Received: 3 August 2021 | Revised: 13 September 2022 | Accepted: 21 November 2022

DOI: 10.1111/desc.13357

RESEARCH ARTICLE

Developmental Science WILEY

Prosodic modulations in child-directed language and their impact on word learning

Jinyu Shi[1] | Yan Gu[1,2] | Gabriella Vigliocco[1]

[1] Department of Experimental Psychology, University College London, London, UK

[2] Department of Psychology, University of Essex, Colchester, UK

Correspondence
Yan Gu, University College London, Department of Experimental Psychology, 26 Bedford Way, London WC1H 0AP, UK.
Email: yan.gu@ucl.ac.uk

Jinyu Shi and Yan Gu are considered as joint first author.

Abstract

Child-directed language can support language learning, but how? We addressed two questions: (1) how caregivers prosodically modulated their speech as a function of word familiarity (known or unknown to the child) and accessibility of referent (visually present or absent from the immediate environment); (2) whether such modulations affect children's unknown word learning and vocabulary development. We used data from 38 English-speaking caregivers (from the ECOLANG corpus) talking about toys (both known and unknown to their children aged 3–4 years) both when the toys are present and when absent. We analyzed prosodic dimensions (i.e., speaking rate, pitch and intensity) of caregivers' productions of 6529 toy labels. We found that unknown labels were spoken with significantly slower speaking rate, wider pitch and intensity range than known labels, especially in the first mentions, suggesting that caregivers adjust their prosody based on children's lexical knowledge. Moreover, caregivers used slower speaking rate and larger intensity range to mark the first mentions of toys that were physically absent. After the first mentions, they talked about the referents louder with higher mean pitch when toys were present than when toys were absent. Crucially, caregivers' mean pitch of unknown words and the degree of mean pitch modulation for unknown words relative to known words (pitch ratio) predicted children's immediate word learning and vocabulary size 1 year later. In conclusion, caregivers modify their prosody when the learning situation is more demanding for children, and these helpful modulations assist children in word learning.

KEYWORDS
child-directed speech, infant-directed prosody, language acquisition, pitch, speaking rate, vocabulary development, word learning

Research Highlights

- In naturalistic interactions, caregivers use slower speaking rate, wider pitch and intensity range when introducing new labels to 3–4-year-old children, especially in first mentions.
- Compared to when toys are present, caregivers speak more slowly with larger intensity range to mark the first mentions of toys that are physically absent.
- Mean pitch to mark word familiarity predicts children's immediate word learning and future vocabulary size.

1 | INTRODUCTION

When talking to children, caregivers use Child-Directed Speech (CDS), characterized by an exaggerated intonation such as slower speaking rate, higher pitch, and larger pitch range (Fernald & Simon, 1984; Fernald et al., 1989; Soderstrom, 2007). A recent meta-analysis of 87 studies shows that such features seem to be found across cultures (Cox et al., 2022). The global prosodic modulations in CDS have been argued to attract children's attention (e.g., Bortfeld & Morgan, 2010; Kuhl, 2007; Segal & Newman, 2015), communicate emotions and attitudes between the caregiver and the child (Fernald, 1989, 1992; Kamiloğlu et al., 2020), and facilitate language acquisition (Cristia, 2013; Hirsh-Pasek et al., 1987; Trainor & Desjardins, 2002). In return, children's responsiveness towards CDS could further reinforce the exaggerated prosody of caregivers (Smith & Trainor, 2008). As children grow older, caregivers gradually shift towards Adult-Directed Speech (ADS) prosody (Leipold et al., 2022; Narayan & McDermott, 2016).

A large body of literature has focused on the differences between CDS and ADS (e.g., Han et al., 2020; Kalashnikova & Kember, 2020; Ma et al., 2011; Tippenhauer et al., 2020; Wang et al., 2021), but research on how prosody is modulated *within* CDS and how such modulations may affect learning remains largely unexplored. In this paper, we focus on caregivers' prosodic adjustments (speaking rate, pitch, and intensity) in communicative contexts where such adjustments could be especially useful, namely, when learning new words and when learning is more difficult because the referent is not physically present. We then assess if the presence of such prosodic modulations predicts children's immediate word learning and their long-term lexical development.

1.1 | General CDS prosody adjustment and lexical development

Child-Directed Speech is a type of hyper-speech (Lindblom, 1990) with exaggerated prosody, where the speakers continuously adjust their signal quality based on their awareness of the information required by the listeners. Specifically, caregivers automatically modify their pitch range, speaking rate, and vowel articulation, which can facilitate children's speech perception and word comprehension (e.g., Cooper & Aslin, 1990; Fernald, 2000; Han et al., 2021; Kuhl et al., 1997; Stern et al., 1983). For example, heightened pitch range and variation in pitch contours in CDS have been shown to support the discrimination between speech sounds in 6- to-7-month-old infants (Trainor & Desjardins, 2002). In addition, 12- to 31-month-old children were found to recognize target words more accurately when the stimuli were presented with a slower speaking rate (Zangl et al., 2005). Song et al. (2010) further demonstrated that 19-month-olds' ability to identify a target picture was improved when presented with typical CDS prosody compared to modified CDS that was two times faster. Interestingly, however, their performance did not differ between typical and monotonous speech with half of the original pitch range. These findings suggest that a slower speaking rate, but not a wider pitch range, can enhance word recognition.

Prosodic adjustments may also help word learning. For example, the use of more exaggerated prosody has been related to better performance in lexical acquisition tasks performed by toddlers (Cristia, 2013; Graf-Estes & Hurley, 2013; Grassmann & Tomasello, 2007; Ma et al., 2011) and increased neural activity (i.e., larger event-related potential responses) in 13-month-olds (Zangl & Mills, 2007). When looking at the effect of prosody on vocabulary, a slower speaking rate in input to 7-months-old was found to correlate with a better expressive vocabulary at 2 years of age (Raneri et al., 2020) and a higher mean pitch was related to larger productive vocabulary in 12- to-14-month-old (Porritt et al., 2014). However, Kalashnikova and Burnham (2018) did not find a correlation between exaggerated pitch in caregivers' speech and infants' expressive vocabulary at 15 and 19 months. In sum, research on how the general adjustments on each dimension of CDS prosody affect lexical development has generated mixed results, and to the best of our knowledge, no studies have looked at the effect of prosodic adjustments on both immediate word learning and long-term vocabulary size.

1.2 | Caregiver's prosody in word learning and displaced contexts

As most studies focus on the relationship between the general prosody of CDS and lexical acquisition (Han, 2019; Han et al., 2020;

Kalashnikova & Burnham, 2018), there is generally a lack of research directly linking the dynamic changes in caregiver's prosody to children's learning and language development. In particular, it is still unclear how caregivers adapt their prosody according to children's familiarity with words and the context of the interaction. If caregivers adapt their speech dynamically based on a child's lexical knowledge, they would use prosodic cues to accentuate the labels for referents that are not accessible to the child, either because the referent and the word are unknown to the child (i.e., what children need to learn), or because the referent is not visible to the child (i.e., situationally more inaccessible). Such amendment within CDS could potentially facilitate word learning. With regards to the question of whether caregivers modulate more on words that are unknown to the child, among the few studies that have examined this issue, Han and colleagues (2020, 2021) asked native Mandarin- and Dutch-speaking caregivers to read a storybook containing five unknown words (e.g., "weasel", "emu") and two known words (e.g., "apple") to their 18- and 24-months old toddlers. Both Dutch and Chinese caregivers slowed down significantly more for the unknown than known words. However, the patterns for pitch measurements were not as consistent across languages and age groups. For Chinese mothers, the unknown words were consistently produced with a higher mean pitch when addressing 18 months old children, whilst the pitch range for unknown words was enlarged for 24 months old children. For Dutch mothers, the mean pitch was raised for known words instead of unknown words. Despite speakers generally tending to use salient prosody to mark important information in the speech stream (Gussenhoven, 2004, 2016), Han et al.'s (2021) results seem to suggest language-specific use of pitch (e.g., Gussenhoven & Chen, 2000; Kitamura et al., 2001; Swerts et al., 2002). Thus, how caregivers prosodically mark unknown words is still unclear[i].

Furthermore, in naturalistic interactions, caregivers talk not only about referents that are present, but also referents that are displaced (Veneziano, 2001). Displacement, namely the ability to refer to and talk about remote objects and events, is considered to be a feature of language (Hockett, 1960). Previous research mostly investigated CDS prosody when the referents of the target words were visually available (Fernald & Mazzie, 1991; Han et al., 2020, 2021; Soderstrom, 2007). No previous study has investigated whether the context of the interaction (i.e., situational accessibility of the referents) may also influence the prosodic marking of words in CDS.

Research on information structure has long established that in a conversation, referents frequently change between three statuses through activation or deactivation: *given* (completely active in a person's focus of attention), *new* (not active at all), or *accessible* (in a person's peripheral attention, not in focus) (Chafe, 1987). Given and accessible information is also more predictable in the context as it has already been introduced to an interlocutor's attention. One way for the referent to be accessible to the listener is to simply be situationally accessible, which is being part of the physical environment (Lambrecht, 1994). Hence, compared to situated contexts, the referents would be less accessible to the children when the conversation happens in a displaced context. This means that more cognitive effort would be required for children to activate or build the association between the labels and the actual objects in displaced contexts. As studies with adults show that English speakers tend to use acoustic prominent expressions (e.g., pitch accent, longer duration, and higher intensity) for referents that are less accessible to the listener (Arnold, 2008), it is likely that English caregivers would also adapt similar prosodic pattern in displaced contexts to compensate for the increase in children's cognitive load.

