AMERICAN PSYCHOLOGICAL ASSOCIATION

Developmental Psychology

ISSN: 0012-1649

2022, Vol. 58, No. 11, 2064–2080
https://doi.org/10.1037/dev0001418

The Ability to Use Contextual Cues to Achieve Phonological Constancy Emerges by 14 Months

Ye Feng[1, 2, 3], René Kager[4], Regine Lai[1], and Patrick C. M. Wong[1, 2]

[1] Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
[2] Brain and Mind Institute, The Chinese University of Hong Kong
[3] Department of Linguistics, Beijing Language and Culture University
[4] Utrecht Institute of Linguistics OTS, Utrecht University

The ability to map similar sounding words to different meanings alone is far from enough for successful speech processing. To overcome variability in the speech signal, young learners must also recognize words across surface variations. Previous studies have shown that infants at 14 months are able to use variations in word-internal cues (i.e., acoustic cues within the target word) to form phonological categories and to learn words. The present study takes into consideration the fact that talker variability can easily lead to acoustic overlap between phonological categories, in which case reliance on word-external cues (i.e., acoustic cues in the context preceding and/or following the target word, also referred to as contextual cues) as a frame of reference is obligatory for successful talker adaptation. In a series of experiments, the present study examines when infants are able to use word-external cues to tune to different talkers for the benefit of word learning. Cantonese-learning 14-month-old, 18-month-old, and 24-month-old infants (N = 258) were tested on the associative learning of Cantonese Tone 1–Tone 3 contrast. Results showed that talker variability that yielded acoustic overlap between the two tonal categories compromised infants' ability to map the contrast onto word meanings. However, when given speaker-matched contextual cues, infants as young as 14 months of age demonstrated a certain degree of talker adaptation which may have subserved their use of phonetic details in novel word learning.

Keywords: contextual cues, talker adaptation, novel word learning, phonological constancy, infancy

Supplemental materials: https://doi.org/10.1037/dev0001418.supp

Two complementary skills are deemed vital to early speech sound processing and vocabulary growth; these are sensitivity to phonological distinctiveness and phonological constancy (Best et al., 2009; Mulak & Best, 2013). The ability to detect phonological distinctiveness refers to the sensitivity to acoustic changes among similar sounding words that are phonologically relevant. For example, to learn minimally contrastive words, like "ball" and "tall," young learners must realize that the change of one sound changes word meaning. However, owing to anatomical vocal tract differences and the way they produce speech, different speakers may have dramatically different acoustic realizations of the same phoneme or similar acoustic patterns for different phonological categories (Nusbaum & Magnuson, 1997). Therefore, young learners must also develop the ability to map acoustically dissimilar sounds onto the same phonological category, that is, to achieve phonological constancy. Previous studies have shown that infants as

This article was published Online First August 29, 2022.

Ye Feng https://orcid.org/0000-0002-2695-422X
René Kager https://orcid.org/0000-0002-5811-839X
Regine Lai https://orcid.org/0000-0003-3764-5892
Patrick C. M. Wong https://orcid.org/0000-0002-6105-5027

This article was not preregistered. Our numeric data are available at the Open Science Framework (https://osf.io/d2thr/; Feng et al., 2022).

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: CRE-2015.410).

This work was supported by 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 Center Limited. Patrick C. M. Wong is the founder of a technological startup company in Hong Kong supported by a Hong Kong government technological startup scheme for universities; the research reported here has no association with the company. All other authors declare no conflict of interest.

This study was completed as part of the first author's doctoral dissertation at The Chinese University of Hong Kong (CUHK). We thank the CUHK–Utrecht University Joint Center for Language, Mind and Brain, as well as the 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.

Correspondence concerning this article should be addressed to Patrick C. M. Wong, Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, G03, Leung Kau Kui Building, Hong Kong, China. Email: p.wong@cuhk.edu.hk

young as 14 months of age are able to use variations in word-internal acoustic cues to achieve such phonological constancy in minimal-pair word learning (Apfelbaum & McMurray, 2011; Höhle et al., 2020; Rost & McMurray, 2009, 2010). However, words are rarely produced in isolation in the real world. What remains unknown is whether infants can capitalize on word-external acoustic cues (i.e., acoustic cues in the context outside of the target word, also referred to as contextual cues) to achieve phonological constancy for the benefit of word learning, which is the focus of the present investigation.

One of the major debates among studies on early phonological distinctiveness is the age at which the ability to learn phonologically distinct word forms emerges. This fundamental ability is often tested via a well-established experimental paradigm called the Switch task (Tsui et al., 2019). In the standard version of the Switch task, infants are presented with two novel objects paired with two nonsense words respectively (i.e., Object A paired with Word A and Object B paired with Word B) so that they are encouraged to take these words as labels of the objects. After infants are habituated to the two word–object pairings, they typically experience two test trials: a Same trial, that is, one of the old pairings, and a Switch trial, that is, old visual and auditory stimuli in a new combination. Critically, if they detect the change in pairing in the Switch trial, their attention will bounce back, which indicates successful associative learning of the two novel words. With this paradigm, it was found that 14-month-olds, but not the younger groups, formed word–object associations (Werker et al., 1998). This was later replicated by Byers-Heinlein et al. (2013) on both monolingual and bilingual populations. These results suggested that 14 months may be the onset age when infants can readily associate novel words with meanings.

Interestingly, this associative word learning ability evident at 14 months does not seem to remain as reliable when the auditory labels are minimally contrastive. The most cited example would be the study by Stager and Werker (1997) who found that 14-month-olds failed to detect the switch between "bih" and "dih" but were able to notice the change in a pure discrimination task and a single word–object association task with phonetically dissimilar labels "lif" and "neem." The failure to learn minimally contrastive words at 14 months was further replicated by Werker et al. (2002) and Pater et al. (2004) with results extending to distinctive objects and different phonetic contrasts. Since then, various proposals have been made concerning the divergent results. The 14-month-olds' performance in the associative word learning task was found to be correlated with their vocabulary size (Werker et al., 2002), influenced by the acoustic properties of the contrast (Curtin et al., 2009; Escudero et al., 2014; Escudero et al., 2018; Escudero & Kalashnikova, 2020) and modulated by the cognitive demands of the task, which may be reduced by using known words (Fennell & Werker, 2003; Swingley & Aslin, 2002) and familiar objects (Fennell, 2012), or presenting the two objects side by side when the target word was called out during the test phase (Yoshida et al., 2009). Converging evidence has shown that infants can only reliably use phonetic details in novel word learning without minimizing task demands by the age of 17 to 18 months (e.g., Escudero et al., 2018; Werker et al., 2002). Similar developmental trajectory has also been found for the learning of native lexical tones. While tone learning infants were found not able to recognize native tones as lexically contrastive at 12 to 13 months (Singh et al., 2016), these young tone learners showed successful use of native tone contrasts in novel word learning at 17 to 18 months (Singh et al., 2014; Singh et al., 2016) and 24 months (Singh et al., 2014).

However, words are never learned from only one talker nor with an acoustically controlled set of stimuli such as those in a lab setting. In the real world, the speech signal is produced by a large number of talkers (e.g., parents, siblings) who introduce a great deal of phonetic variability. For the young listeners, word learning also requires establishing phonological abstraction in the face of variability, that is, to achieve phonological constancy.

Only a handful of studies have examined the development of phonological constancy in early word learning and existing evidence has suggested that infants demonstrate ability to use word-internal cues to achieve phonological constancy across multiple talkers in early word learning. The first attempt to add talker variability into the Switch task was made by Rost and McMurray (2009), who habituated 14-month-old English-learning infants with stimuli produced either by one speaker or by 18 different speakers and tested if they could map the similar sounding novel words /buk/ and /puk/ to different meanings. Everything else being the same, infants in the single talker condition failed to learn the minimal pair, while those who listened to multiple talkers succeeded. Because words were presented in isolation, variability in both indexical cues (talker gender, voice quality, etc.) and phonemically relevant cues (voice onset time of /b/ and /p/) were embedded within the to-be-learned words. Therefore, these results suggested that 14-month-olds were able to capitalize on variability in word-internal cues to achieve phonological constancy and to learn words.

This ability to use word-internal cues to adapt to talker variability in novel word learning was further replicated and delimited in more recent studies. Specifically, it was proposed that, instead of phonemically relevant cues, infants may rely on variability along the noncontrastive dimension, that is, indexical cues (Apfelbaum & McMurray, 2011; Rost & McMurray, 2010), or the relational properties among varying cues within the target word (Höhle et al., 2020) to identify the contrast. Quam et al. (2017) further compared the effect of correlated talker gender (i.e., one word always presented by male speakers and the other always by female speakers) with uncorrelated talker gender (i.e., both words were presented by speakers of both genders). Their results showed that infants were no longer able to use word-internal cues to resolve talker variability with covarying gender information, which may have created more information that can be relevant to the learning task and increased task complexity.

Importantly, one common feature of the above-mentioned studies was the learning of the consonant contrast /b/-/p/ with VOT as the contrastive cue. The stimuli were well controlled such that variations in talker information never interacted with variations in the critical domain. Specifically, there was no overlap of VOT values between categories across different speakers. This acoustic manipulation is far from ecologically valid. In the real world, one speaker's token of /p/ can be physically indistinguishable from another speaker's token of /b/ (e.g., Allen et al., 2003; Clayards, 2018; see Peterson & Barney, 1952; e.g., of acoustic variability of vowels across talkers). For suprasegmentals, studies have also reported overlap among lexical tones in terms of their fundamental frequency (hereinafter referred to as F0) values across speakers (e.g., Peng, 2006; Wong & Diehl, 2003). Converging evidence from adult studies showed that category overlap would give rise to

perceptual ambiguity and listeners had to rely on word-external or contextual cues, that is, acoustic cues in the contexts preceding and following the target word, as a reference frame to adapt to a particular talker's phonetic space and resolve such ambiguity in the perception of vowels (e.g., Ladefoged & Broadbent, 1957), consonants (e.g., Mann, 1980), as well as lexical tones (e.g., Francis et al., 2006; Huang & Holt, 2009, 2012; Leather, 1983; Moore & Jongman, 1997; Wong & Diehl, 2003; Zhang, 2014, 2018; Zhang et al., 2017). A question naturally arises then: When does the ability to capitalize on word-external, contextual cues to achieve phonological constancy emerge?

