Published in final edited form as:
J Exp Child Psychol. 2014 July ; 123: 73–89. doi:10.1016/j.jecp.2014.01.010.

Processing of Lexical-Stress Cues by Young Children

Carolyn Quam[1] and **Daniel Swingley**
Department of Psychology, University of Pennsylvania

Abstract

Though infants learn an impressive amount about their native-language phonological system by the end of the first year of life, after the first year children still have much to learn about how acoustic dimensions cue linguistic categories in fluent speech. The present study investigates what children have learned about how the acoustic dimension of pitch indicates the location of the stressed syllable in familiar words. Preschoolers (2.5–5 years) and adults were tested on their ability to use lexical-stress cues to identify familiar words. Both age groups saw pictures of a bunny and a banana, and heard versions of "bunny" and "banana" in which stress was either indicated normally with convergent cues (pitch, duration, amplitude, and vowel quality), or was manipulated such that only pitch differentiated the words' initial syllables. Adults (n=48) used both the convergent cues, and the isolated pitch cue, to identify the target words as they unfolded. Children (n=206) used the convergent stress cues, but not pitch alone, in identifying words. We discuss potential reasons for children's difficulty exploiting isolated pitch cues to stress, despite children's early sensitivity to pitch in language (e.g., Fernald, 1992). These findings contribute to a view in which phonological development progresses toward the adult state well past infancy.

Keywords

language development; phonological development; prosody; word recognition

Infants begin learning their native language in the first year, passing a series of milestones on their path toward mastery in interpreting the speech signal. They lose discrimination of some non-native consonant and vowel contrasts (Bosch & Sebastián-Gallés, 2003; Polka & Werker, 1994; Werker & Tees, 1984) and improve discrimination of subtle native contrasts (Kuhl et al., 2006; Narayan, Werker, & Beddor, 2010), which enables them to focus on the sound contrasts that their native language uses to differentiate words. They begin to use their language's prosodic properties to guide word-finding (e.g., Nazzi, Jusczyk, & Johnson, 2000; Friederici, Friedrich, & Christophe, 2007; see also Jusczyk, Cutler, & Redanz, 1993; Curtin, Mintz, & Christiansen, 2005; Höhle, Bijeljac-Babic, Herold, Weissenborn, & Nazzi, 2009). And they derive increasingly detailed knowledge of which sounds or syllables tend to

[1]Current affiliation: Department of Psychology, University of Arizona

go together (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Jusczyk, Luce, & Charles-Luce, 1994).

Yet toddlers are still immature in their interpretation of the speech signal. For example, one-year-olds do not consistently interpret phonological distinctions as indicating lexical distinctions, even for contrasts they readily discriminate (Apfelbaum & McMurray, 2011; Fennell & Werker, 2004; Rost & McMurray, 2009; Stager & Werker, 1997; Swingley & Aslin, 2007; Werker & Curtin, 2005). They are also more open-minded than older children about what can constitute a word, apparently accepting gestures and mechanical sounds, for example (Namy, 2001; Namy & Waxman, 1998; Woodward & Hoyne, 1999). Even older children's perception of speech sounds does not align fully with adult models in some phonetic details (e.g., Mayo & Turk, 2005). Thus, while children's (implicit) learning of their language's speech sounds implies significant capacities for analysis of regularities in the speech signal, the accomplishments of the first year are just the first steps in learning to interpret spoken language.

In order to fluently learn and recognize native-language words, children must not only apply their native sound contrasts, but must also disregard, at least at the word level, acoustic variation that their language does *not* use contrastively. But in doing so learners cannot simply discard this sort of variation, because it is often relevant at other levels of linguistic structure. For example, vowel duration in English is not a primary cue to vowel identity, but it partially cues speech rate and the voicing of the following consonant (Dietrich, Swingley, & Werker, 2007; van der Feest & Swingley, 2011) and therefore should not be ignored. The need to attend to non-segmental information also implies that sensitivity to statistical modes in phonetic distributions is unlikely to be sufficient alone as a learning mechanism that can lead to proper interpretation of duration or pitch movement. Children need not only segmental categories, but also interpretive models that assign appropriate communicative roles to phonetic variations.

A number of studies have shown children's ability to insulate lexical interpretations from phonetic variations irrelevant to lexical identity. For example, within the first year, infants improve at recognizing words despite non-phonemic changes in the talker's voice (Houston & Jusczyk, 2000), affect (Singh, Morgan, & White, 2004), and fundamental frequency, perceived as pitch (Singh, White, & Morgan, 2008). English-learning 18-month-olds do not treat large differences in vowel duration as contrastive even when given distributional evidence to the contrary (Dietrich et al., 2007). And in the case of pitch, Quam and Swingley (2010) demonstrated that by age 2.5, when English-learning children were taught a new word that was consistently uttered with one pitch contour, they recognized the word just as well when it was given a very different contour, suggesting that the children did not spontaneously treat pitch contour as a crucial component of a new word. Children appear to rule out pitch contour as lexically relevant in English sometime between 18 and 24 months (Singh, Hui, Chan, & Golinkoff, 2014).

Note that in all of these cases, word recognition is tested and the "right" answer (at least for English learners) requires indifference to the tested nonlexical variable, be it pitch contour, talker's voice, or talker's emotional state. There are relatively few tests of very young

children's appreciation of the *significance* of these phonetic variables in English (though see Creel, 2012). So, taking the case of pitch, although even toddlers have some correct intuitions about what pitch is *not* used for in English, it is less clear what young children think pitch *is* used for.

For example, it seems that it is not until roughly the age of four years that children begin to exploit pitch cues to a talker's emotions as adults do (Quam & Swingley, 2012), even when pitch indicates emotions like "happy" and "sad"—emotions children talk about before two and a half years (Fenson et al., 1994). This delay in pitch interpretation is surprising, given infants' early attention to the pitch characteristics of infant-directed speech (IDS; e.g., Fernald, 1985; 1992; Katz, Cohn, & Moore, 1996). One possible explanation for the delay could be variability in how underlying representations (happy/sad emotions) manifest in surface forms (realizations of pitch cues). The consistency of the relationship between underlying representations and surface forms appears to impact the acquisition trajectory of phonetic features, including pitch contours. For example, the tone patterns of words appear to be learned more quickly in lexical-tone languages like Mandarin than in grammatical-tone languages like Sesotho, where contextual "tone sandhi" effects are pervasive and children appear to learn tone patterns item-by-item (Demuth, 1995). And within Japanese, a pitch-accent language, the variable realization of the phrase-initial rising contour in adult speech appears to account for children's delayed acquisition of this pattern relative to the more consistently realized falling contour (Ota, 2003). Similarly, inconsistency in how pitch indicates the talker's emotions could slow children's learning of these cues (Quam & Swingley, 2012).

The present study investigates a case in which pitch serves a different linguistic function in English: to indicate the stressed syllable in a word. In English, polysyllabic words each contain one syllable with primary lexical stress. Which syllable this is varies from word to word; for example, *table* is trochaic, with stress in the first syllable, while *persist* is iambic, with stress in the second. Lexical stress in English is indicated by four acoustic cues. One is the pitch pattern: when a word has discourse focus or occurs in a prosodically prominent position (e.g., at the end of the utterance), its stressed syllable attracts a pitch accent, typically a pitch peak (Beckman & Edwards, 1994; Hayes, 1995). Stressed syllables in English also tend to be longer in duration and louder than unstressed syllables. Finally, in English, vowel quality is an additional cue to stress: unstressed syllables tend to contain the vowel schwa, whereas stressed syllables vary more in their vowel quality, generally being less centralized in the vowel space (Fry, 1958; Lieberman, 1960). Because the precise use and weighting of these cues differ across languages (Bertinetto, 1980; Berinstein, 1979), children must learn the cue weightings and realizations for their language.

Here, we asked whether children and adults can exploit stress patterns to efficiently recognize words that differ in first-syllable stress (*BUnny* vs. *baNAna*), and whether the pitch component of lexical stress is important in their determination of whether a syllable is stressed. Participants were tested in a restricted (but common) linguistic context in which the stressed syllable reliably contains a pitch peak (i.e., in sentences like "Look at the BUnny/ baNAna"). Because the context demands realization of the pitch cue to lexical stress as a pitch peak, pitch here is relevant to lexical differentiation.