Besides being present in the physical environment, a referent can also become accessible after being mentioned prior in the discourse, which is known as textual accessibility (Lambrecht, 1994). When the speaker first introduces a referent, the "new" concept changes from an inactive state to an active state and becomes "given" as the speaker continues to talk about the referent (Chafe, 1987). The textual accessibility of a referent can greatly impact prosody: in English and many Germanic languages, when a word is mentioned for the first time, it tends to be produced in a pitch-accented manner, but as the speaker mentions the same word multiple times, their production becomes unaccented (Fowler & Housum, 1987). Studies show that second mentions are shorter, quieter, lower-pitched, and less variable in pitch than first mentions (Fisher & Tokura, 1995). In CDS, Tippenhauer et al. (2020) also found that caregivers reduced word duration for repeated mentions, although the extent of the reduction was less in CDS compared to ADS. However, a study by Bortfeld and Morgan (2010) showed that first mentions were longer in duration than second mentions in CDS, but mean pitch and pitch range was not reduced. Since the study was based on 12 caregivers in the same context in a laboratory, where they were given specific words to use, it remains unclear whether similar patterns appear in naturalistic communications. The existing studies probed how textual accessibility affected speakers' prosody independently, but textual accessibility almost always interacts with situational accessibility in real-life communications. How both types of accessibility jointly influence prosody in CDS is unexplored.

1.3 | The current study

Our study investigates how English-speaking caregivers prosodically modulate their speech as a function of word familiarity (i.e., whether the word is known or unknown to the child) and accessibility of referent (i.e., whether the referent is visually present). We then assess whether any of these modulations affect children's learning of new words and their overall vocabulary development. The study uses data taken from the ECOLANG corpus of dyadic communication between English-speaking caregivers and their 3–4 years old children (Vigliocco et al., unpublished). In the corpus, caregivers and their children talked about objects that were either known or unknown to the children under two conditions: the objects were present or absent in the context. The corpus also included measures of children's vocabulary size (concurrent and 1 year later) and their scores in a recognition task of the unknown objects they played with.

We hypothesize that English caregivers adjust their prosody to reduce children's cognitive load when the referent (and label) are unfamiliar to the child, and also when the referent (being familiar or not) is inaccessible because displaced. Therefore, we should observe that they will use more exaggerated prosody for words that are unknown to the child and when the referent is absent from the interaction, especially in the first mentions.

In addition, if these modulations are successful in reducing the child's cognitive load, we should observe that children learn those unknown words that are made more salient by prosodic modulation better than those that are not as salient. Finally, if the caregiver's prosodic behavior captured in the interaction reflects their habitual style of communicating with their children, the effects should be observed not only on the specific unknown words we presented but also on their overall vocabulary size (tested 1 year later).

2 | METHOD

2.1 | Participants

The current study included 38 caregiver-child dyads from the ECOLANG corpus (Vigliocco et al., unpublished). The participants were all native English speakers (including both British and American English) recruited from the wider London area. All the caregivers (37 mothers and one father) gave consent for both themselves and their children (18 girls and 20 boys) to participate in the study. The age of the children ranged from 3 to 4.33 years (mean = 3.58 years, *SD* = 0.38). The mean age of caregivers was 38.45 years (*SD* = 3.73), ranging from 29 to 48 years old. All but one caregiver had received a university degree or above (the median educational level was 4, ranging from 3 to 6, where 1 = GCSE (equivalent to middle school qualification), 6 = doctoral degree). The study obtained ethical approval from University College London.

2.2 | Materials

In the ECOLANG corpus, each dyad included four categories of toys (six toys per category, 24 toys in total), which are comprised of animals, tools, foods, and musical instruments. The toys were selected from an initial set of 98 toys (see a full list of toy labels in the OSF file: https://osf.io/eu8y2). The categories and toys were chosen because they are common and engaging for children. For each category, three of them were known to the child and three were unknown. Prior to the recording session, parents were sent a wordlist in which they were asked to indicate which toys their children already knew (label and concept) without checking back with the children. The assignment of toys to the familiarity conditions was based on these answers. The word frequency of toy labels (van Heuven et al., 2014) was normally distributed $D = 0.057$, $p = 0.89$ ($M = 3.71$, $SD = 0.76$, range = 1.54–5.17). The mean word frequency for unknown words ($M = 3.30$, $SD = 0.69$, range = 1.54–4.76) was significantly lower than for known words ($M = 3.87$, $SD = 0.68$, range = 1.74–5.17), $t\,(137) = 4.82$, $p < 0.001$. To increase the generalizability of our study, we included all our toys without restricting words to a specific stress pattern of the same number of syllables. The mean number of syllables of the unknown words for toys in the interactions is 2.14 ($SD = 0.97$, range = 1–4) and the mean number of syllables of known is 1.93 ($SD = 0.91$, range = 1–4).

A recognition test with 24 target test trials (identification of the unknown objects they had talked about) and four control trials (identification of the known words) was administered for each child to test how much they had learned from the interaction. In each trial, two pictures, either both unknown toys or both known toys (control trials) used in the interaction, were presented on the computer screen side by side. Before each trial, a cartoon puppy appeared on the screen and asked a question such as "Can you help me to find the [pomegranate], where is the [pomegranate]?". The stimuli were pre-recorded by the same female native British English speaker in a sound-proof room at two different time points. The average mean intensity of the stimuli recordings was 66.98 dB (range 64.04–73.28, *SD* = 2.58). The mean intensity for the recordings children heard during the recognition test was 68.03 dB (range 65.66–69.97, *SD* = 1.08). The intensity variations of sound stimuli may be due to the different distances between the speaker's mouth and the microphone. The British Picture Vocabulary Scale 3rd edition (BPVS3) was used to test the children's vocabulary size (Dunn & Dunn, 2009).

2.3 | Design and procedure

The recordings took place at participants' homes. Before the interaction began, the BPVS3 was administered, and caregivers were given pictures of the toys to familiarize themselves with them. During the interaction, the caregiver and the child were sitting 90 degrees from each other around a table. The interaction was videotaped and the speech of the caregiver was also recorded using a clip-on microphone via Audacity at a sample rate of 44.1 kHz, 32 bit.

During the session, the caregiver was asked to talk about the known and unknown toys of each category in a natural way (i.e., how they usually talked to their children) in both toy-present and toy-absent conditions. The sequence of these two conditions was counterbalanced across participants. In the toy-present condition, the experimenter placed six toys from one category (e.g., animal) on the table and left the room. The caregiver and the child talked about and interacted with the toys for 3–4 min. The experimenter then reentered the room, asked the child to help tidy up the toys, and left the room with the toys. In the toy-absent condition, the experimenter asked the caregiver to continue to talk about the toys they just played with (when toy-present was first) or the toys that were about to come (when toy-absent was first) for 3–4 min. Labels of the toys were provided as a reminder to the caregivers. This process was repeated for all four categories, resulting in eight sessions in total for each dyad. The whole recording lasted approximately 30–40 min.

After the interaction, children carried out the recognition test presented in E-prime with a laptop (Acer model, 1366 × 768 pixels). The child was presented with two pictures and asked to point to the target picture that matched the word they heard. The average size of the pictures was about 320 × 394 pixels. The two pictures were vertically centered (Y = 40%) and one horizontally aligned to the left (X = 25%) and the other to the right (X = 75%) on the screen. Children were not under time pressure during the task, so the trial durations were the time children took to make their decision. Usually, children responded immediately after they were presented with the stimuli. After the child made their decisions, the experimenter clicked the mouse to record the response and proceeded to the next trial. A demo video of an example trial of the recognition test can be found in osf.io/8ymrg.

For 32 (out of 38) of the children participating, a second BPVS vocabulary test was administered approximately 1 year (*M* = 1.02 year, *SD* = 0.056) after the interaction (mean age of children = 4.58, *SD* = 0.41, range 4.0–5.4 years).

2.4 | Speech transcriptions and coding

The manually transcribed speech was automatically segmented and aligned to words via the Munich Automatic Segmentation System (MAUS, Poerner & Schiel, 2018) and manually corrected by an expert coder. In total, we obtained 6658 target referents. Segments of the caregivers' speech overlapping with the children's speech or background noise (e.g., instrument playing) were excluded (*N* = 129), resulting in 6529 target referents for analyses. All boundaries of target words were checked and corrected by an experienced coder with Praat (Boersma & Weenink, 2019). An additional coder examined 10.5% of the data (*N* = 683) to access the reliability of the boundaries of target words. This coder agreed on boundaries on 71.4% of the tokens (*N* = 488). For the tokens (*N* = 195) with inconsistent boundaries, there was no significant difference in word duration, *t* (388) = 0.819, *p* = 0.21. Hence, the original coding was judged as reliable.

For each target word we annotated: (1) its familiarity (known or unknown); (2) object presence (present or absent); (3) the number of mentions of the referent, that is, first or subsequent mentions (Arnold, 2008; Lam & Watson, 2010, 2014); (4) positions of the target word in the utterance (initial, medial, final, isolation) (Butler & Frota, 2018; Han et al., 2021; Johnson et al., 2014; Martin et al., 2016); (5) word frequency (Gahl, 2008) of each target word in British National Corpus (BNC) (SUBTLEX-UK data set, van Heuven et al., 2014), as well as in child-directed language from the CHILDES (Sanchez et al., 2019) that was used as robustness checks; (6) number of syllables; (7) utterance type where the target word was produced (including statement, yes-no question, wh-question and single referent); (8) session repetition (given that in the object present and absent conditions the caregivers talked about the same categories of toys (e.g., animals), the first session (e.g., present condition) was coded as new, the subsequent session (e.g., absent condition) was coded as repeated); and (9) session sequence (present first or absent condition first).