Two sources of evidence serve to suggest that infants may be able to exploit contextual cues early on. The first source of evidence comes from findings on infant accent adaption. Apart from talker variability, differences across speech communities also lead to acoustic variations in the language environment and studies on infant speech perception across accents provide valuable insights on the understanding of the development of phonological constancy (Mulak & Best, 2013). Evidence has shown that prior exposure and local sentence context facilitate infants' recognition of familiar words in a new accent. Specifically, although the ability to recognize familiar words in an unfamiliar accent was only observed in toddlers older than 18 months (e.g., Best et al., 2009; Mulak et al., 2013; Potter & Saffran, 2017; van Heugten & Johnson, 2014), a brief exposure to the new accent(s) before testing boosted recognition for 19-month-olds (Potter & Saffran, 2017; White & Aslin, 2011) and even for 15-month-olds (van Heugten & Johnson, 2014). There has also been evidence that presenting the target word at the end of a sentence context was more helpful than prior exposure to the accent for 28-month-olds (van Heugten & Johnson, 2016). Such sentence contexts, also referred to as carrier phrases in word learning studies (e.g., Singh et al., 2016), are introductory phrases or short comments about the target object that resemble what would usually be heard in natural language input, for example, "Look at the *ball*!," "I like the *ball*!" Interestingly, the reported effect of sentence context on accent adaptation coincides with findings that sentence context facilitates 14-month-olds' novel word learning (Fennell & Waxman, 2010). It is possible that the acoustic information in the sentence contexts may provide infants with a reference frame for processing the phonemic information of the target words produced by a particular speaker (van Heugten & Johnson, 2016).

The second source of evidence comes from an interesting discrepancy between two previous studies on cross-talker word recognition in infants, indicating that experimental design with more prior contexts may have helped talker gender adaptation. Specifically, Houston and Jusczyk (2000) found that 7.5-month-old infants could only successfully recognize a familiarized word across talkers of the same gender but not across gender, and by the age of 10.5 months, infants could recognize the same word across talker gender. It seems that talker adaptation across gender is too taxing for the younger group. However, a more recent study reported successful word recognition by both monolingual and bilingual eight-month-olds regardless of the variability in talker gender (Singh, 2018). One possible reason for such discrepancy in the above two studies lies in the subtle difference in experimental design. Singh (2018) presented infants with passages rather than isolated words in the familiarization phase, which inevitably made the familiarization phase longer and provided more contextual information, that is, the preceding and following sounds of the target words produced by the speaker. It is possible that such contexts revealed the speaker's acoustic characteristics and phonetic spaces, hence reinforcing the mental representation of the target word for infants. This suggests that infants may be sensitive to contextual cues in speech perception early on.

However, there are also reasons to think that using external cues as a frame of reference is in general more cognitively demanding and therefore takes longer to develop. For example, research on auditory perception has revealed that eight-month-old infants are able to track patterns of absolute pitches but not of relative pitches, which are more computationally complex and are relied on more by adults (Saffran et al., 1999; Saffran & Griepentrog, 2001). Similarly, relying on word-external cues to adapt to different speakers' phonetic spaces and achieve phonological constancy may require a longer development window than using word-internal cues.

The Present Study

In light of these findings concerning phonological distinctiveness and phonological constancy, the present study examines whether and when infants are able to use word-external, contextual cues in the speech signal to incorporate talker-dependent, acoustically overlapping phonological categories into words. To answer these questions, we conducted a series of minimal-pair word learning experiments in the absence and presence of talker variability. Cantonese level tones Tone 1 (T1, high-level) and Tone 3 (T3, midlevel) were chosen as stimuli because they differ only in relative pitch height, which varies across speakers due to differences in their pitch ranges (see Figure 1; T3 spoken by one talker could have the same F0 level as a T1 spoken by another). In this case, reliance on word-external cues is obligatory for successful talker adaptation.

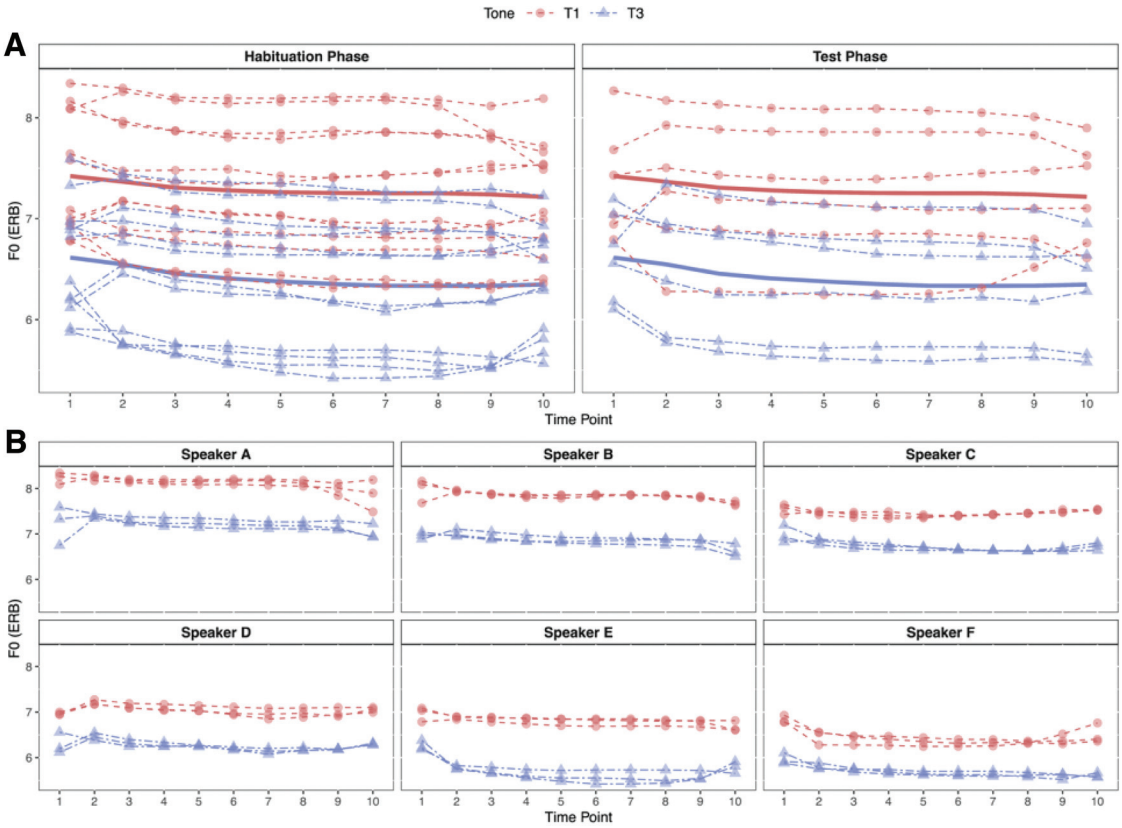
Predictions

Two core predictions were made in the present study. First, it was predicted that talker variability that yields acoustic overlap between phonological categories would impede early word learning. In the present study, overlapping F0 ranges across talkers naturally introduces confusion in lexical tone categories for the infants. Previous research has shown that without contextual cues (e.g., a carrier phrase) to assist with talker adaptation, adult listeners were less able to reliably distinguish Cantonese level tones (e.g., Wong & Diehl, 2003; Zhang et al., 2013, 2017). It was, therefore, predicted that 14-month-old infants would not be able to map the Cantonese T1–T3 contrast onto different meanings in a multiple-speaker set-up without contextual cues. Because none of the previous studies tested associative word learning across multiple speakers on infants older than 14 months, we also included 18-month and 24-month groups in the multiple-talker experiments for a more complete picture of the developmental trajectory. If talker variability that yields acoustic overlap between phonological categories indeed impedes early word learning, the two older groups would not be able to succeed in word learning in a multiple-speaker set-up without contextual cues, either.

The second core prediction in the present study was that infants at the earliest stage of associative word learning (i.e., 14 months) would be able to use word-external cues to tune to different

Figure 1

Pitch Contours of the 36 Tokens Used in the Current Study Separated by Experimental Phase (A) and Speaker (B)



Note. The dashed lines with circles demonstrate F0 (ERB) contours of T1 tokens, and the dash-dotted lines with triangles represent F0 (ERB) contours of T3 tokens. The circles and triangles on the lines are the 10 equidistant time points sampled along each tone contour. The thick solid lines in (A) show the mean F0 (ERB) of all tokens for each tone (T1 above T3). See the online article for the color version of this figure.

talkers' phonetic spaces for the benefit of word learning. The basis of this prediction lies in the observed beneficial effect of sentence context on 14-month-olds' associative word learning (Fennell & Waxman, 2010), toddlers' recognition of accented words (van Heugten & Johnson, 2016), and 8-month-olds' word recognition across talker gender (Singh, 2018). If infants are able to exploit word-external cues to achieve phonological constancy in associative word learning early on, all three age groups would show successful word learning when given contextual cues. If, alternatively, a longer time window is required, it is possible to observe a growing trajectory in the present study where only 18- and 24-month groups (or only the 24-month group) would successfully use contextual cues to achieve phonological constancy for the benefit of word learning.

Method Overview

The above predictions were tested in a series of adapted Switch tasks. In Experiment 1, Cantonese-learning 14-month-old, 18-month-old, and 24-month-old infants were tested on the associative learning of the Cantonese T1–T3 contrast in the face of talker variability without any contextual cues about the speakers' phonetic spaces provided. Such contextual cues were provided in Experiment 2 in the form of a precursor sentence (e.g., "Look here!") preceding the target word. As suggested by Fennell and Waxman (2010), providing a typical naming context in the novel word learning task points infants to the referentiality of the task, which may improve infants' word learning performance. Therefore, two conditions, namely, a speaker-matched condition and a speaker-mismatched condition, were designed in Experiment 2 to tear apart the role of contextual cues from the referential information provided in the precursor sentences. In addition, the associative learning of the same T1–T3 contrast in 14-month-olds was tested with a single-speaker set-up in Experiment 3 to replicate previous findings on phonological distinctiveness with the lexical tone contrast used in the above multiple-talker experiments.

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). Family socioeconomic-status (SES) scores were calculated for each participant by coding parents' educational levels and occupational prestige following the Hollingshead index (Hollingshead, 1975). CCDI (Tardif et al.,

2009), which is the Cantonese version of the MacArthur-Bates Communicative Development Inventories (MCDI; Fenson, 2007), was measured to indicate vocabulary size. CCDI scores for 14-month-olds came from the parental reported scores of the Words and Gestures checklist from CCDI and those for the two older groups were the parental reported scores of the Words and Sentence checklist from CCDI. Across all experiments and conditions in the current study, subjects of the same age groups did not differ in terms of family SES or vocabulary size as indicated by CCDI scores (see Table 1). SES did not differ between age groups across all experiments and conditions either, $F(4, 253) = .302$, $p = .877$.

Participants were recruited mainly via social media (e.g., WhatsApp and Facebook). All infants recruited for the current study were from Cantonese monolingual families in Hong Kong and none was reported to have any prior history of perceptual or neurological disorders. This study was not preregistered. Numeric data is available on Open Science Framework (https://osf.io/d2thr/).

