CHILDREN'S SENSITIVITY TO PITCH VARIATION IN LANGUAGE

Carolyn Quam

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2010

Supervisor of Dissertation

_____
Daniel Swingley
Associate Professor, Psychology

Graduate Group Chairperson

_____
Michael J. Kahana
Professor, Psychology

Dissertation Committee

Delphine Dahan, Associate Professor, Psychology

Daniel Swingley, Associate Professor, Psychology

John Trueswell, Professor, Psychology

Children's Sensitivity to Pitch Variation in Language

2010

Carolyn Quam

DEDICATIONS

This work is dedicated to Brian, who supported me through the ups and downs of graduate school (and edited this document); to my "Little Sister" Queen, who helped me maintain perspective; to my parents, Kathleen and Michael, and my sisters, Diana and Kristin; and, finally, to Alicia Parlette, whose courage inspired me to write.

ACKNOWLEDGEMENTS

Most heartfelt thanks go to my advisor Dan Swingley for his unfailing support and sound advice. Thanks also to current and former members of the Swingley lab, especially Katie Motyka, Jane Park, Allison Britt, Gabriella Garcia, Rachel Weinblatt, Alba Tuninetti, Sara Clopton, Rebecca McCue, and Kristin Vindler Michaelson, for help with participant recruitment, testing, and coding. As lab coordinator, Jane Park provided much appreciated feedback on the experimental design and stimuli for Chapter 3. The experiments in Chapter 3 would not have been possible without the support of John Trueswell and members of his lab, especially Katie McEldoon, Ann Bunger, and Alon Hafri, who helped us run our experiments at Philadelphia preschools. Members of the Institute for Research in Cognitive Science at the University of Pennsylvania, especially John Trueswell, Delphine Dahan, Steve Isard, Lila Gleitman, and Mark Liberman, gave valuable feedback on experimental design. Abigail Cohn at Cornell University helped guide my thinking about the experimental results presented in Chapter 4. The corpus-phonetics project described in Chapter 5 would not have been possible without the advisorship of Jiahong Yuan and technical support from Kyle Gorman. Finally, deepest thanks to the children, parents, preschool administrators and teachers, and Penn undergraduates for their participation in this research. Funding was provided by NSF Graduate Research Fellowship and NSF IGERT Trainee Fellowship grants to C.Q., NSF grant HSD-0433567 to Delphine Dahan and Dan Swingley, and NIH grant R01-HD049681 to Dan Swingley.

The baby, assailed by eyes, ears, nose, skin, and entrails all at once, feels it all as one great blooming, buzzing confusion.

William James, *The Principles of Psychology*

This was a most robust composition, a vigorous music that roused the senses and never stood still a moment. The boy perceived it as light touching various places in space, accumulating in intricate patterns.

E. L. Doctorow, *Ragtime*

One of my purposes was to listen, to hear speech, accent, speech rhythms, overtones and emphasis. For speech is so much more than words and sentences.

John Steinbeck, *Travels with Charley: In Search of America*

ABSTRACT

CHILDREN'S SENSITIVITY TO PITCH VARIATION IN LANGUAGE

Carolyn Quam

Daniel Swingley

Children acquire consonant and vowel categories by 12 months, but take much longer to learn to *interpret* perceptible variation. This dissertation considers children's interpretation of pitch variation. Pitch operates, often simultaneously, at different levels of linguistic structure. English-learning children must disregard pitch at the lexical level—since English is not a tone language—while still attending to pitch for its other functions. **Chapters 1** and **5** outline the learning problem and suggest ways children might solve it. **Chapter 2** demonstrates that 2.5-year-olds know pitch cannot differentiate words in English. **Chapter 3** finds that not until age 4–5 do children correctly interpret pitch cues to emotions. **Chapter 4** demonstrates some sensitivity between 2.5 and 5 years to the pitch cue to lexical stress, but continuing difficulties at the older ages. These findings suggest a late trajectory for interpretation of prosodic variation; throughout, I propose explanations for this protracted time-course.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 1: Introduction—Contextualizing Pitch Within the Study of Phonological Development

Though James (1890) described infants' early sensory experience as a "blooming, buzzing confusion," modern research on auditory and linguistic development suggests that infants rapidly zero in on their parents' speech and learn its regularities. Even fetuses sense the vibrations of their mothers' voices through the amniotic fluid. This early experience leads newborns to prefer their mothers' voices (DeCasper & Fifer, 1980).

If the mother's voice holds special status in the infant world, this solves the infant's most basic language-learning problem: identifying language out of all the auditory stimuli in the environment. But the next problem is even more daunting: sorting out the highly complex and multilayered structure of the native language. Language learning is not just about linking sound-forms (words) to meanings in the world. Though extracting meaning from speech is the child's ultimate goal, recognizing and differentiating words in fluent speech requires knowing the sounds that compose words, the category-boundaries between those sounds, and how coarticulation, when sounds combine, can shift those boundaries. Speech sound, or phoneme, categories must be learned, because they vary across languages. Above the word level, words are organized into phrases, whose meanings are determined by the order of words and by the interaction of words and intonation. How does the infant break into this complex system?

Though they may not yet comprehend much of what they are hearing, newborns are processing a complex acoustic stimulus that contains information about categories

including speech sounds, word boundaries, phrase boundaries, talker's voice characteristics, emotions, and lexical stress. The infant must take in this complex, multi-layered, continuously varying acoustic signal, find patterns in it, and attribute those patterns to the correct sources.

Adding complexity, many of the linguistic categories in the signal are cued simultaneously by multiple acoustic dimensions. Phrase boundaries are cued by pauses, intonation, and lengthening of the syllable just before the boundary; lexical stress is cued by duration, amplitude, pitch, and vowel quality. How does the child figure out that the intonational cue to a phrase boundary should be bound with the syllable-lengthening and pause cues that occur near the phrase boundary, while the pitch cue to lexical stress should be bound with the duration, vowel quality, and amplitude cues? As reviewed in the following paragraphs, learners must first be able to identify distributional patterns in the input; they must then use these distributions to form linguistic categories like phonemes or intonational phrases. Forming these categories and recognizing them in fluent speech is not trivial; it requires identifying the relevant cues to each category and weighting each cue appropriately. The next section reviews infants' sensitivity to distributions of sounds.

1.1 The Amazing Pattern-Detecting Infant!

Over the first year of life, infants become proficient at exploiting patterns to learn categories. By 10 months (but not at 4 and 7 months), infants can learn a visual (animal) category defined by correlations between body-part attributes (Younger and Cohen, 1983). They also exploit the particular distribution formed by the correlations between

attribute values. Ten-month-olds exposed to a bimodal distribution of attribute values learn two categories of animals, whereas those familiarized to a unimodal distribution learn only one (Younger, 1985; see also Mareschal & Quinn, 2001, for a review of how the distributions of exemplars in familiarization affect the breadth of infants' categories).