Both age groups were tested in two conditions. In the *convergent-cues condition*, participants heard words in which pitch, amplitude, duration, and vowel quality all converged to indicate the location of the stressed syllable, just as they ordinarily do in English in this intonational/pragmatic context. Children might be expected to exploit these convergent cues to stress in word recognition, given infants' early attentiveness to the stress properties of their language (Friederici, Friedrich, & Christophe, 2007), and evidence that early word learners encode stress in their word representations (Curtin, 2010) and can learn minimal stress pairs (Curtin, 2009). On the other hand, children might not use stress effectively given its relatively low functional load compared with, for example, lexical tone in Mandarin Chinese (Cooper, Cutler, & Wales, 2002; Cutler, Dahan, & Donselaar, 1997), pitch accent in Japanese (Shibata & Shibata, 1990; in Sekiguchi & Nakajima, 1999), or even lexical stress in Spanish and Dutch (Hochberg, 1988; Cooper et al., 2002). Even so, we expected that children would indeed exploit the words' stress patterns in differentiating the test words.

In the *pitch-only condition*, participants heard words in which only isolated pitch cues indicated the location of stress. This condition speaks to the question of whether preschool children can flexibly adapt their cue weights to capitalize on the locally informative cue. Again, given infants' early sensitivity to pitch patterns in language (Fernald, 1985; 1992; Katz, Cohn, & Moore, 1996), we might predict that children would exploit isolated pitch cues. However, even if children make use of convergent stress cues, there are two reasons to think they may not use pitch alone.

First, children's cue weights may differ from adults'. Children sometimes over-rely on the most reliable or accessible cue to the exclusion of others. In one example from word segmentation, 7.5-month-olds over-rely on the tendency for words to begin with a stressed syllable, missegmenting "guitar is" as the strong-weak nonword "taris" (Jusczyk, Houston, & Newsome, 1999; see also Houston, Santelmann, & Jusczyk, 2004). At 15 months, children also appear to over-rely on the most salient or reliable acoustic cue to vowel contrasts (the first formant, F1), causing them to fail to discriminate vowel pairs that do not differ primarily in F1 (Curtin, Fennell, & Escudero, 2009). Children may overweight a particular cue because they find it easier to attend to, either because of nonadultlike general auditory processing (Mayo & Turk, 2005) or a nonadultlike phonetic system (Nittrouer & Lowenstein, 2007; Nittrouer, 1996). Among the cues to stress, pitch may be less reliable than the others, because of its use in signaling other linguistic features, such as intonational phrasing and sentential focus. Children might also need multiple cues to a category. Seidl (2007; see also Seidl & Cristià, 2008) has found that, even though pitch is a necessary cue for infants' clause-boundary segmentation, infants are not able to exploit pitch alone to segment clauses, but instead need two convergent cues.

Second, children might fail to use pitch as a cue in isolation because of difficulty flexibly adjusting perceptual weights to capitalize on the locally informative cue. Children exhibit more difficulty than adults adjusting their phonetic cue weights to compensate for effects of contextual variation such as noise or talker variability (e.g., Hazan & Barrett, 2000; Nittrouer, Miller, Crowther, & Manhart, 2000). While naturalistic lexical stress is indicated with four convergent cues, in the pitch-only condition listeners could best succeed by

adjusting their cue weights to down-weight amplitude, duration, and vowel-quality cues relative to pitch. Thus, difficulty flexibly adjusting cue weights could prevent children from capitalizing on isolated pitch cues in our task.

Given that we might predict late development of the ability to exploit either convergent or isolated-pitch cues to lexical stress, we tested children at a broad range of ages: 2.5 through 5 years. This enabled us to investigate whether developmental change occurs across the preschool years in children's ability to exploit stress cues during word recognition. Before testing children, in Experiment 1 we tested a group of adults in both conditions, to confirm the manipulation and provide a quantitative assessment of the mature state, particularly with respect to interpretation of the isolated pitch cue.

Experiment 1

We tested 48 adults' recognition of familiar words under two between-subjects conditions. In the *convergent-cues* condition, all cues jointly indicated stress. In the *pitch-only* condition, an isolated pitch cue indicated the stressed syllable, and all other stress cues were neutralized.

Method

Participants—We included 48 adults (20 women) in the analysis, all native speakers of English. All but two participants were undergraduates or very recent university graduates, assumed to be between 18 and 23 years of age, recruited primarily through a pool of students in introductory psychology courses. Eight more participated but were excluded: six for not being native English speakers, one for equipment failure, and one because his glasses interfered with the eye-tracking. Because it was likely that adults' responses to isolated pitch cues to stress would be less marked than their responses to convergent stress cues, we tested twice as many participants in the pitch-only condition (32) as in the convergent-cues condition (16).

Apparatus and Procedure—We used a language-guided looking procedure to investigate whether adults could exploit convergent vs. isolated-pitch cues to lexical stress during recognition of familiar words. Because adults participated in essentially the same experiment as the children in Experiment 2, experimental trials included only the words/objects *bunny* and *banana*, and the auditory stimuli were presented in a child-directed voice. To make this experience less odd, adult participants were told before the study that they would be helping to calibrate an experiment designed for young children.

Experimental (bunny/banana) and filler (other familiar-word) trials alternated. There were 16 trials of each of these types, making 32 trials total. In each trial, two pictures appeared on the computer screen; two seconds later, recorded sentences, referring to one of the two pictures, played from speakers on both sides of the screen. Of the 16 experimental trials, there were 4 each of correctly stressed "BUnny" trials (e.g., "Look at the BUnny"), misstressed "buNNY" trials, correctly stressed "baNAna" trials, and misstressed "BAnana" trials; these four target words were intermixed throughout the experiment. Eight attention-getting videos (e.g., an expanding and contracting star, or brightly colored shapes moving

around) were evenly spaced throughout, a manipulation used for both age groups but intended for maintaining children's attention in Experiment 2. For adults, there were also four filler trials (not eye-tracked) presented between each of the coded trials, so that adults saw five trials for every one trial children saw. These extra filler trials were intended to render the purpose of the experiment less obvious to adults. Because of these extra trials, the adult experiment was about 25 minutes long.

We used the Eyelink eye-tracking system to automatically code participants' eye-movements. The eye-tracker was an Eyelink CL (SR Research Ltd.), with an average accuracy of 0.5° and a sampling rate (from one eye) of 500Hz. The EyeLink eye-event detection system is based on an internal heuristic saccade detector. A blink is defined as a period of saccade-detector activity with the pupil data missing for three or more samples in a sequence. A fixation event is defined as any period that is not a blink or saccade.

The eye-tracking camera was mounted to the bottom of a 34.7 × 26.0 cm LCD computer screen. Before the experiment, we calibrated the eyetracker to the participant. First, a round sticker with a black-and-white target symbol on it was placed on the participant's forehead just above one of their eyebrows. Then the experimenter, viewing a live video of the participant's face on the computer monitor, checked that the eye-tracker had located the target symbol and the participant's pupil and corneal reflection (CR). The eye-tracker used the locations of the target symbol, the pupil, and the corneal reflection to compute the location of the subject's fixations to the screen. Once the target and pupil/CR were identified, the experimenter began the automated five-point calibration procedure, which involved drawing the participant's gaze to the four corners of the screen and the center in turn using a bulls-eye pattern. Once the calibration and validation were completed satisfactorily, the experiment began. During the experiment, if the eye-tracker lost the location of the pupil/CR, the participant was recalibrated between trials.

Auditory Stimuli—Experimental trials used the word-pair "bunny" and "banana" for four reasons. First, the words differ in their stress patterns—"bunny" has a stressed first syllable, while "banana" has an unstressed first syllable (and a stressed second syllable). Second, the vowel in the first syllable of "bunny" (IPA:/ʌ/), is acoustically similar to schwa (IPA:/ə/), making it easier for us to neutralize the vowel-quality contrast between stressed and unstressed first syllables so as to isolate the pitch cue in the pitch-only condition. Third, both bunnies and bananas are readily picturable. Finally, "bunny" and "banana" constitute the only word pair in most 2-year-olds' vocabularies that fit all three of these criteria.[2] We restricted ourselves to a word pair of this sort rather than simply misstressing words with unmatched partners because previous eyetracking work with 3-year-olds (de Bree, van Alphen, Fikkert, & Wijnen, 2008) did not show sensitivity to misstressings of iambic words in an unmatched-pair design. In principle, a paired design would be expected to be particularly sensitive because, for example, "buNNY" with an unstressed first syllable not only fails to match *bunny*, but also initially provides a good match to *banana*.

[2]Pilot testing also included the pair "button"/"balloon", but the L-coloring of the first vowel in "balloon" eliminated any ambiguity between the first syllables of the words.