Experiment 1

The aim of Experiment 1 was to investigate if Cantonese learning 14-, 18-, and 24-month-old infants would be able to associatively learn a pair of nonwords differentiated only by lexical tone (i.e., /pi1/ and /pi3/) in the face of talker variability. We used an adapted version of the classic Switch task such that auditory stimuli presented in the habituation and test phase were produced by six speakers with varying pitch ranges instead of only one, thus introducing talker variability into the word learning task. Tokens were naturally produced by human speakers instead of synthesized into perfect distributions to maintain the naturalness as much as possible. As shown in Figure 1, there existed a large overlap between the two tonal categories in the F0 contours across speakers in natural speech. In this case, the indexical cue (pitch range) was mixed with the phonological cue (relative pitch height). Without contextual cues about the speakers' pitch ranges, it would be difficult to achieve phonological constancy. Therefore, it is predicted that infants will not be able to use phonetic details in novel word learning in this case.

Method

Participants

Seventy Cantonese-learning infants were included in this experiment: 24 fourteen-month-olds (mean age = 426 days; range = 392–450 days; 12 girls), 22 eighteen-month-olds (mean age = 537 days; range = 510–567 days; 13 girls), and 24 twenty-four-month-olds (mean age = 718 days; range = 690–747 days; 12 girls). An additional 10 infants who participated in the experiment were excluded from the analysis due to caregiver interference ($N = 1$), equipment failure ($N = 1$), fussiness ($N = 2$), failure to complete the task ($N = 2$), failure to reach the habituation criterion ($N = 3$), and experimenter error ($N = 1$).

Power analyses included in the present study were conducted with G*Power (Faul et al., 2007). For a main effect of trial type in a multiple-talker novel word learning task, a priori sample size estimates were computed using results from Rost and McMurray (2009; Cohen's $dz = .61$). A minimum sample size of infants within each group would be 19 (one-tailed, which is justified as the Switch task makes directional predictions for longer looking to the switch than the same trial) to detect a similar effect, with alpha threshold at .05 and 80% power.

Stimuli

Auditory Stimuli. The speech stimuli consisted of a minimal pair of Cantonese nonwords differing only in lexical tones, that is, the CV syllable /pi/ carrying Cantonese T1 (high-level) and T3 (midlevel). The nonwords were chosen to ensure that they were phonotactically legal lexical forms in Cantonese but did not correspond to any real word in Cantonese. All recordings were conducted in a sound-attenuated booth. Six female native speakers of Cantonese read the stimuli in a lively child-directed manner. Three tokens of each tone were selected per speaker based on their recording quality, which resulted in a total of 36 tokens. The VOT of the consonant was manipulated to 80 ms for all tokens to avoid unnecessary confusion. This value is in line with Cantonese speakers' average VOT values for the same consonant, that is, 77 ms (Lisker & Abramson, 1964). Each token was normalized to 500 ms in length and 70 dB in volume. The average acoustic measurements of the vowels are presented in Table 2, and the pitch contours of the tokens are demonstrated in Figure 1 . The F0 values were calculated using the "prosodypro" script (Xu, 2013) with a sampling rate of 100Hz. Following Wang et al. (2021), the pitch contours of the tokens and the means in Figure 1 were the equivalent-rectangular-bandwidth-rate (ERB) converted from 10 time-normalized F0 values in hertz sampled along each tone contour using the formula of Glasberg and Moore (1990, p. 114; [21.4*log10(.00437*F0 + 1)]). ERB as a psychoacoustic measure is believed to best characterize F0 changes from the perspective of

Table 1
Results From Cross-Experiment Comparisons on SES and CCDI of the 14-Month-Olds, 18-Month-Olds, and 24-Month-Olds

	14-month-old		18-month-old		24-month-old	
Experiment	SES (*SD*)	CCDI (*SD*)	SES (*SD*)	CCDI (*SD*)	SES (*SD*)	CCDI (*SD*)
1	47.33 (12.02)	226.83 (95.66)	43.98 (11.01)	89.29 (106.63)	45.54 (9.33)	367.78 (225.91)
2 (condition A)	45.29 (9.95)	252.42 (80.89)	47.19 (11.67)	80.01 (75.97)	47.05 (7.79)	281.36 (210.38)
2 (condition B)	47.29 (10.33)	231.04 (95.70)	46.33 (8.20)	82.44 (88.54)	47.73 (9.20)	340.17 (209.22)
3a	47.65 (9.31)	234.29 (113.38)	—	—	—	—
3b	47.21 (8.06)	216.04 (106.11)	—	—	—	—
F value	0.213	0.432	0.58	0.064	0.37	0.95
Pr (>*F*)	0.931	0.786	0.562	0.938	0.691	0.391

Note. SES = socioeconomic status; CCDI = Cantonese Communicative Development Inventory. For Experiment 2, condition A refers to the speaker-matched condition, and condition B means the speaker-mismatched condition.

Table 2
Average Acoustic Measurements and Standard Deviations (in Italics) of the Vowels of the 36 Tokens Used in the Current Study, Separated by Speakers

		F0 (Hz)		F1 (Hz)		F2 (Hz)	
Lexical tone	Speaker	*M*	*SD*	*M*	*SD*	*M*	*SD*
T1 (/pi1/)	A	320	5	351	6	2,907	131
	B	302	3	385	24	2,346	47
	C	281	1	304	5	2,961	21
	D	261	4	417	12	2,665	44
	E	246	2	322	17	2,928	46
	F	227	4	275	5	2,810	23
	Average	273	33	342	51	2,767	226
T3 (/pi3/)	A	270	4	377	33	2,756	121
	B	250	5	368	31	2,826	27
	C	243	1	388	22	2,931	45
	D	222	1	401	2	2,747	30
	E	193	3	317	14	2,732	51
	F	193	1	284	7	2,790	16
	Average	229	29	356	47	2,797	85

speech perception (Hermes & van Gestel, 1991; Liu et al., 2007). Stimuli strings in the habituation phase were constructed by concatenating two tokens for each lexical tone from six speakers into 18-s strings at 1-second intervals, while the third token from these speakers was used to form stimuli strings in the test phase. In this way, each trial during habituation (maximum 24 trials, 12 per tone) consisted of 12 tokens of either T1 or T3 (2 tokens × 6 speakers) and each trial in the test phase (two trials total, one for each tone) consisted of two repetitions of six tokens (1 token × 6 speakers × 2 repetitions). The order of tokens within each trial was randomized individually. A token of another Cantonese nonword /nu/ carrying T2 (high-rising) normalized to the same amplitude (70 dB) and length (500 ms) produced by a male native speaker of Cantonese was used as pre- and posttest stimulus.

Visual Stimuli. The visual stimuli were three novel objects used in Singh et al. (2016) and Singh et al. (2018). Each object was paired with an auditory stimulus (/nu2/, /pi1/, or /pi3/) as the visual referent of that speech sound (see Table 3).

Apparatus and Procedure

The experiment was conducted in a testing booth covered with white curtains such that no obvious visual attraction was present in the sight of our young participants except the screen, which was placed approximately 70 cm in front of the participants. Hidden behind the screen was a speaker (Bose SoundLink Color II), through which the auditory stimuli were presented. The experiment was administered by a trained experimenter from an adjacent control room using Habit2 (Oakes et al., 2019) on a Macintosh computer. The infants' looking behavior was recorded through a hidden camera mounted above the screen and transferred online to the experimenter's computer. During the experiment, the infant was seated on the lap of his or her caregiver, who was required to wear a sound-proof headphone (Bose QC II) listening to masking music and was instructed not to interact with the baby throughout the experiment unless necessary.

Each experiment consisted of four phases, namely, a pretest phase (one trial), a habituation phase (six to 24 trials), a test phase (two trials), and a posttest phase (one trial). Before each trial, there was an attention-getter to make sure that the infant was directed toward the screen. During each trial, the visual and auditory stimuli were synchronized, and the infants' looking behavior was logged and coded online through Habit2 by an experimenter who was blind to the trial type presented. Once the infant had oriented to the attention-getter, the experimenter initiated the trial by pressing a button. During each trial, the experimenter indicated looking and nonlooking behavior from the infant by pressing and releasing a second button. The duration of each trial throughout all experimental phases was infant-controlled. The trial ended when the infant looked away for 2 seconds, or when the looking time reached the maximum trial length of 18 seconds. The trial continued if the look-away time was less than 2 seconds. If the infants' looking time within the 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.

We used the classic Switch design with two word–object pairings (e.g., Stager & Werker, 1997; Experiment 1). As illustrated in Table 3, the experiment began with a one-trial pretest, consisting of repeated presentations of a nonword /nu/ in Cantonese T2 (high rising tone) produced by a male native speaker of Cantonese and it was paired with object C. During the habituation phase, the infant was shown repetitions of two word–object pairings. Each trial contained only one word–object pairing, that is, either /pi/ in Tone 1 paired with object A or /pi/ in Tone 3 with object B. Following Werker et al. (2002), there were six blocks of stimuli in the habituation phase presented in a random order, and every block contained two trials of each word–object pairing in a different order (ABAB, ABBA, BABA, BAAB, AABB, BBAA). This phase continued until infants reached the preset habituation criterion: a 50% decrease in the total looking time during three consecutive trials compared with the total looking time in the longest three habituation trials. Therefore, infants went through a minimum of six trials and a maximum of 24 trials during the habituation phase. The test phase consisted of two trials, one same trial and one switch trial. In the same trial, infants were presented with one of the word–object pairings they had listened to in the habituation phase, while in the switch trial they experienced a switch in the word–object pairing, that is, either object A paired to T3 or object B to T1. The order of the two trials was counterbalanced. Note that in this paradigm, successful word learning would be indicated by an increase in looking

Table 3
A Demonstration of Visual and Auditory Stimuli Used Throughout the Procedures

Phases	Visual Stimuli	Auditory stimuli	
		Experiment 1 and 3a	Experiment 2 and 3b
Pretest (1 trial)		/nu2/, /nu2/, [. . .], /nu2/	/nu2/, /nu2/, [. . .], /nu2/
Habituation (6−24 trials)		/pi1/, /pi1/, [. . .], /pi1/	/tai2 ni1 dou6/. /pi1/. /tai2 m4 tai2 dou3/? /pi1/. /nei5 tai2 ha5/. /pi1/. /hou2 leng3 a1/! /pi1/. /hou2 dak1 yi3 a1/! /pi1/. /me1 lai4 ga3/? /pi1/.
		/pi3/, /pi3/, [. . .], /pi3/	/tai2 ni1 dou6/. /pi3/. /tai2 m4 tai2 dou3/? /pi3/. /nei5 tai2 ha5/. /pi3/. /hou2 leng3 a1/! /pi3/. /hou2 dak1 yi3 a1/! /pi3/. /me1 lai4 ga3/? /pi3/.
Test			
Same (1 trial)		/pi1/, /pi1/, [. . .], /pi1/	/tai2 ni1 dou6/. /pi1/. /tai2 m4 tai2 dou3/? /pi1/. /nei5 tai2 ha5/. /pi1/. /hou2 leng3 a1/! /pi1/. /hou2 dak1 yi3 a1/! /pi1/. /me1 lai4 ga3/? /pi1/.
Switch (1 trial)		/pi3/, /pi3/, [. . .], /pi3/	/tai2 ni1 dou6/. /pi3/. /tai2 m4 tai2 dou3/? /pi3/. /nei5 tai2 ha5/. /pi3/. /hou2 leng3 a1/! /pi3/./hou2 dak1 yi3 a1/! /pi3/. /me1 lai4 ga3/? /pi3/.
Posttest (1 trial)		/nu2/, /nu2/, [. . .], /nu2/	/nu2/, /nu2/, [. . .], /nu2/