2.5 | Prosodic measurements

In this study we did not investigate the phonology of intonation but focused on the phonetic aspect of prosody. A Praat script was used to extract duration (ms), F0 (Hz, pitch) and intensity (dB) of target referents to obtain the following measurements:

Speaking rate: the mean number of syllables per second. The value of speaking rate was log-transformed before the analysis.

Pitch: mean F0 and F0 range (F0 maximum–F0 minimum). We manually checked each F0 value obtained from Praat and corrected pitch errors and mis-tracked points. F0 values were transformed to semitones from observed Hertz values with 50 Hz as the reference, using the formula Semitone = 12* log2(target Hertz/50) (Tang et al., 2017). F0 range was calculated in semitones using the formula of 12*log2(F0 maximum/F0 minimum).

Intensity: mean intensity and intensity range (intensity maximum–intensity minimum).

2.6 | Unknown words recognition and vocabulary measurements

The recognition result of one participant was excluded as 37.5% of the target trials were missing. For the rest of the participants, all but two (missed one and four trials out of 24 target trials) completed the recognition test. The BPVS3 was coded for each child to generate a raw score for the concurrent vocabulary size and a raw score for the vocabulary size 1 year later.

2.7 | Data analysis

1. *Do caregivers adjust their speaking rate, pitch and intensity?*

We used linear mixed-effects models in the R environment (R Core Team, 2021) to assess which factors influence caregivers' speaking rate, pitch, and intensity of target words (dependent variables). Familiarity of the label (known/unknown), presence or absence of the object, and the number of mentions of the target word (first/subsequent) were included as fixed effects, as well as their two-way and three-way interactions. Additionally, the fixed effect of age of the child (in months), word position in the utterance, target word utterance type, target word frequency in English, number of syllables of the target words, session repetition and present/absent session sequence were included in the model as control variables. To account for the individual differences of participants and item differences of target words (e.g., different stress patterns), we included participants and words as two grouping variables for the maximal random effects structure of the models. For both participants and words, the structure consists of a random intercept, and random slopes of presence/absence, word familiarity, and their interaction. The formula of our regression model is:

$$Cu\,e_{ij} = X_{ij}\beta + S_{ij}s_i + W_{ij}w_j + \varepsilon_{ij}$$

where vector Cue_{ij} represents the cue responses of individual i to referent j

$$\text{Matrix } X_{ij} = (1\ PrAb_{ij}\ Familiarity_{ij}\ NuMenBin_{ij}\ PrAb_{ij} * Familiarity_{ij}\ PrAb_{ij} * NuMenBin_{ij}$$

$$Familiarity_{ij} * NuMenBin_{ij}\ PrAb_{ij} * Familiarity_{ij} * NuMenBin_{ij}\ initial_{ij}\ medial_{ij}$$

$$isolation_{ij}\ sessionRepeat_{ij}\ yn_{ij}\ wh_{ij}\ SingleRef_{ij}\ Age\_month_i\ Syllables_j\ frequency_j)$$

$$\text{Fixed effects : Vector } \beta = (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_{17})'$$

$$\text{Matrix } S_{ij} = (1\ PrAb_{ij}\ Familiarity_{ij}\ PrAb_{ij} * Familiarity_{ij})$$

$$\text{Random effects : Vector } s_i = (s_{i0} \quad s_{i1} \quad s_{i2} \quad s_{i3})'$$

$$\text{Matrix } W_{ij} = (1\ PrAb_{ij}\ Familiarity_{ij}\ PrAb_{ij} * Familiarity_{ij})$$

$$\text{Random effects : Vector } w_j = (w_{j0} \quad w_{j1} \quad w_{j2} \quad w_{j3})'$$

ε_{ij} is the error term

We constructed maximal mixed-effects models in the first instance. When the maximal model did not converge, we ran the models with different optimizers that allow convergence or reduced the model complexity by removing the correlation between slope and intercept or random slope of interaction. Data of all independent variables were mean-centered before analyses. Data files in .csv format and all analysis scripts and output can be found at: https://osf.io/ykcvf/.

2. *Do prosodic modulations predict learning?*

We used multiple logistic regressions to assess the effect of caregivers' speaking rate, pitch, and intensity cues on children's immediate unknown word recognition and long-term vocabulary (BPVS3) size. The variables investigated were caregivers' speaking rate, mean pitch, pitch range, mean intensity, and intensity range of unknown words (Han, 2019), and the degree of modifications (i.e., differences between known and unknown words) for these measurements. For the degree of modification, a ratio was computed for each caregiver's speaking rate, mean pitch, pitch range, mean intensity and intensity range, respectively. For example, the speaking rate ratio = (The mean speaking rate for known referents)/(The mean speaking rate for unknown referents). We reasoned that the ratios capture the relative salience of a word in the overall speech (e.g., accentuating the unknown words while deaccenting the known words). For the analysis related to recognition test results, caregivers' average number of mentions for the unknown words, children's concurrent vocabulary size and age, as well as the other related prosodic cues were controlled for. When looking at children's long-term vocabulary development, we included children's concurrent vocabulary size as a control variable to account for the initial individual differences.

3 | RESULTS

3.1 | Caregivers adjust their speaking rate, pitch and intensity

The first set of research questions addressed in this study was if caregivers would use prosodic cues to mark: (a) word familiarity and (b) toys presence (absent vs. present), as well as whether such cues displayed any differences as a function of the textual accessibility–measured by mentions (1st vs. subsequent).

Table 1 presents the mean of prosodic measurements of both known and unknown target referents, and referents in the absent and present conditions. On average, known target words (M = 7.01 times, SD = 3.93) were mentioned to a similar extent as the unknown target words (M = 7.30 times, SD = 4.43). However, target referents were mentioned a larger number of times in the absent condition (M = 4.05 times, SD = 2.77) compared to present condition (M = 3.11 times, SD = 2.52). Figure 1 shows the mean for the prosodic measurements of the words' first and subsequent mentions splitting up into different

TABLE 1 The mean and standard deviation of the prosodic measures per each condition

	Known (SD)*N* = 3199	Unknown (SD)*N* = 3330	Present (SD)*N* = 2833	Absent (SD)*N* = 3696
Speaking rate	4.00 (1.77)	3.86 (1.76)	3.96 (1.74)	3.91 (1.79)
Mean F0 (ST)	24.62 (5.32)	25.86 (5.00)	25.02 (5.23)	24.53 (5.09)
F0 range (ST)	6.71 (5.15)	8.15 (5.27)	7.63 (5.30)	7.30 (5.22)
Mean intensity (dB)	65.63 (7.19)	65.16 (6.60)	66.02 (6.57)	64.91 (7.10)
Intensity range (dB)	26.22 (7.64)	28.90 (7.42)	27.71 (7.61)	27.49 (7.67)

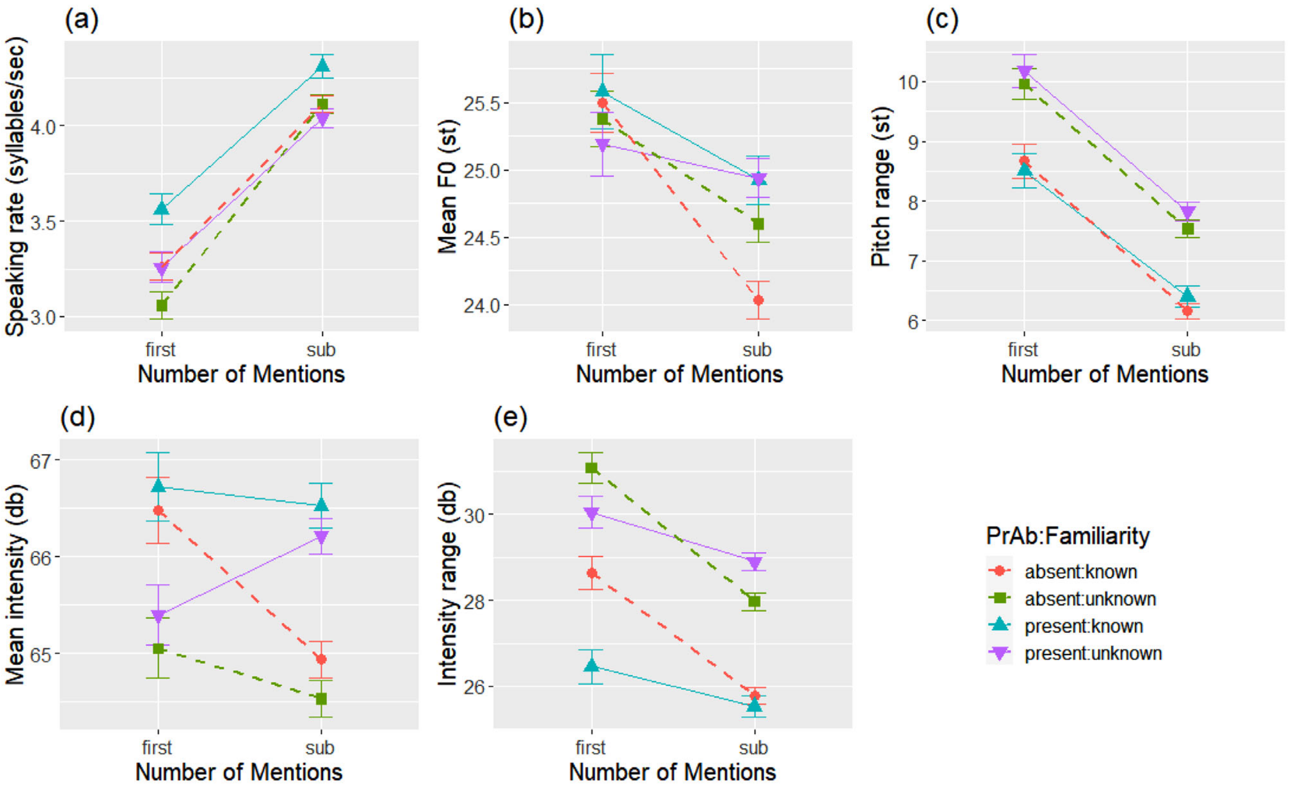

FIGURE 1 Average values of (a) speaking rate, (b) mean pitch, (c) pitch range, (d) mean intensity and (e) intensity range for caregivers' production of target referents under each condition based on the number of mentions (first vs. subsequent). Error bars represent the standard error of the mean.

conditions. The model results for the different prosodic measures are summarized in Table 2.

