Received: 23 July 2019 | Revised: 2 January 2020 | Accepted: 1 May 2020

DOI: 10.1111/desc.12997

PAPER

Developmental Science WILEY

The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning

Mira L. Nencheva[1] | Elise A. Piazza[1,2] | Casey Lew-Williams[1]

[1]Department of Psychology, Princeton University, Princeton, NJ, USA

[2]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

Correspondence
Mira L. Nencheva, Department of Psychology, Princeton University, Princeton, NJ 08540, USA.
Email: nencheva@princeton.edu

Funding information
National Institute of Child Health and Human Development, Grant/Award Number: R01HD095912 and R03HD079779; Princeton University

Abstract

Young children have an overall preference for child-directed speech (CDS) over adult-directed speech (ADS), and its structural features are thought to facilitate language learning. Many studies have supported these findings, but less is known about processing of CDS at short, sub-second timescales. How do the moment-to-moment dynamics of CDS influence young children's attention and learning? In Study 1, we used hierarchical clustering to characterize patterns of pitch variability in a natural CDS corpus, which uncovered four main word-level contour shapes: 'fall', 'rise', 'hill', and 'valley'. In Study 2, we adapted a measure from adult attention research—pupil size synchrony—to quantify real-time attention to speech across participants, and found that toddlers showed higher synchrony to the dynamics of CDS than to ADS. Importantly, there were consistent differences in toddlers' attention when listening to the four word-level contour types. In Study 3, we found that pupil size synchrony during exposure to novel words predicted toddlers' learning at test. This suggests that the dynamics of pitch in CDS not only shape toddlers' attention but guide their learning of new words. By revealing a physiological response to the real-time dynamics of CDS, this investigation yields a new sub-second framework for understanding young children's engagement with one of the most important signals in their environment.

KEYWORDS

attention, child-directed speech, learning, pitch, pupillometry, synchrony

1 | INTRODUCTION

Successful communication requires attention to people and the complex signals they generate. Adapting attention from one moment to the next allows speakers and listeners to produce and perceive cues in real time and to adjust their responses in ways that facilitate an efficient exchange of information (MacDonald, 2013a, 2013b). In order to break into these dynamics of communication and learn from their caregivers, young children have to process and adapt to the rich, multidimensional information embedded in child-directed speech (CDS; McMurray, 2016; Potter & Lew-Williams, 2019). While it is well-known that young children prefer listening to CDS over adult-directed speech (ADS; Cooper & Aslin, 1990; Fernald & Kuhl, 1987; Werker & McLeod, 1989), less is understood about their processing of and learning from specific features of CDS, such as its characteristically variable prosody, across both shorter and longer timescales. In the current set of studies, we explored how prosodic cues–specifically, the dynamics of pitch—affect young children's in-the-moment engagement with CDS, and in turn, how these differences in attention to pitch contours affect learning of new words.

Learning in early childhood occurs in social contexts, and many experiments and theories have prioritized social cues in enabling the development of human cognition (Kuhl, 2007; Tomasello, 1992; Vygotsky, 1978). For example, Kuhl (2007) hypothesized that social interaction facilitates learning by modulating children's attention and arousal during key moments. Brand, Baldwin, and Ashburn

(2002) highlight that when interacting with infants, caregivers modify various aspects of their communication, including speech, gesture and action, in ways that optimize the infants' attention and emphasize important units (in speech or action). Recent studies on the dynamics of attention provide a promising approach for understanding what 'key moments' might mean in the context of early communication. This work shows that two interaction partners fluctuate over time in their responses to incoming signals; sometimes they converge, likely reflecting both partners' responses to a shared external stimulus, and sometimes they diverge, likely reflecting one partner's response to the other partner's behavior (Butler, 2015). Importantly, this variation is modulated by the level of engagement exhibited by both interaction partners at specific moments across time (Kang & Wheatley, 2017). CDS, as one of the most important information signals in childhood, plays a crucial role in modulating children's attention, affect, and learning during everyday interactions with caregivers (Saint-Georges et al., 2013). While the mechanism through which CDS facilitates attention and learning is unclear, caregivers' exaggerated expression of emotional states and intentions when using CDS (Trainor, Austin, & Desjardins, 2000) and dynamic adaptation to children's own states (Smith & Trainor, 2008) might play a role. CDS prosody, therefore, may be more effective than ADS in increasing children's arousal and attention to their environment, allowing them to process information more fully and efficiently (Kaplan, Bachorowski, & Zarlengo-Strouse, 1999).

Caregivers alter their prosody in a range of consistent ways that may support such engagement (Fernald, 1992; Grieser & Kuhl, 1988). CDS prosody is characterized by higher fundamental frequency, increased pitch variability, exaggerated and repetitive intonation contours, slower rate of speech, and distinct spectral timbre (Fernald, 1992; Fernald & Simon, 1984; Gleitman, Newport, & Gleitman, 1984; Grieser & Kuhl, 1988; Piazza, Iordan, & Lew-Williams, 2017), as well as increased emotional expressivity (Fernald & Kuhl, 1987; Scherer, Banse, Wallbott, & Goldbeck, 1991). Children show a preference for listening to CDS over ADS (Cooper & Aslin, 1990; Fernald, 1985; Kaplan, Goldstein, Huckeby, Owren, & Cooper, 1995; ManyBabies Consortium, 2019; Pegg, Werker, & McLeod, 1992; Werker & McLeod, 1989), and a stronger cortical tracking of speech rate in response to CDS (vs. ADS) prosody (Kalashnikova, Peter, Di Liberto, & Lalor, 2018), supporting the idea that CDS prosody more successfully engages and sustains children's attention than ADS prosody. Consistent with the hypothesis that increased attention to a stimulus improves learning (Posner, Snyder, & Solso, 2004), CDS prosody seems to facilitate multiple aspects of learning, such as word segmentation in 8-month-olds (Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989; Thiessen, Hill, & Saffran, 2005), familiar word recognition in 19-month-olds (Song, Demuth, & Morgan, 2010), and novel word learning in 17- and 21-month-olds (Graf Estes & Hurley, 2013; Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011).

On a much shorter timescale, caregivers use moment-to-moment variability in pitch in consistent ways to convey information during interactions with children. Caregivers' pitch contours adapt to the child's attentional state in the moment and reflect caregivers' communicative intent (Fernald, 1989; Katz, Cohn, & Moore, 1996; Kitamura & Burnham, 2003; Papoušek, Papoušek, & Symmes, 1991; Stern, Spieker, & MacKain, 1982). Adult raters distinguish better between utterances expressing approval, prohibition, and bids for attention based on pitch contours in CDS than in ADS, suggesting a potential role for CDS prosody in guiding the child's attentional and affective state during caregiver–child interactions (Fernald, 1989). Caregivers also use emphatic stress to highlight new or important words in a sentence, characterized by a hill-shaped (or "rise-fall") contour (Aslin, 1993; Fernald & Mazzie, 1991), and local increases in pitch and loudness during target words facilitate learning in 2-year-olds (Grassmann & Tomasello, 2007). These studies highlight the presence of short-timescale effects of prosodic variation in CDS on children's processing and learning from the speech signal. As one of the defining features of CDS, pitch variation provides an optimal case study for the fluctuations of children's attention and learning in response to dynamically changing signals in the environment.

How can we measure young children's dynamic attention to CDS on a moment-to-moment timescale? Recent adult work has used the extent of temporal coupling between incoming perceptual information and physiological responses to distinguish between states of high versus low attention (Kang, Huffer, & Wheatley, 2014; Smallwood et al., 2011). According to Smallwood et al. (2011), moments of high time-locked attention are characterized by strong coupling to external perceptual information, such that the responses of the perceptual system are closely aligned in time to changes in the perceived signal, whereas in moments of decreased attention, our perceptual systems are decoupled from external perceptual information, as activity in the perceptual systems is not temporally aligned to a present but not attended to external stimulus. Pupil dilation responses are driven by the activity of neural structures related to arousal and attention; in particular, the release of norepinephrine (NE) from neurons in the locus coeruleus (LC; Rajkowski, 1993). Episodes of decreased attention are associated with a decrease

Research highlights

- Child-directed speech (CDS) elicits higher pupil size synchrony than adult-directed speech (ADS), indicating greater moment-to-moment attention to CDS
- Word-level pitch variation in natural CDS follows four broad types of contours: 'fall', 'rise', 'hill' (inverted U-shape), and 'valley' (U-shape)
- Word-level pitch contours modulate toddlers' in-the-moment attention, such that 'hill' contours are associated with high synchrony, 'valley' contours with low synchrony
- Toddlers' pupil size synchrony (a dynamic measure of attention) during exposure to a novel word predicts their learning

in fast, time-locked responses (referred to as phasic responses) in the LC-NE system (Mittner et al., 2014), compared to episodes of sustained attention in which increased phasic LC responses help optimize task performance (Aston-Jones & Cohen, 2005). During moments of attention, pupils dilate in response to task-relevant stimulus presentation (Kahneman & Beatty, 1966). Two previous studies (Kang et al., 2014; Smallwood et al., 2011) have estimated how attention to a task affects time-locked pupil responses following the presentation timing of important cues. In line with their predictions, pupil size fluctuations were only coupled to the temporal dynamics of the task during moments of increased attention to stimuli, but not during moments of decreased attention. Given evidence that time-locked responses to auditory information are present in adults and young children, pupil size might be a good measure of attention to speech. Frith (1981) showed that in the absence of light changes, adults' pupils responded to short bursts of sound. Similarly, the pupils of both adults and 14-month-old infants responded to 800-millisecond odd-ball sounds embedded in a sequence of repeated sounds (Wetzel, Buttelmann, Schieler, & Widmann, 2016).