Infants also put their distributional-learning abilities to use in learning sound categories. Infants learn the categories /d/ and unaspirated /t/ when familiarized to a bimodal distribution of voice-onset time (VOT; Maye, Werker, & Gerken, 2002), but not when familiarized to a unimodal distribution. They also appear to learn allophonic variants—versions of the same sound realized differently as a result of context—by exploiting distributions. English-learning infants lose discrimination of the allophones aspirated /t/ (which occurs syllable or word initially) versus unaspirated /t/ (which occurs medially, e.g., after /s/ as in 'stand') by 10–12 months (Pegg & Werker, 1997). White, Peperkamp, Kirk, & Morgan (2008) demonstrated that 8.5- and 12-month-olds could learn phonological alternations from distributional information. Infants were familiarized to a voicing alternation in either stop consonants or fricatives that was conditioned by the voicing of the preceding determiner; voicing in the other type of segment varied freely. In test, both age groups preferred to listen to sequences of stimuli that, based on their familiarization, could be interpreted as alternating forms of a single word, over sequences that (though no more acoustically distinct) were alternations of two unrelated words. Only 12-month-olds appeared to have grouped the voiced and unvoiced sounds into one functional category, however.

Peperkamp, Le Calvez, Nadal, & Dupoux (2006) argue that to learn allophonic variants, children must integrate distributional learning with higher-level linguistic

constraints that prevent them from learning spurious allophones (unrelated sounds that occur in near-complementary distribution). According to Peperkamp et al.'s (2006) model, children track distributions of sounds, identify two sounds that occur in complementary distributions (like the two versions of /t/), and treat them as allophonic variants if two conditions are met. First, the two sounds must be phonetic neighbors; second, the allophone (unaspirated /t/) must be more similar to its context (e.g., initial /s/) than the default segment (aspirated /t/) is (Peperkamp, Le Calvez, Nadal, & Dupoux, 2006; see also Wilson, 2006, for a proposal that infants are biased to learn patterns that are phonetically natural). Similar integrations of distributional learning and prior linguistic knowledge have also been proposed for acquisition of semantic categories like noun versus verb (Braine, 1987).

Infants can also rapidly track the transitional probabilities between syllables and use this distributional information to infer word boundaries. Saffran, Aslin, & Newport (1996) created an artificial language in which the transitional probability between syllable pairs was a perfect cue to word boundaries; syllable pairs occurring within the same word were 100% likely to co-occur, while syllable pairs that straddled a word boundary had a much lower transitional probability (since at the word boundary several different syllables could come next). They found that 8-month-olds could extract words from continuous speech using only these transitional probabilities between syllables (see also Aslin, Saffran, & Newport, 1998).

Young infants also demonstrate knowledge of the distributional properties of their *native* language. Because trochaic (strong-weak) words are more common than iambic words in English, English-learning infants already have a bias to extract trochaic words

from a continuous speech stream by 7.5 months (Jusczyk, Houston, & Newsome, 1999). Infants also prefer to listen to high-probability sequences of sounds by 9 months (Jusczyk, Luce, & Charles-Luce, 1994); and they exploit allophonic cues to word boundaries by 10.5 months (Jusczyk, Hohne, & Bauman, 1999) and prosodic cues to word boundaries by 13 months (Christophe, Gout, Peperkamp, & Morgan, 2003).

1.2 Applying Distributional Learning to Form and Recognize Sound Categories

Infants are adept at identifying and tracking distributions in the speech they hear. But how do they translate these patterns into sound categories? Tracking distributions over syllables (as Jusczyk, 1993, argues that infants do) would provide the infant listener with many examples of voiced and unvoiced syllable onsets (/da/, /ba/, /gu/, /ti/, /pe/, /kuh/, etc.). These examples would be bimodally distributed in voice-onset time (VOT), suggesting that VOT is a relevant cue for differentiating syllables. After detecting a bimodal distribution, the infant could increase the attention weight of the VOT dimension for future speech perception. This selective attention would enable the infant to better differentiate words like “doll” and “tall” that differ in VOT.

Presumably by tracking these types of distributions, infants learn their language’s consonant and vowel inventory in their first 12 months, as evidenced by their loss of discrimination for some nonnative constrasts (e.g., Polka & Werker, 1994; Bosch & Sebastián-Gallés, 2003; Werker & Tees, 1984) and enhanced discrimination for some native contrasts (e.g., Kuhl et al., 2006; Kuhl, Conboy, Padden, Nelson, & Pruitt, 2005; Narayan, Werker, & Beddor, 2010). This early perceptual reorganization does not

translate to adult-like interpretation of speech, however. Though differentiating /d/ versus /t/ requires only attending to the VOT dimension, *recognizing* a /d/ or a /t/ and differentiating it from the set of all consonants in the language requires attending to multiple acoustic dimensions. In addition, as discussed in **Chapter 2**, interpreting these sounds in natural-language contexts requires accommodating a large amount of variation on acoustic dimensions both relevant and irrelevant to the identification of phonemes (e.g., Hazan & Barrett, 2000).

Consonants are typically described linguistically as varying on three phonetic dimensions: their place of articulation (where the articulators, e.g., the tongue, produce the sound), their manner of articulation (in stop consonants, air is completely obstructed, whereas in fricatives, air is compressed through a narrow opening), and their voicing (the time between when the obstruction in the mouth is released and when the vocal folds begin to vibrate). For example, to produce a /b/, the lips join together to block air from leaving the mouth, making /b/ a labial stop consonant. As the lips open to release the obstruction and allow air to move out of the mouth to produce sound, the vocal folds quickly begin to vibrate, making /b/ a voiced consonant. Phonetic descriptions like "voiced" and "labial" have corresponding acoustic signatures that the listener must detect in the speech input in order to identify a /b/ (voiced, labial stop) and discriminate it from an /s/ (unvoiced, alveolar fricative), for example.

Vowels are characterized phonetically by the location of the tongue in the mouth: its height and "backness" (orientation on the front-back dimension). Vowel height can also be described in terms of the openness of the jaw; when the tongue is high in the mouth the jaw is relatively closed, and a low tongue allows the jaw to be relatively open.

Acoustically, height/openness translates to the frequency of the first formant (F1; the first band of energy above the fundamental frequency in the sound spectrum), while backness translates to the second formant, F2. A closed front vowel like /i/ has a low F1 and a high F2, while an open back vowel like /ɑ/ has a high F1 and a low F2.

Correctly identifying a consonant or a vowel thus requires integrating multiple acoustic dimensions. Adding further complexity, the realization of a sound is heavily affected by its phonetic context. A back vowel (e.g., /ɑ/) produced after an alveolar consonant (e.g., /t/) typically has a higher F2 than a back vowel produced after a velar consonant (e.g., /k/), because the tongue does not have enough time to move from the front of the mouth, after producing /t/, to the back of the mouth to fully realize /ɑ/. These effects of coarticulation are especially strong in rapid speech, where there is little time for the tongue to move between articulatory targets. Because of coarticulation, even within a single speaker there is no invariant mapping between sounds and their acoustic realizations. Add large differences *between* speakers in the fundamental frequency, amplitude, and spectral characteristics with which they produce sounds (even without dialect differences, speakers' unique vocal tracts create differences in sound realizations), and it seems miraculous that infants ever correctly identify speech sounds.

These additional sources of complexity might explain why, despite infants' apparent sophistication with native-language sound categories, older children struggle to appropriately interpret variation they can readily perceive. They are less likely than adults to treat a subtle change in a consonant or vowel as signaling a new word (e.g., Nazzi, 2005; Pater, Stager, & Werker, 2004; Stager & Werker, 1997; Swingley & Aslin, 2007; White & Morgan, 2008). Before 18–20 months, they are also still willing to treat

gestures, pictograms, and nonverbal sounds as words (Namy, 2001; Namy & Waxman, 1998; Woodward & Hoyne, 1999). Both these sources of evidence suggest that knowing the native-language sound categories is only the first step in phonological development; the next step involves learning to apply that knowledge to interpret words.