3.1.1 | Familiarity

Results from the mixed effect models showed that unknown words had a significantly wider pitch range than known words (β = 0.58, p = 0.008, 95% CI [0.17, 1.02]), regardless of toy presence and the number of mentions. For speaking rate and intensity range, there were significant interactions between word familiarity and the number of mentions. The breakdown of the interactions revealed that the differences between unknown and known words mainly appeared in the first mentions (speaking rate: β = –0.038, p = 0.00072, 95% CI [–0.060, –0.17]; intensity range: β = 1.15, p = 0.032[ii], 95% CI [0.11, 2.22]) whereas differences became only marginally (speaking rate: β = –0.018, p = 0.063, 95% CI [–0.038, 0.001]) or not significant (intensity range: β = 0.35, p = 0.47, 95%CI [–0.59, 1.32]) in subsequent mentions. As for the mean pitch and mean intensity, there were no significant main effects of word familiarity or any two/three-way interactions (all p > 0.20). In short, caregivers used a wider pitch range (generally), as well as a slower speaking rate and larger intensity range (first mentions) to mark word familiarity.

3.1.2 | Accessibility

There were significant two-way interactions between toy-presence and the number of mentions (first vs. subsequent) in speaking rate, mean intensity, and intensity range. For both speaking rate and

TABLE 2 Results of analyses on different prosodic measures

	Speaking rate	Mean pitch	Pitch range	Mean intensity	Intensity range
Presence.ct	0.019**	0.38	−0.067	1.14***	−0.23
Familiarity.ct	−0.023*	0.12	0.58**	0.19	0.52
NumMentions.ct	0.074***	−0.84***	−2.04***	−1.43***	−1.92***
Age.ct	0.003.	−0.097	−0.18***	−0.22	−0.16
Frequency.ct	0.033**	0.024	−0.28.	0.75.	−0.71
Initial.ct	0.10***	1.92***	−0.84**	2.71***	0.27
Medial.ct	0.11***	0.46***	−1.79***	1.68***	−2.33***
Isolation.ct	−0.022**	−0.28	−0.48.	−0.35	1.14**
Syllables.ct	0.16***	0.001	0.83***	0.65*	−0.24
Yes-no Question.ct	0.029***	1.46***	−0.14	0.35*	0.21
Wh-question.ct	0.030***	−0.31	−0.74***	−0.71***	−0.30
Single referent.ct	−0.035***	1.02***	0.58**	1.15***	0.88**
Session repetition.ct	0.010.	0.037	−0.25	0.66*	−0.30
Presence.ct: NumMen.ct	−0.024** First: 0.037*** Sub: 0.013.	0.37 First: 0.11 Sub: 0.47.	0.15	1.43*** First: 0.21 Sub: 1.50***	1.85*** First: −1.42*** Sub: 0.21
Familiarity.ct: NumMen.ct	0.020** First: −0.038*** Sub: −0.018.	0.24	−0.12	0.28	−0.89* First:1.15* Sub: 0.35
Presence.ct: Familiarity.ct	−0.004	−0.32	0.14	0.26	0.69
Presence.ct: Familiarity.ct:NumMen.ct	−0.014	−0.41	−0.15	−0.75	−0.17

Note: The number stands for an estimate,$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $.p < 0.1$

intensity range, the breaking down of the interactions revealed that, the significant differences between toy-present and toy-absent conditions were mainly in the first mentions (participants spoke slower, $\beta = 0.037$, $p < 0.001$, 95%CI [0.021, 0.053], with wider intensity range, $\beta = -1.42$, $p < 0.001$, 95%CI [−2.11, −0.74] in toy-absent condition) but became only marginally significant (speaking rate: $\beta = 0.013$, $p = 0.053$, 95%CI [0.0001,0.026]) or not significant in the subsequent mentions (intensity range: $\beta = 0.21$, $p = 0.42$, 95%CI [−0.29, 0.72]). However, for mean intensity, the difference between toy-present and toy-absent was only significant in subsequent mentions ($\beta = 1.50$, $p < 0.001$, 95%CI [0.93, 2.07]) instead of the first mention ($\beta = 0.21$, $p = 0.55$, 95%CI [−0.48, 0.90]), such that referents mentioned in the toys-present condition were spoken with significantly higher mean intensity. Similarly, the mean pitch of the words in the toys-present condition tended to be higher in the subsequent mentions only (marginally significant, $\beta = 0.47$, $p = 0.076$, 95%CI [−0.03,0.97]). There was no difference in pitch range between toy-present and absent conditions and there were no significant interactions between children's familiarity and the toy-presence condition or any three-way interactions across the cues (all $p > 0.10$). In general, compared to the toy-present condition, caregivers spoke slower with a larger intensity range when the toys were mentioned in the absent condition for the first time, whereas after the first mention, caregivers talked about the referents louder with a higher pitch when the toys were present in the interaction.

Furthermore, the number of mentions was significant itself: After the first mention, there was a rapid increase in speaking rate ($\beta = 0.074$, $p < 0.001$, 95% CI [0.067, 0.082]) and a sharp decrease in mean pitch ($\beta = -0.84$, $p < 0.001$, 95% CI [−1.09, −0.58]), pitch range ($\beta = -2.04$, $p < 0.001$, 95% CI [−2.32, −1.76]), mean intensity ($\beta = -1.43$, $p < 0.001$, 95% CI [−1.72, −1.14]) and intensity range ($\beta = -1.92$, $p < 0.001$, 95% CI [−2.27, −1.57]).

Caregivers' prosody was also affected by several control variables, especially the type of utterances and word position of a target referent. For example, comparing to statements, yes-no questions are produced faster ($\beta = 0.029$, $p < 0.001$, 95% CI [0.021, 0.038]) with a higher mean pitch ($\beta = 1.46$, $p < 0.001$, 95% CI [1.16, 1.75]) and higher mean intensity ($\beta = 0.35$, $p = 0.040$, 95% CI [0.017, 0.69]), and wh-questions had a faster speaking rate ($\beta = 0.030$, $p < 0.001$, 95% CI [0.019, 0.041]), smaller pitch range ($\beta = -0.74$, $p < 0.001$, 95% CI [−1.13, −0.35]), as well as a lower mean intensity ($\beta = -0.71$, $p < 0.001$, 95% CI[−1.12, −0.30]). Single labels were produced with slower speaking rate ($\beta = -0.035$, $p < 0.001$, 95% CI [−0.046, −0.023]), higher mean pitch ($\beta = 1.02$, $p < 0.001$, 95% CI [0.62, 1.42]) and mean intensity ($\beta = 1.15$, $p < 0.001$, 95% CI [0.70, 1.60]), wider pitch range ($\beta = 0.58$, $p = 0.008$, 95% CI [0.15, 1.01]) and intensity range ($\beta = 0.88$, $p = 0.001$, 95% CI [0.34, 1.42]). Notably, referents in utterance final position had a slower speaking rate, wider pitch range, but lower mean pitch and mean intensity than those in utterance initial and medial positions (all $p < 0.05$). Additionally, there were significant influences of BNC English word frequency on speaking rate ($\beta = 0.033$, $p = 0.0012$, 95% CI [0.014, 0.053]), but the influence was not significant for other cues ($p > 0.05$). The results were generally robust using CHILDES word frequency except in speaking rate CHILDES word frequency was no longer significant whereas word familiarity became more significant (See details in Appendix A, Table A1). As for the influence of children's age, there were two interesting results: first, caregivers talked about known toys significantly faster for older children ($\beta = 0.0030$, $p = 0.047$, 95% CI [0.0001, 0.006]), but their speaking rate for unknown words was not significantly faster ($\beta = 0.0015$, $p = 0.37$, 95% CI [−0.002, 0.005]). Second, caregivers used a wider pitch range for younger children ($\beta = -0.18$, $p < 0.001$, 95% CI [−0.27, −0.09]). In sum, results from control factors showed that the prosody of a word was more exaggerated when it was mentioned for the first time, produced at the final position of an utterance or in isolation, and when addressing younger children. For details see the output of regression models in OSF link.

3.2 | Prosodic modulations predict learning

On average, children answered a proportion of 0.83 test trials correctly in the recognition test (range: 0.625–1.0, $SD = 0.11$). The mean raw BPVS score children received at the time of the interaction is 61.95 (range = 30–87, $SD = 13.44$), and the mean score is 82.81 when collected 1 year after the interaction (range = 65–99, $SD = 8.31$).