If time-locked attention increases an individual's coupling to a given stimulus (Kang et al., 2014; Smallwood et al., 2011), then one might expect that the pupil responses of *multiple* individuals should become more synchronous *to each other* at times of increased attention. That is, there may be a signature of similarity in how people process a particular external stimulus across time. Measures of inter-subject synchrony allow us to quantify the degree to which different moments in rich naturalistic stimuli elicit shared responses across a group of listeners (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004; Lerner, Honey, Silbert, & Hasson, 2011; Schmälzle, Häcker, Honey, & Hasson, 2015; Schmälzle, Häcker, Renner, Honey, & Schupp, 2013; Stephens, Silbert, & Hasson, 2010). This approach reduces individual-specific noise and isolates stimulus-specific responses that are consistent across subjects, therefore providing insights into which stimulus properties reliably drive the dynamics of attention during comprehension. For instance, adult listeners show higher synchrony in pupil size during the most salient narrative moments in a story (Kang & Wheatley, 2017). Additionally, listeners with higher (vs. lower) overall levels of neural synchrony with a storyteller comprehend the story better (Stephens et al., 2010).

In the present experiments, we adapted the logic of pupil size synchrony from these adult studies on narrative comprehension for the purpose of studying toddlers' moment-to-moment engagement with fluctuations of prosody in CDS. In Study 1, we sought to identify common patterns of natural pitch variation used by caregivers during individual words. Using hierarchical clustering, we identified four common word-level pitch contours in natural CDS, which we term 'falls', 'rises', 'hills', and 'valleys' based on their shape. In Study 2, we used measures of pupil size synchrony, a time-locked measure of the reliability of attentional engagement across observers (Kang & Wheatley, 2017), to assess toddlers' engagement with and processing of pitch variation in a children's storybook. We did so at two timescales of granularity. First, we examined toddlers' overall engagement with the story spoken in CDS versus ADS, and predicted that they would show higher pupil size synchrony when listening to CDS prosody compared to ADS prosody. Then, we specifically examined toddlers' processing of the word-level contour types identified in Study 1 using both the CDS children's story and carefully controlled sentences. We predicted that toddlers would show higher pupil size synchrony for contours that sound more natural and are often associated with emphasis on novel or important words (Aslin, 1993; Fernald & Mazzie, 1991).

In Study 3, we aimed to understand if toddlers' in-the-moment engagement with pitch contours impacts the learning of new words from CDS. To do so, we presented novel objects, and toddlers heard novel words with contours from Study 2 that had elicited the highest versus lowest pupil size synchrony. At test, toddlers looked at two novel objects and we assessed their eye movements and pupil size synchrony in response to labels for the target object. We predicted that toddlers would more successfully learn novel words presented in contours that had elicited higher (vs. lower) pupil size synchrony across participants. Furthermore, independent of contour type, we predicted that higher synchrony during exposure to a novel word would predict more accurate looking toward that target object at test. By examining sub-second changes in toddlers' engagement with CDS, this investigation advances what is currently known about toddlers' learning from an important information signal in their environment.

2 | STUDY 1: IDENTIFYING COMMON WORD-LEVEL PITCH CONTOURSIN NATURAL CHILD-DIRECTED SPEECH

Previous work has suggested that sentence-level pitch contours in CDS carry communicative information (Fernald, 1989). Even over the course of individual words, caregivers use variation in pitch to highlight new or important information in the discourse (Aslin, 1993; Fernald & Mazzie, 1991). In Study 1, we used hierarchical clustering to characterize the most common word-level pitch contours in naturally occurring CDS.

2.1 | Method

Nouns from two corpora of natural CDS from the CHILDES database (MacWhinney, 2000), one of a mother addressing a 6–12-month old infant (Soderstrom, Blossom, Foygel, & Morgan, 2008) and one of two mothers addressing 24–30-month-old children (Weist & Zevenbergen, 2008), were identified using the accompanying part-of-speech markings in the %mor lines in CHAT transcripts. After identifying all nouns in the corpus, we manually time-stamped the onset and offset of each noun, and extracted pitch contours using PRAAT (Boersma & Weenink, 2019). Nouns were removed from analyses if they were masked by noise or concurrent speech from another speaker (N = 1,390), or if they contained fewer than 10 timepoints of pitch data (N = 87, mean

duration = 92.9 ms). We used Tukey's running median smoothing filter (Tukey, 1977) to reduce noise and help ensure the quality of the pitch measurements. The nouns in the analyses varied in their duration [duration percentiles: 25 (240 ms); 50 (310 ms); 75 (400 ms)]. To facilitate clustering of pitch contours of different durations, we used the same number of equally spaced pitch measurements (30) for each noun. First, we sampled the 10 most informative points from the pitch of each word nonlinearly. To do so, we passed over the full pitch time series of each noun and selected the 10 points with slope closest to zero (usually a local maximum or minimum, or an inflection point). This allowed us to sample an equal number of points from each word and capture the local extremes of the time series with higher density (see Figure S1). Then, in order to get equally spaced observations, we used a linear interpolation function based on the most informative 10 points to create the same number of equally spaced measurements across nouns (30). Finally, the pitch time series corresponding to each noun was z-scored to equalize the overall amount of pitch variability across nouns. Dynamic time-warping distance (DTW; Tormene, Giorgino, Quaglini, & Stefanelli, 2009) was used to quantify the difference between the pitch time-series of 932 nouns from the first corpus and 1,128 from the second corpus, and create two separate dissimilarity matrices (one for each corpus). Each dissimilarity matrix was used in a hierarchical clustering analysis (Montero & Vilar, 2014), yielding clusters of similar contour types.

To determine an optimal number of clusters and to maximize the distance between clusters, we performed the clustering analysis for 2, 3, 4, 5, 6, 7, and 8 clusters. For each, we mapped the clusters in the first corpus to the closest clusters in the second corpus based on the average DTW distance between the pitch contours of the nouns in the two corpora. Only the analyses for 2, 4 and 6 clusters resulted in one-to-one mappings between the two corpora. Others did not generalize consistently; for example, when we split each corpus into three clusters, the first, second, and third cluster in one corpus mapped most closely to the second, third, and second clusters in the other corpus, respectively, which suggests that the three clusters did not clearly group together into true underlying categories. Next, we computed the distance between each cluster and the remaining clusters to estimate their distinctiveness The mean distance between clusters was highest for four clusters, suggesting more reliable results with four clusters compared to 2 or 6 (see Figure S2).

2.2 | Results and discussion

The hierarchical clustering analysis yielded four clusters of noun-level contours (Figure 1). The same four clusters emerged from both the first corpus (6–12 months) and the second corpus (24–30 months). We refer to these contours as follows: 'fall' (shown in green), characterized by a continuous decrease over time in pitch; 'rise' (blue), characterized by a continuous rise over time in pitch; 'hill' (cyan) characterized by an increase and then a decrease in pitch; and 'valley' (red), characterized by a decrease and then an increase in pitch. In the infant-directed corpus there were 315 falls (average range = 116.60 Hz, *SD* = 94.28 Hz), 191 rises (average range = 151.94 Hz, *SD* = 93.88 Hz), 167 hills (average range = 137.11 Hz, *SD* = 99.73 Hz), and 259 valleys (average range = 153.07 Hz, *SD* = 105.11 Hz). In the child-directed corpus there were 246 falls (average range = 92.47 Hz, *SD* = 86.38 Hz), 348 rises (average range = 125.40 Hz, *SD* = 86.34 Hz), 431 hills (average range = 136.78 Hz, *SD* = 101.47 Hz), and 103 valleys (average range = 111.19 Hz, *SD* = 97.43 Hz).

To quantify the robustness of the four clusters, we performed a twofold cross-validation between the two corpora (adapted from Tibshirani & Walther, 2005). Each corpus was treated as onefold; at each of the two iterations, one corpus served as training data and the other corpus as left-out test data. We quantified the distance between the pitch contour of each noun in the left-out corpus and the contours of the nouns in each of the clusters in the training corpus. The label of the closest cluster (with the smallest average DTW distance to the noun) was then compared against the cluster label of the noun in the original clustering analysis (described above). Both clusters yielded correct classification reliably above chance (25%), and classification of toddler-directed held-out data were better (76.22%) than classification accuracy on infant-directed held-out data (36.95%). Excluding the flattest 50% of contours (determined by the standard deviation in pitch over the course of the noun) improved classification (80% accuracy classifying toddler-directed held-out data, and 48% accuracy classifying infant-directed held-out data), but it did not eliminate the difference between the two corpora. This difference is likely due to the larger number of speakers in the toddler-directed corpus. While the number of nouns is comparable across the two corpora, the infant corpus includes only one speaker, whereas the toddler corpus includes two. Thus, it is possible that the cluster labels (which we used to determine "correct" classification) in the infant corpus were overfitted to that specific speaker, while the clusters in the toddler corpus generalized better during cross-validation.

3 | STUDY 2: PROCESSING OF MOMENT-TO-MOMENT DYNAMICS OF CDS PROSODY

In Study 2, we used the four contours that emerged from Study 1 to examine effects of CDS prosody on the temporal dynamics of toddlers' attention. We did this in two distinct ways. We first assessed broad effects of CDS versus ADS prosody on pupil size synchrony. Then, we compared toddlers' pupil size synchrony during words that matched the fall, rise, hill, and valley pitch contours.