Note. The order of the two test trials and the switched tone were counterbalanced across subjects. Experiment 1 and 3a, as well as Experiment 2 and 3b, differed in the number of speakers during the habituation and test phases. Gloss on the Cantonese precursor sentences can be found in Figure 3. The color images for visual stimuli were adapted from "Limits on Monolingualism? A Comparison of Monolingual and Bilingual Infants' Abilities to Integrate Lexical Tone in Novel Word Learning" by L. Singh, F. L., Poh, and C. S. Fu, 2016, *Frontiers in Psychology*, *7*, Article 667, p. 5, CC BY, and "Novel Word Learning in Bilingual and Monolingual Infants: Evidence for a Bilingual Advantage" by L. Singh, C. S. Fu, Z. W. Tay, and R. M. Golinkoff, 2018, *Child Development*, *89*(3), p. 7. Copyright 2017 by the Society for Research in Child Development. Adapted with permission. See the online article for the color version of this table.

time to the switch trial compared with that to the same trial. The experiment ended with a one-trial posttest, which was the same as the pretest. The purpose of including a posttest is to ensure that the young participant is still attentive toward the end of the task so that any possible failure of discrimination during the test phase is not caused by fatigue or total loss of attention, thus avoiding the type II error.

For a random selection of 10 (of 253 total) subject videos (recorded infants' looking behavior during the task), frame-by-frame offline coding was conducted with ELAN (Sloetjes & Wittenburg, 2008) by a different experimenter who was blind to the experimental design. Reliability was evaluated by calculating the Pearson's correlation between trial-by-trial total looking times coded online and offline for the ten subjects. Correlations were high for all of them (mean correlation coefficient: .990; range: .971–.999; all p values $< .001$).

Results

Descriptive statistics for all dependent variables can be found in Table 4 for each experiment and age group. 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). We computed p values using the lmerTest package (Kuznetsova et al., 2017), and pairwise comparisons using Tukey's HSD test were conducted with the emmeans package (Lenth, 2020) where appropriate. The dependent variable for all LME models in this study was infants' looking time (in milliseconds), and a random intercept was specified for subject. Independent variables defined for Experiment 1 were trial type and Age Group.

A preliminary analysis was conducted comparing infants' looking time in the last habituation trial to that in the posttest trial to assess whether infants recovered to the posttest (e.g., Byers-Heinlein et al., 2013; Singh et al., 2016). A likelihood ratio test indicated a main effect of trial type, $\chi^2(1) = 166.05$, $p < .001$, with looking time in the posttest trial significantly longer than that in the last habituation trial by 10751.7 ms ($\beta = 10752$, $SE = 1007$; $t = 10.680$). No effect of Age group, $\chi^2(2) = 3.25$, $p = .196$, nor interaction between Age group and trial type, $\chi^2(2) = .20$, $p = .903$, was significant, indicating that null effects in the test phase were not the product of general fatigue as infants from all the three age groups were engaged throughout the task.

Table 4
Descriptive Statistics for Dependent Variables by Experiment (Exp.) and Age Group (Means and Standard Deviations)

Age group	Mean fixation to same trial (*SD*)	Mean fixation to switch trial (*SD*)	Mean fixation to the last habituation trial (*SD*)	Mean fixation to posttest trial (*SD*)	No. trials to habituation (*SD*)	Total habituation time (*SD*)
Exp. 1						
14 months	6,858 (4,957)	6,165 (4,634)	4,680 (2,780)	15,432 (3,528)	9.96 (3.24)	107,056 (50,414)
18 months	7,527 (5,187)	6,315 (3,809)	3,917 (1,967)	14,126 (4,549)	10.32 (3.97)	125,829 (59,375)
24 months	8,561 (5,747)	8,475 (6,110)	5,164 (3,527)	15,362 (4,127)	11.29 (4.74)	141,985 (84,114)
Exp. 2 (speaker-matched condition)						
14 months	6,773 (3,676)	11,895 (5,168)	4,336 (3,012)	16,168 (3,281)	11.67 (3.75)	148,040 (54,876)
18 months	4,963 (4,600)	9,225 (5,088)	4,291 (2,553)	15,230 (3,696)	10.54 (3.83)	130,592 (56,362)
24 months	5,353 (3,323)	9,969 (5,081)	4,796 (3,563)	16,988 (1,951)	12.27 (4.81)	154,572 (82,260)
Exp. 2 (speaker-mismatched condition)						
14 months	6,910 (4,743)	7,787 (4,997)	4,183 (1,759)	16,063 (3,399)	11.75 (4.81)	152,182 (79,994)
18 months	7,924 (4,698)	9,225 (5,553)	4,772 (1,777)	16,808 (2,859)	11.92 (4.59)	154,578 (67,356)
24 months	8,698 (5,876)	7,302 (6,093)	5,267 (2,566)	17,106 (2,431)	11.73 (3.21)	156,464 (62,383)
Exp. 3a						
14 months	6,977 (4,895)	7,148 (4,686)	5,067 (3,159)	16,811 (2,528)	10.5 (4.38)	136,232 (71,763)
Exp. 3b						
14 months	5,361 (4,440)	13,716 (5,206)	5,470 (4,323)	16,851 (2,082)	13.0 (4.82)	173,224 (75,501)

Infants' looking performances during the habituation phase were compared (see Table 4 for descriptive statistics). Results revealed that infants' total habituation time did not differ across age groups, $F(2, 67) = .684$, $p = .508$. Similarly, the number of habituation trials did not differ across age groups, $F(2, 67) = 1.628$, $p = .204$. Thus, there was no general bias in attention across age groups.

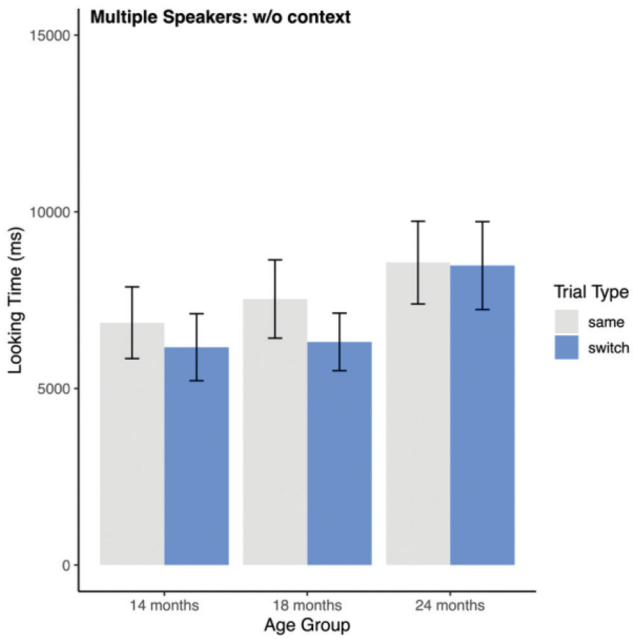
The critical analyses aimed to assess infants' looking time in the same versus switch trials during the test phase (see Table S1 in the online supplemental materials for the model summary). Mean looking time separated by age group is displayed in Figure 2. Likelihood ratio tests revealed no significant main effect of trial type, $\chi^2(1) = .88$, $p = .347$, Age group, $\chi^2(2) = 3.14$, $p = .208$, or interaction between the two factors, $\chi^2(2) = .42$, $p = .810$. Pairwise comparisons confirmed that none of the age groups showed significantly different looking behaviors during same and switch trials (14 months: $\beta = 693$, $SE = 1228$; $t = .564$; 18 months: $\beta = 1212$; $SE = 1282$; $t = .945$; 24 months: $\beta = 86$, $SE = 1228$; $t = .070$).

Results of Experiment 1 suggested that talker variability that yielded acoustic overlap in phonological categories may have impeded lexical integration of the contrast for infants. Owing to differences in speakers' pitch ranges, talker variability inevitably led to variations in the F0 height (the contrastive domain) of the Cantonese level tones used in the present study. The fact that 14-month-olds failed to learn the T1–T3 contrast in this experiment was in line with the previous finding that 14-month-old infants failed to learn novel words when variability was manipulated to exist in the contrastive domain (i.e., VOT for the /b/–/p/ contrast) (Rost & McMurray, 2010). Even the two older groups (18- and 24-month groups) who have demonstrated more reliable word learning ability in previous studies (e.g., Escudero et al., 2014, 2018; Singh et al., 2014, 2016) failed in our Experiment 1, suggesting that even 2-year-old infants were not able to resolve the category ambiguity across speakers to reach phonological constancy and learn the minimal pair without contextual cues. This was not entirely unexpected though, given the previous findings that even adult listeners were less able to reliably distinguish Cantonese level tones across speakers without contextual cues (e.g., Wong & Diehl, 2003; Zhang et al., 2013, 2017). Experiment 2 was designed to provide the required cues.