Convergent-cues condition: For the convergent-cues condition, the first author produced stimuli in which she allowed all four cues to stress to covary between the two versions of each word. This meant that the first syllable of "BAnana" and "BUnny," in addition to containing a pitch peak, differed in three other ways from the first syllables of "baNAna" and "buNNY." It was longer in duration, higher in amplitude (the mean amplitude of the whole word was normalized to 70 dB, but relative differences between the syllables were maintained), and contained a/ʌ/vowel rather than schwa (see Figure 1, top; only one of the two tokens is depicted for each stimulus). We used two tokens of each word (e.g., two "BUnny" tokens); this was partly to reduce boredom in the child version (Experiment 2), and also to reduce the likelihood that participants could memorize the acoustic values of particular stimuli (e.g., the precise pitch values of "BAnana" versus "BUnny") to anticipate which word they were hearing.

Pitch-only condition: To test English-speaking adults' and children's ability to exploit the pitch cue to stress, for the pitch-only condition, we used Praat *Pitch Resynthesis* (Boersma & Weenink, 2008) to isolate the pitch cue to the stress contrast between "bunny" and "banana," holding the other three cues—amplitude, duration, and vowel quality—constant (see Figure 1, bottom). In simple declarative or Wh-question contexts like the sentences we used ("Look at the bunny/banana." and "Where's the bunny/banana?"), the stressed syllable of a word with utterance focus is indicated with a pitch peak during the stressed syllable (Hayes, 1995). In these simple sentence contexts, therefore, a word with a stressed first syllable will have an earlier pitch peak than a word with a stressed second syllable.

To isolate the pitch cue, the first author first recorded tokens of each word with stress on the first or second syllable (e.g., BUnny and buNNY) and also a "neutrally stressed" version of each word, in which she attempted to produce comparable values for amplitude, duration, and vowel quality in the first and second syllables, so that stress was not clearly on either syllable (though note that it is not possible to produce fully "neutral" stress, so the long first-syllable duration might have suggested first-syllable stress when listening incrementally). Mean amplitude between syllables of the "neutrally stressed" tokens was equalized using the "Shape Volume" function in the acoustic-editing software Goldwave.

Next, we superimposed the pitch contour from each of the stressed versions onto the neutrally stressed tokens, using Praat *Pitch Resynthesis* (Boersma & Weenink, 2008; see Streeter, 1978, for a similar procedure). Thus, the stimuli presented to listeners were two versions of the same recorded tokens, with superimposed pitch patterns taken from either trochaic or iambic recordings. This method enabled us to ensure that the two versions of *bunny* and of *banana* differed, to the extent possible, in only their pitch contours. Acoustic measurements for stimuli in both conditions are summarized in Table 1. (Note that the *Pitch Resynthesis* process isolates pitch as much as possible, but some small differences in amplitude and spectral content, which can be seen in Table 1, are unavoidable if stimuli are to sound like speech). Because of the complexity of creating these resynthesized stimuli (for instance, to maximize the naturalness of the resynthesis process, syllable durations had to be similar between the trochaic, iambic, and "neutrally" stressed original recordings), only one token of each experimental stimulus (e.g., "BUnny") was used, and was spliced into the two

carrier phrases, “Look at the [BUnny]” and “Where is the [BUnny].” Waveforms and pitch tracks for the resulting stimuli are shown in Figure 1 (bottom).

Visual Stimuli—Visual stimuli were color photographs on gray backgrounds. There were two different *banana* photos and two *bunny* photos (see examples in Figure 2), as well as two versions of each of the filler pictures. In pilot testing, participants (especially children) had a strong bias to fixate the *bunny* in bunny/banana trials, so for the experiments reported here we reduced the size and contrast of the *bunny* photos, and increased the size and brightness of the *banana* photos. This reduced (but did not eliminate) children’s baseline (before target-word) preference for the *bunny* object.

Data reduction—The 500-Hz output files of the EyeLink system were converted to ASCII format and condensed into target- and competitor-fixation proportions in 50-millisecond time-bins using custom Python scripts created by Sarah Creel. The result of this processing was information, for each trial, about the proportion of time each participant was fixating each location (target, distracter, other, offscreen, or lost data) during each 50-millisecond (ms) time bin.

Results and Discussion

Figure 3 plots adults’ fixations of the target picture over time in both the convergent-cues and pitch-only conditions. Time on the x-axis is displayed relative to target-word onset. Gaze proportions in Figure 3 are averaged within each 50-ms timebin, and numbers on the x-axis represent the end of each time bin. Zero on the x-axis thus represents the time bin from 50–0 ms before the onset of the target word. The ambiguous region, “bun,” ended at 610 ms for all stimuli (the first syllable, “bu…”, ended at roughly 450 ms; see Table 1 for details). By considering the time-course of participants’ responses, we can get a sense of how they integrated the information about first-syllable stress—which in “misstressed” trials should mislead them to fixate the distractor object—with the later-arriving segmental information starting in the second syllable (the “nny” in *bunny* vs. the “nana” in *banana*).

Convergent-cues condition—Gaze responses indicated that adults were very sensitive to convergent misstressings of the words, as expected. For both words, adults identified the target picture straightforwardly when stress cues were consistent with the word (“BUnny” or “baNAna”), but their target fixation dipped substantially when stress was inconsistent (“buNNY” or “BAnana”), before increasing to asymptote at 100% once segmental information was unambiguous (i.e., once adults had heard either the “nny” in bunny or the “nana” in banana). The effect of misstressing occurred faster for bunny than for banana (the “BUnny” and “buNNY” lines diverge sooner than the two *banana* lines).

Pitch-only condition—Looking-time measures revealed less sensitivity to mistressings of the pitch cue alone than of all four, convergent cues. Nevertheless, for both words, target fixation was higher in correctly stressed trials than in misstressed trials. When adults heard “BAnana,” with a pitch pattern appropriate for *bunny*, their target-fixation dipped as they looked over at the bunny picture, just as we found for convergent stress cues. When adults heard “buNNY,” with a pitch pattern appropriate for *banana*, their response was uncertain

(target fixation revealed a middling response, lower than when pitch signaled *bunny*, but still not so much as to favor *banana*). This could be because, for adults, pitch is asymmetric: high pitch definitely indicates a stressed syllable, whereas the absence of high pitch does not clearly indicate an unstressed syllable. A plausible additional explanation is that the pitch cue was competing with a long syllable duration, so adults were responding to cue *conflict* with uncertainty (whereas for BUnny, the pitch and syllable duration were convergent cues indicating a stressed syllable). Either way, the pitch manipulation affected listeners' interpretation of both words by drawing fixations in the predicted direction, though not always to the degree one might expect if pitch were the only cue to stress that listeners were accustomed to exploiting.

To compute inferential statistics on the looking-time data, we first averaged target-fixation proportions across the time window of 200–2000 ms. post–noun onset. The start of this window is the earliest adults can initiate an eye-movement response (Hallett, 1986). In Experiment 2, the time window began slightly later, at 350 ms., based on prior findings that before 367 ms., children are unlikely to be responding to the target word (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998; Swingley & Aslin, 2000; we used 350 ms. here because data were binned in 50 ms. chunks). The end of the window was chosen because after 2000 ms., both adults and children have usually finished identifying and fixating the target picture. Table 2 summarizes mean target fixations for this time window. We also calculated participants' asymptotic performance (the window of 2000–3000 ms.) to estimate their ultimate choice of target picture. Among adults, performance in this window was uniformly high: over 97% even on the misstressed, naturalistic-stress trials. We will see in Experiment 2 that children varied more than adults on this measure.

We next conducted an analysis of variance (ANOVA) by-subjects, in which the dependent variable was target-fixation proportion, and the predictors were the word ("bunny" or "banana"), the cue-type (convergent-cues or pitch-only), the pronunciation (correct or misstressed), and their interactions. Pronunciation exerted a significant effect on target fixation ($F(1,46) = 41.6$, $p < .001$), which was higher in response to correctly stressed words (*M*, 80.0%, *SD*, 8.6%) than misstressed words (*M*, 69.8%, *SD*, 9.1%; with a large effect size: paired Cohen's $d = 0.84$). There was also a significant interaction between pronunciation and cue-type ($F(1,46) = 12.6$, $p < .001$). In follow-up t-tests, convergent-cue and pitch-only participants did not differ significantly in target fixation on misstressed trials (convergent-cues: 67.7%; pitch-only: 70.9%), but convergent-cue participants showed significantly higher target-fixation in *correctly stressed* trials (convergent-cues: 85.8%; pitch-only: 77.1%; unpaired $t(36.3) = 4.0$, $p < .001$; with a large effect size: pooled Cohen's $d = 1.14$). Thus, when all cues indicated that the correct syllable was stressed (i.e., on correctly stressed trials in the convergent-cues condition), participants responded quickly and accurately. However, when cues were more mixed in pitch-only "baNAna" and "BUnny" trials—where amplitude, duration, and vowel-quality cues had been neutralized—participants were relatively slow in identifying the target words.