1.3 Speech Simultaneously Conveys Categories at Multiple Levels of Structure

While the infant is tracking distributions of consonant and vowel sounds, the linguistic signal is also conveying information at other levels. Prosody, or speech rhythm, allows the speaker and listener to organize consonants and vowels into a rhythmic structure that is thought to be easier to perceive and produce fluently. Segmental information (/d/, /t/, /i/, etc.) is organized into syllables, which are either stressed (given primary or secondary stress) or unstressed. Syllables are integrated into prosodic words (or prosodic feet), each of which can contain only one syllable with primary stress. Prosodic words are then organized into phonological phrases and intonational phrases, which contain multiple words (e.g., "[the little dog] [was running fast]" is an intonational phrase containing two phonological phrases, indicated with brackets; Christophe, Peperkamp, Pallier, Block, & Mehler, 2004).

Structure at each of these levels is conveyed through, again, combinations of acoustic cues. For example, children must learn to integrate multiple probabilistic cues to identify word boundaries. A stressed syllable suggests a word boundary, since English words tend to begin with stressed syllables (though iambic—weak-strong—words are also fairly frequent). Phonotactic probabilities also help indicate word boundaries; some

sounds and sound combinations are restricted to certain positions in words or syllables. For example, in English the sound /ŋ/ (as in si**ng**i**ng**) cannot occur at syllable onsets (the way it does in Filipino; Narayan, Werker, & Beddor, 2010), so in the English phrase "singing in the rain," English phonotactics help the listener identify the word boundary between "singing" and "in," rather than segmenting "singi" "ngin." The allophonic realizations of phonemes provide another cue to word onsets (Jusczyk, Hohne, & Bauman, 1999), as do prosodic cues to word boundaries (Christophe, Gout, Peperkamp, & Morgan, 2003). Finally, the transitional probabilities between syllable combinations can also suggest the presence of a word boundary (Saffran, Aslin, and Newport, 1996).

Phonological and intonational phrases are also indicated through the convergence of multiple cues. Phonological-phrase boundaries are characterized by pre-boundary lengthening, a single melodic contour per phrase, exaggerated realization of the first phoneme in the phrase, and reduced coarticulation between segments that straddle the boundary. Intonational-phrase boundaries are marked by pauses, lengthening of the final syllable or word (Turk & Shattuck-Hufnagel, 2007), pitch declination, and a resetting of pitch to a higher value at the start of the next phrase (see Christophe, Peperkamp, Pallier, Block, & Mehler, 2004, for a review).

These facts about structure above the segmental level suggest that at *all* levels of linguistic structure, recognizing linguistic categories requires integrating multiple cues. In all these cases, the proper *weighting* of each acoustic cue must also be applied. Not only do linguistic categories vary across languages (e.g., English has more vowel categories than does Spanish), but, even when two languages have roughly equivalent phonological categories, subtle phonetic differences require listeners to learn different weights for the

acoustic cues to those categories. For example, lexical stress can be cued by duration, amplitude, pitch, and vowel quality. Italian relies more heavily on duration than on the other cues (Bertinetto, 1980), while the Mayan language K'etchi appears not to use duration at all to indicate stress, putatively because it uses duration to contrast words (Berinstein, 1979). English has been argued to rely more heavily on pitch than on intensity or duration (e.g., Morton & Jassem, 1965). In English (unlike, e.g., Spanish), vowels in unstressed syllables are heavily reduced compared to vowels in stressed syllables. As discussed in **Chapter 4**, these differences in the cues and cue weighting used to convey stress in different languages require the child to learn the weights of different acoustic/phonetic cues from experience with speech.

Given that it conveys multiple acoustic cues to multiple linguistic categories simultaneously, the speech signal is dizzyingly complex. How might the child sort out these different levels of structure and identify and properly weight the cues to categories at each level? The child's task has two components: integrating and properly weighting cues from different acoustic dimensions that indicate the *same* linguistic category (e.g., stressed syllables) and distinguishing cues to *different* linguistic categories that are cued by the same acoustic dimension (e.g., duration cues to lexical stress versus phonological-phrase boundaries versus intonational-phrase boundaries).

The following sections review work on cue weighting and integration, with the goal of addressing how children identify and learn to properly weight phonetic cues. We first summarize research on how adult learners form new categories by shifting attention weights; these studies also reveal causes of adults' suboptimal cue weighting. We then address children's cue weighting, and reasons it often differs from adults'.

1.4 Cue Weighting in Adults' Category Learning: Mechanisms for Category Formation

In order to successfully learn new categories, adults must identify the dimensions that best separate the categories and selectively attend to those dimensions. For example, in a task that involves judging a letter string as legal or illegal, if errors can only occur in the beginning of the string, adults can learn to focus only on that region (Haider & Frensch, 1996). This selective attention minimizes within-category differences and maximizes between-category differences (Kruschke, 1992). The learner can then use the relevant dimensions (or feature distributions; Fried & Holyoak, 1984) to classify new instances.

Category-learning models (e.g., Fried & Holyoak, 1984; Nosofsky, 1986, 1992; Kruschke, 1992) have often assumed that category learning must rely on existing dimensions rather than creating or attending to new features. Goldstone (1998), by contrast, proposed three ways that learners can create or attend to *new* dimensions (though see Francis & Nusbaum, 2002, *General Discussion*, for caveats regarding the difficulty of defining a "new" dimension). First, after sufficient exposure to members of the same category, stimulus imprinting creates specialized detectors to allow those stimuli or their features to be processed. Second, dimension differentiation allows stimuli and features that had once been indistinguishable to be discriminated (Goldstone, 1998). For example, Goldstone and Steyvers's (2001) participants learned to perceptually differentiate two dimensions in order to attend to the relevant one and ignore the irrelevant one. Finally, unitization fuses previously psychologically separate dimensions

to facilitate performance in a task that requires attending to both dimensions simultaneously (Goldstone, 1998).

1.5 Cue Weighting in Adults' Category Learning: Constraints on Learning

In acquiring new perceptual categories, adults often struggle to learn the optimal weighting of cues, whether because they cannot properly integrate multiple relevant cues (Ashby, Queller, & Berretty, 1999), they fail to deweight an irrelevant dimension (Francis, Baldwin, & Nusbaum, 2000), or their perceptual biases trump local evidence of the diagnosticity of cues (Holt & Lotto, 2006). Adults also struggle to modify their attention weights to different dimensions when the relevance of the dimensions changes (Goldstone & Steyvers, 2001). Adults expect exemplars of a new category to be normally distributed on the dimensions relevant for categorization, so they find it much easier to learn a category with a unimodal distribution than one with a bimodal distribution (Flannagan, Fried, & Holyoak, 1986). (Note that this describes the *within-category* structure, so is not at odds with findings by Maye, Werker, & Gerken, 2002, and others that infants can learn two categories from a bimodal distribution.) After learning one category with either a multimodal or a skewed distribution, however, adults are better able to learn a second nonnormally distributed category (Flannagan et al., 1986), suggesting that they flexibly shift their expectations about category structure.