3.1 | Method

3.1.1 | Participants

Thirty full-term monolingual English-learning 2-year-old toddlers (15 female, *M* = 27.5 months, range = 24.8–30.6 months) with no known

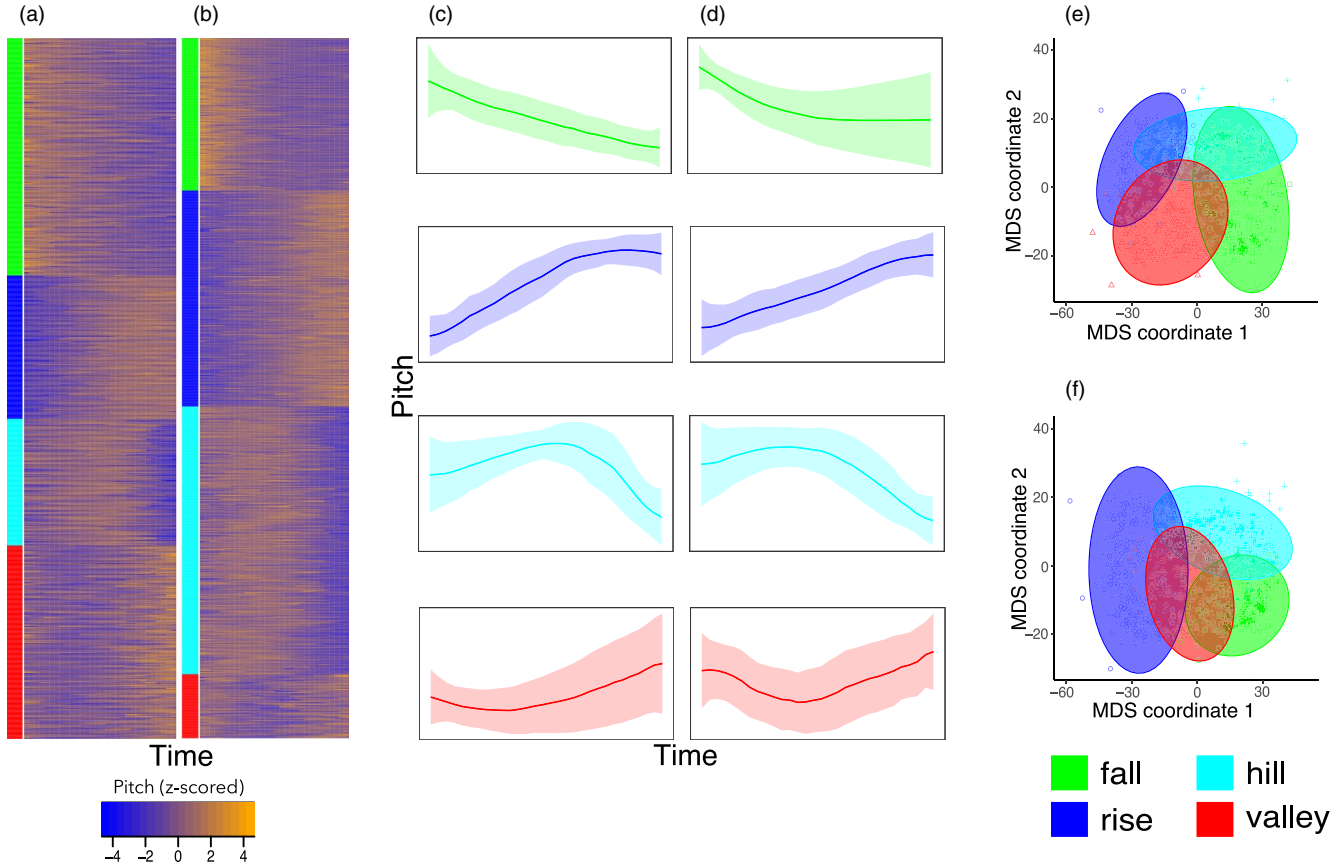

FIGURE 1 (a) Pitch contours of 932 nouns from a speaker addressing a 6–12-month-old infant, with the pitch of each noun over time represented as a row in the heatmap; a colored side bar represents which of the four clusters the noun belongs to; (b) Pitch contours of 1,128 nouns for two speakers addressing 24–30-month-old toddlers; (c) Average pitch contours for nouns in each cluster of the infant-directed corpus; (d) Average pitch contours for nouns in each cluster of the toddler-directed corpus; (e) Multidimensional scaling (MDS) of the four clusters of nouns in the infant-directed corpus in two-dimensions; (e) MDS of the four clusters of nouns in the toddler-directed corpus.

vision or hearing impairments were included in the final analyses in Study 2. Five additional toddlers participated in the study but were excluded due to fussiness (n = 3), unwillingness to sit on their caregiver's lap (n = 1), or problems detecting the toddler's pupil (n = 1). A legal guardian provided informed consent for the participation of each child. The Princeton University Institutional Review Board approved all protocols.

3.1.2 | Stimuli

A female native speaker of English recorded two versions of a children's story, *The Little Mouse Who Lost Her Squeak* (Robaard, 2015), once using natural CDS prosody and once using natural ADS prosody. Additionally, the speaker recorded simple sentences containing four target nouns (bear, bunny, kitty, and piggy), each with the four contours described in Study 1 (fall, rise, hill and valley), as well as a flat baseline contour. To ensure that the contours of the recorded nouns faithfully reflected the four contour types from Study 1, we clustered the pitch contours of recorded nouns with the contours of nouns from the two CHILDES corpora. This clustering procedure correctly assigned all target nouns to the appropriate contour cluster (fall, rise, hill, and valley). Mean pitch did not significantly differ between the four original contours, but was lower for the baseline flat contour (see Figure S3). For each contour, two of the target nouns were spliced into the final position of a sentence frame, while the other two were spliced in the middle position ("Look at the [target noun]" and "This [target noun] looks cool" respectively). To avoid acoustic artifacts associated with co-articulation, the target nouns in the final sentences were recorded along with "the" preceding them and this phrase was normed to be 0.80 s for all target words before splicing with the sentence frame using PRAAT (Boersma & Weenink, 2019). Similarly, target nouns in sentence-medial position were normed to 0.65 s. Additionally, all nouns and sentence frames were normed to the same loudness (60db).

Isoluminant visual stimuli were paired with the audio during the experiment in order to keep toddlers' gaze on the screen without affecting pupil size. The story trials were paired with three screensaver video clips, each modified to be isoluminant in order to avoid differences in pupil size due to the luminance of the stimulus (De Groot & Gebhard, 1952). Sentences were paired with isoluminant drawings of the target animals (bear, bunny, kitty, or piggy), which were animated to rotate slightly from side to side.

3.1.3 | Design and procedure

An EyeLink 1,000 Plus eye tracker recorded participants' pupil sizes at a sampling rate of 500 Hz while they sat on their caregiver's lap and listened to the story and the sentences. Speech was presented via two speakers positioned in front of the child, and isoluminant visual stimuli were viewed on a 17-inch monitor. Caregivers wore opaque sunglasses to avoid influencing the toddlers' behavior during the experiment.

The experiment consisted of two alternating types of blocks: story trials and sentence trials. Participants saw the first story block (ADS or CDS, counterbalanced for order) followed by the first sentence block, then a brief attention-getting video. Then, they saw the second story block (with the opposite speech register) and the second sentence block (Figure 2).

Each block of story trials consisted of three 15-s segments (splitting the original story into three parts). A different isoluminant video clip accompanied each story segment. The trial started with 500 ms of silence, plus an additional delay until the participant looked at the screen, after which the recording was played. Another 500 ms of silence followed the audio-recording at the end of the trial.

Each block of sentence trials consisted of ten sentences. There were two sentences for each of the five contour types (fall, rise, hill, valley, and flat), one in sentence-final and one in sentence-medial position. A blank gray screen preceded each trial onset and remained on the screen for a random inter-trial interval (ITI) (M = 1.5 s, range = 1–2 s), plus an additional delay until the participant looked at the screen. During each sentence trial, an isoluminant drawing representing the target was displayed for the entire duration of the trial. The accompanying sentence recording started 500 ms after the image appeared on the screen and ended 500 ms before the image disappeared. Sentence order in each block was randomized.

3.1.4 | Pre-processing of pupil size

Raw pupil size values were converted from arbitrary units to percent change from the average pupil size of the trial. In order to identify artifacts due to partial blinks, variation in pupil size was estimated with a sliding 0.05-s window. Any 0.05-s intervals with a range of change exceeding 15% were removed as artifacts (Merritt, Keegan, & Mercer, 1994).

3.1.5 | Computing pairwise pupil size synchrony

To estimate the temporal alignment of pupil responses between participants, we transformed a *distance* measure (DTW, the same measure used for computing distance between pitch contours in Study 1) quantifying the dissimilarity between two time series, into *synchrony*, an inversely related measure of the similarity between two time series. Because the analysis time window varied, we used a DTW distance measure that was normalized for duration, which eliminated effects of duration differences on synchrony. To convert from distance to synchrony, we subtracted the distance results from the maximum distance between two participants on a trial in the current experiment.

To compare synchrony for CDS and ADS, we computed pairwise synchrony over the pupil size time window spanning the three 15-s story segments for all possible pairs of participants. We excluded segments in which more than half of the data were missing, and interpolated gaps smaller than 100 ms using Stineman interpolation (Johannesson, Bjornsson, Grothendieck, & Johannesson, 2018), which yielded the most accurate interpolation of all the techniques we tested (see Figure S4).