3.2.1 | Recognition accuracy

First, we ran logistic regression analyses of children's word recognition results on the prosodic measures for unknown words. In addition to the control variables of children's age, concurrent vocabulary size and the average number of mentions of the unknown words by caregivers, we included all prosodic cues as independent variables except unknown mean intensity (highly correlated with both unknown pitch range ($r = 0.60$, $p < 0.001$) and intensity range ($r = 0.73$, $p < 0.01$), see a correlation matrix in Appendix B, Figure B1). There was no multicollinearity between variables in the model (all VIFs < 1.8). We used the step.model function in R to find the best-fit model, the results of which showed that only the mean pitch of unknown words ($\beta = 0.096$, $p = 0.0043$, 95% CI [0.02, 0.16]) and children's concurrent vocabulary size ($\beta = 0.023$, $p = 0.008$, 95% CI [0.01, 0.04]) were significantly correlated with the unknown word recognition outcome. Figure 2a plotted the predicted proportion of correct responses based on the mean pitch for unknown words. An independent analysis with unknown mean intensity revealed that it was not a significant predictor of recognition outcome ($p = 0.76$).

Second, for the analyses of the degree of modification (ratio between known and unknown words), we included ratios of all prosodic cues (see a correlation matrix in Appendix C, Figure C1) and control variables. The best fit model only consisted of mean pitch ratio ($\beta = -6.14$, $p = 0.0049$, 95% CI [−10.40, −1.85]), intensity range ratio ($\beta = -3.36$, $p = 0.0448$, 95% CI [−6.67, −0.09]) and children's concurrent vocabulary size ($\beta = 0.02$, $p = 0.0047$, 95% CI [0.006, 0.34]. There was no multicollinearity between the three variables (all VIFs < 1.08). Effects were robust when we independently analysed the mean pitch ratio ($\beta = -6.48$, $p = 0.0026$, 95% CI [−10.70, −2.25]) and the intensity range ratio ($\beta = -3.77$, $p = 0.023$, 95% CI [−7.05, −0.52]). Figure 2b plotted the predicted proportion of correct responses based on the mean pitch ratio. All details of analyses can be found in the OSF file. To summarize, caregivers' use of a higher pitch for unknown words, and greater adjustment in pitch and intensity range between known and unknown words facilitated immediate recognition of unknown words[iii].

3.2.2 | Predicting vocabulary size 1 year later

Furthermore, we examined whether prosodic cues of unknown words and the modification between known and unknown words (ratio) may have a long-term impact on children's vocabulary size 1 year later. For prosodic measures of the unknown words, none of the cues were significant. However, for the ratios, results of the best fit model revealed that caregivers' mean pitch ratio predicted children's vocabulary size after 1 year ($\beta = -83.26$, $p = 0.0125$, 95% CI [−147.07, −19.44]), while controlling for children's concurrent vocabulary score ($\beta = 0.25$, $p = 0.0165$, 95% CI [0.05, 0.45]), speaking rate ratio ($\beta = 11.72$, $p = 0.0996$, 95% CI [−2.38, 25.82] and intensity ratio ($\beta = 95.16$, $p = 0.16$, 95% CI [−38.64, 229.56]). Figure 2c plotted the predicted long term vocabulary size based on the mean pitch ratio. There was no multicollinearity between these variables (VIFs < 1.24). When pitch ratio was analyzed independently the effect was also significant ($\beta = -79.01$, $p = 0.0177$, 95% CI [−38.64, 229.56]. These results showed that the greater adjustment in mean pitch for the unknown words was also related to long-term vocabulary gains.

4 | DISCUSSION

The current study investigated how caregivers' prosody in their input to 3-to-4-year-old children varies according to contextual factors and whether such modulations support word learning at the moment and in the long term. We first included three primary dimensions (i.e., speaking rate, pitch, and intensity) of prosody and conducted comparisons based on children's familiarity with the target words, the availability of the referent, and the textual accessibility of the words. Second, we analyzed whether caregivers' prosodic modulations in speaking rate and pitch for unknown words predict children's immediate learning of these words and long-term vocabulary size.

4.1 | Caregivers' prosodic adjustments

4.1.1 | Word familiarity

We found that caregivers used a larger pitch range, slower speaking rate and wider intensity range (at least for first mentions) for

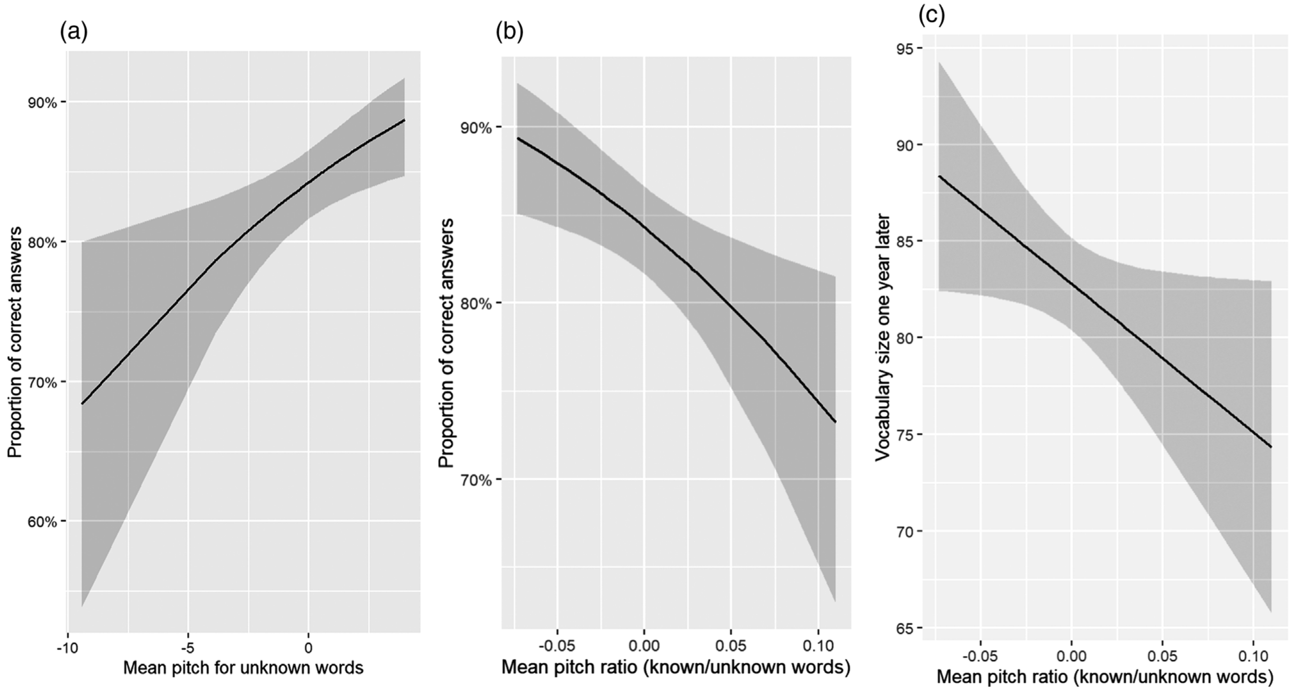

FIGURE 2 Model prediction of recognition test outcome is based on: (a) mean pitch and (b) mean pitch ratio. Model prediction of children's vocabulary size 1 year later is based on mean pitch ratio (c).

words unknown to their children compared to words that the child already knew. Crucially, the effect of word familiarity still held even when the factor of word frequency had been controlled for, indicating that these prosodic modulations do not only reflect speakers' processes (i.e., less frequent words take longer to retrieve and produce), but also the dynamic adjustments that caregivers make according to their perception of children's lexical knowledge (e.g., Han et al., 2020, 2021).

One may question whether these results are affected by repetition of the objects across the two sessions. When an object was repeatedly talked about in the interactions, children may have already learned a previously unknown word just from listening to their caregivers telling them. While it is impossible to track the live status of familiarity of each unknown word at the moment of each mention, if the label was already learnt at the time of mention, this would reduce the difference between known and unknown words, hence our work provides a conservative test of the effect of familiarity.

We failed to observe an effect of familiarity for mean pitch, which is an aspect of prosody commonly associated with word learning (Soderstrom, 2007). This result contrasts with the findings reported in Han et al. (2020), where caregivers produced words that were unfamiliar to children using higher mean pitch. A more detailed inspection of the data revealed that the absence of such a main effect could be explained by the large individual differences in caregivers' modulation of mean pitch. To better understand this, we divided caregivers into two groups according to their pitch ratio of known/unknown words (see Figure 3): One with ratio < 1 (23/38) and the other with ratio > 1 (15/38). Reanalyzing the pitch data using the same mixed-effects models showed that the ratio < 1 group had a significantly higher mean pitch for unknown words than known words ($\beta = 0.64$, $p = 0.007$) whereas the ratio > 1 group had a significantly lower mean pitch for unknown words than known words ($\beta = -0.64$, $p = 0.026$)[iv]. Such differences did not seem to be motivated by children's age as age was not a significant predictor for mean pitch in our sample. A large variability in prosody between caregivers was also found in previous studies by Narayan and McDermott (2016), who reported that the general trend for the exaggerated prosody of CDS was not always present in each caregiver.