Experiment 2

The aim of Experiment 2 was to test whether and when infants would be able to rely on word-external contextual cues to adapt to talker differences and achieve phonological constancy for the benefit of word learning. Given the successful use of carrier phrases to assist accent adaptation (van Heugten & Johnson, 2016) and novel word learning (Fennell & Waxman, 2010; Singh et al., 2016;

Figure 2
Mean Fixation Times to the Visual Stimulus for Same and Switch Trials in the Test Phase in Experiment 1 Divided by Age Group (Error Bars: SEM)

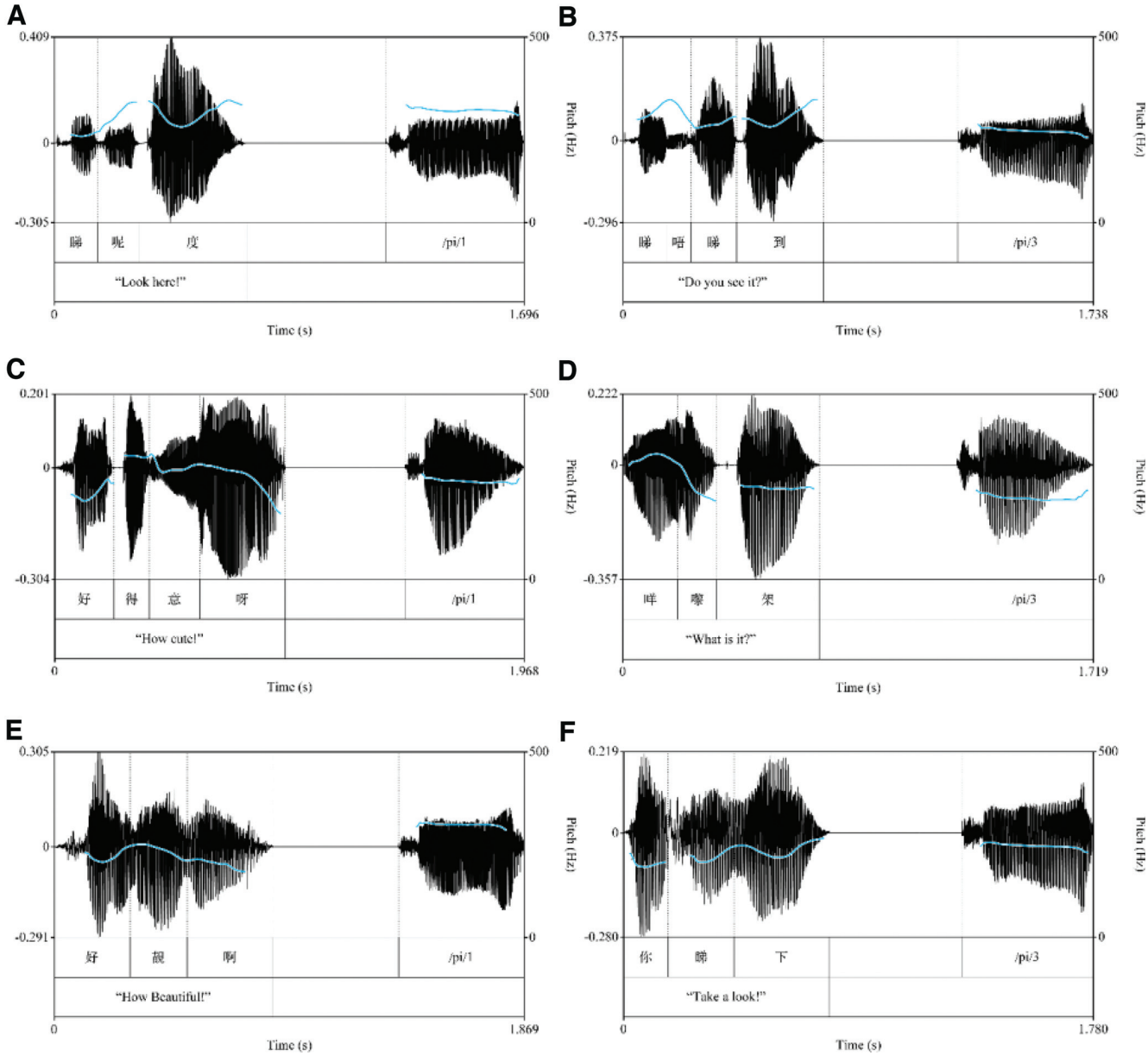


Note. See the online article for the color version of this figure.

Yoshida et al., 2009), we used a similar stimuli structure, a precursor sentence preceding the target word (e.g., "Look here! /pi1/." See Figure 3 for more example stimuli), to provide contextual cues about different speakers' tone space in the Switch task to see if infants could demonstrate context-dependent talker adaptation. Importantly, to clarify whether it is the contextual cues about speakers' phonetic information or the referential information in the precursor sentences that may come into play, we compared word learning results in two conditions: one speaker-matched condition where the precursor sentences and the targets were produced by the same speakers, thus providing phonetic information about the talkers in the contexts, the other being a speaker-mismatched condition where precursor sentences and the targets were read by different speakers and only referential cues were provided.

Figure 3

Examples of the Final Stimuli With Sound Waves (in Black), Pitch Tracks (in Blue), Text Grids and Glosses



Note. (A) and (B) are example stimuli of the speaker-matched condition, where the precursor sentences and the target novel words were from the same speaker (Speaker B in these examples). (C) and (D) are example stimuli of the speaker-mismatched condition, where precursor sentences from Speaker A are mismatched with targets from Speaker D who has a lower mean F0. In this case, the target T1 sounds more like a T3 (3C) and the target T3 sounds lower than normal (3D). (E) and (F) are example stimuli of the speaker-mismatched condition, where precursor sentences from Speaker F are mismatched with targets from Speaker B who has a higher mean F0. Therefore, the target T1 sounds higher than normal (3E) and the target T3 sounds closer to a T1 (3F). See the online article for the color version of this figure.

Method

Participants

One hundred forty Cantonese-learning infants were included in this experiment, and they were randomly assigned to the two conditions. For the speaker-matched condition, there were 24 fourteen-month-olds (mean age = 428 days; range = 393–447 days; 12 girls), 24 eighteen-month-olds (mean age = 543 days; range = 512–569 days; 12 girls), and 22 twenty-four-month-olds (mean age = 716 days; range = 692–745 days; 10 girls). For the speaker-mismatched condition, there were 24 fourteen-month-olds (mean age = 432 days; range = 392–449 days; 12 girls), 24 eighteen-month-olds (mean age = 540 days; range = 512–569 days; 12 girls), and 22 twenty-four-month-olds (mean age = 711 days; range = 691–746 days; 11 girls). An additional 27 infants who participated in the experiment were excluded from the analysis owing to caregiver interference ($N = 3$), equipment failure ($N = 1$), fussiness ($N = 4$), failure to complete the task ($N = 11$), failure to reach the habituation criterion ($N = 5$), and experimenter error ($N = 3$).

For a main effect of trial type, the same sample size estimates were used as Experiment 1 (19 per group) based on Rost and McMurray (2009; Cohen's $dz = .61$). For an interaction between trial type (same vs. switch) and condition (speaker-matched vs. speaker-mismatched), effect size estimates were drawn from the Trial Type (same vs. switch) × Condition (name- vs. exclaim training) interaction reported in Fennell and Waxman (2010, Experiment 2, $f = .44$), which led to an estimate of seven infants per group to detect a similar effect, with alpha threshold at .05 and 80% power.

Stimuli

Visual stimuli were identical to Experiment 1. Auditory stimuli for this experiment consisted of six precursor sentences and two target novel words. The two nonwords were the same as those used in Experiment 1 (see Table 2 and Figure 1 for the acoustic measurements and pitch contours of the tokens). The six precursor sentences were produced by the same six speakers as those who produced the target novel words, which made a total of 36 phrases with a mean length of 841 ms ($SD = 136$ ms). The same set of short sentences was used to carry tokens of both tones in both habituation and test phases. Each token of the target novel words was spliced onto a precursor sentence with a 500-ms interval to form the final stimuli (two for the habituation phase and the third for test; see Figure 3, e.g., of the final stimuli). Within each trial, the six precursor sentences, together with the attached target novel word, were presented once, each by a different speaker in a random order. Unlike the carrier phrases used in Fennell and Waxman (2006) and Singh et al. (2016), which were sentences with the final word left out (requiring the target word to form a complete sentence), we designed short sentences that were complete on their own so that participants would not experience an unnaturally sudden change of speaker in the middle of a sentence in the speaker-mismatched condition.

In the speaker-matched condition, the precursor sentence and the target were produced by the same speaker so that the precursor sentence would provide contextual cues for the perception of the target novel word, whereas in the speaker-mismatched condition, the speaker who produced the precursor sentence was different from the one who read the target novel word, resembling the scenario where two caregivers introduce a new toy to the infant, one directing the infant's attention and the other calling out the name of the toy. The total number of speakers the infant experienced within a trial was the same (i.e., 6) for both conditions. The mismatching followed a scheme that two speakers would never be paired with each other if they were close in mean pitch height, so that they were less likely to be confused as the same person, as illustrated in Table 5. To ensure that no contextual cues were provided in the speaker-mismatched condition, the order of the stimuli within each trial was pseudorandomized such that the target novel word in one stimulus and the precursor sentence in the following stimuli were never from the same speaker. Two additional native speakers of Cantonese who did not participate in the stimuli recording and were unaware of the experimental design listened to the final stimuli in both conditions. They were in agreement that all stimuli sounded natural.

Apparatus and Procedure

The apparatus and procedure were exactly the same as those in Experiment 1. Details of the stimuli presentations throughout the procedure are demonstrated in Table 3.

Results

For Experiment 2, the full model included trial Type, Age group and condition as fixed factors with random intercepts specified for Subject. First, infants' looking time in the last habituation trial was compared with that in the posttest trial to check recovery. A likelihood ratio test revealed a main effect of trial type, $\chi^2(1) = 480.60$, $p < .001$, with looking time in the posttest trial significantly longer than that in the last habituation trial by 11,831.42 ms ($\beta = 11831$, $SE = 813$, $t = 14.549$). No effect of age group, $\chi^2(2) = 5.19$, $p = .075$, or condition, $\chi^2(1) = 1.46$, $p = .228$, was found, nor was there any significant interaction between age group and trial type, $\chi^2(2) = .45$, $p = .799$, or between condition and trial type, $\chi^2(1) = .16$, $p = .689$, or among the three, $\chi^2(2) = .86$, $p = .651$. Therefore, there was no general fatigue throughout the experiment across age groups or conditions.

Then, infants' looking performances during the habituation phase for the two between-subjects conditions were compared. Results revealed that infants' total habituation time did not differ across age groups, $F(2, 134) = .423$, $p = .656$, or conditions, $F(1, 134) = .796$, $p = .374$, and there was no interaction between the two factors, $F(2, 134) = .378$, $p = .686$. Similarly, the number of habituation trials did not differ across age groups, $F(2, 134) = .389$, $p = .678$, or conditions, $F(1, 134) = .210$, $p = .648$, and there

Table 5
An Illustration of the Speaker Mismatching Scheme in Experiment 2

Speaker for context	Speaker for novel word
A	D
B	E
C	A
D	F
E	C
F	B

Note. Speakers were ordered from A to F in terms of their mean pitch height.

was no interaction between the two factors, $F(2, 134) = .618$, $p = .541$, either. Thus, there was no general bias in attention across age groups or conditions.