Experience with the native language can also interfere with adults' learning of new categories, because the cue weighting that was appropriate for differentiating the native-language (L1) categories is inappropriate for learning the second-language (L2) categories. Native Mandarin-speaking adults appear to apply their Mandarin tone and

vowel categories in producing English lexical stress, for example. They produce higher fundamental frequency on stressed syllables than do English speakers, and inconsistently reduce unstressed vowels (Zhang, Nissen, & Francis, 2008). Presumably because Mandarin has fewer vowels than English, Mandarin speakers also use a more compressed vowel space than English speakers when producing English vowels (Chen, Robb, Gilbert, & Lerman, 2001). Speakers of Thai, which has fewer fricative categories than English, are more heavily affected by the discriminability of English fricative contrasts than are English speakers (Luksaneeyanawin, Burnham, Francis, & Pansottee, 1997). Speakers of French, which does not have contrastive stress, exhibit "persistent stress deafness" when learning Spanish (Dupoux, Sebastián-Gallés, Navarrete, & Peperkamp, 2008). Finally, Japanese speakers struggle to learn the English /r/-/l/ distinction. In both perception and production, they tend to over-rely on the second formant (F2), which differentiates Japanese /l/ versus /w/, and under-rely on F3, which is the primary differentiator of English /r/ versus /l/ (Lotto, Sato, & Diehl, 2004; Iverson et al., 2003).

Training with nonnative speech can help listeners reweight acoustic cues to differentiate nonnative speech categories. Francis and Nusbaum (2002) found that, with training, English listeners learned to attend to a new acoustic dimension to differentiate a 3-way Korean stop-consonant contrast, though they over-weighted this new cue. Adaptive training, which moves from highly exaggerated contrasts to more subtly contrasting tokens as participants' discrimination improves, is especially helpful for second-language learners. Adaptive training with feedback improved Japanese speakers' discrimination of the English /r/-/l/ distinction (e.g., McClelland, Fiez, & McCandliss, 2002; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002). Though they

continued to over-rely on F2, Japanese speakers learning English did begin to integrate F3—the primary cue to the /r/-/l/ contrast—into their productions (Lotto, Sato, & Diehl, 2004).

Learning trajectories for second languages (L2s) suggest that adults initially apply their native categories to discriminate L2 sounds if the native categories are somewhat similar acoustically to the new categories (as proposed by Best, 1994; see also Best, McRoberts, & Sithole, 1988). Applying existing categories to a new categorization task leads to misperceptions when the native categories are not identical to the new categories (as with Japanese learners of English), but it appears to speed up category learning (e.g., Fei-Fei, Fergus, & Perona, 2006). With training, L2 learners begin to integrate a new dimension into their categorizations, but still struggle to fine-tune the weights of each dimension.

Escudero and Boersma (2004) propose that adult learners of a new language only have particular difficulty acquiring new sound categories (relative to infant L1 learners) when contrasting them requires attending to a *previously used* dimension—one used to contrast sounds in the native language—in a new way. If the new contrast relies on a dimension that has *not* been used in the native language (a "blank slate" dimension; Escudero & Boersma, 2004), adults should apply the same learning mechanism children do: detect a bimodal distribution and use it to define two categories. Escudero and Boersma (2004) present evidence from Spanish learners of English /i/ versus /I/ vowel categories. The Scottish English /i/ versus /I/ contrast relies primarily on the first formant (F1), so Spanish learners of that contrast assimilated it to their /i/-/e/ vowel categories, which also contrast in spectral information. In the southern British dialect, by contrast,

the /i/-/I/ vowels contrast in both F1 and duration. In this case, Spanish learners relied on the dimension that Spanish does *not* use contrastively, distinguishing the vowels based only on duration. A few learners who had had more exposure to English appeared to begin integrating the two cues for better performance.

Training can also cause adults to show speech-like effects in categorizing *nonspeech* stimuli. Typically, coarticulation effects are found only with real speech. Aravamudhan, Lotto, and Hawks (2008) found, however, that training with sine-wave stimuli that mimic the properties of real speech led adults to incorporate context effects into their categorizations, the way they would with real speech. Aravamudhan et al. (2008) argued that training enabled listeners to establish perceptual categories for sounds, which were necessary to incorporate these context effects. Lotto and colleagues have also demonstrated that bird species can learn properties of speech like the vowel categories /i/ and /I/ (for which birds even show the perceptual-magnet effect; Kluender, Lotto, Holt, & Bloedel, 1998), and the correlation between voice-onset time and fundamental frequency (Holt, Lotto, & Kluender, 2001). These findings suggest that speech perception relies on general-purpose categorization mechanisms that are not specific to language, contradicting articulatory theories of speech perception (Bailey & Summerfield, 1980; Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989; Fitch, Halwes, Erickson, & Liberman, 1980), which have argued that speech perception is a highly specialized process that attends to the underlying articulatory gesture rather than the multiple acoustic cues that result from it.[1]

[1] Articulatory theories were inspired by findings that perceptual difficulties with nonnative speech categories did not carry over to *nonspeech* stimuli (e.g., Japanese speakers, who typically fail to exploit the F3 dimension to discriminate and produce English /r/ vs. /l/, perform just as well as English listeners at

1.6 Cue Weighting in Children's Category Learning

In order to become competent speech recognizers, infants must not only track regularities in the input, but must fine-tune the weights of different acoustic/phonetic cues. As discussed, adults can learn new categories by attending to previously relevant dimensions or by creating or attending to new ones. Their weighting of perceptual dimensions is not always optimal, however; it is subject to human perceptual biases as well as interference from existing categories like the sound categories of the native language.

There are several reasons children's weighting of perceptual cues might differ from adults'. First, infants appear to begin life with language-universal acoustic or phonetic weights that they must then adjust to match the native language sound categories. From birth, infants appear to possess rudimentary sound categories that are neither specific to language (Jusczyk, Pisoni, Walley, & Murray, 1980) nor unique to humans (Kuhl, 1981; Kuhl and Padden, 1982; 1983). This suggests that speech processing, a specialized system, initially relies on a more general perceptual categorization mechanism (Holt, Lotto, & Diehl, 2004). Exposure to the native language tunes vowel categories (Polka & Werker, 1994) and consonant categories (Werker & Tees, 1984) to the sound categories in the input by 6 and 12 months, respectively.

The precise nature of the reorganization process in the first year is debated. Kuhl's Native Language Magnet (NLM) theory (Kuhl et al., 2008; Kuhl, 2000; Grieser &

---

using F3 to discriminate nonspeech stimuli; Miyawaki et al., 1975), and by the phenomena of cue trading (in which weakness or ambiguity in one cue to phoneme identity is offset by a strong signal from the other cue; Bailey & Summerfield, 1980; Best, Morongiello, & Robson, 1981) and perceptual equivalence (in which adults cannot discriminate two ambiguous stimuli even though they are *acoustically* highly discriminable, differing strongly on two acoustic dimensions; Fitch et al., 1980; Best et al., 1981).

Kuhl, 1989; see also Boersma, Escudero, & Hayes, 2003) assumes that category prototypes drive perceptual reorganization and explain the perceptual warping effects whereby within-category stimuli near the prototype are less discriminable than stimuli farther from the prototype (but see Guenther & Gjaja, 1996, and Lotto, Kluender, & Holt, 1998, for alternative explanations of perceptual warping effects).