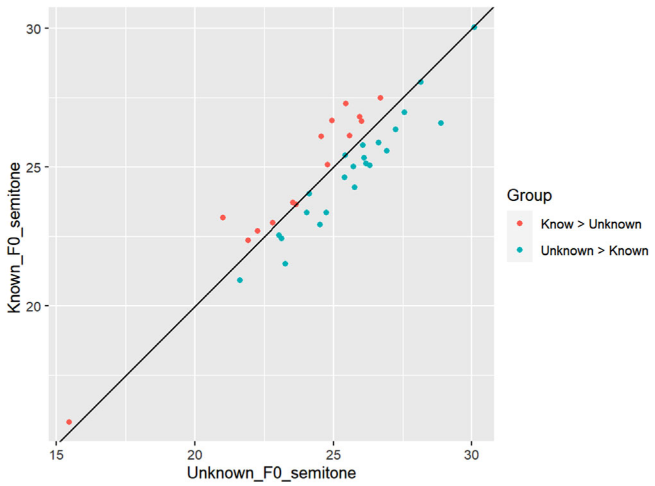

FIGURE 3 The individual differences in caregivers' mean pitch. The line represents mean pitch ratio of known/unknown words = 1. Caregivers below the line produced known words with higher pitch and those above the line produced unknown words with higher pitch.

4.1.2 | Physical accessibility

The effect of toy presence revealed interesting patterns in two different ways. On the one hand, caregivers produced the label for a toy significantly slower and with a larger intensity range when the toy was absent from the environment than when the toy was on the table (mainly for the first time), showing that they indeed use speaking rate and intensity range to compensate for the physical inaccessibility. On the other hand, contrary to our prediction, for mean intensity and mean pitch, there were no differences between the toy-present and absent conditions in the first mentions. Instead, caregivers tend to produce words with lower mean intensity and mean pitch in the toy absent condition for the non-first mentions.

At first glance, such differences in loudness and mean pitch are quite unexpected. It could be that increased noise during play with toys created a need for louder speech, as caregivers may recognize the more optimal learning moment when the toy is present and optimize their speech (higher mean intensity; mean pitch) to ensure their child can take advantage of this moment. However, this would be difficult to reconcile with results that differences only happened after the first mentions and also be difficult to reconcile to the pattern of speaking rate and intensity range. Alternatively, changes in caregivers' emotional state across different contexts (affecting their prosody (Kamiloğlu et al., 2020), especially in terms of loudness and pitch) might be responsible for these results. Previous studies have demonstrated that positive emotions with high arousal, including happiness, interest, joy, and amusement, consistently led to increased mean intensity and pitch in a speaker's voice compared to neutral vocalizations (Kao & Lee, 2006; Laukka et al., 2016; Rao et al., 2013). In our study, it is possible that caregivers were more engaged when the toys were present than when they were absent, leading to the difference in prosody. Nevertheless, the emotional effect may have been masked in the first mentions, when caregivers were likely to increase loudness and pitch and slow down speaking rate to mark physical inaccessibility of toys. Consequently, their mean intensity and mean pitch for the first mentions in toy-absent conditions could be as high as that in the toy-present condition (speaking rate and intensity range were even more salient in the absent condition). After the first mentions, with the increased accessibility of referents, the effect of emotion on prosody in the toy-present condition became more prominent.

4.1.3 | First versus subsequent mentions

In our study, we used the number of mentions (first vs. subsequent) for each word as an index of information structure and found a robust and consistent reduction in all five aspects of prosody when caregivers repeated the words. These results are consistent with past research showing that the accessibility of a referent can greatly impact word duration and that speakers tend to produce a word with a pitch accent when it is mentioned for the first time, but deaccent the same word in subsequent mentions (e.g., Fisher & Tokura, 1995; Fowler & Housum, 1987). Even in child-directed speech, caregivers reduced word duration for second mentions (Bortfeld & Morgan, 2010), although the extent of the reduction has been shown to be smaller in CDS than ADS (Tippenhauer et al., 2020).

However, our results for pitch are inconsistent with the findings of Bortfeld and Morgan's (2010) study, in which caregivers did not reduce the mean pitch or pitch range significantly in the second mentions compared to the first mentions. It has been argued that caregivers' use of sing-song pattern (exaggerated pitch in repeated mentions) keeps infants interested, rather than merely easing processing load. Our study differs from Bortfeld and Morgan's in at least three aspects: first, the age of children in our study (36–52 months) is much older than the 9-month-old infants in theirs. Particularly, we show that caregivers significantly increase their speaking rate for known words (but not for unknown words) and decrease their pitch range generally for older children. This suggests that caregivers take the children's age into account in dynamically adjusting their prosodic modulations. Second, we studied semi-naturalistic interactions, whereas the other was an experiment carried out in a laboratory where mothers saw a cue card and were instructed to use the nouns and verbs provided. Third, the number of observations is different. We had 38 caregivers of about 6500 observations, whereas Bortfeld and Morgan's (2010) findings were based on 12 caregivers of 648 observations. In short, the purpose and effect of caregivers' prosodic enhancement and reduction in response to lexical familiarity and accessibility may be influenced by various other factors such as children's age, experiment environment, and study design, etc.

4.2 | Impact on learning

Caregivers in our sample displayed individual differences in the use of mean pitch to mark word familiarity. If prosodic modulations on unknown words can potentially facilitate learning, such individual differences affected children's short-term retention of the words and long-term vocabulary size.

Children's recognition results for the unknown words were significantly predicted by both mean pitch of the unknown words and the extent to which caregivers varied their mean pitch of unknown words relative to known words (mean pitch ratio). Specifically, the higher the mean pitch for unknown words, and the greater the adjustment in mean pitch for unknown words relative to known words, the better the immediate learning results are. Crucially, mean pitch ratio also predicted children's vocabulary score 1 year after the interaction, and such prediction still held even when the initial differences in children's vocabulary sizes were accounted for. These results indicated that the use of high(er) mean pitch for unknown words, particularly its relative exaggeration has potential facilitation effects on children's lexical acquisition[v]. Although mean pitch for known and unknown words are highly correlated ($r = 0.92$, $p < 0.001$), when the unknown words are made more salient in the caregiver's speech via pitch adjustments, they attract children's attention and consistent modulation is beneficial for children's learning of words (e.g., Nencheva et al., 2021). However, using pitch invariantly or in a reverse pattern (more emphasis on known words than unknown words) may not help to support word learning as children might find it hard to utilize the prosodic pattern as a cue. Therefore, our study provides new insight into what specific aspect of the prosodic modulation relates to children's lexical development.

Furthermore, there was a relationship between intensity range ratio and unknown word recognition, but this result needs to be interpreted with caution as it was not robust in the long-term prediction. We did not find evidence for a significant correlation between children's word learning results and the ratio of speaking rate and pitch range as reported by Han (2019). This may be due to the different measurements used. Han (2019) focused on the difference between CDS and ADS, whilst we captured how much the unknown words were highlighted with each prosodic cue. In addition, since all the children in our study are above the age of three, they may not rely on speaking rate and pitch range as much as younger infants. For example, Raneri et al. (2020) found that caregivers' speaking rate to infants at seven months but not at 2 years predicted children's expressive vocabulary size, and Song et al. (2010) showed that 19-month-olds' word recognition did not improve with a wider pitch range. In addition, instead of being passive receivers, children themselves have started to take an active role in the interaction in our study. Their increasing participation can not only elicit different responses from the caregivers (Smith & Trainor, 2008), but also influence children's own language development, which may mask the effect of some prosodic cues. In future work, it would be interesting to explore the potential effect of children's production on caregivers' prosody, and study children's word learning based on the interaction between caregivers and children.

Additionally, we found that sentence type and word position affected the prosody but we did not study the extent to which these factors might have impacted learning. A recent study has shown that sentence type can affect word learning and vocabulary (Dong et al., 2021) while results on the word position seem to be mixed (e.g., Keren-Portnoya, et al., 2019; Lew-Williams et al., 2011). Future research should examine the joint impact of multiple factors on learning together and quantify the respective role and weight of each cue.

5 | CONCLUSIONS

The current study focused on the prosodic pattern of CDS in different contexts and its effect on children's word learning. We found that caregivers dynamically modify their prosody to mark word familiarity and accessibility. When speaking about unknown toys for the first time, they use a slower speaking rate, wider pitch range, and larger intensity range. When talking about toys absent from the environment for the first time, they use a slower speaking rate and wider intensity range. In addition, there is large individual variability in caregivers' use of mean pitch to mark word familiarity, and such differences predict children's word learning and vocabulary size. Our findings support the idea that speakers are aware of the listener's mental model and are constantly amending their speech based on such awareness. The findings provided a more comprehensive understanding of the prosodic qualities of CDS and how these affect word learning in childhood.

AUTHOR CONTRIBUTION

Jinyu Shi: speech transcription and annotation; data preparation; contribution to analysis; preparation of initial draft and revisions. Yan Gu: conceptualization; data collection and preparation; formal analysis; supervision of Jinyu Shi and other coders' annotations; data analysis and visualization; writing of the initial manuscript and revisions. Gabriella Vigliocco: conceptualization; development of the protocol for ECOLANG data collection; supervision; contribution to the writing and editing of the manuscript.

ACKNOWLEDGEMENTS

We are grateful for the hard work of members, student assistants and interns of the ECOLANG project. We particularly thank Ricarda Brieke, Beata Grzyb and Margherita Murgiano for their help with the data collection and preparation. We thank Mingtong Li and Shiyu Dong for their work on the intercoder reliability and pitch error correction. We thank Mengru Han for her useful input on the study and Yeqiu Zheng for her endless help with the statistics. We also thank the four reviewers whose valuable and constructive feedback helped improve this manuscript. The work reported here was supported by a European Research Council Advanced Grant (ECOLANG, 743035), a Royal Society Wolfson Research Merit Award (WRM\R3\170016) to Gabriella Vigliocco, and a Rubicon Grant from the Netherlands Organization for Scientific Research (NWO) (019.182SG.023) to Yan Gu.