In sum, adults were sensitive to mispronunciations of the pitch cue to stress for both "bunny" and "banana," though they responded more strongly to convergent stress cues. In

Experiment 2, we next tested preschool-aged children's ability to exploit convergent vs. pitch-only cues to the location of stress.

Experiment 2

Method

Participants—Children between the ages of 2.5 and 5 years were tested. 206 children were included in the analyses: 100 in the convergent-cues condition, and 106 in the pitch-only condition (102 girls in total; mean age 4 years, 7 days, *SD,* 1 year, 27 days; the sample included 50 two-year-olds, 49 three-year-olds, 51 four-year-olds, and 50 five-year-olds, roughly equally distributed across the two conditions). Children were recruited via letters sent to parent addresses from a commercial database, and by word of mouth. Thirty-eight more children were excluded from the analysis: nineteen for inattentiveness, ten because they were hearing a language other than English at home more than 30% of the time or were exposed to a tone language (which could increase sensitivity to the pitch cue to stress), four for reported language or cognitive delays, two for talking/screaming over the auditory stimuli, one because of distracting noise in the room (from a younger sibling), and two because parents reported their children did not understand either the word "bunny" or "banana." Children were deemed inattentive if, in more than half (2) of the trials in each trial type (e.g., "BUnny" or "buNNY"), they failed to fixate the pictures for at least 300 milliseconds during the analyzed time window, 350–2000 milliseconds after noun onset.

Apparatus and Procedure—The procedure was very similar to that used with adults. The differences were that some children sat on their parents' laps, and the experiment was only about five minutes long, containing 32 trials and the attention-getting videos.

Results and Discussion

We tested children across a wide age range (2.5 to 5.0 years) in order to track the developmental timing of any relevant changes in phonetic interpretation over this period. In fact, in an analysis of covariance described below, children's performance within the analyzed time window (350–2000 ms. after noun onset) did not show important developmental change with regard to the variables of interest (number of cues to stress, and whether the word was correctly stressed vs. misstressed), though children as a group did respond differently from adults. Still, some age differences were noticeable in our measure of asymptotic performance (the analysis window 2000–3000 ms after the target word's onset). In experimental trials overall, 4- to 5-year-olds reached an asymptote of 76.5% target fixation (*SD,* 15.6%), whereas 2.5- to 3-year-olds reached 70.1% (*SD,* 14.7%). Thus, to further investigate effects of age on performance, we included analyses of covariance (ANCOVAs) investigating children's responses both in the primary analysis window (350–2000 ms.) and in this "asymptotic" window (2000–3000 ms.).

Convergent-stress condition—We predicted that children, like adults, would be sensitive to misstressings of "bunny" and "banana" when all four cues were allowed to covary in the typical English fashion, though this ability might vary across age (see analyses of covariance below). Plots of children's target fixation over time (Figure 4, left) indicate

that indeed, children fixated the target picture more when the word was correctly stressed than when it was misstressed. This was true for both words, in spite of an overall preference, beginning before target-word onset, to fixate the *bunny* picture. Children responded slightly earlier to misstressing of *bunny* than misstressing of *banana*, as adults had in Experiment 1.

Pitch-only condition—Two aspects of the pitch-only responses (Figure 4, right) are most salient. First, children again showed a bias, which preceded target-word onset, to fixate the *bunny* picture. Second, children showed little sensitivity to the pitch cues. They appeared to detect the anomalous pitch in "bunny" late in the time window; target fixation in response to the correctly stressed version of "bunny" began to exceed the misstressed version around 1000 ms post–target onset. However, this effect was numerically small and did not increase with age, and there was no such effect for "banana."

In order to evaluate whether children's age affected their sensitivity to mispronunciations of convergent vs. pitch cues, we conducted an analysis of covariance (ANCOVA) by-subjects in which the dependent variable was target-fixation proportion, the continuous predictor was age, and the categorical predictors were the word ("bunny" or "banana"), the cue-type (convergent-cues or pitch-only), and the pronunciation (correct or misstressed). Interactions of all predictors were also included.

There was a significant main effect of cue-type ($F(1,202) = 15.7$, $p < .001$), reflecting higher overall target fixation in response to convergent cues (*M*, 61.5%, *SD*, 9.5%) than the isolated pitch cues (*M*, 56.1%, *SD*, 10.1%; with a medium effect size: pooled Cohen's $d = 0.54$). Pronunciation also exerted a significant effect on target fixation ($F(1,202) = 46.2$, $p < .001$), which was higher in response to correctly stressed words (*M*, 63.1%, *SD*, 14.6%) than misstressed words (*M*, 54.4%, *SD*, 13.9%; with a small-to-medium effect size: paired $d = 0.43$). Cue-type interacted significantly with pronunciation ($F(1,202) = 28.7$, $p < .001$). Follow-up t-tests revealed that mispronunciations decreased children's fixation of the target picture in the convergent-cues condition (paired $t(99) = 8.6$, $p < .001$; with a large effect size: paired $d = 0.86$), but not in the pitch-only condition (see Table 2 for means). As with adults, the effect of cue-type appeared only in *correctly stressed* trials (unpaired $t(204.0) = 6.7$; $p < .001$; with a large effect size: pooled $d = 0.94$); the two participant groups did not differ in misstressed trials. There was also a significant effect of target word ($F(1,202) = 20.2$, $p < 001$; with a small effect size: paired $d = 0.31$), reflecting children's overall preference for *bunny* (*M*, 63.0%, *SD*, 18.1%) over *banana* (*M*, 54.5%, *SD*, 16.1). Pronunciation also interacted with target word ($F(1,202) = 11.3$, $p < .001$), reflecting a larger mispronunciation effect for "bunny" than for "banana" (see Table 2 for means), but the mispronunciation effect was significant for both words (*bunny*: paired $t(205) = 6.6$, $p < .001$; with a medium effect size: paired $d = 0.55$; *banana:* paired $t(205) = 3.1$, $p < .005$; with a small effect size: paired $d = 0.25$).

Children's age was not correlated with their target-fixation proportions during the primary time window. This is somewhat surprising because children generally improve with age in picture-fixation assessments of word recognition (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998; Fernald, Perfors, & Marchman, 2006), and speed increases between kindergarten and adulthood (Sekerina & Brooks, 2007). This prompted our analysis of

asymptotic performance in a later time window, where we evaluated whether children would show some developmental improvement in performance independently of their attempts to interpret the first syllable in our word pair. To this end we conducted a second ANCOVA in which asymptotic performance was the dependent variable (defined as target-fixation proportions averaged over 2000–3000 ms. after noun onset). As before, the predictors were the word (bunny/banana), the cue-type (convergent-cues/pitch-only), and the pronunciation (correct or misstressed). In this ANCOVA, we did find a main effect of the continuous variable age ($F(1,202) = 8.4$, $p < .005$), though it did not interact with our other factors of interest (all $p > .29$). A Pearson's correlation test indicated that age was positively correlated with overall asymptotic performance ($r = .21$, $p < .005$).[3]

In sum, both children and adults exploited convergent stress cues to identify familiar words. Both age groups responded more strongly to convergent misstressings than pitch-only mispronunciations, and in both cases this difference appeared in correctly stressed trials. However, whereas adults as a group used pitch to guide their identification of which word they were hearing, children did not. There was no difference across age in children's responses to misstressings of either convergent cues or pitch alone, though age was correlated with overall asymptotic performance in the task.

One possible reason why children might have been unable to exploit the pitch cues could be that in fact, contrary to our assumptions, pitch is an unreliable cue to stress in child-directed speech. To confirm the utility of pitch in contexts like those we tested here, we examined pitch patterns in bunny and banana in the Providence corpus of parental speech (Demuth, Culbertson, & Alter, 2006), available in the CHILDES database (MacWhinney, 2000). Because of strong influences of context on pitch realizations, we restricted the analysis to words in utterance-final position, and to words not in yes/no questions (for words in isolation, yes/no question was inferred using utterance prosody).