In contrast, Jusczyk's (1993) Word Recognition and Phonetic Structure Acquisition (WRAPSA) model does not posit category prototypes, arguing instead that phonetic learning involves reweighting acoustic dimensions in the input to focus on those that are meaningful in the native language. In WRAPSA, the auditory system automatically extracts certain acoustic features from the signal, regardless of the native-language structure. As the infant's native-language experience accumulates, the auditory-analysis system tweaks the weights of different features to highlight those that the language is using meaningfully. The resulting de-emphasis on irrelevant features (as opposed to the category reorganization assumed in NLM-e) causes the reduced discriminability of nonnative contrasts. Best (1994) shares this view, saying that "rather than causing a sensory-neural loss of sensitivity to nonnative distinctions, the native language most likely promotes an adjustment of attention to language-particular, linguistic characteristics of speech signals" (p. 173). The retention of irrelevant dimensions helps explain findings that adults are sensitive to talker's-voice characteristics that are irrelevant to the word-recognition task (e.g., Goldinger, 1996; Luce & Lyons, 1998; Palmieri, Goldinger, & Pisoni, 1993), while young children are even more susceptible to interference from irrelevant dimensions (Singh, White, & Morgan, 2008; Houston & Jusczyk, 2000; Singh, Morgan, & White, 2004).

Another reason children's cue weights differ from adults' is that children sometimes overweight cues that they find particularly salient. Overweighting of salient cues can occur for several reasons. For one, children sometimes have difficulty integrating multiple cues the way adults do, instead depending heavily on the most reliable or accessible cue. With age, children become better able to integrate diverse sources of information. For example, as discussed in **Chapter 4**, English-learning infants initially overrely on English's trochaic stress bias as a cue to word boundaries (e.g., Jusczyk, Houston, & Newsome, 1999), later integrating other cues to word boundaries, like phonotactic probabilities, allophonic distributions, and prosody.

Children also 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). For example, when 15-month-olds were habituated to words that differed in their vowels (e.g., "deet" vs. "doot"), and then tested on their detection of changes to the word-object pairings, they detected some changes but not others. This inconsistency appeared to stem from infants' reliance on the first formant (F1), which led them to successfully differentiate the /i/–/I/ contrast. They failed to differentiate /i/–/u/ or /I/–/u/, however, which require relying more heavily on other formants (F2 and F3; Curtin, Fennell, & Escudero, 2009). F1 may be more salient to infants because it contains more acoustic energy (Lacerda 1993, 1994), or children may have learned that F1 is generally a more reliable cue than F2 or F3 (Curtin et al., 2009).

In the Developmental Weighting Shift (DWS) model (Nittrouer, Miller, Crowther, & Manhart, 2000; Nittrouer, 2002), experience listening to and speaking the native

language leads children to shift the weights of different phonetic dimensions (in this sense it is compatible with WRAPSA; Jusczyk, 1993). The DWS model specifically argues that children are initially biased to attend to dynamic cues like formant transitions, that represent change over time in the vocal tract, over static cues like duration of silent gaps or of voicing. Nittrouer argues that this overreliance on formant transitions may initially serve to facilitate gross segmentation of syllables, and, "as the child becomes more skilled with a native language, the more static and/or more detailed components of the signal come to be weighted more heavily" (Nittrouer, 1996).

Evidence for the DWS comes from Nittrouer and colleagues' findings that children weight formant transitions more heavily than adults do relative to static cues like voicing duration or silent-gap duration. Nittrouer (2002) had 4-, 6-, and 8-year-olds and adults categorize ambiguous fricative-plus-vowel stimuli for which formant transition frequencies had been manipulated continuously while noise spectra indicated either /s/ or /sh/. With increasing age, listeners increased the weight of spectral information, and reduced the weight assigned to formant transitions. There was no effect of age on weighting of cues to an /f/–/θ/ contrast, for which formant transitions are the primary cue.

Other evidence suggests limits to the generalizability of the DWS hypothesis, however. Mayo and Turk (2004, 2005) have found that children's weighting of dynamic versus static cues depends on the particular phonemic contrast. In their studies, 3- to 7-year-olds weighted transitions more heavily than adults did for /saI/–/shaI/, and weighted transitions over other cues for /ta/–/da/. This was not true for /de/–/be/ and /ti/–/di/, however, suggesting that children's cue weighting differs with segmental context. Mayo and Turk (2005) also found that the physical distinctiveness of a cue to a particular

contrast affects children's reliance on that cue. For contrasts in which the formant transitions were spectrally distinct (had "extensively changing transitions"), like /no/–/mo/, children and adults weighted transitions to the same degree. For contrasts with less distinct transitions, like /ni/–/mi/, both children and adults weighted transitions less heavily, but younger children had particular difficulty relying on the less distinct transition cue.

Mayo and Turk (2005) suggest that children's sensitivity to cue distinctiveness reflects effects of general auditory processing on children's cue weighting, as opposed to the specifically *phonetic* preference for dynamic cues advocated by Nittrouer (2000, 2002). Nittrouer (1996) did find that children's auditory sensitivities for both fricative-noise spectra and F2-onset frequencies were less sharp than adults', raising the possibility that general auditory sensitivity was a factor in their cue weighting differences. Still, she argued that auditory-sensitivity differences could not fully explain developmental changes in cue weighting. Children rely more heavily on formant cues than adults do for categorizing ambiguous sounds, even though children's auditory sensitivity to formant transitions is inferior to adults' (Nittrouer, 1996). In addition, children do not show the same overreliance on formant cues when categorizing *nonspeech* stimuli (Nittrouer & Lowenstein, 2007). Both these findings suggest that auditory sensitivity limitations cannot fully explain children's nonadultlike cue weighting.

A final and crucial way in which children's cue weighting may differ from adults' concerns the need to flexibly adjust perceptual weights to compensate for effects of context (e.g., coarticulation of segments or talker variability). The reliability of formant-transition versus noise-spectrum cues to fricative identity depends on the context, so

adults and, to a lesser extent, 7-year-olds take context into account in their cue weighting. However, 5-year-olds apply a more rigid cue-weighting strategy that is not context sensitive (Nittrouer, Miller, Crowther, & Manhart, 2000). Adults adjust their weights for context even when stimuli are presented in reverse order from how they were produced, so they seem to be adjusting weights based on their accumulated experience with the effects of segmental order rather than on local order effects in the signal (they have "learned not to look to the vocalic portion of the syllable for information about fricative identity in [vowel-fricative] syllables because no information is usually available"; Nittrouer et al., 2000). This suggests that adjusting phonetic weights to the context requires accumulating sufficient experience with speech sounds. Children are also less flexible than adults at recognizing familiar phonemes despite limited acoustic cue information, even at age 12 (Hazan & Barrett, 2000), suggesting difficulty adapting cue weights to the particular context that continues well into childhood.

Learning to account for contextual effects in weighting phonetic cues may be especially difficult when cues to multiple categories are conveyed on the same acoustic dimension. For example, the duration of the preceding vowel is an important cue to the voicing of final stop consonants in English, but vowel duration also cues the emotions of the speaker and the ends of intonational phrases. The listener must therefore interpret vowel duration in conditions under which it may vary due to the emotion of the talker, the position of the word in the utterance, and the speech rate (Dietrich, Swingley, & Werker, 2007). Learning to account for these effects of context appears to take much longer than the early perceptual reorganization that establishes the inventory of native-

language sounds (Cohn, submitted; Nittrouer, Miller, Crowther, & Manhart, 2000; Hazan & Barrett, 2000).

1.7 Cue Differentiation and Integration in the Learning of Pitch Categories

Like duration, fundamental frequency, perceived as pitch, simultaneously conveys information at many levels of structure. This means that from a single acoustic dimension, and while processing speech as it unfolds in time, children must disentangle pitch cues to lexical stress, phrase boundaries, the speaker's emotions, characteristics of the speaker's voice, etc., and attribute each cue to the proper level of linguistic structure. Since many linguistic categories are indicated by a bundle of several cues, pitch cues must also be integrated with cues from other acoustic dimensions, and properly weighted, for successful recognition of each linguistic category.