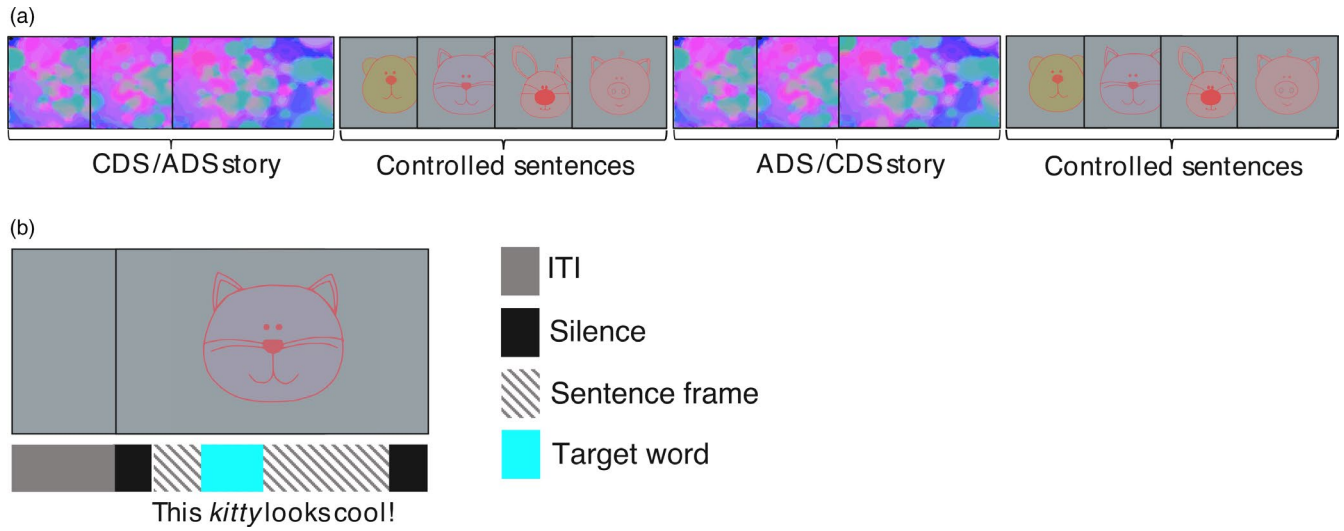

FIGURE 2 Experimental procedure for Study 2. (a) Example block order. Participants were presented with two story blocks (either CDS or ADS first), during which they saw an isoluminant filler video while the story was playing. The story blocks were interleaved with two controlled sentence blocks, during which isoluminant referents were presented along with the sentence recordings. (b) Example sentence trial, with the target word *kitty* in a hill contour. The sentence was preceded and followed by 500 ms of silence, and the isoluminant image remained on the screen for the entire duration of the trial.

In order to compare how different CDS word-level contours affect pupil size synchrony, we first categorized words into types. We extracted and pre-processed the pitch time series of all words as described above (see Study 1 methods), resulting in 63 words from the story. We computed the distance between each word from the story and each noun within the four clusters in Study 1. Each of the 63 story words was assigned to the closest cluster, i.e., the cluster with the smallest average distance. Words in the bottom quartile of pitch variability of all story words, measured by the standard deviation in pitch, were labeled as flat. This yielded four words with fall contour, seven with rise, 15 with hill, 2 with valley, and 35 with flat. We only explored the effects of pitch contours in the CDS story, because a similar analysis for ADS yielded flat contours for nearly all nouns (with the exception of two hills and one rise). We computed the pairwise synchrony (normalized for duration) over the pupil size time series interval spanning the duration of the target noun in each of the recorded sentences (0.80 s for nouns in sentence-final position and 0.65 s for nouns in sentence-medial) and each of the clustered words (average duration of 0.30 s) from the CDS story trials. Due to the short duration of the nouns, segments with any amount of missing data were excluded from this analysis.

3.1.6 | Model selection

For most analyses, we used mixed-effect models, controlling for participant-level variation. To test the effects of each predictor of interest, we created a *null* model, with random intercept per participant (or participant pair), and no fixed or random effect for the predictor of interest (e.g., speech register). We then created two models by adding the predictor of interest to the null model: a smaller model including only fixed effect for the predictor of interest, and a larger model including a fixed and a random participant-level effect for the predictor of interest. If both models converged, we used the `anova` function for model selection in R to select the better model (using a cut-off of $p < .05$). If the more complex model converged and outperformed the smaller model, we selected it as the *predictor* model; otherwise, we selected the smaller model as the final *predictor* model. We then compared the *predictor* to the *null* model to test whether the predictor of interest significantly improved our predictions using the `anova` function. For mixed-effect models, effect size was computed as Cohen's f^2 adapted from Selya, Rose, Dierker, Hedeker, and Mermelstein (2012). According to Cohen's (1988) guidelines, $f^2 \geq 0.02$, $f^2 \geq 0.15$, and $f^2 \geq 0.35$ represent small, medium, and large effect sizes respectively.

3.2 | Results and discussion

3.2.1 | Effect of speech register on pupil size synchrony

We first analyzed story trials and found that pupil size synchrony was significantly higher during the CDS story compared to the ADS story [Welch's *t*-test: $t(775.64) = 4.69$, $p < .001$, $d \sim 0.3$], indicating higher moment-to-moment engagement with the speech signal for CDS (Figure 3a). To further probe the effect of speech register, controlling for participant-level variation, we used a linear mixed model (Bates, Mächler, Bolker, & Walker, 2014). The predictor model, including speech register as a fixed and a random effect per participant, was significantly better than the null model [χ^2 (3) = 53.13, $p < .001$, $f^2 \sim 0.2$], confirming the effect of speech register on synchrony. Variance in pupil size did not significantly differ between the ADS and the CDS story [Welch's *t*-test: $t(107.6) = 1.40$, $p \sim 0.16$, $d \sim 0.2$], suggesting that the synchrony effect was not driven by the range of pupil size responses, but rather by their timing. These results show, for the first time, that toddlers' pupil sizes become entrained to external stimuli. The key finding is that toddlers showed a more consistent, shared response during processing of CDS than of ADS.

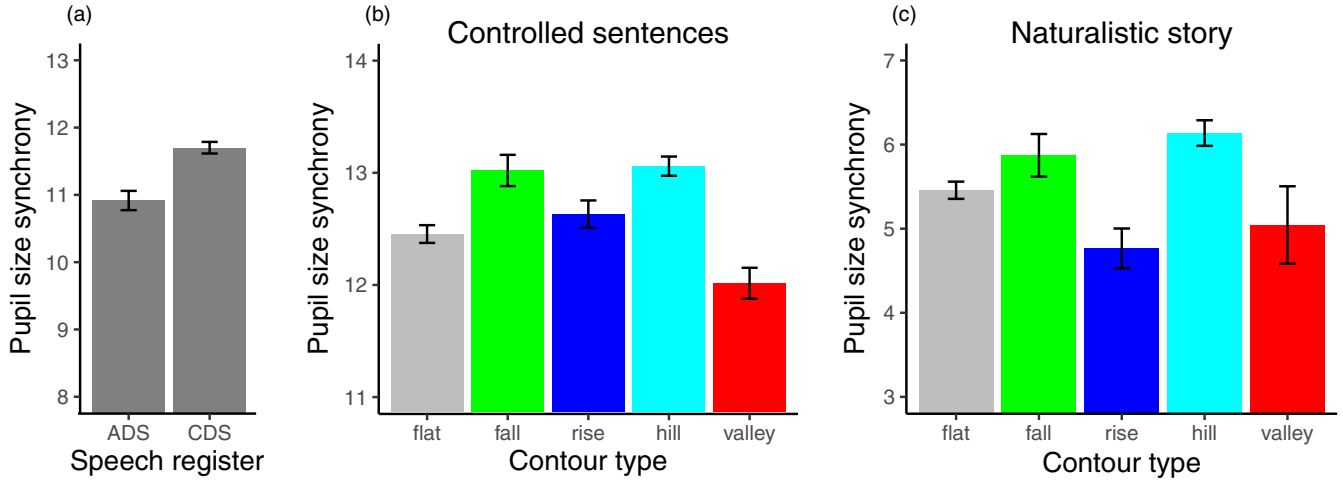

FIGURE 3 (a) Effect of speech register on pupil size synchrony in story trials; Effect of noun contour type on pupil size synchrony during word presentation in controlled sentence trials (b) and story trials (c). See Table S1 for post-hoc *t*-test comparisons between synchrony in response to different contour types in the controlled sentences and the story.

3.2.2 | Effect of word contour type on pupil size synchrony

In separate analyses of sentence trials and CDS story trials, we estimated the effect of a word's contour type (fall, rise, hill, valley, or flat) on pupil size synchrony during the span of the word. Synchrony over the course of the target word systematically differed by the word contour type in the two blocks of controlled sentence trials (Figure 3b) and in the story words (Figure 3c). Hills and falls elicited higher synchrony than the remaining contour types, with hills showing on average the highest synchrony and valleys the lowest (See Figure 3 and Table S1). We ran two identical linear mixed-effects models testing the effect of contour type on synchrony - one for words in the story and one for nouns in the controlled sentences. Only the smaller predictor models (see Model selection) converged. The predictor model was better than the null model, both in the controlled sentence words [χ^2 (4) = 58.48, p < .001, f^2 ~ 0.1] and the story words [χ^2 (4) = 39.54, p < .001, f^2 ~ 0.004], suggesting that contour type predicted synchrony.