Twenty-one tokens of *banana(s)* and forty-nine tokens of *bunny/-ies* were included in the corpus analysis. Each token was hand-segmented into syllables using Praat (Boersma & Weenink, 2008), and a Praat script calculated pitch means, amplitude means, and durations for the first and second syllables. For each of these acoustic dimensions, a difference score was calculated to reflect the first-syllable value minus the second-syllable value. Analyses showed that pitch tended to fall from the first syllable of *bunny* to the second, and rise from the first syllable of *banana* to the second (Figure 5).

Linear mixed-effects logistic regression models were used to assess the power of pitch, amplitude, and duration to predict whether the word was *bunny* or *banana*. As there were five mother-child pairs in the corpus, mother-child pair was included as a random effect, and

[3]The asymptotic ANCOVA also revealed a main effect of pronunciation ($F(1,202) = 14.2$, $p < .001$; with a small effect size: paired $d = 0.21$), indicating higher asymptotic target-fixation in normal-stress trials (*M*, 75.5%, *SD*, 18.5%) than in misstressed trials (*M*, 71.2%, *SD*, 19.0%). Pronunciation also interacted with cue-type ($F(1,202) = 6.2$, $p < .05$). As before, mispronunciations decreased children's fixation of the target picture in the convergent-cues condition (paired $t(99) = 3.19$, $p < .005$; with a small effect size: $d = 0.32$; correctly stressed: *M*, 77.4%, *SD*, 18.7%; misstressed: *M*, 70.6, *SD*, 18.8%), but not in the pitch-only condition (correctly stressed: *M*, 73.8%, *SD*, 18.2%; misstressed: *M*, 71.7%, *SD*, 19.4%). Finally, there was a three-way interaction between word, condition, and pronunciation ($F(1,202) = 7.4$, $p < .01$), indicating that the effect of mispronunciations in the convergent-cues condition was significant only for *bunny* trials ($t(99) = 4.85$, $p < .001$; with a medium effect size, paired $d = .61$).

the full model included pitch-difference, duration-difference, and amplitude-difference as fixed effects. In the full model, only pitch-difference significantly predicted the word ($z = 2.72$, $p < .01$). The full model's performance was only marginally better than a model that included only pitch-difference as a fixed effect (Chisq(2) = 4.83, $p = .089$; full model comparison details are available from the authors). Thus, in this sample of child-directed speech from five American mothers, pitch appeared to be a better predictor of word stress than either amplitude or duration, at least for *bunny* and *banana* realized in salient prosodic contexts similar to those of our test sentences. This does not imply that pitch as a cue should be easy to learn, but it suggests that pitch could be a useful cue for differentiating iambic and trochaic words.

General Discussion

The two experiments described here showed differences between preschoolers and adults in their ability to exploit an isolated cue to lexical stress. Across the two experiments, we found that both adults and 2.5- to 5-year-old children exploited convergent cues to lexical stress in recognizing *bunny* and *banana*. Participants took longer to identify the target picture when the stress of the first syllable violated their expectations (as in "buNNY" and "BAnana") than when stress matched their expectations (as in "BUnny" and "baNAna"). The fact that even 2.5-year-old children exploited lexical stress when it was indicated by all four convergent cues is noteworthy and convergent with other recent findings by Curtin (2009, 2010; and with the predictions of Peperkamp, 2004). We might have predicted a more protracted acquisition pattern for lexical stress in English, given its low functional load relative to lexical stress in other languages. Whereas a Spanish-learning preschooler might know 'PApa' (potato) and 'paPÁ,' (dad), an English-learning preschooler is less likely to know minimal pairs like 'REcord' (the noun) vs. 'reCORD' (the verb). Lexical stress in English is also limited in its scope in that it cannot minimally contrast monosyllabic words as lexical tone does (Beckman & Edwards, 1994). Thus, it is perhaps surprising that by 2.5 years, English learners can use lexical stress as efficiently as 5-year-olds to predict the word they are hearing, differentiating between pseudo–minimal pairs like "bunny" and "banana."

Adults were able to exploit pitch cues presented in the presence of neutralized (or at least ambiguous) duration, amplitude, and vowel-quality cues, to recognize "bunny" and "banana," but children were not. Though age was positively correlated with asymptotic performance in the task, it did not predict responsiveness to mispronunciations of stress.

There are three plausible—and not mutually exclusive—explanations for children's insensitivity to isolated pitch cues in this task. The first possibility is that children's cue weighting or cue integration is different from adults'. It could be that children weight other cues (duration, intensity, or vowel quality) more strongly than pitch. Reliance on other cues could come about because the pitch cue to lexical stress interacts with sentence intonation, leading to variability in its realization (it has a high pitch target in "neutral" contexts like statements, but a low target in yes/no questions). The multiple functions of pitch in English (e.g., marking focus, conveying the speaker's emotions, etc.) also lead to ambiguity in how a pitch peak should be attributed. Both these factors likely reduce the reliability of the pitch cue.

A second possible reason why children did not exploit pitch cues to stress is that the manipulation of the pitch cue in the presence of static amplitude, duration, and vowel-quality cues introduced unavoidable cue conflict that might have posed particular difficulty for children. We attempted to remove the information value of the other three stress cues by recording a "neutrally" stressed version of "bunny" and "banana" and then superimposing the pitch contour from first- and second-syllable-stressed versions of each word onto the neutral token. However, there is no such thing as neutral stress in English, meaning that some amount of bias toward trochaic or iambic stress could not be avoided. In our stimuli, words were produced with a slow, exaggerated speech style, making it likely that the first syllable of the word (putatively the most important for recognition, since listeners process the words incrementally; Creel, Aslin, & Tanenhaus, 2006; Swingley, 2009) had duration and vowel-quality cues more consistent with its being stressed than unstressed (amplitude was controlled). Thus, when the pitch cue indicated second-syllable stress (in "baNAna" and "buNNY"), children might have struggled to resolve the conflict between the pitch cue and duration and vowel quality. By contrast, adults overcame this cue conflict more easily, to successfully exploit the locally informative pitch cues.

Our results suggest that both explanations might be relevant in describing children's behavior. Children's pitch-mispronunciation effects for both words were weak to nonexistent, and target fixation was low overall relative to the convergent-cues condition (see Figure 4 and Table 2), suggesting that children struggled to access the pitch cue in general, even when it converged with the other cues—when it indicated first-syllable stress. In other words, they did not reliably fixate the *bunny* more in response to the initial portions of "BUnny" and "BAnana" (the latter would have shown up as reduced target fixation relative to "baNAna") than in response to "buNNY" and "baNAna." However, target fixation did tend to be higher for "BUnny" than for "buNNY"; this might reflect the fact that "BUnny" was the most highly convergent of all words; all four cues to stress and the segmental content of the entire word all pointed to the word "bunny," whereas "BAnana," though it had convergent stress cues in the first syllable (pointing to *bunny*), had divergent segments in the second and third syllables ("nana"). Children's modest success in "BUnny" trials relative to the other three target words suggests that the task might have been difficult because of additive effects of (1) the subtlety of the pitch cue and (2) the conflict between it and the other cues to stress (which, again, was unavoidable given that "neutral" stress does not exist, so that information from the other cues could be weakened and made invariant, but not entirely extinguished).

A third and final difference between adults and children might be adults' greater ability to flexibly shift the weights of different cues to adapt to the particular context. In the pitch-only condition, pitch was the most reliable cue in the local context, in the sense that it was varying in the presence of static amplitude, duration, and vowel-quality cues. Of course, it was *unreliable* in the sense that half the time it indicated the wrong word (in "buNNY" and "BAnana"), but adults may have shifted their cue weightings to attend primarily to pitch, the cue that was varying across sentences. Flexibly shifting cue weights is crucial for identifying linguistic categories in noise and across different contexts, and has been demonstrated to develop over a protracted time course (Nittrouer, Miller, Crowther, & Manhart, 2000; Hazan

& Barrett, 2000; Cohn, 2011), which could explain why even 5-year-olds did not exploit the isolated pitch cue in our task. Of course, even adults performed best when all stress cues converged as in natural speech, but, crucially, in the pitch-only condition they were able to capitalize on the locally informative cue.

Despite evidence that young infants are highly sensitive to pitch, and despite the early acquisition of consonant and vowel categories, we have found that correct *interpretation* of discriminable pitch exhibits a protracted learning course. Children learn to rule out pitch as lexically contrastive in English by 24 months (Singh, Hui, Chan, & Golinkoff, 2014; Quam & Swingley, 2010), but they seem to take longer to learn to exploit pitch when it *is* relevant in English as a component of ensembles of phonetic cues to meaning. Children do not exploit pitch cues to emotions until around age 4 (Quam & Swingley, 2012), and the present study indicates that they struggle to exploit pitch cues to lexical stress even at age 5.