As discussed further in **Chapter 5**, the need to integrate pitch cues with other cues to a given linguistic category might actually help address the issue of disentangling pitch cues to *different* categories. Children are not learning pitch categories in isolation; they are simultaneously exposed to structure on multiple acoustic dimensions. Temporal co-occurrence of cues is a strong indication that they may indicate the same linguistic category. The temporal co-occurrence of the pitch cue to lexical stress with the amplitude, duration, and vowel-quality cues, for example, likely helps children identify the pitch cue to stress and incorporate it into their representation of a stressed syllable. This would facilitate children's recognition of stressed syllables, by allowing them to attribute pitch peaks, when occurring synchronously with the other cues, to the category

stressed syllable. And accounting for the pitch cue to stress would in turn help children differentiate that cue from pitch cues to other linguistic categories. In this way, cue binding across acoustic dimensions should help children interpret pitch cues, and the learning problem should get smaller and more manageable over time, as more acoustic variation is attributed to the correct linguistic categories.

Though this iterative learning process seems plausible, the evidence presented in this dissertation suggests that children still take several years to correctly interpret various types of pitch variation. This long time-course could have several (not mutually exclusive) explanations. First, children might struggle to interpret pitch cues to categories like stressed syllables and *happy* or *sad* speech because several cues typically combine to indicate these categories. Because of this redundancy, any particular cue—like pitch—may not need to be very reliably realized in the signal. In some cases, children may have to learn the other cues to a linguistic category before they notice the co-occurrence between the pitch cue and the other cues.

The second possible explanation for the late acquisition of pitch cues relates to differences in children's versus adults' cue weighting for category learning. Because categories like stressed syllables and *happy* speech are indicated by multiple converging cues, children may initially rely on a more holistic recognition strategy that requires multiple cues for successful recognition (see also Seidl & Cristià, 2008) and that is less flexible in response to context than adults' cue weighting (Cohn, submitted; Nittrouer, Miller, Crowther, & Manhart, 2000; Hazan & Barrett, 2000). Children's difficulty coping with contextual variability may also explain why pitch categories appear to be particularly difficult for children to learn. The single acoustic dimension of fundamental

frequency (perceived as pitch) is simultaneously cuing multiple categories, which leads to unavoidable distortion as different pitch cues interact in speech (Xu, 1994).

1.8 Overview of the Present Research

This dissertation concerns how the child learns to identify pitch cues, attribute them to the proper linguistic categories, and bind them with cues from other acoustic dimensions to recognize those linguistic categories in fluent speech. In the following chapters, we consider the learning task for English learners, who must rule out pitch as lexically contrastive (because they are not learning a tone language), but must continue to attend to pitch as a cue to phrase boundaries, yes/no questions, emotional content, lexical stress, and other categories. In **Chapter 2**, we show that 2.5-year-old English learners, like adults, do not treat a consistent, stereotyped pitch contour as relevant at the word level. The finding that young children, and adults, do not appear to store pitch-contour information in the lexicon relates to a debate over the nature of word representations. It is consistent with either the idea that representations are abstracted away from the acoustics of each production of the word, or that acoustic dimensions are weighted to emphasize those that are relevant to differentiating the native-language sounds. It is not consistent with an extreme exemplar model, in which word representations consist of all the rich acoustic detail of every experienced token.

In ongoing, related research, we are investigating whether children's interpretations of potentially lexical pitch change across development. Given the evidence that the native-language phonology constrains children's interpretations of perceptible variation as they develop (e.g., Dietrich, Swingley, & Werker, 2007), we

predicted that younger children would be more open-minded about interpreting pitch as relevant at the lexical level. Our findings appear to support this prediction, suggesting that 18-month-olds are willing to learn two words contrasting in *either* pitch contour or vowel quality (Quam & Swingley, in progress).

We are currently probing whether 18-month-olds truly consider these two dimensions to be equally relevant, by presenting toddlers with word-object pairs for which the words, like before, contrast in either vowel (*vihdo* versus *vahdo*) or pitch contour (rise-fall versus low fall). Now, however, we are introducing variability on the other dimension. In the vowel-contrast case, *vihdo* and *vahdo* are each pronounced with four different pitch contours, while in the pitch-contrast case, the rise-fall and the low fall words are each pronounced with four different vowels. We propose that under these conditions, 18-month-olds may learn the vowel contrast but not the pitch contrast. In other words, when forced to choose between treating vowel or pitch variation as relevant, they may weight vowel information more heavily, because it is phonological in English.

While learning to disregarding pitch variation to recognize words across tokens, the English learner must still attend to pitch for functions including conveying emotional content and cuing lexical stress. In **Chapter 3**, we demonstrate that children do not successfully exploit pitch cues to the emotions *happy* versus *sad* until age 4–5. Though this late trajectory seems surprising given infants' well-established sensitivity to the pitch characteristics of infant-directed speech (reviewed in **Chapters 2–4**), it emphasizes the distinction between perception/discrimination and interpretation of perceptible variation. Interpreting another person's emotional expression appears to be more difficult than the lower-level perceptual sensitivities to emotional prosody demonstrated by infants.

In **Chapter 4**, we investigate children's developing ability to exploit the pitch cue to lexical stress. Adults and 2.5–5-year-olds saw pictures of a bunny and a banana, and heard versions of "bunny" and "banana" that were either correctly stressed ("BUnny" and "baNAna") or misstressed ("buNNY" or "BAnana"), with the location of stress marked using only pitch. We found a developmental change between 2.5 years and adulthood in use of pitch cues to lexical stress; 3- to 5-year-olds showed some sensitivity to the pitch mispronunciations, but not as reliably as adults did.

To begin exploring whether preschool children and adults are better at exploiting lexical stress in word recognition when *all* cues are provided, we next tested adults with naturally produced stress mispronunciations. We found that adults were significantly more sensitive to stress mispronunciations when all four cues (pitch, duration, amplitude, and vowel quality) were mispronounced than when the pitch cue alone was manipulated. We predict that preschoolers will exploit converging cues to lexical stress, and that their difficulty exploiting the isolated pitch cue reflects differences in their weighting of the pitch cue relative to adults (see Seidl, 2007; Seidl & Cristià, 2008), as well as their ability to flexibly shift weights to match the reliability of cues in the particular context.

The most striking finding that runs through all three chapters is the *late development* of learning how to attribute pitch variation. We are finding that children rule out lexical tone in English between 18 and 30 months and learn to interpret pitch cues to emotions and stressed syllables substantially later. Why these late acquisition trajectories relative to other linguistic categories? We know that after learning phonetic categories, children struggle to adapt their phonetic weights to account for contextual variation (e.g., Nittrouer, Miller, Crowther, & Manhart, 2000; Hazan & Barrett, 2000). Learners also

seem to have more difficulty acquiring a new contrast using a dimension already being used for other categories (Escudero & Boersma, 2004), so it is possible that the early importance of pitch for pragmatic and communicative functions temporarily blocks its use for learning linguistic categories. Children learn tone categories as early as they learn segment categories (Mattock & Burnham, 2006), however. More likely, the functions of pitch considered here have protracted acquisition trajectories because they require interpreting perceptible variation. Learning to properly interpret perceptible variation appears to follow a later time-course than the early loss of discrimination for nonnative sounds (Dietrich, Swingley, & Werker, 2007; Cohn, submitted).