The critical analyses aimed to assess infants' looking time in the same versus switch trials during the test phase in the speaker-matched and speaker-mismatched conditions (see Table S2 in the online supplemental materials for the model summary). Mean looking times separated by age group and condition are displayed in Figure 4. A main effect of trial type, $\chi^2(1) = 24.26$, $p < .001$) was observed through likelihood ratio tests, with looking time in the switch trial longer than that in the same trial by 5,122.2 ms ± 1180.9 (standard errors). An interaction of trial Type × Condition was also found to be significant, $\chi^2(1) = 19.72$, $p < .001$. No effect of age group, $\chi^2(2) = .52$, $p = .770$, or condition, $\chi^2(1) = .01$, $p = .934$, was observed. There was no significant interaction of Trial Type × Age Group, $\chi^2(2) = 1.57$, $p = .456$, and no three-way interaction of Trial Type × Age Group × Condition, $\chi^2(2) = 1.67$, $p = .435$, either. Pairwise comparisons revealed that the switch effect was only shown in those infants participating in the speaker-matched condition ($\beta = 4667$; $SE = 692$, $t = 6.743$). No effect of trial type was found in the speaker-mismatched condition ($\beta = 261$, $SE = 692$, $t = .377$). Detailed comparisons by age group are shown in Table 6.

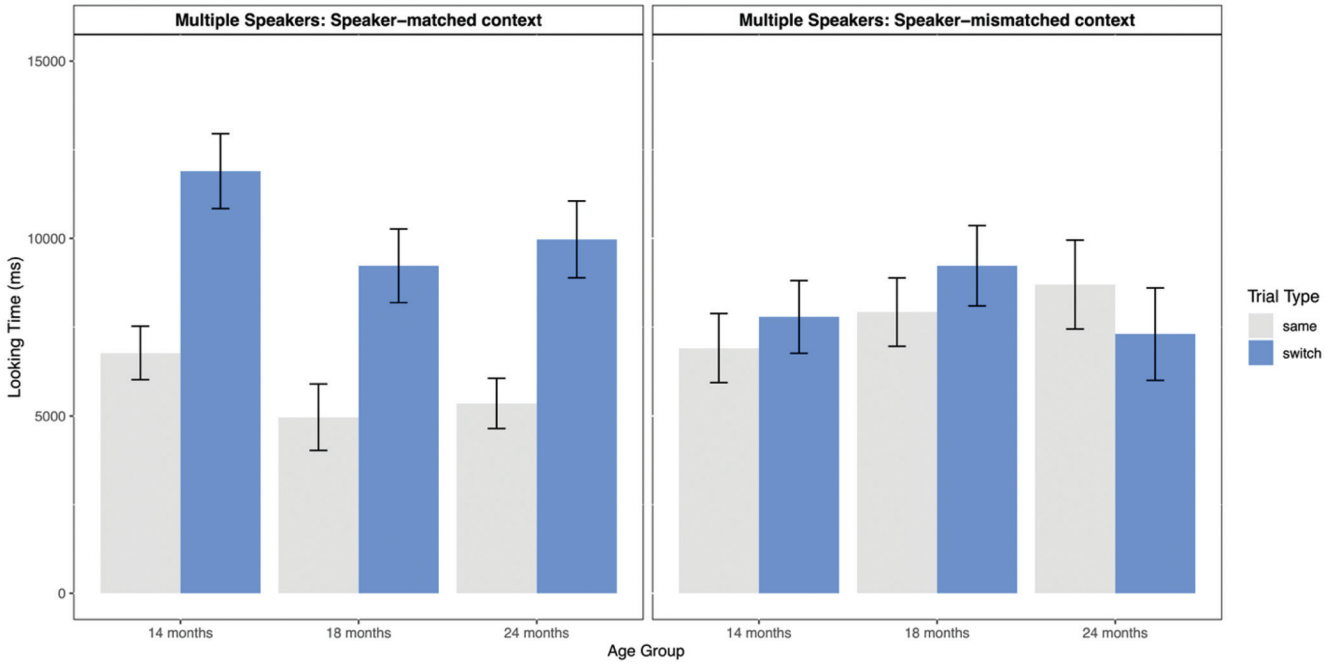
Results of Experiment 2 suggested that when given contextual cues, infants from all age groups (14 to 24 months) demonstrated a certain degree of talker adaptation. Specifically, infants only succeeded in the speaker-matched condition. This was in line with the literature on adult talker adaptation which reported that the ability to identify the same stimuli improved drastically when a precursor sentence from the same speaker was provided to the listeners (e.g., Wong & Diehl, 2003; Zhang et al., 2013, 2017). Infants relied on the phonetic information provided in contextual cues to track and adapt to different speakers' phonetic spaces (tone spaces in this case) and extract the relative pitch height of the target tone produced by each speaker. And the ability to achieve phonological constancy may have subserved their use of phonetic details in novel word learning. In contrast, in the speaker-mismatched condition, no such contextual cues can be found and used as a frame of reference to determine the relative pitch height of the target novel words. All that remained was the same referential information in the precursor sentences, which, as indicated by the null result, was not enough to facilitate word learning in the multiple-talker set-up.

Experiment 3a

The aim of Experiment 3 was to replicate previous findings on phonological distinctiveness with the Cantonese T1–T3 contrast used in the above multiple-talker experiments. Recall that infants at 14 months of age did not show reliable learning of minimally contrastive words when presented in isolation (e.g., Pater et al., 2004; Stager & Werker, 1997; Werker et al., 2002), but they appeared to be able to do so when the target words were buried in sentence frames (e.g., Fennell & Waxman, 2010). However, none of the previous word learning studies tested the Cantonese level tone contrast used in the present experiments. To ensure that the current findings with the multiple-talker set-up are not attributable to any specific characteristics of the stimuli used and thus likely generalizable to the perception and learning of other phonemic categories, we tested 14-month-old infants on the associative learning of the same Cantonese T1–T3 contrast when intertalker variability was not involved.

Figure 4

Mean Fixation Times to the Visual Stimulus for Same and Switch Trials in the Test Phase in Experiment 2 Divided by Age Group With Speaker-Matched Condition on the Left and Speaker-Mismatched Condition on the Right (Error Bars: SEM)



Note. See the online article for the color version of this figure.

Table 6
Detailed Results of the Statistical Analysis for Each Age Group From the Two Conditions in Experiment 2

Age group	Contrast	Estimate β (*SE*)	*df*	*t* value	Pr (>|*t*|)
Speaker-matched condition					
14 months	same – switch	−5,122 (1,181)	146	−4.337	<0.001***
18 months	same – switch	−4,262 (1,181)	146	−3.609	<0.001***
24 months	same – switch	−4,616 (1,233)	146	−3.742	<0.001***
Speaker-mismatched condition					
14 months	same – switch	−877 (1,181)	146	−0.743	0.459
18 months	same – switch	−1,302 (1,181)	146	−1.102	0.272
24 months	same – switch	1,396 (1,233)	146	1.132	0.260

*** $p < .001$.

In Experiment 3a, we adopted the classic Switch task where the target novel words were produced by a single speaker and presented in isolation. On the basis of previous findings, it was predicted that 14-month-olds will not be able to show reliable use of phonetic details in word learning.

Method

Participants

Twenty-four 14-month-old infants were included in this experiment (mean age = 423 days; range = 391–446 days; 12 girls). An additional six infants who participated in the experiment were excluded from the analysis owing to failure to reach the habituation criterion ($N = 4$) and experimenter error ($N = 2$).

Effect size estimates were obtained from the results reported in Curtin et al. (2009) for the learning of the /i/–/I/ contrast at 15 months in a Switch task (Cohen's $dz = .72$). Based on this effect size, an estimated sample size of 14 infants would be sufficient to detect a similar effect, with alpha threshold at .05 and 80% power.

Stimuli

Visual stimuli were identical to previous experiments in this study. Auditory stimuli were the same as those used in Experiment 1, except that in Experiment 3a, stimuli were presented by only one of the six speakers (counterbalanced across participants).

Apparatus and Procedure

The apparatus and procedure were the same as those in Experiment 1. Details of the stimuli presentations throughout the procedure are demonstrated in Table 3.

Results

Because Experiment 3a and 3b focused on the 14-month group, statistical analyses no longer included age group as a fixed factor. For both experiments, the full model included trial type as a fixed factor with random intercepts specified for Subject and Speaker.

As stated earlier, recovery was checked first to safeguard against the type II error. Results showed a significant elevation of fixation to the posttest trial compared with the last habituation trial, $\chi^2(1) = 80.90$, $p < .001$, $\beta = 11744$, $SE = 826$, $t = 14.218$, indicating a successful recovery of attention.

Of primary interest to our analyses was whether infants would demonstrate different looking behaviors during the same and switch trials, that is, notice the switch in the test phase when the stimuli were produced by one single speaker but without contextual support. A likelihood ratio test indicated that the model including trial type did not provide a better fit for the data than a model without it, $\chi^2(1) = .10$, $p = .747$ (see Table S3 in the online supplemental materials for the model summary). As shown in Figure 5, looking times were on average quite similar between the two types of trials ($\beta = 171$, $SE = 542$, $t = .316$).

Consistent with previous findings on the associative learning of minimal pairs distinguished by other phonemic contrasts, results of Experiment 3a showed that even in the absence of talker variability 14-month-olds failed to associate the Cantonese T1–T3 contrast with different meanings when presented in isolation. This confirmed that infants' early sensitivity to phonological distinctiveness was not reliable at 14 months of age.

Experiment 3b

In Experiment 3b, we provided the contextual cues of the speaker by adding precursor sentences produced by the same speaker into the task with the hope that infants could get a clearer idea of the speaker's acoustic properties (pitch range in this case) to form tone categories.

Method

Participants

Twenty-four 14-month infants were included in this experiment (mean age = 422 days; range = 391–448 days; 12 girls). An additional six infants who participated in the experiment were excluded from the analysis owing to fussiness ($N = 3$), experimenter error ($N = 2$), and premature birth ($N = 1$).

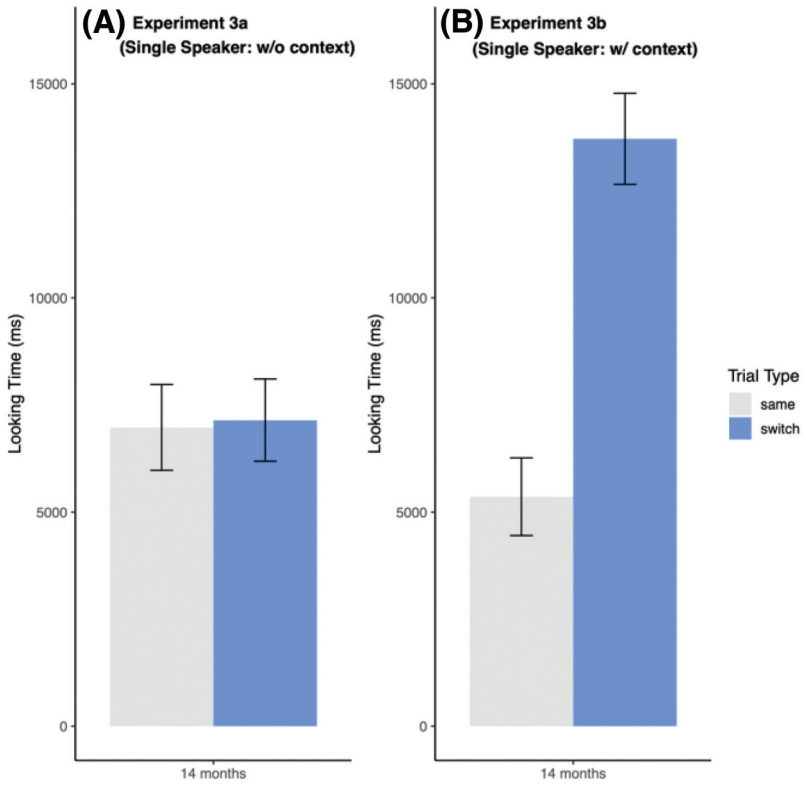
Based on effect sizes obtained from Fennell and Waxman (2010, Experiment 1, Cohen's $dz = .52$), a minimum sample size of infants would be 25 to detect a similar effect, with alpha threshold at .05 and 80% power. Because data collection was stopped before reaching this sample threshold owing to the outbreak of COVID-19, a post hoc sensitivity analysis was carried out to determine the achieved power for this Experiment 3b. Results indicated that a sample size of 24 would yield 80% power to detect a medium-size effect (Cohen's $dz = .52$).