CONFLICT OF INTEREST

We have no conflicts of interest to disclose

DATA AVAILABILITY STATEMENT

Data files in .csv format and all analysis scripts can be found at: https://osf.io/ykcvf/. The links provided by the author to their available data connect to the relevant actual data as described in the Data Availability Statement .

ORCID

Yan Gu https://orcid.org/0000-0001-6093-3919

ENDNOTES

[i] There are several caveats to Han's studies. First, only a few items were used questioning the generalizability of the findings. Second, the modulations of speaking rate could have been driven by caregivers' own familiarity with the words instead of children's lexical knowledge. This is because known words are usually of higher frequency than unknown words, and high-frequency words are produced faster than less frequent words (Baker & Bradlow, 2009; Gahl, 2008). Finally, the number of times each referent was produced and their position in the sentences were not controlled for, even though they can significantly influence prosodic modulations (e.g., Arnold, 2008; Fowler & Housum, 1987; Lam & Watson, 2010, 2014).

[ii] Since we have analysed five different prosodic cues of the sample, results for the significance level that is larger than $p = .01$ may be interpreted with caution.

[iii] Based on the suggestion from one of our reviewers, we also examined the correlation between the situational accessibility and learning outcomes. The analyses related to the degree of modifications on the prosodic cues in the absent condition revealed that only the speaking rate ratio significantly predicted recognition outcome in an independent model ($\beta = 1.78$, $p = .045$), but it is no longer significant ($\beta = 1.46$, $p = .11$) when we controlled for children's concurrent vocabulary ($\beta = 0.19$, $p = .004$) and it is not a significant predictor of long term vocabulary size ($\beta = -9.64$, $p = .50$). None of the other cues is significant (all $p > .5$) (see OSF for full results). Another reviewer asked whether the variation in the sound volume of the stimuli in the recognition test might have affected the results. There was no correlation between mean intensity of the sound stimuli and recognition outcome ($\beta = -0.06$, $p = .47$). All results were unchanged when the mean intensity of stimuli was additionally controlled for.

[iv] One reviewer asked whether the children that listened to the ratio < 1 group were better/faster learners than the children that listened to the ratio > 1 group. We think our analyses using continuous scaling ratio is more accurate than splitting participants into binary categories. To satisfy the review, we did a sensitivity analysis with splitting the ratio into two categories (ratio < 1 and ratio > 1). The results were consistent: Children that listened to the ratio < 1 group had significant higher scores both in the recognition test ($\beta = .78$, $p < 0.001$) and BPVT outcome one year later ($t(30) = 2.58$, $p < .008$) than the children that listened to the ratio > 1 group. The differences in BPVT outcome one year later between two groups remained to be significant ($\beta = 6.51$, $p = .016$) even after controlling for the concurrent BPVT baseline.

[v] One concern is that 'unknown' does not capture the dynamic status of familiarity during the experiment (i.e., becomes known after repeated mentions). We used pitch ratio of the first mentions between known and unknown words as a predictor for vocabulary size one year later, and the result was still significant, $\beta = -54.62$, $p = .026$. Of course, future research should corroborate our findings in a laboratory setting using stimuli with more controlled vowels and stress patterns.

REFERENCES

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, *23*(4), 495–527. https://doi.org/10.1080/01690960801920099

Baker, R. E., & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, *52*(4), 391–413. https://doi.org/10.1177/0023830909336575

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by compute (6.1.08) [Computer software]*.

Bortfeld, H., & Morgan, J. L. (2010). Is early word-form processing stressfull? How natural variability supports recognition. *Cognitive Psychology*, *60*(4), 241–266. https://doi.org/10.1016/j.cogpsych.2010.01.002

Butler, J., & Frota, S. (2018). Emerging word segmentation abilities in European Portuguese-learning infants: New evidence for the rhythmic unit and the edge factor. *Journal of Child Language*, *45*(6), 1294–1308.

Chafe, W. (1987). Cognitive constraints on information flow. *Coherence and Grounding in Discourse*, *11*, 21–51.

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*(5), 1584–1595. https://doi.org/10.2307/1130766

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, 1–20. https://doi.org/10.1038/s41562-022-01452-1

Cristia, A. (2013). Input to language: The phonetics and perception of infant-directed speech. *Language and Linguistics Compass*, *7*(3), 157–170. https://doi.org/10.1111/lnc3.12015

Dong, S., Gu, Y., & Vigliocco, G. (2021). The impact of child-directed language on children's lexical development. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 1444–1450. Retrieved from https://escholarship.org/uc/item/38X9h9h4

Dunn, D. M., & Dunn, L. M. (2009). National foundation for educational research in England and Wales, & GL assessment (Firm). *The British picture vocabulary scale. GL assessment*.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, *60*(6), 1497–1510. https://doi.org/10.2307/1130938

Fernald, A. (1992). Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. The Adapted Mind: Evolutionary Psychology and the Generation of Culture, 391–428.

Fernald, A. (2000). Speech to infants as hyper speech: Knowledge-driven processes in early word recognition. *Phonetica*, *57*(2–4), 242–254. https://doi.org/10.1159/000028477

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221. https://doi.org/10.1037/0012-1649.27.2.209

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, *20*(1), 104–113. https://doi.org/10.1037/0012-1649.20.1.104

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. https://doi.org/10.1017/S0305000900010679

Fisher, C., & Tokura, H. (1995). The given-new contract in speech to infants. *Journal of Memory and Language*, *34*(3), 287–310.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*(5), 489–504. https://doi.org/10.1016/0749-596X(87)90136-7

Gahl, S. (2008). Time" and "Thyme" are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496.

Graf-Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5), 797–824. https://doi.org/10.1111/infa.12006

Grassmann, S., & Tomasello, M. (2007). Two-year-olds use primary sentence accent to learn new words. *Journal of Child Language*, *34*(3), 677–687.

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

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

Gussenhoven, C., & Chen, A. (2000). Universal and language-specific effects in the perception of question intonation. In B. Yuan, T. Huang, & X. Tang (Eds.), *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)* (pp. 91–94). China Military Friendship Publish.

Han, M. (2019). *The role of prosodic input in word learning: a Cross-linguistic investigation of Dutch and mandarin Chinese Infant-directed speech*. LOT Netherlands Graduate School of Linguistics.

Han, M., de Jong, N. H., & Kager, R. (2020). Pitch properties of infant-directed speech specific to word-learning contexts: A cross-linguistic investigation of mandarin Chinese and Dutch. *Journal of Child Language*, *47*(1), 85–111. https://doi.org/10.1017/S0305000919000813

Han, M., de Jong, N. H., & Kager, R. (2021). Language specificity of infant-directed speech: Speaking rate and word position in word-learning contexts. *Language Learning and Development*, *17*(3), 221–240. https://doi.org/10.1080/15475441.2020.1855182

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*(3), 269–286.

Johnson, E. K., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: Utterance level prosody enables word form extraction by 6-month-olds. *PLoS ONE*, *9*(1), e83546.

Kalashnikova, M., & Burnham, D. (2018). Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions. *Journal of Child Language*, *45*(5), 1035–1053. https://doi.org/10.1017/S0305000917000629

Kalashnikova, M., & Kember, H. (2020). Prosodic cues in infant-directed speech facilitate young children's conversational turn predictions. *Journal of Experimental Child Psychology*, *199*, 104916. https://doi.org/10.1016/j.jecp.2020.104916

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, *27*(2), 237–265.

Kao, Y., & Lee, L. (2006). Feature analysis for emotion recognition from mandarin speech considering the special characteristics of Chinese language. *Ninth International Conference on Spoken Language Processing*.

Keren-Portnoy, T., Vihman, M., & Fisher, R. L. (2019). Do infants learn from isolated words? An ecological study. *Language Learning and Development*, *15*(1), 47–63.

Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development*, *24*(4), 372–392. https://doi.org/10.1016/S0163-6383(02)00086-3

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., & Chistovich, L. A. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686.

Lam, T. Q., & Watson, D. G. (2010). Repetition is easy: Why repeated referents have reduced prominence. *Memory & Cognition*, *38*(8), 1137–1146. https://doi.org/10.3758/MC.38.8.1137

Lam, T. Q., & Watson, D. G. (2014). Repetition reduction: Lexical repetition in the absence of referent repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 829–843. https://doi.org/10.1037/a0035780

Lambrecht, K. (1994). *Information structure and sentence form: topic, focus, and the mental representations of discourse referents*. Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511620607

Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, *111*(5), 686.

Leipold, S., Abrams, D. A., & Menon, V. (2022). Mothers adapt their voice during children's adolescent development. *Scientific Reports*, *12*(1), 951. https://doi.org/10.1038/s41598-022-04863-2

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323–1329.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling*. (pp. 403–439). Springer. https://doi.org/10.1007/978-94-009-2037-8_16

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant- and adult-directed speech. *Language Learning and Development*, *7*(3), 185–201. https://doi.org/10.1080/15475441.2011.579839

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition*, *156*, 52–59. https://doi.org/10.1016/j.cognition.2016.07.015

Narayan, C. R., & McDermott, L. C. (2016). Speech rate and pitch characteristics of infant-directed speech: Longitudinal and cross-linguistic observations. *The Journal of the Acoustical Society of America*, *139*(3), 1272–1281.