Based on these data alone, we cannot conclusively identify the cause of children's failure to exploit isolated pitch cues to stress. We have identified three possible—not mutually exclusive—explanations: differences in baseline cue-weighting strategies, inability to resolve conflict between the pitch cue and the other three cues in the stimuli, and inability to flexibly shift cue weights to capitalize on the locally informative cue.

The late developmental time-course found here for exploiting pitch cues to stress contrasts with the evidence that young infants are highly sensitive to pitch, that infants learn consonant and vowel categories by 12 months, and that children of the same age do exploit convergent stress cues. It emphasizes, however, that detecting patterns of sounds in language (consonants, vowels, and prosodic structure) is just a first step in phonological development. Children also must learn how their language weaves together these patterns of sound to convey meaning across several levels of linguistic analysis. In learning to properly attribute acoustic variation, children must cope with ambiguity in the assignment of acoustic cues to categories (e.g., whether a pitch peak indicates a stressed syllable, a focused word, or the speaker's excitement (cf. Dietrich et al., 2007) and variability in the realization of cues introduced by linguistic context, environmental noise, and other factors. Evidence from these experiments and others (e.g., Hazan & Barrett, 2000; Nittrouer, Miller, Crowthers, & Manhart, 2000; Quam & Swingley, 2012) suggests that this learning process continues well into childhood.

References

Apfelbaum KS, McMurray B. Using variability to guide dimensional weighting: Associative mechanisms in early word learning. Cognitive Science. 2011; 35:1105–1138. [PubMed: 21609356]

Beckman, M.; Edwards, J. Articulatory evidence for differentiating stress categories. In: Keating, PA., editor. Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III. Cambridge: Cambridge University Press; 1994. p. 7-33.

Berinstein AE. A cross-linguistic study on the perception and production of stress. University of California Working Papers in Phonetics. 1979; 47:1–59.

Bertinetto PM. The perception of stress by Italian speakers. Journal of Phonetics. 1980; 8:385–395.

Boersma, P.; Weenink, D. Praat: doing phonetics by computer (Version 5.0.30) [Computer program]. 2008. Retrieved from http://www.praat.org/

Bosch L, Sebastián-Gallés N. Simultaneous bilingualism and the perception of a language-specific vowel contrast. Language and Speech. 2003; 46:217–244. [PubMed: 14748445]

Cohn, A. Features, segments, and the sources of phonological primitives. In: Clements, GN.; Ridouane, R., editors. Where Do Features Come From?. Amsterdam: John Benjamins; 2011. p. 15-41.

Cooper N, Cutler A, Wales R. Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. Language and Speech. 2002; 45:207–228. [PubMed: 12693685]

Creel SC. Preschoolers' use of talker information in on-line comprehension. Child Development. 2012; 83:2042–2056. [PubMed: 22803622]

Creel SC, Aslin RN, Tanenhaus MK. Acquiring an artificial lexicon: Segment type and order information in early lexical entries. Journal of Memory and Language. 2006; 54:1–19.

Curtin S. Twelve-month-olds learn word-object associations differing only in stress patterns. Journal of Child Language. 2009; 36:1157–1165. [PubMed: 19281635]

Curtin S. Young infants encode lexical stress in newly encountered words. Journal of Experimental Child Psychology. 2010; 105:376–385. [PubMed: 20089259]

Curtin S, Fennell C, Escudero P. Weighting of vowel cues explains patterns of word object associative learning. Developmental Science. 2009; 12:725–731. [PubMed: 19702765]

Curtin S, Mintz TH, Christiansen MH. Stress changes the representational landscape: Evidence from word segmentation. Cognition. 2005; 96:233–262. [PubMed: 15996560]

Cutler A, Dahan D, van Donselaar W. Prosody in the comprehension of spoken language: A literature review. Language and Speech. 1997; 40:141–201. [PubMed: 9509577]

de Bree, E.; van Alphen, P.; Fikkert, P.; Wijnen, F. Metrical stress in comprehension and production of Dutch children at risk of dyslexia. Proceedings of the 32nd Boston University Conference on Language Development; 2008. p. 60-71.

Demuth, K. Problems in the acquisition of tonal systems. In: Archibald, J., editor. The Acquisition of Non-linear Phonology. Hillsdale, N.J: Lawrence Erlbaum Associates; 1995. p. 111-134.

Demuth K, Culbertson J, Alter J. Word-minimality, epenthesis, and coda licensing in the acquisition of English. Language & Speech. 2006; 49:137–174. [PubMed: 17037120]

Dietrich C, Swingley D, Werker JF. Native language governs interpretation of salient speech sound differences at 18 months. Proceedings of the National Academy of Sciences of the USA. 2007; 104:16027–16031. [PubMed: 17911262]

Fennell CT, Werker JF. Infant attention to phonetic detail: Knowledge and familiarity effects. Proceedings of the 28th annual Boston University conference on language development. 2004; 1:165–176.

Fenson L, Dale PS, Resnick JS, Bates E, Thal DJ, Pethick SJ. Variability in early communicative development. Monograph of the Society for Research in Child Development. 1994; 59:174–9. (serial no. 242).

Fernald A. Four-month-old infants prefer to listen to motherese. Infant Behavior and Development. 1985; 8:181–195.

Fernald, A. Meaningful melodies in mothers' speech to infants. In: Papousek, H.; Jurgens, U.; Papousek, M., editors. Nonverbal vocal communication: Comparative and developmental approaches. Cambridge: Cambridge University Press; 1992. p. 262-282.

Fernald A, Perfors A, Marchman VA. Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. Developmental Psychology. 2006; 42:98–116. [PubMed: 16420121]

Fernald A, Pinto JP, Swingley D, Weinberg A, McRoberts GW. Rapid gains in speed of verbal processing by infants in the second year. Psychological Science. 1998; 9:72–75.

Friederici AD, Friedrich M, Christophe A. Brain responses in 4-month-old infants are already language specific. Current Biology. 2007; 17:1208–1211. [PubMed: 17583508]

Fry D. Experiments in the perception of stress. Language and Speech. 1958; 1:205–213.

Hallett, PE. Eye movements. In: Boff, KR.; Kaufman, L.; Thomas, JP., editors. Handbook of Perception and Human Performance. New York: Wiley; 1986. p. 10-1-10–112.

Hayes, B. Metrical Stress Theory: Principles and Case Studies. Chicago, IL: University of Chicago Press; 1995.

Hazan V, Barrett S. The development of phonemic categorization in children aged 6–12. Journal of Phonetics. 2000; 28:377–396.

Hochberg JG. Learning Spanish stress: Developmental and theoretical perspectives. Language. 1988; 64:683–706.

Höhle B, Bijeljac-Babic R, Herold B, Weissenborn J, Nazzi T. Language specific prosodic preferences during the first half year of life: Evidence from German and French infants. Infant Behavior and Development. 2009; 32:262–274. [PubMed: 19427039]

Houston DM, Jusczyk PW. The role of talker-specific information in word segmentation by infants. Journal of Experimental Psychology: Human Perception and Performance. 2000; 26:1570–1582. [PubMed: 11039485]

Houston DM, Santelmann LM, Jusczyk PW. English-learning infants' segmentation of trisyllabic words from fluent speech. Language and Cognitive Processes. 2004; 19:97–136.

Jusczyk PW, Cutler A, Redanz NJ. Infants' preference for the predominant stress patterns of English words. Child Development. 1993; 64:675–687. [PubMed: 8339688]

Jusczyk PW, Friederici AD, Wessels JMI, Svenkerud VY, Jusczyk AM. Infants' sensitivity to the sound patterns of native language words. Journal of Memory and Language. 1993; 32:402–420.

Jusczyk PW, Houston DM, Newsome M. The beginnings of word segmentation in English-learning infants. Cognitive Psychology. 1999; 39:159–207. [PubMed: 10631011]

Jusczyk PW, Luce PA, Charles-Luce J. Infants' sensitivity to phonotactic patterns in the native language. Journal of Memory and Language. 1994; 33:630–645.

Katz GS, Cohn JF, Moore CA. A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. Child Development. 1996; 67:205–217. [PubMed: 8605829]

Kuhl PK, Stevens E, Hayashi A, Deguchi T, Kiritani S, Iverson P. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Developmental Science. 2006; 9:F13–F21. [PubMed: 16472309]

Lieberman P. Some acoustic correlates of word stress in American English. Journal of the Acoustical Society of America. 1960; 32:451–454.

MacWhinney, B. The CHILDES Project: Tools for analyzing talk. 3. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.