Stimuli

Visual stimuli were identical to previous experiments in this study. Auditory stimuli were the same as those used in Experiment 2 (speaker-matched condition), except that in Experiment 3b,

Figure 5

Fixation Times to the Visual Stimulus for Same and Switch Trials in the Test Phase in Experiment 3a (A) and Experiment 3b (B; Error Bars: SEM)



Note. See the online article for the color version of this figure.

stimuli were presented by only one of the six speakers (counterbalanced across participants).

Apparatus and Procedure

The apparatus and procedure were the same as those in Experiment 2. Details of the stimuli presentations throughout the procedure were demonstrated in Table 3.

Results

Again, successful recovery was confirmed with infants' looking time in the posttest trial significantly longer than that in the last habituation trial, $\chi^2(1) = 65.76$, $p < .001$, $\beta = 11381$, $SE = 979$, $t = 11.621$.

Primary analysis focused on whether infants looked differently to the same and switch trials during the test phase when contextual cues were provided (see Table S4 in the online supplemental materials for the model summary). Indeed, a main effect of trial type was found, $\chi^2(1) = 28.65$, $p < .001$. Examination of the summary output for the full model indicated that looking times were on average an estimated 8354.79 ms longer in the switch trial compared with the same trial ($\beta = 8355$; $SE = 1245$; $t = 6.709$). Mean looking times are displayed in Figure 5.

Results of Experiment 3b showed that 14-month-olds successfully mapped the Cantonese T1–T3 contrast with different word meanings when the target novel words were preceded by a precursor sentence, in line with previous findings on the faciliatory role of sentence frames in 14-month-olds' learning of other phonemic contrasts (e.g., Fennell & Waxman, 2010). Together, results of Experiment 3 replicated previous findings on phonological distinctiveness, showing that even when the stimuli were produced by a single speaker, 14-month-old infants cannot reliably integrate Cantonese T1–T3 contrast into different word meanings unless contextual cues were provided, similar to previous results with segmental contrasts. Therefore, the current findings with the multiple-talker set-up (Experiments 1 and 2) were not attributable to any specific characteristics of the stimuli used.

General Discussion

In a series of minimal-pair word learning experiments, the present study examined when infants are able to use word-external cues to tune to different talkers for the benefit of word learning. Results showed that when talker variability yielded acoustic overlap between two tonal categories, infants failed to map the contrast onto word meanings at 14, 18, and 24 months. However, when given speaker-matched contextual cues, infants as young as 14 months of age demonstrated successful word learning across multiple talkers. The facilitatory effect of the precursor sentences was eliminated when they were speaker-mismatched with the target word. The present study provides evidence for the first time that infants as young as 14 months are able to use word-external cues

to adapt to different talkers' phonetic spaces and integrate talker-dependent, acoustically overlapping phonological categories into word meanings.

Experiment 1 tested 14-, 18- and 24-month-old Cantonese-learning infants on their ability to integrate the Cantonese T1–T3 contrast into novel word meanings when the stimuli were produced by six speakers and thus inevitably yielded overlap in F0 contours between the two tonal categories. Although other indexical cues like voice quality also affect the perception of pitch height (Kuang & Liberman, 2018), there is no conclusive evidence that an unknown speaker's pitch range can be estimated accurately without contextual cues (compare Honorof & Whalen, 2005, with Bishop & Keating, 2012). Indeed, none of the age groups exhibited successful word learning in the face of talker variability, which suggested that infants under 2 years of age were not able to resolve the category ambiguity brought by talker differences without word-external cues, and their failure in reaching phonological constancy accompanied their failure in achieving phonological distinctiveness. These results are in line with previous findings that adult listeners were less accurate in word identification when listening to multiple speakers (e.g., Verbrugge et al., 1976; Wong & Diehl, 2003). It is worth noting that the present study does not make the argument that talker variability impairs early word learning or speech perception. After all, 14-month-olds still failed to learn minimal pairs without talker variability (Experiment 3a). Instead, we argue that young listeners are unable to reliably resolve phonological ambiguity induced by talker differences without word-external cues, similar to adults. When word-external cues in the context preceding the target nonwords were provided as in Experiment 2, infants from all the three age groups demonstrated successful talker adaptation in the speaker-matched condition, suggesting that young listeners are capable of using contextual cues as a frame of reference to form phonological categories and learn words as young as 14 months. In contrast, when the indexical information was mismatched, all groups failed in the word learning task. This suggested that the effect of contextual cues in talker adaptation was irrelevant to the referential information carried in the precursor sentence. Experiment 3 replicated previous findings on the development of phonological distinctiveness with the Cantonese T1–T3 contrast. Fourteen-month-old infants were consistently found to have difficulty with lexically integrating phonetic contrasts such as /b/-/d/ (Pater et al., 2004; Stager & Werker, 1997; Werker et al., 2002), /b/-/p/ (Pater et al., 2004; Rost & McMurray, 2009), /i/-/u/ (Curtin et al., 2009; Escudero et al., 2014), Mandarin T2–T3 (rising-dipping), or a less similar T1–T3 (level-dipping; 12–13 months, Singh et al., 2016) when presented in isolation. The same results were found with Cantonese level tones in the present study. Despite the absence of talker-induced overlap in F0 contours between the two tonal categories, infants at the earliest developmental stage of associative word learning (14 months) did not show reliable learning of the Cantonese T1–T3 contrast unless contextual cues were provided. This replication suggests that our findings of context-dependent talker adaptation in the multiple-talker experiments did not result from any specific properties of the Cantonese level tones and thus are likely generalizable to the perception and learning of other phonemic categories.

The ability to use word-external contextual cues to adapt to different talkers for the benefit of word learning emerges early on. As discussed, adult listeners capitalize on preceding or following sentence contexts to adapt to talker variability in speech perception (e.g., Francis et al., 2006; Huang & Holt, 2009, 2012; Ladefoged & Broadbent, 1957; Leather, 1983; Mann, 1980; Moore & Jongman, 1997; Wong & Diehl, 2003; Zhang, 2014, 2018; Zhang et al., 2017). Current findings suggest that unlike the protracted process in learning to use an external frame of reference in other cognitive domains, the ability to dynamically map the ambiguous signals onto the intended categories based on external cues emerges at the earliest stages of associative word learning in human infants. This is also in line with previous findings that 14-month-old infants are sensitive to contextual factors in their perception of phonological categories. Specifically, infants have been found to be able to discriminate a phonetic contrast only when it is placed in a particular phonological context (see Fort et al., 2017; for an example of successful discrimination of [f]-[v] in [ofne]-[ovne] but not in [ofbe]-[ovbe]; see also Eimas & Miller, 1991 and Levitt et al., 1988, e.g., of other contextual constraints). However, it is worth noting that the contextual effect revealed in the current study extends that of a particular phonological context for a particular phoneme. It demonstrates infants' ability to adapt to different phonetic spaces with word-external contextual cues in a talker-dependent manner, and it subserves word learning. Although young learners' ability to accommodate accent variation in familiar word recognition was found to improve significantly as a function of vocabulary spurt in the second year of life (van Heugten et al., 2015), the present study did not observe significant difference across the three age groups. One explanation for the lack of age effect is that the development of this ability is relatively stable, not subject to prosodic changes in the language input or the young learner's vocabulary growth with age. It is also possible that context-dependent talker adaptation is not as cognitively complex as accent adaptation, as accented speech typically varies along multiple dimensions (Schmale et al., 2012). Therefore, no observable age difference was found in the present study.

Early phonetic representations are detailed in nature and can be adjusted online. According to our design of the multiple-talker experiments, infants were exposed to six to 24 habituation trials before entering the test phase. Each trial contained one of the two word–object pairings, and there were six stimuli within a trial, each from a different speaker. That means infants were experiencing changes of speaker-specific phonetic properties within a trial, and changes of tonal categories across trials. Their successful learning performance in the speaker-matched condition in Experiment 2 suggested a dynamic adjustment of phonetic representations online in a speaker-dependent or context-dependent manner. Intriguingly, the young listeners appeared capable of readjusting the tonal representation from one speaker to another within the length of a trial, just like adults (Dahan et al., 2008; Zhang, 2018). And these representations are linguistic in nature as they are readily available to be used to construct word forms.

It is important to note that the beneficial effect of the context observed in the current study did not result from attentional factors. Neither infants' total habituation time nor the number of trials they had before reaching habituation criterion differed across age groups or between the two conditions. The similarity in attention resources between conditions also speaks against the results being explained by naturalness of the conditions. If the speaker-mismatched condition sounded noticeably unnatural compared with the speaker-matched condition, infants would be expected to exhibit different looking behaviors during habituation.

Likewise, the results could not be explained by the referential status of the context. In fact, the purpose of including a speaker-mismatched condition is to disentangle the role of phonetic information and referential information in the precursor sentences. As mentioned previously, such an account of the role of referential information in the task was proposed by Fennell and Waxman (2010). Successful learning of the /din/-/bin/ contrast by 14-month-olds was observed when the target word was embedded in introductory phrases or when a name-training was provided with objects familiar to the infants ("car," "kitty," "shoe") but not when an exclaim-training was presented where the same familiar objects were paired with exclamations ("whee," "wow," "yay"). It was argued that 14-month-olds were able to use phonetic details in word learning only when they were clearly informed of the referential status of the novel words. However, there existed apparent differences in phonetic information presented between the two training conditions. For example, there were four consonants in the name-training words (/k/, /r/, /t/, /ʃ/), but only two in the exclaim-training words (/w/, /j/). It is possible that the extra consonant information in the name-training words provided infants with more information about the speaker's phonetic space and helped them form more robust lexical categories. In our Experiment 2, the exact same precursor sentences were used in the two conditions, except that in the speaker-mismatched condition, the tone space indicated in the precursor sentence did not correspond to the tone space of the target nonword. What remained was the same referential information. The distinct results between the two conditions suggest that referential cues alone are not sufficient to boost the 14-month-olds' ability to use phonetic details in minimal pair word learning.