Chapter 5 discusses the particular challenge of interpreting pitch variation (e.g., knowing whether a pitch peak signals a lexical tone, lexical stress, sentence focus, or excitement), and offers some possibilities for how children might solve this learning problem. It also discusses the importance of phonetic corpus analyses, which offer another perspective on the acquisition task by focusing on how phonetic categories are realized in speech to children. In ongoing work (Quam, Yuan, & Swingley, 2008; Quam, Yuan, Swingley, & Wang, in progress), we are using phonetic-corpus methods to ask how the input to children might help them identify pitch categories. Integrating phonetic corpus analyses with the experimental methods described above can help us paint a picture of how children converge on adult-like processing of speech.

Chapter 2: Phonological Knowledge Guides Two-year-olds' and Adults' Interpretation of Salient Pitch Contours in Word Learning

Carolyn Quam and Daniel Swingley[2]

Abstract

Phonology provides a system by which a limited number of types of phonetic variation can signal communicative intentions at multiple levels of linguistic analysis. Because phonologies vary from language to language, acquiring the phonology of a language demands learning to attribute phonetic variation appropriately. Here, we studied the case of pitch-contour variation. In English, pitch contour does not differentiate words, but serves other functions, like marking yes/no questions and conveying emotions. We show that, in accordance with their phonology, English-speaking adults and two-year-olds do not interpret salient pitch contours as inherent to novel words. We taught participants a new word with consistent segmental and pitch characteristics, and then tested word recognition for trained and deviant pronunciations using an eyegaze-based procedure. Vowel-quality mispronunciations impaired recognition, but large changes in pitch contour did not. By age two, children already apply their knowledge of English phonology to interpret phonetic consistencies in their experience with words.

[2] This chapter has been published in the *Journal of Memory and Language* as Quam and Swingley (2010a); see **References** for details.

2.1 Introduction

To acquire the phonology of their native language, children must learn to assign appropriate interpretations to various sorts of phonetic variation. This learning process begins early in development. During the first year of life, infants home in on their native language's consonant and vowel categories, becoming better at discriminating some acoustically difficult native contrasts (Kuhl, et al., 2006; Narayan, Werker, & Beddor, 2010) and worse at discriminating pairs of similar sounds that the native language groups into one category (Bosch & Sebastián-Gallés, 2003; Polka & Werker, 1994; Werker & Tees, 1984). By rendering irrelevant segmental distinctions difficult to discriminate, these developmental changes preclude certain linguistic errors. For example, an English-learning child who no longer readily perceives distinctions between dental and alveolar stop consonants is unlikely to mistakenly interpret dental and alveolar realizations of a word-initial /t/ as signaling two separate words.

There is more to phonological interpretation than the categorization of speech sounds, however. A great deal of phonetic variation that is readily perceptible may convey meaning at one or more levels of linguistic structure, in ways that are not universal across languages or retrievable from low-level distributional information in the signal. For example, vowel duration in American English serves functions like helping to signal prosodic boundaries (e.g., Salverda, Dahan, & McQueen, 2003; Turk & Shattuck-Hufnagel, 2000) and lexical stress (Lieberman, 1960), but generally provides only a secondary cue to identification of the vowel itself (e.g., Hillenbrand, Clark, & Houde, 2000). By contrast, many languages, like Japanese and Finnish, have distinct pairs of vowels that differ primarily in duration, so that identifying the exact vowel requires

evaluating its duration. Because vowel duration is informative about *something* in all languages, the learner's task is to discover its function in her particular language–not simply whether it can be ignored altogether (Dietrich, Swingley, & Werker, 2007).

Most research on early perceptual development in phonology has been concerned with the changing *discriminability* of native and nonnative speech-sound contrasts, but *interpretation* of the sounds in words likely follows a different developmental course, at least for some phonological features—and may be governed by different learning principles. Two lines of evidence suggest that one- and two-year-olds are still figuring out how to apply their phonological categories in interpreting new words. First, young children do not consistently interpret single phonological-feature changes as indicating lexical distinctions (e.g., Nazzi, 2005; Pater et al., 2004; Stager & Werker, 1997). For example, Stager and Werker (1997) habituated 14-month-olds to the words *bih* and *dih*, paired with two different objects (in the "Switch" procedure). Despite substantial training with the words, infants apparently failed to connect the words to the objects; they did not look longer when the taught word-object pairings were violated than when they were maintained. The same age group succeeded with the dissimilar words *lif* and *neem*, and 17-month-olds succeeded with the similar-sounding words (Werker, Fennell, Corcoran, & Stager, 2002; see also Fennell, Waxman, & Weisleder, 2007; Fennell, 2006; Thiessen, 2007; Yoshida, Fennell, Swingley, & Werker, 2009). Phonetic similarity also appears to play a stronger role among young children, relative to adults, in determining whether they treat phonological changes as indicating separate words—even when it is clear that children can *perceive* the phonological changes. Swingley and Aslin (2007) and White and Morgan (2008) found that 1.5-year-olds, upon viewing a familiar object (like a car)

and a novel object, did not assume that a novel phonological neighbor of the familiar object's label (such as *gar*) referred to the novel object, though they did make this inference with more phonologically distinct nonwords, and did show some sensitivity to the mispronunciations. Swingley and Aslin's (2007) participants also showed much worse performance in learning novel words that were phonological neighbors of familiar words than in learning nonneighbors.

The second line of evidence that young children are still learning how to apply their phonological categories to word learning comes from findings that they appear to be more open-minded than older children about what they will treat as a word. Children under 18 months sometimes interpret noisemaker sounds, melodies, and gestures as words, while older toddlers do not. Namy (2001) successfully taught 17-month-olds gestures, sounds, and pictograms as object-category labels by embedding the symbols in familiar labeling routines. Namy and Waxman (1998) similarly found that 18-month-olds were willing to interpret both gestures and novel words as category labels, but found that 26-month-olds were reluctant to learn gestures as category labels, and required more practice with gestures before they would do so. Finally, Woodward and Hoyne (1999) found that 13-month-olds could learn the pairing of a new toy with either a novel word or a noisemaker sound, while 20-month-olds did not. These findings suggest fundamental changes around 18–20 months of age in children's expectations about how their language uses sound for reference (see also Roberts, 1995 and Fulkerson & Haaf, 2003).

As discussed, correct interpretation of phonological variation in word learning appears to follow a more protracted developmental course than the learning of language-specific phonetic categories. The present study further investigates children's

interpretations of potentially relevant acoustic variability, focusing on interpretation of highly salient pitch contours. Pitch is a particularly interesting dimension of variation that English learners must interpret at appropriate levels of structure. In English, pitch varies systematically at the phrasal level (e.g., to mark yes/no questions, convey intonational meaning, and demarcate phrases), but it cannot contrast words. Since pitch is not contrastive in English, we might expect a particular word, like "good," to vary greatly in its pitch realization across tokens, because the pitch realization is not constrained by an underlying lexical tone. In English infant-directed speech, however, frequent words like "good" and "no" exhibit some consistency in their pitch patterns across tokens, probably because they tend to occur with particular pragmatic meanings and in stereotyped lexical contexts (Quam, Yuan, & Swingley, 2008). English-learning children must learn to interpret this pitch consistency at the phrasal level rather than the word level, even though it is potentially ambiguous between the two. Here, we address whether English-learning toddlers correctly avoid attribution of pitch regularities to the word level when learning a novel word.