To test whether the effects of contour type in the controlled sentences replicate in the story, we created a combined model, including nouns from both contexts, with random intercepts and fixed effects for contour type, context type (story vs. sentence), and critically, the interaction between the two. The interaction between the effects of contour type and word source (controlled sentences vs. CDS story) did not improve the prediction of synchrony compared to a null model [χ^2 (4) = 2.42, p ~ 0.66, f^2 ~ 0.0002]. This suggests that the effect of contour type on synchrony was replicated across different contexts. Indeed, the significance of the effect of contour type was preserved when combining nouns from both contexts [χ^2 (8) = 44.99, p < .001, f^2 ~ 0.003] in spite of the lower baseline synchrony for story words compared to controlled sentence words [Welch t-test: t(9,769.3) = −77.3, p < .001, d ~ 0.9].

The results of Study 2 suggest that the dynamics of pitch variation affect toddlers' moment-to-moment engagement with speech. The overall higher level of pupil size synchrony elicited by CDS compared to ADS suggests that this is true across global prosodic variation, and our analyses of local pitch variation suggest that the pitch contours of individual words impact toddlers' engagement in the moment. Critically, in both the controlled sentences and the naturalistic children's story, we observed the highest synchrony for words with hill-shaped contours and the lowest synchrony for words with valley-shaped contours, likely reflecting caregivers' use of hill-shaped contours to highlight important words (Aslin, 1993; Fernald & Mazzie, 1991). We next aimed to investigate if these dynamic changes in pitch contours impact not only toddlers' engagement with but their learning from CDS.

4 | STUDY 3: IMPACT OF MOMENT-TO-MOMENT ENGAGEMENT WITH PROSODY ON LEARNING FROM CDS

In Study 3, we explored whether or not toddlers' in-the-moment engagement with pitch contours affects novel word learning from CDS. To do so, we first compared learning for novel words introduced with a hill contour, which elicited the highest synchrony in Study 2, and words introduced with a valley contour, which elicited the lowest synchrony in Study 2. Then, we assessed if toddlers' pupil size synchrony during exposure to a novel word predicted their recognition of that word at test.

4.1 | Method

4.1.1 | Participants

A new group of 32 full-term monolingual English-learning toddlers between the ages of 24 and 30 months (11 female, M = 27.11, range = 24.5–30.8 months) with no known vision or hearing impairments were included in the final analysis in Study 3. An additional 11 toddlers participated in the study but were excluded due to fussiness (n = 4), unwillingness to sit on caregiver's lap (n = 2), a computer error (n = 2), parent speech during experiment (n = 1), or having fewer than 1 codable test trials per contour type (n = 2).

4.1.2 | Stimuli

Novel words were introduced with a hill- or valley-shaped contour during training. Similar results would likely be observed with words with a 'fall' contour, which also elicited high synchrony. However, 'hills' and 'valleys' were chosen because their pitch contours follow approximately opposite shapes. We recorded four novel words (dax, fep, coro, seebu) with hill and valley contours. The novel word recordings were correctly assigned to the intended cluster when clustered with the CHILDES nouns from Study 1. Each novel word was spliced with three different training sentence frames ("This [novel word] is neat", "Look at this [novel word]", "What a pretty [novel word]"). Each novel word appeared consistently only with a hill or valley contour during training, and the pairing between novel word and contour type was counterbalanced across participants, such that there was one 1-syllable word (dax, fep) and one 2-syllable word (coro, seebu) per contour type.

To avoid any influence of high- versus low-synchrony contours on performance during test trials, each novel word appeared at test with two different contours that were not used during training (i.e., rise or fall). We recorded all four novel words with rise and fall contours and spliced these recordings with 2 different testing phase prompts, one with a rise contour ("Can you find the [novel word]?") and the other with a fall contour ("Where is the [novel word]?"). All novel word recordings were normed to the same duration (0.80 s) and loudness (55db) in PRAAT for both training and test, and the sentence frames were matched for loudness.

The audio recordings were paired with isoluminant stimuli. During the training phase, each novel word was consistently paired with one isoluminant novel object. To keep the trials engaging, at each repetition of the novel word, the isoluminant object moved

within a small range in one of three directions (left-right, up-down, or a slight rotation side to side). During the testing phase, two isoluminant novel objects (target and distracter) appeared side by side. Target side was determined following one of two pseudo-random orders, such that the same side did not appear more than two times in a row. Pairings between novel objects and novel words were counterbalanced across participants.

4.1.3 | Procedure

During the experiment, toddlers sat on their caregiver's lap and an EyeLink 1,000 Plus eye tracker recorded their pupil size and eye movements. As in Study 2, caregivers wore opaque sunglasses to avoid influencing the toddlers' behavior. The experiment consisted of two phases: a training phase, during which participants learned novel word-referent pairs, and a testing phase, during which toddlers were tested on their knowledge of label-referent pairs.

In the training phase, the participants were exposed to sentences containing novel words. Each novel word was presented concurrently with a unique isoluminant object (see Figure 4). Two of the novel words were consistently introduced with a hill contour and the remaining two with a valley contour, counterbalanced across participants. Each sentence trial started with a short, variable-length ITI (min: 0.25 s, max: 0.75 s, mean: 0.5 s), or until the toddlers looked at the screen (whichever took longer), during which a blank screen was displayed. The target object then appeared on the screen without any audio for 1.5 s or until the toddlers looked at the object (whichever took longer), after which the sentence recording was played. The trial ended after 1.5 s of silence with the object on the screen. The training phase was structured in two main blocks. In each block, the toddlers had three back-to-back exposures to each of the four novel word-object pairs, yielding 12 trials in each block, or 24 total. In each sequence of three trials, a given novel word was embedded in each of the three testing sentence frames. In each block, the two words appearing in a hill contour appeared sequentially, followed by the two words appearing in a valley contour, counterbalanced across participants. An attention-getting clip was presented after every two blocks.

In the testing phase, the toddlers saw two of the isoluminant objects side by side and heard a prompt to find the object labeled by the novel word. Each test trial started with an ITI (see description of training phase), followed by a displaying of the target and distracter isoluminant objects. The objects remained on the screen without any audio for 1.5 s, or until the participant looked at the screen (whichever took longer), after which the prompt sentence played. The target word in the test sentences followed one of two types of different contours (rises and falls), and each word appeared equally with both. Following the test sentence, the two objects remained on the screen for another 3 s without any audio. There were 16 test trials, and a short attention-getting clip appeared after every four trials. Trials followed one of two pseudo-random orders, such that there were no more than two back-to-back repetitions of any word.

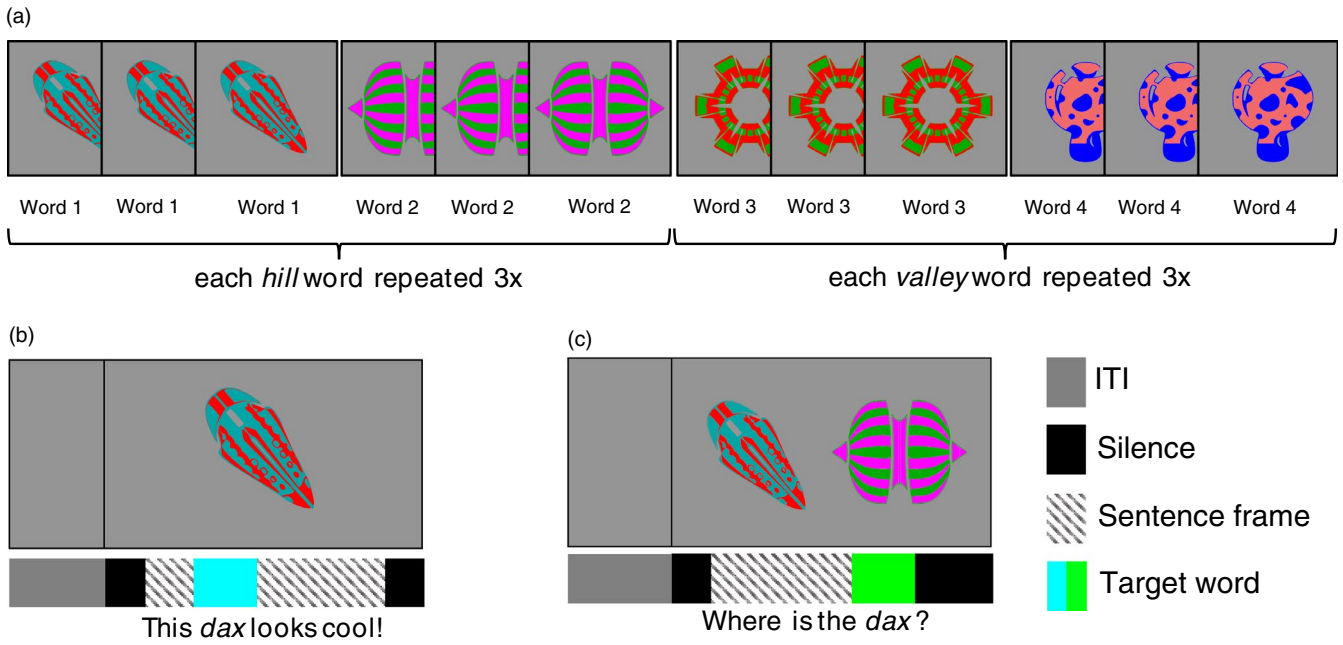

FIGURE 4 Study 3 experimental procedure. (a) Example block from the training phase, consisting of three back-to-back repetitions of each novel word. In each block, the two hill words were presented sequentially, followed by the two valley words (or vice versa). Each word was paired with a unique isoluminant object during training. (b) Example training trial, with the target novel word "dax" in a hill contour. The sentence was preceded and followed by 1.5 s of silence, and the isoluminant image remained on the screen for the entire duration of the trial. (c) Example test trial prompting the novel word "dax"; all test sentences used a fall contour. Two isoluminant images were presented side by side, one of which was associated with the novel word during training. The sentence was preceded by 1.5 s of silence and followed by 3 s of silence, and the isoluminant images remained on the screen for the entire duration of the trial.