Nevertheless, our conclusions should not be taken to imply that pointing to the referential status of the task is not useful. After all, a clear referential status of the words may ease the memory retrieval and processing load for infants, similar to using well-known words (e.g., "ball" and "doll"; Fennell & Werker, 2003; Swingley & Aslin, 2002) and familiar objects (Fennell, 2012), or presenting the two objects side by side when the target word was called out during the test phase (Yoshida et al., 2009). One limitation of the current design is that the speaker-matched condition in Experiment 2 did consist of both phonetic information and the referential cues, so it remained unknown whether it is the combination of the two that took effect. Future studies may be able to test this with reversed speech or pure tones as context though it may be extremely artificial.

Another limitation is that the Switch task only revealed whether or not infants associated the target word with the object without revealing how they perceived each presented target. Hence it is unclear whether infants perceived the tones differently according to the contextual cues provided as adults do, for example, mistaking a T1 as T3 if the context shows higher pitch range, or simply failed to align the tokens with exemplars of any category.

In conclusion, the current findings suggest that although talker variability could yield acoustic overlap between phonological categories which impedes early word learning, infants as young as 14 months are able to tune to different talkers' phonetic spaces based on speaker-matched word-external cues for the benefit of novel word learning. These results complement previous findings that have shown infants' ability at 14 months to incorporate variations in word-internal cues to achieve phonological constancy in the learning of minimal-pair contrasts. Collectively, these results indicate that infants are ready to deal with talker variability to learn words in the real world shortly after the first birthday.

References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552. https://doi.org/10.1121/1.1528172

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. https://arxiv.org/abs/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

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America*, *132*(2), 1100–1112. https://doi.org/10.1121/1.4714351

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

Clayards, M. (2018). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica*, *75*(1), 1–23. https://doi.org/10.1159/000448809

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

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, *108*(30), 710–718.

Eimas, P. D., & Miller, J. L. (1991). A constraint on the discrimination of speech by young infants. *Language and Speech*, *34*(3), 251–263. https://doi.org/10.1177/002383099103400303

Escudero, P., Best, C. T., Kitamura, C., & Mulak, K. E. (2014). Magnitude of phonetic distinction predicts success at early word learning in native and non-native accents. *Frontiers in Psychology*, *5*, Article 1059. https://doi.org/10.3389/fpsyg.2014.01059

Escudero, P., & Kalashnikova, M. (2020). Infants use phonetic detail in speech perception and word learning when detail is easy to perceive. *Journal of Experimental Child Psychology*, *190*, Article 104714. https://doi.org/10.1016/j.jecp.2019.104714

Escudero, P., Mulak, K. E., Elvin, J., & Traynor, N. M. (2018). "Mummy, keep it steady": Phonetic variation shapes word learning at 15 and 17 months. *Developmental Science*, *21*(5), Article e12640. https://doi.org/10.1111/desc.12640

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

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* [Data file]. Open Science Framework. https://doi.org/10.17605/OSF.IO/D2THR

Fennell, C. T. (2012). Object familiarity enhances infants' use of phonetic detail in novel words. *Infancy*, *17*(3), 339–353. https://doi.org/10.1111/j.1532-7078.2011.00080.x

Fennell, C. T., & Waxman, S. (2006). Infants of 14 months use phonetic detail in novel words embedded in naming phrases. *Proceedings of the*

30th annual Boston University conference on language development (pp. 178–189). Citeseer.

Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development*, *81*(5), 1376–1383. https://doi.org/10.1111/j.1467-8624.2010.01479.x

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, *46*(2–3), 245–264. https://doi.org/10.1177/00238309030460020901

Fenson, L. (2007). *MacArthur-Bates communicative development inventories*. Brookes Publishing Company.

Fort, M., Brusini, P., Carbajal, M. J., Sun, Y., & Peperkamp, S. (2017). A novel form of perceptual attunement: Context-dependent perception of a native contrast in 14-month-old infants. *Developmental Cognitive Neuroscience*, *26*, 45–51. https://doi.org/10.1016/j.dcn.2017.04.006

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

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138. https://doi.org/10.1016/0378-5955(90)90170-T

Hermes, D. J., & van Gestel, J. C. (1991). The frequency scale of speech intonation. *The Journal of the Acoustical Society of America*, *90*(1), 97–102. https://doi.org/10.1121/1.402397

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.

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America*, *117*(4), 2193–2200. https://doi.org/10.1121/1.1841751

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), 1570–1582. https://doi.org/10.1037/0096-1523.26.5.1570

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, *125*(6), 3983–3994. https://doi.org/10.1121/1.3125342

Huang, J., & Holt, L. L. (2012). Listening for the norm: Adaptive coding in speech categorization. *Frontiers in Psychology*, *3*, Article 10. https://doi.org/10.3389/fpsyg.2012.00010

Kuang, J., & Liberman, M. (2018). Integrating voice quality cues in the pitch perception of speech and non-speech utterances. *Frontiers in Psychology*, *9*, 2147. https://doi.org/10.3389/fpsyg.2018.02147

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. https://doi.org/10.1121/1.1908694

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*, *11*(4), 373–382. https://doi.org/10.1016/S0095-4470(19)30836-8

Lenth, R. (2020). emmeans: Estimated marginal means, aka least-squares means. R Package Version 1.5.2-1. https://CRAN.R-project.org/package=emmeans

Levitt, A., Jusczyk, P. W., Murray, J., & Carden, G. (1988). Context effects in two-month-old infants' perception of labiodental/interdental fricative contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 361–368. https://doi.org/10.1037/0096-1523.14.3.361

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, H. M., Tsao, F. M., & Kuhl, P. K. (2007). Acoustic analysis of lexical tone in Mandarin infant-directed speech. *Developmental Psychology*, *43*(4), 912–917. https://doi.org/10.1037/0012-1649.43.4.912

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412. https://doi.org/10.3758/BF03204884

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America*, *102*(3), 1864–1877. https://doi.org/10.1121/1.420092

Mulak, K. E., & Best, C. T. (2013). Development of word recognition across speakers and accents. *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence* (pp. 242–269). IGI Global. https://doi.org/10.4018/978-1-4666-2973-8.ch011

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

Nusbaum, H., & Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. Mullenix (Eds.), *Talker variability in speech processing* (pp. 109–132). Academic Press.

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), 1943–1952. https://doi.org/10.3758/s13428-019-01244-y

Pater, J., Stager, C., & Werker, J. (2004). The perceptual acquisition of phonological contrasts. *Language*, *80*(3), 384–402. https://doi.org/10.1353/lan.2004.0141

Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of Mandarin and Cantonese. *Journal of Chinese Linguistics*, *34*(1), 134–154.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

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., Knight, S., & Gerken, L. (2017). The distribution of talker variability impacts infants' word learning. *Laboratory Phonology*, *8*(1), Article 1. https://doi.org/10.5334/labphon.25

R Core Team. (2016). *R: A language and environment for statistical computing*. R Fundation for Statistical Computing.

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

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74–85. https://doi.org/10.1037/0012-1649.37.1.74

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science*, *15*(6), 732–738. https://doi.org/10.1111/j.1467-7687.2012.01175.x

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., Fu, C. S. L., Tay, Z. W., & Golinkoff, R. M. (2018). Novel word learning in bilingual and monolingual infants: Evidence for a bilingual advantage. *Child Development*, *89*(3), e183–e198. https://doi.org/10.1111/cdev.12747

Singh, L., Hui, T. J., Chan, C., & Golinkoff, R. M. (2014). Influences of vowel and tone variation on emergent word knowledge: A cross-linguistic investigation. *Developmental Science*, *17*(1), 94–109. https://doi.org/10.1111/desc.12097

Singh, L., Poh, F. L., & Fu, C. S. (2016). Limits on monolingualism? a comparison of monolingual and bilingual infants' abilities to integrate lexical tone in novel word learning. *Frontiers in Psychology*, *7*, Article 667. https://doi.org/10.3389/fpsyg.2016.00667

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. https://archive.mpi.nl/tla/elan

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

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–484. https://doi.org/10.1111/1467-9280.00485

Tardif, T., Fletcher, P., Liang, W., & Kaciroti, N. (2009). Early vocabulary development in Mandarin (Putonghua) and Cantonese. *Journal of Child Language*, *36*(5), 1115–1144. https://doi.org/10.1017/S0305000908009185

Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the switch task. *Developmental Psychology*, *55*(5), 934–950. https://doi.org/10.1037/dev0000699

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), 340–350. https://doi.org/10.1037/a0032192

van Heugten, M., Krieger, D. R., & Johnson, E. K. (2015). The developmental trajectory of toddlers' comprehension of unfamiliar regional accents. *Language Learning and Development*, *11*(1), 41–65. https://doi.org/10.1080/15475441.2013.879636

van Heugten, M., & Johnson, E. K. (2016). Toddlers' word recognition in an unfamiliar regional accent: The role of local sentence context and prior accent exposure. *Language and Speech*, *59*(3), 353–363. https://doi.org/10.1177/0023830915600471

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *The Journal of the Acoustical Society of America*, *60*(1), 198–212. https://doi.org/10.1121/1.381065

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

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), 1289–1309. https://doi.org/10.1037/0012-1649.34.6.1289

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

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. 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)

Xu, Y. (2013). ProsodyPro — A tool for large-scale systematic prosody analysis. In *Tools and resources for the analysis of speech prosody* (pp. 7–10). Laboratoire Parole et Langage. https://discovery.ucl.ac.uk/id/eprint/1406070

Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, *12*(3), 412–418. https://doi.org/10.1111/j.1467-7687.2008.00789.x

Zhang, C. (2014). *Perceptual normalization of inter-and intra-talker variations in tone categorization*. Chinese University of Hong Kong.

Zhang, C. (2018). Online adjustment of phonetic expectation of lexical tones to accommodate speaker variation: A combined behavioural and ERP study. *Language, Cognition and Neuroscience*, *33*(2), 175–195. https://doi.org/10.1080/23273798.2017.1376752

Zhang, C., Peng, G., & Wang, W. S.-Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. *Brain and Language*, *126*(2), 193–202. https://doi.org/10.1016/j.bandl.2013.05.010

Zhang, K., Wang, X., & Peng, G. (2017). Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism. *The Journal of the Acoustical Society of America*, *141*(1), 38–49. https://doi.org/10.1121/1.4973414

Received May 11, 2021
Revision received April 28, 2022
Accepted May 17, 2022 ■