Mayo C, Turk A. The influence of spectral distinctiveness on acoustic cue weighting in children's and adults' speech perception. Journal of the Acoustical Society of America. 2005; 118:1730–1741. [PubMed: 16240831]

Namy LL. What's in a name when it isn't a word? 17-month-olds' mapping of nonverbal symbols to object categories. Infancy. 2001; 2:73–86.

Namy LL, Waxman SR. Words and gestures: Infants' interpretations of different forms of symbolic reference. Child Development. 1998; 69:295–308. [PubMed: 9586206]

Narayan CR, Werker JF, Beddor PS. The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. Developmental Science. 2010; 13:407–420. [PubMed: 20443962]

Nazzi T, Jusczyk PW, Johnson EK. Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. Journal of Memory and Language. 2000; 43:1–19.

Nittrouer S. Discriminability and perceptual weighting of some acoustic cues to speech perception by 3-year-olds. Journal of Speech & Hearing Research. 1996; 39:278–297. [PubMed: 8729917]

Nittrouer S, Lowenstein JH. Children's weighting strategies for word-final stop voicing are not explained by auditory sensitivities. Journal of Speech, Language, and Hearing Research. 2007; 50:58–73.

Nittrouer S, Miller ME, Crowther CS, Manhart MJ. The effect of segmental order on fricative labeling by children and adults. Perception & Psychophysics. 2000; 62:266–284. [PubMed: 10723207]

Ota M. The development of lexical pitch accent systems: An autosegmental analysis. Canadian Journal of Linguistics. 2003; 48:357–383.

Peperkamp SA. Lexical exceptions in stress systems: Arguments from early language acquisition and adult speech perception. Language. 2004; 80:98–126.

Polka L, Werker JF. Developmental changes in perception of nonnative vowel contrasts. Journal of Experimental Psychology: Human Perception and Performance. 1994; 20:421–435. [PubMed: 8189202]

Quam C, Swingley D. Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. Journal of Memory and Language. 2010; 62:135–150. [PubMed: 20161601]

Quam C, Swingley D. Development in children's interpretation of pitch cues to emotions. Child Development. 2012; 83:236–250. [PubMed: 22181680]

Rost GC, McMurray B. Speaker variability augments phonological processing in early word learning. Developmental Science. 2009; 12:339–349. [PubMed: 19143806]

Seidl A. Infants' use and weighting of prosodic cues in clause segmentation. Journal of Memory and Language. 2007; 57:24–48.

Seidl A, Cristià A. Developmental changes in the weighting of prosodic cues. Develomental Science. 2008; 11:596–606.

Sekerina IA, Brooks PJ. Eye movements during spoken word recognition in Russian children. Journal of Experimental Child Psychology. 2007; 98:20–45. [PubMed: 17560596]

Sekiguchi T, Nakajima Y. The use of lexical prosody for lexical access of the Japanese language. Journal of Psycholinguistic Research. 1999; 28:439–454.

Shibata T, Shibata R. Accent ha douongo wo donoteido benbetsu shiuruka: Nihongo, eigo, cyugokugo no baai. [Is word accent significant in differentiating homonyms in Japanese, English and Chinese?]. Mathematical Linguistics. 1990; 17:317–327. (in Japanese). Cited in Sekiguchi, T., & Nakajima, Y. (1999). The use of lexical prosody for lexical access of the Japanese language. *Journal of Psycholinguistic Research, 28*, 439–454.

Singh L, Morgan JL, White KS. Preference and processing: The role of speech affect in early spoken word recognition. Journal of Memory and Language. 2004; 51:173–189.

Singh L, Hui TJ, Chan C, Golinkoff RM. Influences of vowel and tone variation on emergent word knowledge: A cross-linguistic investigation. Developmental Science. 2014; 17:94–109. [PubMed: 24118787]

Singh L, White KS, Morgan JL. Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. Language Learning and Development. 2008; 4:157–178.

Stager CL, Werker JF. Infants listen for more phonetic detail in speech perception than in word-learning tasks. Nature. 1997; 388:381–382. [PubMed: 9237755]

Streeter LA. Acoustic determinants of phrase boundary perception. Journal of the Acoustical Society of America. 1978; 64:1582–1592. [PubMed: 739094]

Swingley D. Onsets and codas in 1.5-year-olds' word recognition. Journal of Memory and Language. 2009; 60:252–269. [PubMed: 20126290]

Swingley D, Aslin RN. Spoken word recognition and lexical representation in very young children. Cognition. 2000; 76:147–166. [PubMed: 10856741]

Swingley D, Aslin RN. Lexical competition in young children's word learning. Cognitive Psychology. 2007; 54:99–132. [PubMed: 17054932]

van der Feest SVH, Swingley DS. Dutch and English listeners' interpretation of vowel duration. Journal of the Acoustical Society of America. 2011; 129:EL57–63. [PubMed: 21428468]

Werker JF, Curtin S. PRIMIR: A developmental framework of infant speech processing. Language Learning and Development. 2005; 1:197–234.

Werker JF, Tees RC. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development. 1984; 7:49–63.

Woodward AL, Hoyne KL. Infants' learning about words and sounds in relation to objects. Child Development. 1999; 70:65–77. [PubMed: 10191515]

- We examine 2.5–5-year-old children's knowledge of how pitch cues lexical stress.
- We track participants' eyes as they match words (correctly/misstressed) to objects.
- Children at 2.5-years-old exploit naturalistic stress cues to recognize words.
- However, only adults exploit isolated pitch cues in the same task.
- We discuss potential causes of this protracted developmental time-course.

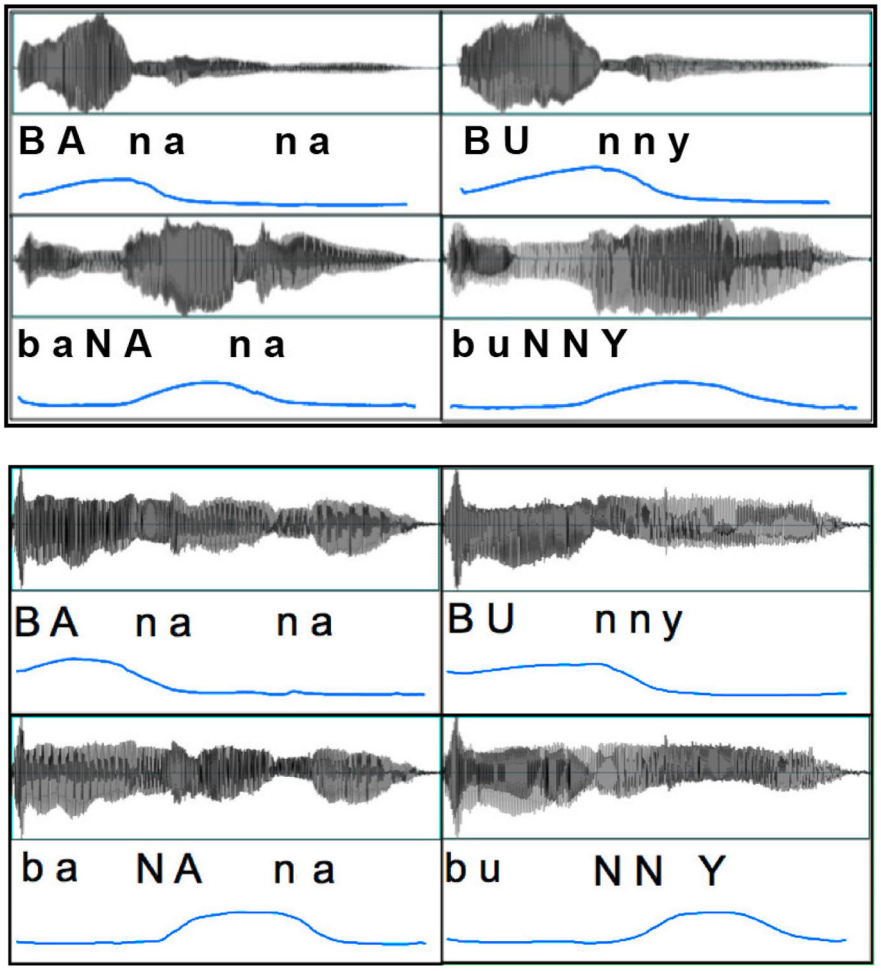
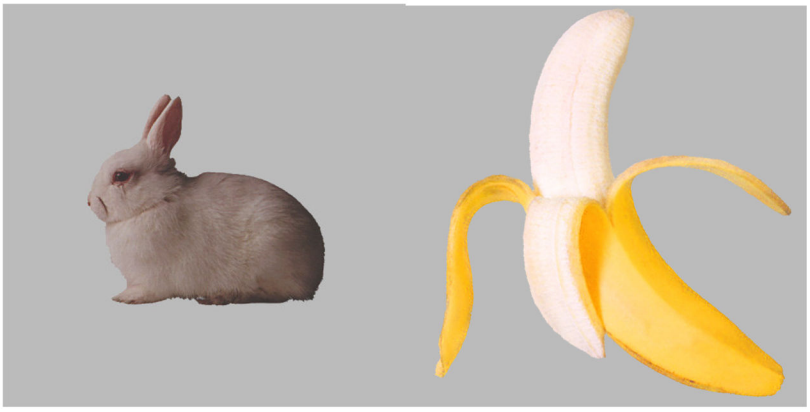


Figure 1.