4.1.4 | Pupil size and gaze position data pre-processing

We analyzed the pupil size time series data from the training trials using the same pre-processing steps and exclusion criteria as the sentence trials in Study 2. Then, we extracted the gaze position for the test trials. We excluded trials if more than half of the total data were missing or if a window of missing data was longer than 0.5 s.

4.2 | Results and discussion

We conducted three types of comparisons to evaluate whether or not pupil size synchrony during exposure to novel words predicts toddlers' learning.

4.2.1 | Comparing pupil size synchrony for hill and valley words during training

To confirm the difference in synchrony between novel words introduced with hill and valley contours during the training phase, we calculated the pairwise synchrony (normalized for duration; see Study 2) over the pupil size time series interval spanning the duration of the novel word in each sentence (0.80 s). Replicating Study 2, synchrony was significantly higher for hills than for valleys [Welch's *t*-test: $t(5,191.4) = -2.02$, $p < .05$, $d \sim 0.1$]. The same relationship held after accounting for participant variation; a predictor linear mixed-effects model with random and fixed effects for contour type performed better than the corresponding null model [χ^2 (3) = 431.86, $p < .001$, $f^2 \sim 0.14$].

4.2.2 | Learning of novel words presented in high- and low-synchrony pitch contours

Next, we compared the proportion of time toddlers spent looking to target novel objects during test trials for objects that had been presented in a hill versus valley contour during training trials (Figure 5b). Only the horizontal (*x*-axis) dimension of gaze position was used for data collected from test trials, because the calibration process only included fixation points along this axis. For each trial, we calculated the proportion of looks to the correct object, defined as the number of samples with gaze position within the bounds of the target, divided by the number of samples with gaze position within the bounds of any object (target or non-target). First, we tested whether toddlers' learning performance was above chance. A one-sample *t*-test showed that overall learning of novel words was significantly above chance for novel words presented with a hill contour [Welch's *t*-test: $t(156) = 1.93$, $p \sim 0.05$, $d \sim 0.15$], and at chance for novel words presented with a valley contour [Welch's *t*-test: $t(154) = 0.53$, $p \sim 0.59$, $d \sim 0.04$] (see Figure 5b). However, a mixed-effect beta regression predictor model (Brooks et al., 2017) with random participant intercepts and a fixed effect for contour type during exposure did not predict learning better than the null model [χ^2 (1) = 0.56, $p \sim 0.45$, $f^2 \sim 0.12$]. Thus, while toddlers showed above-chance learning of novel labels presented with a hill contour, contour type alone did not significantly predict learning.

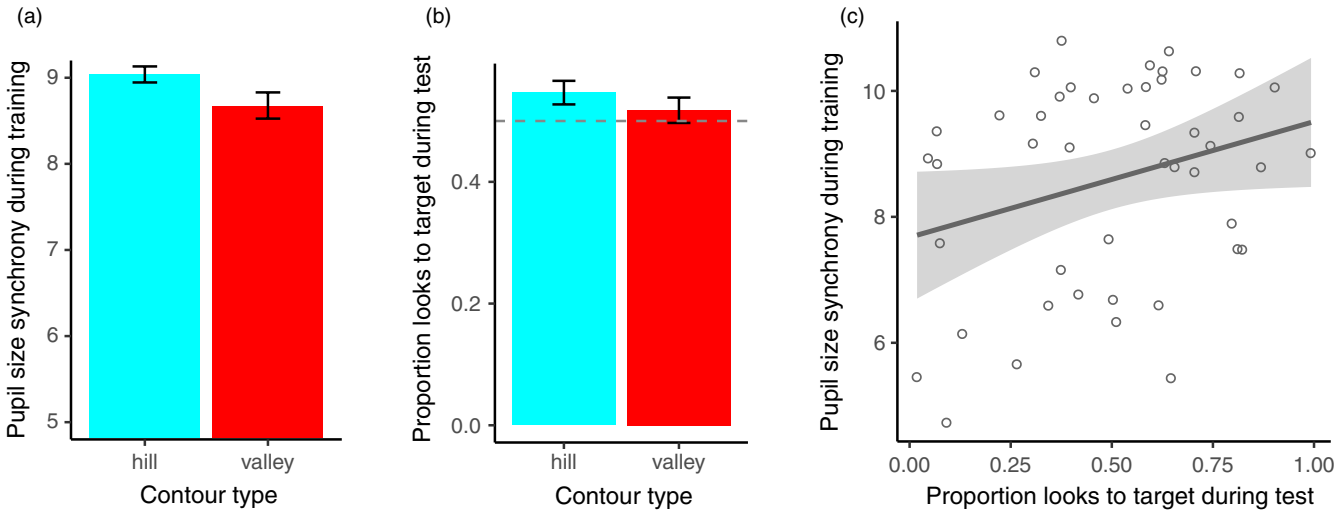

FIGURE 5 (a) Pairwise synchrony of pupil size time series spanning novel nouns during training, presented with a hill or a valley contour, respectively; (b) Proportion of looks to the labeled object during test trials for novel words that were introduced with a hill or a valley contour during training; (c) Correlation between pupil synchrony for a novel word during training (irrespective of its contour) and proportion of looks to the labeled object when the same word was presented at test. Each point's y-coordinate is the average synchrony for each novel word for all pairs that a given participant was a part of. Therefore, each participant contributed 4 points to this plot. The x coordinate is the average proportion of looks to the labeled object for that word for the given participant at test.

4.2.3 | Predicting learning at test from pupil size synchrony during training

Although hill contours, on average, elicited higher synchrony than valley contours, comparing average learning for the two contour types omits valuable information about each toddler's moment-to-moment engagement during training. To understand if moment-to-moment engagement with speech predicts novel word learning, we tested whether or not a given participant's proportion of looks to a target object at test could be predicted from their average pupil size synchrony for that word during training (Figure 5c).

To do so, we computed an average measure of the synchrony of each toddler for each novel word during training. We averaged the pairwise synchrony estimates (from the first analysis in Study 3) between each toddler's pupil response during all six exposures to the novel word (0.80 s) and the pupil responses of all remaining participants during the same novel word. Next, to compute a participant-level measure of comprehension, we averaged each participant's proportion of looks to the target for that word during the four test trials. Using these two measures, we performed a mixed-effects beta regression predicting toddlers' accuracy at test from synchrony during training. The predictor model with random participant-level intercepts and a fixed effect for synchrony predicted learning better than the corresponding null model [$\chi^2(1) = 9.99$, $p < .01$, $f^2 \sim 0.2$]. This indicates that toddlers' pupil size synchrony when hearing a novel word predicted their encoding and retention of that word at test. Together, the analyses from Study 3 reveal that toddlers' moment-to-moment engagement with the pitch contours of novel words impacts their learning, with particular benefits for words heard in higher-synchrony (hill-shaped) contours.

5 | GENERAL DISCUSSION

From the beginning of infancy and through early childhood, young children learn from and communicate with a dynamically changing speech signal. Much prior work has documented the characteristic ways in which CDS differs from ADS, and we know that prosodic characteristics of CDS rapidly change from one moment to the next. Critically, all aspects of language learning and communication occur over short, moment-to-moment variations in information signals. Study 1 used hierarchical clustering to identify four common word-level pitch contour types ('falls', 'rises', 'hills', and 'valleys') from natural speech to English-learning infants and toddlers. The next studies characterized the effects of sub-second variation in prosody on English-learning toddlers' attention and learning by adapting a pupil size synchrony measure from prior work with adults. Study 2 demonstrated higher pupil synchrony for CDS compared to ADS among two-year-olds, as well as higher synchrony during words presented with 'hill' and 'fall' pitch contours, compared to those presented with 'valley' or 'rise' contours. In Study 3, toddlers only showed above-chance learning when words were presented in contours that had elicited high synchrony (the 'hill' contour), suggesting that toddlers' attention is enhanced during moments of prosodic variation that are likely to carry useful information in their native language, such as pitch peaks that mark emphasis in CDS (Aslin, 1993; Fernald & Mazzie, 1991). Regardless of contour type, synchrony during the presentation of a novel word at training predicted learning of that word, indicating that pupil synchrony may serve as a signature of toddlers' attention to incoming information. This is the first study to use pupil synchrony on a shorter timescale, and we demonstrated that toddlers do in fact process speech in a consistent, shared, time-locked manner. Their pupils exhibit meaningful fluctuations in attention and learning, opening the door for future research on other attentional phenomena that unfold at short timescales.

