01181.x>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv:1406.5823*.
- Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. (2009). Development of phonological constancy: Toddlers' perception of native-and Jamaican-accented words. *Psychological Science*, 20(5), 539–542. <https://doi.org/10.1111/j.1467-9280.2009.02327.x>
- Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer* [Computer program]. <http://www.praat.org/>. <https://cir.nii.ac.jp/crid/1571417124269710464>
- Byers-Heinlein, K., Fennell, C. T., & Werker, J. F. (2013). The development of associative word learning in monolingual and bilingual infants. *Bilingualism: language and cognition*, 16(1), 198–205. <https://doi.org/10.1017/S1366728912000417>
- 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>
- Ciocca, V., & Lui, J. (2003). The development of the perception of Cantonese lexical tones. *Journal of Multilingual Communication Disorders*, 1(2), 141–147. <https://doi.org/10.1080/1476967031000090971>
- Curtin, S., Fennell, C., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word–object associative learning. *Developmental Science*, 12(5), 725–731. <https://doi.org/10.1111/j.1467-7687.2009.00814.x>
- Dar, M., Keren-Portnoy, T., & Vihman, M. (2018). An order effect in English infants' discrimination of an Urdu affricate contrast. *Journal of Phonetics*, 67, 49–64. <https://doi.org/10.1016/j.wocn.2017.12.002>
- Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968), 303–306. <https://doi.org/10.1126/science.171.3968.303>
- Feng, Y., Kager, R., Lai, R., & Wong, P. C. M. (2022). The ability to use contextual cues to achieve phonological constancy emerges by 14 months. *Developmental Psychology*, 58(11), 2064–2080. <https://doi.org/10.1037/dev0001418>
- Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America*, 119(3), 1712–1726. <https://doi.org/10.1121/1.2149768>
- Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The role of single talker acoustic variation in early word learning. *Language Learning and Development*, 11(1), 66–79. <https://doi.org/10.1080/15475441.2014.895249>

- Götz, A., Yeung, H. H., Krasotkina, A., Schwarzer, G., & Höhle, B. (2018). Perceptual reorganization of lexical tones: Effects of age and experimental procedure. *Frontiers in Psychology*, *9*, 477. <https://doi.org/10.3389/fpsyg.2018.00477>
- Harrison, P. (2000). Acquiring the phonology of lexical tone in infancy. *Lingua*, *110*(8), 581–616. [https://doi.org/10.1016/S0024-3841\(00\)00003-6](https://doi.org/10.1016/S0024-3841(00)00003-6)
- Höhle, B., Fritzsche, T., Meß, K., Philipp, M., & Gafos, A. (2020). Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning. *Developmental Science*, *23*(5), e12950. <https://doi.org/10.1111/desc.12950>
- Hollingshead, A. B. (1975). *Four factor index of social status*. Yale University [Unpublished manuscript].
- Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570. <https://doi.org/10.1037/0096-1523.26.5.1570>
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23. <https://doi.org/10.1006/cogp.1995.1010>
- Jusczyk, P. W., Pisoni, D. B., & Mullenix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, *43*(3), 253–291. [https://doi.org/10.1016/0010-0277\(92\)90014-9](https://doi.org/10.1016/0010-0277(92)90014-9)
- Kalashnikova, M., Singh, L., Burnham, R., Cannistraci, R., Chen, H., Ng, B. C., Dos Santos, M., Dwyer, A., Feng, Y., Gisvold, A. K., Gustavsson, L., Hui, O. S., Hay, J., Kager, R., de Klerk, M., Lai, R., Liu, L., Marklund, E., Nazzi, T., Schwarz, I.-C., Tsao, F.-M., Wong, P. C. M., & Woo, P.-J. (2023). The Development of tone categories in infancy: Evidence from a cross-linguistic, multi-lab report. *Developmental Science*, e13459. <https://doi.org/10.1111/desc.13459>
- Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, *66*(6), 1668–1679. <https://doi.org/10.1121/1.383639>
- Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*(2–3), 263–285. [https://doi.org/10.1016/S0163-6383\(83\)80036-8](https://doi.org/10.1016/S0163-6383(83)80036-8)
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). <https://doi.org/10.18637/jss.v082.i13>
- Lenth, R. (2023). *Emmeans: Estimated marginal means, aka least-squares means (R package version 1.8.5.) [Computer software]*. <https://CRAN.R-project.org/package=emmeans>.
- Leung, M.-T., Law, S.-P., & Fung, S.-Y. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 500–505. <https://doi.org/10.3758/BF03195596>
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431. <https://doi.org/10.1037/h0020279>
- Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422. <https://doi.org/10.1080/00437956.1964.11659830>
- 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>
- Luthra, S. (2024). Why are listeners hindered by talker variability? *Psychonomic Bulletin & Review* *31*(1), 104–121. <https://doi.org/10.3758/s13423-023-02355-6>
- 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>
- Miller, N. E., & Dollard, J. (1941). *Social learning and imitation*. Yale University Press.
- Mok, P. P. K., Li, V. G., & Fung, H. S. H. (2020). Development of phonetic contrasts in Cantonese tone acquisition. *Journal of Speech, Language, and Hearing Research*, *63*(1), 95–108. https://doi.org/10.1044/2019_JSLHR-19-00152
- Mok, P. P. K., Zuo, D., & Wong, P. W. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language variation and change*, *25*(3), 341–370. <https://doi.org/10.1017/S0954394513000161>