Nencheva, M. L., Piazza, E. A., & Lew-Williams, C. (2021). The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning. *Developmental Science*, *24*(1), e12997. https://doi.org/10.1111/desc.12997

Poerner, N., & Schiel, F. (2018). A web service for pre-segmenting very long transcribed speech recordings. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). https://www.aclweb.org/anthology/L18-1452

Porritt, L. L., Zinser, M. C., Bachorowski, J.-A., & Kaplan, P. S. (2014). Depression diagnoses and fundamental frequency-based acoustic cues in maternal infant-directed speech. *Language Learning and Development*, *10*(1), 51–67. https://doi.org/10.1080/15475441.2013.802962

Raneri, D., Von Holzen, K., Newman, R., & Bernstein Ratner, N. (2020). Change in maternal speech rate to preverbal infants over the first two years of life. *Journal of Child Language*, *47*(6), 1263–1275. https://doi.org/10.1017/S030500091900093X

Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, *16*(2), 143–160.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(4), 1928–1941. https://doi.org/10.3758/s13428-018-1176-7

Segal, J., & Newman, R. S. (2015). Infant preferences for structural and prosodic properties of infant-directed speech in the second year of life. *Infancy*, *20*(3), 339–351. https://doi.org/10.1111/infa.12077

Smith, N. A., & Trainor, L. J. (2008). Infant-Directed speech is modulated by infant feedback. *Infancy*, *13*(4), 410–420. https://doi.org/10.1080/15250000802188719

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Song, J. Y., Demuth, K., & Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *The Journal of the Acoustical Society of America*, *128*(1), 389–400. https://doi.org/10.1121/1.3419786

Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language*, *10*(1), 1–15. https://doi.org/10.1017/S0305000900005092

Swerts, M., Krahmer, E., & Avesani, C. (2002). Prosodic marking of information status in Dutch and Italian: A comparative analysis. *Journal of Phonetics*, *30*(4), 629–654.

Tang, P., Xu Rattanasone, N., Yuen, I., & Demuth, K. (2017). Phonetic enhancement of mandarin vowels and tones: Infant-directed speech and Lombard speech. *The Journal of the Acoustical Society of America*, *142*(2), 493–503.

Tippenhauer, N., Fourakis, E. R., Watson, D. G., & Lew-Williams, C. (2020). The scope of audience design in child-directed speech: Parents' tailoring of word lengths for adult versus child listeners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(11), 2163–2178. https://doi.org/10.1037/xlm0000939

Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*(2), 335–340. https://doi.org/10.3758/BF03196290

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Veneziano, E. (2001). Displacement and informativeness in child-directed talk. *First Language*, *21*(63), 323–356. https://doi.org/10.1177/014272370102106306

Vigliocco, G., Gu, Y., Motamedi, Y., Grzyb, G., Brekelmans, M., Murgiano, M., Brieke, R., & Perniss, P. [unpublished] The ECOLANG corpus of dyadic interactions between caregivers and their 2–4 year-old child and between two adults.

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

Zangl, R., Klarman, L., Thal, D., Fernald, A., & Bates, E. (2005). Dynamics of word comprehension in infancy: Developments in timing, accuracy, and resistance to acoustic degradation. *Journal of Cognition and Development*, *6*(2), 179–208.

Zangl, R., & Mills, D. L. (2007). Increased brain activity to infant-directed speech in 6-and 13-month-old infants. *Infancy*, *11*(1), 31–62.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97.

How to cite this article: Shi, J., Gu, Y., & Vigliocco, G. (2022). Prosodic modulations in child-directed language and their impact on word learning. *Developmental Science*, e13357. https://doi.org/10.1111/desc.13357

APPENDIX A

TABLE A1 Results of analyses on different prosodic measures with CHILDES frequency

	Speaking rate	Mean pitch	Pitch range	Mean intensity	Intensity range
Presence.ct	0.019**	0.38	−0.069	1.14***	−0.22
Familiarity.ct	−0.031**	0.055	0.72***	0.053	0.79.
NumMentions.ct	0.074***	−0.84***	−2.04***	−1.43***	−1.92***
Age.ct	0.003.	−0.098	−0.19***	−0.22	−0.16
CHILDESfrequency.ct	0.010	−0.26	−0.29	−0.047	0.33
Initial.ct	0.10***	1.92***	−0.84**	2.71***	0.27
Medial.ct	0.11***	0.46***	−1.79***	1.68***	−2.33***
Isolation.ct	−0.022**	−0.27	−0.47	−0.35	1.14**
Syllables.ct	0.15***	−0.034	0.87***	0.42	0.020
Yes-no Question.ct	0.029***	1.46***	−0.15	0.36*	0.21
Wh-question.ct	0.030***	−0.31.	−0.74***	−0.71***	−0.31
Single referent.ct	−0.035***	1.02***	0.58**	1.17***	0.88**
Session repetition.ct	0.010.	0.039	−0.24	0.66*	−0.30
Presence.ct: NumMen.ct	−0.024** First: 0.037*** Sub: 0.013.	0.35 First: 0.11 Sub: 0.46.	0.15	1.43*** First: 0.22 Sub: 1.50***	1.86*** First: −1.42*** Sub: 0.22
Familiarity.ct: NumMen.ct	0.020** First: −0.046*** Sub: −0.026**	0.24	−0.12	0.28	−0.88* First:1.44** Sub: 0.63
Presence.ct: Familiarity.ct	−0.004	−0.32	0.13	0.28	0.60
Presence.ct: Familiarity.ct:NumMen.ct	−0.014	−0.41	−0.16	−0.75	−0.17

Note: The number stands for an estimate, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, .$p < 0.1$

APPENDIX B

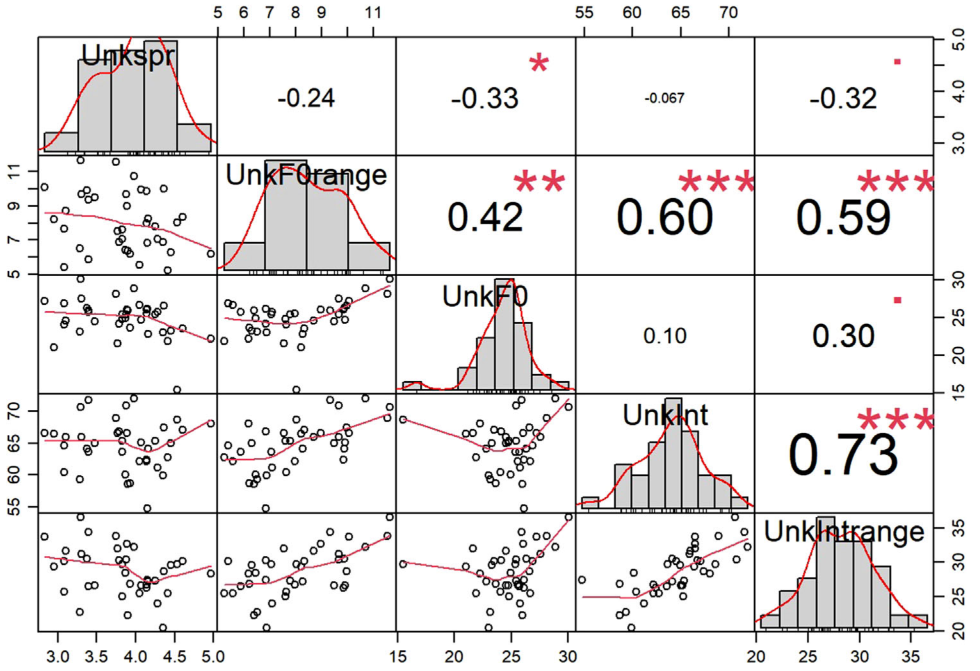

FIGURE B1 Correlation matrix between unknown mean pitch, pitch range, speaking rate, intensity range and mean intensity.

APPENDIX C

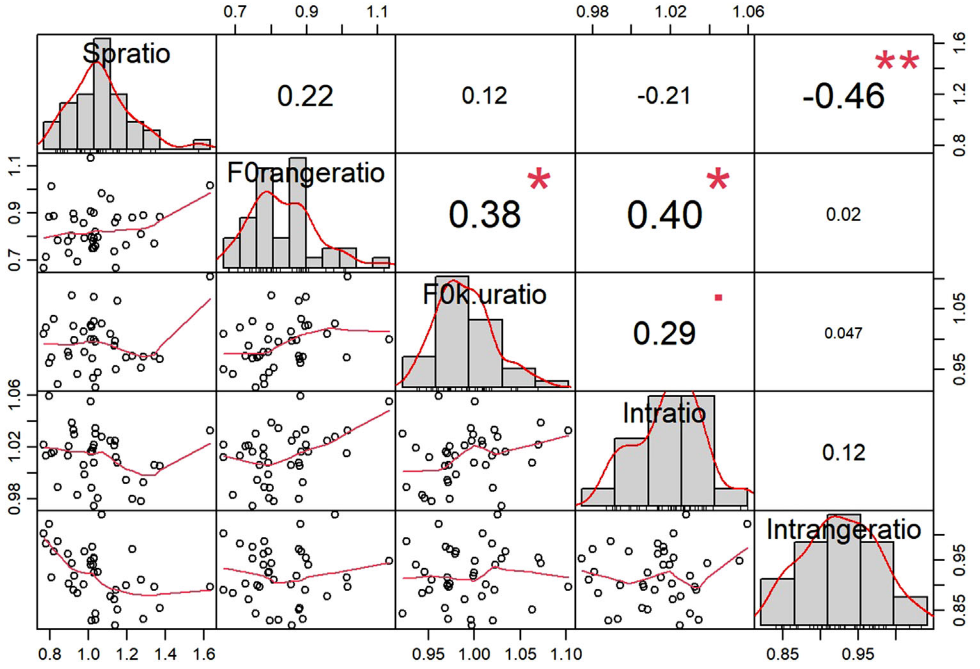

FIGURE C1 Correlation matrix between cue ratios (known/unknown).