By showing differences in the moment-to-moment dynamics of toddlers' attention during processing CDS prosody and ADS prosody, our research extends prior work showing young children's preference for listening and attending to CDS, compared to ADS (Cooper & Aslin, 1990; Fernald, 1985; Fernald & Kuhl, 1987; Pegg et al., 1992; Werker & McLeod, 1989) and stronger cortical tracking of speech rate for CDS compared to ADS (Kalashnikova et al., 2018). Our results suggest that CDS prosody drives not only an overall preference for the speech signal and higher engagement but yields a more consistent, time-locked pupillary response across child-listeners, compared to ADS prosody. Kang et al. (2014) and Smallwood et al. (2011) show that pupils respond to external stimuli in a time-locked way during moments of increased attention but not during moments of decreased attention. This suggests that the difference in synchrony between CDS and ADS might reflect greater attention to the speech stream in CDS.

There are two primary interpretations of this boost in the consistency of pupil responses across participants for CDS compared to ADS. First, synchrony may have resulted from higher endogenous attention to CDS. Over time, children might learn to associate features of CDS with higher value and relevance, because this signal is consistently used by caregivers to address the child. Therefore, upon detecting somebody's use of CDS prosody, they might tune in and maintain attention for the remainder of the utterance or conversational turn. This heightened attention, in turn, would automatically support better encoding of information. Alternatively, the trademark prosodic and emotional properties of CDS (Fernald & Kuhl, 1987; Grieser & Kuhl, 1988; Scherer et al., 1991) might drive exogenous attention to the speech stimulus. That is, specific cues in CDS may optimally capture young children's attention from moment to moment. Indeed, there is evidence that specific acoustic and positive/negative emotional cues elicit a pupil response in adults and 12-month-old infants (Frith, 1981; Geangu, Hauf, Bhardwaj, & Bentz, 2011; Partala & Surakka, 2003; Wetzel et al., 2016). We cannot currently pinpoint the endogenous or exogenous locus of the effects observed in Study 1, but the two may work in tandem to generate young children's higher pupil size synchrony in response to CDS.

In addition to addressing global engagement with CDS, the current investigation furthers our understanding of how toddlers process in-the-moment changes in CDS prosody. Variability in prosody

is one of the core features that defines CDS (Fernald & Kuhl, 1987; Grieser & Kuhl, 1988). Our results uncover toddlers' real-time tracking of patterns of variation, specifically in response to four word-level contour types, both when they were embedded in controlled sentences and when they occurred in a naturalistic children's story. From prior literature, we know that hill-shaped contours are commonly used by caregivers to highlight important words (Aslin, 1993; Fernald & Mazzie, 1991) and rise and hill contours may be used to capture or maintain an infant's attention and elicit a smile (Stern et al., 1982). Consistent with these prior findings, we found the highest synchrony in pupil response during familiar and novel words presented with a hill-shaped contour.

What gives rise to these differences in toddlers' engagement with different contour types? There are two main explanations for the connection between their increased engagement with hill contours and caregivers' use of this contour for emphasis. One possibility is that toddlers learn from regularities in caregivers' input and thereby associate 'hill' contours with relevant or engaging information. That is, they may learn that their caregivers use hill contours to mark important content. Support for such a learning mechanism could come from studies on cross-language variation in the contour types used to mark important words, plus evidence that young children engage differently with these contour types across languages. For instance, the use of pitch contours for emphasis might be different in tonal languages, where pitch contours are constrained by the tonal structure of the lexical unit. Chen and Gussenhoven (2008) showed that speakers of Standard Chinese exaggerate lexically defined pitch contours to mark emphasis, using relatively higher pitch at pitch peaks and lower pitch during pitch valleys.

The other possibility is that *caregivers* learn. Caregivers may adapt their input based on young children's preference—organic or otherwise—for the hill contour, and therefore use it when they want to direct children's attention toward something of relative importance. Transactional models of development suggest that young children play a central role in shaping their input, and caregivers adapt to children's preferences and needs (Sameroff, 2009; Sameroff & Chandler, 1975). Indeed, we know that caregivers change their input to young children as children's vocabulary changes (Schwab, Rowe, Cabrera, & Lew-Williams, 2018). This type of adult adaptation to the child is present even on a shorter timescale. Stern et al. (1982) showed that the infant's current state affected the mothers' pitch contours. For example, when the infant was gazing away from the mother, mothers used predominantly 'rise' contours, whereas when the infant was gazing directly at the mother, they used hill-shaped bell contours. Smith and Trainor (2008) directly tested how caregivers adapt their prosody in response to infant feedback, via manipulations of the feedback caregivers received. When mothers received positive reinforcing feedback from their infants in moments when their pitch was naturally increasing, they adopted higher overall pitch in their CDS; this was not the case for mothers who did not receive appropriately timed reinforcement. These findings suggest that the infants' behavior guides caregivers' use of prosody. The most likely possibility is that these two potential mechanisms work together to define toddlers' increased attention to hill contours. Through months and years of communicative interactions, caregivers may exhibit certain prosodic regularities, children may form increasingly precise preferences, and caregivers in turn may adapt their input via feedback from children.

The current research also speaks to how sub-second variations in CDS prosody can affect learning through modulations of the child's attention in the moment they occur. In Study 3, we showed that toddlers learned novel words presented in hill contours, which elicited high synchrony in Study 2, but not those presented in valley contours, which elicited relatively lower synchrony. There are, again, two main explanations that our data cannot disentangle. First, hill contours may improve word learning because they sound more natural, perhaps due to frequency in the input, and are therefore easier to process. We collected adult ratings of contour naturalness, and found that they rated words presented in hill contours as more natural than those presented in valley contours, both in the controlled sentence stimuli and the naturally occurring instances of these contours in the CDS story (see Figure S5). Furthermore, hills were approximately four times more prevalent in the corpus of CDS directed at 2-year olds (431 hills vs. 103 valley). However, hills were somewhat less prevalent than valleys in the infant-directed corpus (167 hills vs. 259 valleys), consistent with recent findings showing more even use of various contour types in infant-directed speech (Graf Estes & Zellou, 2019). Alternatively, caregivers may use hill contours for words that they want to receive special emphasis, e.g., to highlight a word as different from the rest of a sentence (Aslin, 1993; Fernald & Mazzie, 1991). Previous research supports the idea that emphasized words within CDS are learned better (Grassmann & Tomasello, 2007). Furthermore, consistent with previous research showing higher neural synchrony among adults who more successfully remembered a narrative (Simony et al., 2016), we found that increased synchrony during exposure to a novel label predicted better memory for the label-referent association at test. This suggests that fluctuations in attention might drive differences in toddlers' learning. Further research is needed, including analyses of additional corpora and experiments with additional languages, to determine if emphasis or naturalness contribute more to toddlers' enhanced learning from certain contours over others.

6 | CONCLUSION

This investigation introduces pupil size synchrony as a new approach for quantifying moment-to-moment attention in young children. In addition to this methodological contribution, pupil size synchrony highlights the value of examining CDS at two distinct timescales. On a longer timescale, we showed that the global properties of CDS engage toddlers' attention more effectively than ADS, and on a shorter timescale, we showed that toddlers track the pitch contours of important words. Critically, we found that toddlers' in-the-moment attention to pitch dynamics predicts their word learning. By studying fluctuations in young children's attention and learning from

CDS (even at sub-second timescales), we will be able to understand how rapidly unfolding cognitive processes interact with dynamically changing communicative signals.

ACKNOWLEDGMENTS

We thank the participating families and the members of the Princeton Baby Lab. Additionally, we thank Ting Qian for helpful advice on the statistical methods used in the paper. This work was supported by grants from the National Institute of Child Health and Human Development to C.L.W. (R01HD095912, R03HD079779), and a Princeton University C.V. Starr Fellowship to E.A.P.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at https://osf.io/5sq8h/?view_only=ecddeecb5bd04a2a9a715f657e3b25c5

ORCID

Mira L. Nencheva https://orcid.org/0000-0003-4854-4608
Elise A. Piazza https://orcid.org/0000-0001-6729-8559
Casey Lew-Williams https://orcid.org/0000-0002-8781-4458

REFERENCES

Aslin, R. N. (1993). Segmentation of fluent speech into words: Learning models and the role of maternal input. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 305–315). Dordrecht: Springer, Netherlands.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.08, retrieved 5 December 2019 from http://www.praat.org/.

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for "motionese": Modifications in mothers' infant-directed action. *Developmental Science*, *5*(1), 72–83. https://doi.org/10.1111/1467-7687.00211

Brooks, M., Kristensen, K., Benthem, K., Magnusson, A., Berg, C., Nielsen, A., … Bolker, B. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Butler, E. A. (2015). Interpersonal affect dynamics: It takes two (and time) to tango. *Emotion Review: Journal of the International Society for Research on Emotion*, *7*(4), 336–341. https://doi.org/10.1177/1754073915590622

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

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

De Groot, S. G., & Gebhard, J. W. (1952). Pupil size as determined by adapting luminance. *JOSA*, *42*(7), 492–495. https://doi.org/10.1364/JOSA.42.000492

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181–195. https://doi.org/10.1016/S0163-6383(85)80005-9

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. In J. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind* (pp. 391–428). Oxford: Oxford University Press..

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*(3), 279–293. https://doi.org/10.1016/0163-6383(87)90017-8

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. https://doi.org/10.1037/0012-1649.20.1.104

Frith, C. D. (1981). The effects of sound on pupil size and the pupil light reflex. *Personality and Individual Differences*, *2*(2), 119–123. https://doi.org/10.1016/0191-8869(81)90006-4

Geangu, E., Hauf, P., Bhardwaj, R., & Bentz, W. (2011). Infant pupil diameter changes in response to others' positive and negative emotions. *PLoS One*, *6*(11), e27132. https://doi.org/10.1371/journal.pone.0027132

Gleitman, L. R., Newport, E. L., & Gleitman, H. (1984). The current status of the motherese hypothesis. *Journal of Child Language*, *11*(1), 43–79. https://doi.org/10.1017/S0305000900005584

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy: The Official Journal of the International Society on Infant Studies*, *18*(5), 797–824. https://doi.org/10.1111/infa.12006

Graf Estes, K., & Zellou, G. (2019). "Prosodic variation in infant-directed speech", presented at the Society for Research in Child Development. Baltimore, MD.