Waveforms and pitch tracks for one of the two tokens of each of the auditory stimuli used in the convergent-cues condition (top) and for the stimuli used in the pitch-only condition (bottom).



"Where is the BUnny? That's pretty."

Figure 2. Example photographs used in both experiments, with example sentences
"Where is the BUnny? That's pretty."

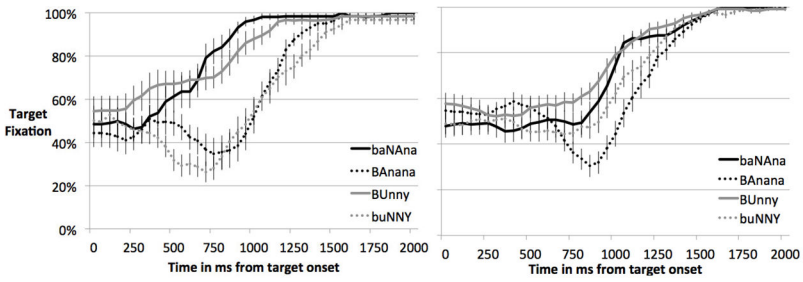


Figure 3.

Adults' target fixation over time in response to convergent stress cues (left) and isolated pitch cues (right).

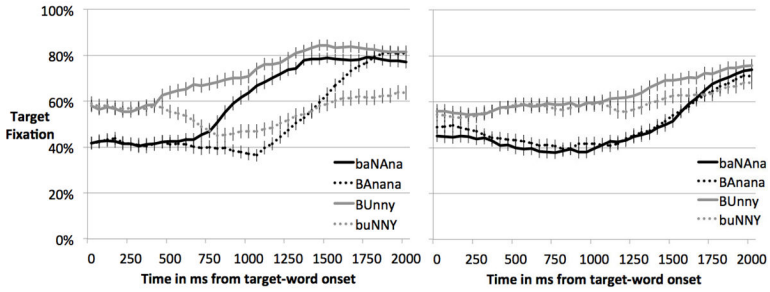


Figure 4.

Children's fixation of the target picture over time in response to convergent stress cues (left) and isolated pitch cues (right).

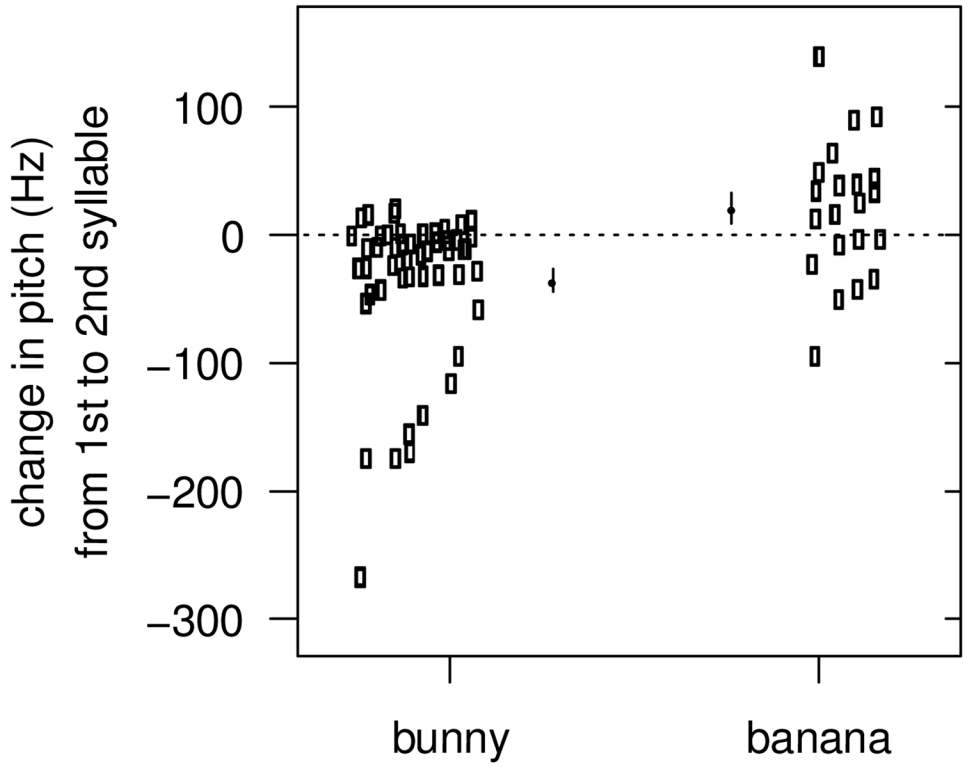


Figure 5. Change in pitch (in Hz) from the first to the second syllable for tokens of *bunny* and *banana* in the corpus analysis

Means and standard errors are indicated with the filled circles and vertical lines.

Table 1

Acoustic measurements for the first and second syllables of each target word used in the convergent-cues (averaged over the two tokens of each word) and pitch-only conditions.

	Pitch mean (SD)	Pitch max	Intensity	Duration	F1/F2
			Convergent cues – first syllable		
baNAna	201.5 Hz (10.1)	235.8 Hz	68.1 dB	0.26 sec	626.4/1751.5 Hz
BAnana	371.6 (49.9)	424.3	75.5	0.40	975.9/1420.5
BUnny	374.4 (52.6)	431.6	74.5	0.40	920.5/1610.5
buNNY	195.4 (4.9)	204.6	69.3	0.17	601.4/1748.6
			Convergent cues – second syllable		
baNAna	303.3 Hz (80.8)	404.6 Hz	73.0 dB	0.49 sec	726.7/1749.0 Hz
BAnana	257.2 (65.5)	409.2	65.6	0.46	751.2/2111.6
BUnny	250.3 (69.1)	414.0	62.1	0.69	435.5/2597.8
buNNY	260.2 (66.2)	358.8	70.1	0.89	412.9/2241.8
			Pitch only – first syllable		
baNAna	200.8 Hz (2.9)	209.4 Hz	70.6 dB	0.49 sec	891.9/1606.2 Hz
BAnana	387.4 (29.4)	420.5	72.1	0.49	862.8/1668.4
BUnny	368.4 (20.7)	394.2	69.8	0.42	987.4/1749.2
buNNY	200.4 (4.9)	215.2	70.0	0.43	798.0/1741.3
			Pitch only – second syllable		
baNAna	332.6 Hz (71.2)	395.3 Hz	69.9 dB	0.52 sec	850.6/1779.4 Hz
BAnana	222.1 (30.9)	317.3	69.8	0.52	818.2/2008.2
BUnny	233.1 (58.9)	394.1	68.4	0.75	511.2/2658.7
buNNY	292.7 (72.2)	387.3	67.6	0.74	464.2/2738.6

Table 2

Mean target-fixation (with standard deviations) in the two conditions—convergent (all) cues and pitch-cue only—for adults (Experiment 1) and children (Experiment 2).

	Correctly Stressed			Misstressed		
Word	**BUnny**	**baNAna**	**All**	**buNNY**	**BAnana**	**All**
Adults (Experiment 1)						
All cues	85.4% (12.0)	86.1% (9.1)	**85.8%** (6.6)	65.7% (9.9)	69.8% (13.3)	**67.8%** (8.0)
Pitch	78.9% (11.0)	75.3% (13.3)	**77.1%** (8.1)	72.3% (12.8)	69.5% (12.9)	**70.8%** (9.6)
Children (Experiment 2)						
All cues	74.5% (18.8)	64.4% (17.4)	**69.5%** (13.0)	54.0% (19.8)	53.4% (17.7)	**53.6%** (13.5)
Pitch	64.3% (22.4)	50.1% (19.7)	**57.1%** (13.5)	59.6% (26.5)	50.7% (21.2)	**55.2%** (14.2)