- Mulak, K. E., & Best, C. T. (2013). Development of word recognition across speakers and accents. In *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence* (pp. 242–269). IGI Global.
- Mulak, K. E., Best, C. T., Tyler, M. D., Kitamura, C., & Irwin, J. R. (2013). Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words in a non-native regional accent. *Child Development*, 84(6), 2064–2078. <https://doi.org/10.1111/cdev.12087>
- Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science*, 6(2), 136–142. <https://doi.org/10.1111/1467-7687.00263>
- Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, 109(3), 1181–1196. <https://doi.org/10.1121/1.1348009>
- Novitskiy, N., Maggu, A. R., Lai, C. M., Chan, P. H. Y., Wong, K. H. Y., Lam, H. S., Leung, T. Y., Leung, T. F., & Wong, P. C. M. (2022). Early development of neural speech encoding depends on age but not native language status: Evidence from lexical tone. *Neurobiology of Language*, 3(1), 67–86. https://doi.org/10.1162/nol_a_00049
- Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. *Talker Variability in Speech Processing*, 109–132.
- 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, 1943–1952. <https://doi.org/10.3758/s13428-019-01244-y>
- Polka, L., & Bohn, O.-S. (2003). Asymmetries in vowel perception. *Speech Communication*, 41(1), 221–231. [https://doi.org/10.1016/S0167-6393\(02\)00105-X](https://doi.org/10.1016/S0167-6393(02)00105-X)
- Polka, L., & Bohn, O.-S. (2011). Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics*, 39(4), 467–478. <https://doi.org/10.1016/j.wocn.2010.08.007>
- Potter, C. E., & Saffran, J. R. (2017). Exposure to multiple accents supports infants' understanding of novel accents. *Cognition*, 166, 67–72. <https://doi.org/10.1016/j.cognition.2017.05.031>
- Quam, C., Clough, L., Knight, S., & Gerken, L. (2021). Infants' discrimination of consonant contrasts in the presence and absence of talker variability. *Infancy*, 26(1), 84–103. <https://doi.org/10.1111/inf.12371>
- R Core Team. (2016). *R: The R project for statistical computing*. <https://www.r-project.org/>
- Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12(2), 339–349. <https://doi.org/10.1111/j.1467-7687.2008.00786.x>
- Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, 15(6), 608–635. <https://doi.org/10.1111/j.1532-7078.2010.00033.x>
- Shi, R., Gao, J., Achim, A., & Li, A. (2017). Perception and representation of lexical tones in native Mandarin-learning infants and toddlers. *Frontiers in Psychology*, 8, 1117. <https://doi.org/10.3389/fpsyg.2017.01117>
- Singh, L. (2018). He said, she said: Effects of bilingualism on cross-talker word recognition in infancy. *Journal of Child Language*, 45(2), 498–510. <https://doi.org/10.1017/S0305000917000186>
- Singh, L., Poh, F. L. S., & Fu, C. S. L. (2016). Limits on monolingualism? A comparison of monolingual and bilingual infants' abilities to integrate lexical tone in novel word learning. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00667>
- Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388(6640), 381–382. <https://doi.org/10.1038/41102>
- 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>
- Tardif, T., & Fletcher, P. (2008). *Chinese communicative development inventories: user's guide and manual*, Peking University Medical Press.
- Tsao, F.-M. (2008). The effect of acoustical similarity on lexical-tone perception of one-year-old Mandarin-learning infants. *中華心理學刊*, 50(2), 111–124.
- Tse, J. K.-P. (1978). Tone acquisition in Cantonese: A longitudinal case study. *Journal of Child Language*, 5(2), 191–204. <https://doi.org/10.1017/S0305000900007418>
- van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General*, 143(1), 340. <https://doi.org/10.1037/a0032192>

- Wang, L., Kalashnikova, M., Kager, R., Lai, R., & Wong, P. C. M. (2021). Lexical and Prosodic pitch modifications in Cantonese infant-directed speech. *Journal of Child Language*, **48**(6), 1235–1261. <https://doi.org/10.1017/S0305000920000707>
- Wang, L., & Wong, P. C. M. (2024). Age-related changes in lexical tones and intonation in Cantonese infant-directed speech: A longitudinal study. *Journal of Child Language*, 1–24. <https://doi.org/10.1017/S030500924000333>
- Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, **39**(4), 703–728. <https://doi.org/10.1017/S0142716418000152>
- Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word–object associations by 14-month-old infants. *Developmental Psychology*, **34**(6), 1289. <https://doi.org/10.1037/0012-1649.34.6.1289>
- Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, **1**(2), 197–234. <https://doi.org/10.1080/15475441.2005.9684216>
- Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, **3**(1), 1–30. https://doi.org/10.1207/S15327078IN0301_1
- Werker, J. F., & Yeung, H. H. (2005). Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences*, **9**(11), 519–527. <https://doi.org/10.1016/j.tics.2005.09.003>
- White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, **14**(2), 372–384. <https://doi.org/10.1111/j.1467-7687.2010.00986.x>
- Wong, P., & Leung, C. T.-T. (2018). Suprasegmental features are not acquired early: Perception and production of monosyllabic Cantonese lexical tones in 4- to 6-year-old preschool children. *Journal of Speech, Language, and Hearing Research*, **61**(5), 1070–1085. https://doi.org/10.1044/2018_JSLHR-S-17-0288
- Wong, P. C., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, **46**(2), 413–421. [https://doi.org/10.1044/1092-4388\(2003\)034](https://doi.org/10.1044/1092-4388(2003)034)
- Xu Y. (2013). *ProsodyPro—A tool for large-scale systematic prosody analysis*. <https://discovery.ucl.ac.uk/id/eprint/1406070/>
- 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>