Grassmann, S., & Tomasello, M. (2007). Two-year-olds use primary sentence accent to learn new words. *Journal of Child Language*, *34*(3), 677–687. https://doi.org/10.1017/S0305000907008021

Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, *24*(1), 14. https://doi.org/10.1037/0012-1649.24.1.14

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640.

Johannesson, T., Bjornsson, H. and Icelandic Met. Office, Grothendieck, G., & Johannesson, M. T. (2018). Stinepack: Stineman, a consistently well behaved method of interpolation. R package version 1.4. https://CRAN.R-project.org/package=stinepack

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585.

Kalashnikova, M., Peter, V., Di Liberto, G.M., Lalor, E.C., & Burnham, D. (2018). Infant-directed speech facilitates seven-month-old infants' cortical tracking of speech. *Scientific reports*, *8*(1), 1–8. https://doi.org/10.1038/s41598-018-32150-6.

Kang, O. E., Huffer, K. E., & Wheatley, T. P. (2014). Pupil dilation dynamics track attention to high-level information. *PLoS One*, *9*(8), e102463. https://doi.org/10.1371/journal.pone.0102463

Kang, O., & Wheatley, T. (2017). Pupil dilation patterns spontaneously synchronize across individuals during shared attention. *Journal of Experimental Psychology. General*, *146*(4), 569–576. https://doi.org/10.1037/xge0000271

Kaplan, P. S., Bachorowski, J.-A., & Zarlengo-Strouse, P. (1999). Child-directed speech produced by mothers with symptoms of depression fails to promote associative learning in 4-month-old infants. *Child Development*, *70*(3), 560–570. https://doi.org/10.1111/1467-8624.00041

Kaplan, P. S., Goldstein, M. H., Huckeby, E. R., Owren, M. J., & Cooper, R. P. (1995). Dishabituation of visual attention by infant- versus adult-directed speech: Effects of frequency modulation and spectral composition. *Infant Behavior and Development*, *18*(2), 209–223. https://doi.org/10.1016/0163-6383(95)90050-0

Katz, G. S., Cohn, J. F., & Moore, C. A. (1996). A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Development*, *67*(1), 205–217. https://doi.org/10.1111/j.1467-8624.1996.tb01729.x

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy: The Official Journal of the International Society on Infant Studies*, *4*(1), 85–110. https://doi.org/10.1207/S15327078IN0401_5

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

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

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

MacDonald, M. C. (2013a). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. *The Emergence of Language* (pp. 195–214). Psychology Press.

Macdonald, M. C. (2013b). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226. https://doi.org/10.3389/fpsyg.2013.00226

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (third edition): Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics*, *26*(4), 657. https://doi.org/10.1162/coli.2000.26.4.657

ManyBabies Consortium (2019). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52.

McMurray, B. (2016). Language at three timescales: The role of real-time processes in language development and evolution. *Topics in Cognitive Science*, *8*(2), 393–407. https://doi.org/10.1111/tops.12201

Merritt, S. L., Keegan, A. P., & Mercer, P. W. (1994). Artifact management in pupillometry. *Nursing Research*, *43*(1), 56–59. https://doi.org/10.1097/00006199-199401000-00012

Mittner, M., Boekel, W., Tucker, A. M., Turner, B. M., Heathcote, A., & Forstmann, B. U. (2014). When the brain takes a break: A model-based analysis of mind wandering. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(49), 16286–16295.

Montero, P., & Vilar, J. A. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software*, *62*(1), 1–43.

Nelson, D. G. K., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, *16*(1), 55–68. https://doi.org/10.1017/S030500090001343X

Papoušek, M., Papoušek, H., & Symmes, D. (1991). The meanings of melodies in motherese in tone and stress languages. *Infant Behavior and Development*, *14*(4), 415–440. https://doi.org/10.1016/0163-6383(91)90031-M

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Pegg, J. E., Werker, J. F., & McLeod, P. J. (1992). Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behavior and Development*, *15*(3), 325–345. https://doi.org/10.1016/0163-6383(92)80003-D

Piazza, E. A., Iordan, M. C., & Lew-Williams, C. (2017). Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology: CB*, *27*(20), 3162–3167.e3. https://doi.org/10.1016/j.cub.2017.08.074

Posner, M. I., Snyder, C. R., & Solso, R. (2004). Attention and cognitive control. *Cognitive Psychology: Key Readings*, 205–223.

Potter, C. E., & Lew-Williams, C. (2019). Infants' selective use of reliable cues in multidimensional language input. *Developmental Psychology*, *55*(1), 1–8. https://doi.org/10.1037/dev0000610

Rajkowski, J. (1993). Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. Soc. Neurosc. Abstract, Washington, DC.

Robaard, J. (2015). *The little mouse who lost her squeak*, New York, NY: . Little bee books.

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., … Cohen, D. (2013). Motherese in interaction: At the cross-road of emotion and cognition? (A Systematic Review). *PLoS One*, *8*(10), e78103. https://doi.org/10.1371/journal.pone.0078103

Sameroff, A. (2009). The transactional model. In A. Sameroff (Ed.), *The transactional model of development: How children and contexts shape each other*, (Vol. 290; pp. 3–21). Washington, DC: American Psychological Association, xiv.

Sameroff, A. J., & Chandler, M. J. (1975). Reproductive risk and the continuum of caretaking casualty. *Review of Child Development Research*, *4*, 187–244.

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, *15*(2), 123–148. https://doi.org/10.1007/BF00995674

Schmälzle, R., Häcker, F. E. K., Honey, C. J., & Hasson, U. (2015). Engaged listeners: Shared neural processing of powerful political speeches. *Social Cognitive and Affective Neuroscience*, *10*(8), 1137–1143. https://doi.org/10.1093/scan/nsu168

Schmälzle, R., Häcker, F., Renner, B., Honey, C. J., & Schupp, H. T. (2013). Neural correlates of risk perception during real-life risk communication. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(25), 10340–10347. https://doi.org/10.1523/JNEUROSCI.5323-12.2013

Schwab, J. F., Rowe, M. L., Cabrera, N., & Lew-Williams, C. (2018). Fathers' repetition of words is coupled with children's vocabularies. *Journal of Experimental Child Psychology*, *166*, 437–450. https://doi.org/10.1016/j.jecp.2017.09.012

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology*, *3*, 111. https://doi.org/10.3389/fpsyg.2012.00111

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*, 12141. https://doi.org/10.1038/ncomms12141

Smallwood, J., Brown, K. S., Tipper, C., Giesbrecht, B., Franklin, M. S., Mrazek, M. D., … Schooler, J. W. (2011). Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PLoS One*, *6*(3), e18298. https://doi.org/10.1371/journal.pone.0018298

Smith, N. A., & Trainor, L. J. (2008). Infant-directed speech is modulated by infant feedback. *Infancy: The Official Journal of the International Society on Infant Studies*, *13*(4), 410–420. https://doi.org/10.1080/15250000802188719

Soderstrom, M., Blossom, M., Foygel, R., & Morgan, J. L. (2008). Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, *35*(4), 869–902. https://doi.org/10.1017/S0305000908008763

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

Stephens G. J., Silbert L. J., Hasson U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, *107*, (32), 14425–14430. http://dx.doi.org/10.1073/pnas.1008662107.

Stern, D. N., Spieker, S., & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology*, *18*(5), 727–735. https://doi.org/10.1037/0012-1649.18.5.727

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy: The Official Journal of the International Society on Infant Studies*, *7*(1), 53–71. https://doi.org/10.1207/s15327078in0701_5

Tibshirani, R., & Walther, G. (2005). Cluster Validation by prediction strength. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *14*(3), 511–528. https://doi.org/10.1198/106186005X59243

Tomasello, M. (1992). The social bases of language acquisition. *Social Development*, *1*(1), 67–87. https://doi.org/10.1111/j.1467-9507.1992.tb00135.x

Tormene, P., Giorgino, T., Quaglini, S., & Stefanelli, M. (2009). Matching incomplete time series with dynamic time warping: An algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine*, *45*(1), 11–34. https://doi.org/10.1016/j.artmed.2008.11.007

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, *11*(3), 188–195. https://doi.org/10.1111/1467-9280.00240

Tukey, J. W. (1977). *Exploratory data analysis*, Reading, MA: . Addison-Wesley Publishing Company.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children*, *23*(3), 34–41.

Weist, R. M., & Zevenbergen, A. A. (2008). Autobiographical memory and past time reference. *Language Learning and Development: The Official Journal of the Society for Language Development*, *4*(4), 291–308. https://doi.org/10.1080/15475440802293490

Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology*, *43*(2), 230–246. https://doi.org/10.1037/h0084224

Wetzel, N., Buttelmann, D., Schieler, A., & Widmann, A. (2016). Infant and adult pupil dilation in response to unexpected sounds. *Developmental Psychobiology*, *58*(3), 382–392. https://doi.org/10.1002/dev.21377

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Nencheva ML, Piazza EA, Lew-Williams C. The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning. *Dev Sci*. 2021;24:e12997. https://doi.org/10.1111/desc.12997