2.1.1 The Curious Case of Pitch Variation

Pitch is relevant at the lexical level in some but not all languages. In tone languages, words with very different meanings can differ only in their tone. For example, in Thai, *khaa* means *a grass* when pronounced with a mid tone, *to kill* when pronounced with a low tone, and *leg* when rising (Gandour, 1978). All of the world's languages—tone and nontone alike—convey meaning through phrasal intonation (e.g., the English phrase "oh, great" can mean very different things depending on its intonation). What makes tone languages special is that they use pitch contrastively, to distinguish words.

There is mixed evidence about whether tone categories are clarified in infant-directed speech (IDS) or distorted by the exaggerated pitch patterns typical of IDS. Papousek and Hwang (1991) found that Mandarin speakers reduced or even neglected tone information in order to produce simple intonation contours to two-month-old infants. In contrast, Liu, Tsao, and Kuhl (2007) found that tones in Mandarin IDS to 10- to 12-month-olds were not distorted by the sweeping pitch patterns of IDS, and were in fact *exaggerated* in a manner comparable to the exaggeration of vowel categories found in IDS (Burnham, Kitamura, & Vollmer-Conna, 2002). This difference could arise because parents' speech needs to convey different information to children of different ages: intonational meaning to younger infants and tone and segmental information to older infants (Kitamura & Burnham, 2003; Stern, Spieker, Barnett, & MacKain, 1983.) Thai speakers appear to exaggerate intonation contours in speech to children from birth to 12 months, without causing much distortion of tones (Kitamura, Thanavishuth, Burnham, and Luksaneeyanawin, 2002). Even in speech to two-month-olds, however, Mandarin speakers appear to expand their pitch range and raise their pitch mean *less* than speakers of nontone languages, though they still produce the same intonational meanings (M. Papousek, H. Papousek, & Symmes, 1991; see also Kitamura et al., 2002).

Recent research has asked whether the acquisition of tone contrasts parallels that of consonant and vowel categories. The perceptual reorganization by which infants become worse at discriminating nonnative sound contrasts, but maintain good discrimination of native contrasts, occurs as early as six months for vowels: English-learning six-month-olds fail to discriminate some German vowel contrasts (Polka & Werker, 1994), and Spanish learners fail to discriminate the Catalan /□/-/e/ contrast by

eight months (Bosch & Sebastian-Galles, 2003). The reorganization is evident slightly later for consonants: while six-month-old English learners easily discriminate Hindi and Salish consonant contrasts, twelve-month-olds fail to do so (Werker & Tees, 1984).

Perceptual reorganization for tone seems to follow a similar trajectory; recent studies suggest that infants learning tone languages develop adult-like tone perception within the first year. Mattock and Burnham (2006) found that English learners failed to discriminate Thai tones by nine months, but Chinese learners—who were acquiring a tone language—did not undergo the same worsening of discrimination with age. Harrison (2000) tested English-learning and Yoruba-learning six- to eight-month-old infants' perception of Yoruba tones. The Yoruba-learning infants were more sensitive than the English learners to changes in fundamental frequency (f0), but only in the region surrounding a tone boundary (190 versus 210 Hz). This response aligned with that of adult native speakers of Yoruba, providing evidence that the infants were already responding in an adult-like way to the tone contrasts.

Adults' perception of tones also suggests that listeners are shaped by their native-language structure. Mandarin speakers perceive Mandarin tones quasi-categorically, apparently assimilating the tones to linguistic categories, while French speakers perceive them continuously (suggesting French speakers perceive the tones psychophysically versus linguistically; Halle, Chang, & Best, 2004). Finally, there is evidence that tones, like other speech sounds, form classifiable clusters. An unsupervised learning algorithm can learn the four tone categories of Mandarin from pitch movement in syllables extracted from fluent speech (Gauthier, Shi, & Xu, 2007).

Evidence from children's productions suggests that the reliability of the realization of tones affects their age of acquisition. Hua and Dodd (2000) found early acquisition of tones in Putonghua (Modern Standard Chinese, a variety of Mandarin). For children between the ages of eighteen months and 4.5 years, tone errors were rare relative to consonant and vowel errors. The distribution of production errors across the age groups suggested that Putonghua-learning children acquire tones first, then vowels and syllable-final consonants, then syllable-initial consonants. In another language, Sesotho, words' surface forms often diverge from their underlying tones because of pervasive tone sandhi. Demuth (1995) found a slower, more item-specific acquisition of tone in Sesotho than had been found for lexical tone languages. This suggests that the reliability of the mapping between underlying tone and surface form has a large impact on the speed of acquisition of a tone (see also Ota, 2003).

Beyond acquisition of tones, we can ask how perception and interpretation of pitch cues to other levels of structure develop. In English, pitch demarcates phrase boundaries (Gussenhoven, 2004), marks yes/no questions (with a terminal rise), and cues lexical stress, e.g., helping distinguish the noun *PERmit* from the verb *perMIT* (Fry, 1958; for reviews, see Ladd, 2008, and Gussenhoven, 2004, Chapter 2). Because of contrastive stress pairs like these, there is a sense in which pitch can help contrast words in English. But in these cases other correlated cues to stress, including vowel quality, vowel duration, and amplitude, contribute strongly to the contrast. Cutler and Clifton (1984) found that adults were slower to identify words when the acoustic cues to stress were naturally produced to stress the wrong syllable. This mispronunciation effect occurred even when the unstressed vowel was unreduced—meaning the vowel-quality

cue was essentially neutralized—but the effect was greater when the unstressed vowel was reduced. It is not yet known whether listeners can exploit an *isolated* pitch cue to stress in word recognition.

Pitch also conveys highly complex intonational meanings in adult-directed speech, through particular, stereotyped contours. The ToBI transcription system (Pierrehumbert, 1980; Beckman, Hirschberg, & Shattuck-Hufnagel) was developed to characterize different intonation contours in English as a series of High and Low tones, and has led to the identification of certain, fairly reliably realized intonational meanings. For example, the 'fall-rise' or 'rise-fall-rise' pattern conveys uncertainty or incredulity in some sentential contexts (Ward & Hirschberg, 1985; Hirschberg & Ward, 1992), while the 'continuation rise' contour can convey that the speaker is about to continue talking (Bolinger, 1989).

For very young infants who have not begun learning words, the meaning of caregivers' speech is carried entirely by prosodic characteristics, particularly intonation. The distinctive pitch characteristics of infant-directed speech (IDS) complement the infant's developing auditory system; the higher f0 mean and wider f0 range make the speech more interesting and easier for the infant to tune in to (Fernald, 1992). Infants prefer listening to IDS over adult-directed speech (ADS; Fernald, 1985), a preference driven primarily by IDS's pitch characteristics (Fernald & Kuhl, 1987; Katz, Cohn, & Moore, 1996). Some pragmatic functions of speech are expressed more clearly in IDS than in ADS; listeners are more successful at identifying the pragmatic functions of content-filtered IDS utterances than comparable ADS utterances (Fernald, 1989). Considering the clarity of intonational meaning in IDS, it is not surprising that infants can

categorize utterances from different emotional classes before they know many words (Moore, Spence, & Katz, 1997).

Despite the relevance of pitch at nonlexical levels of structure, and the clear importance of pitch in parental communication to infants, the English-learning child must learn to disregard intonational pitch as a lexically contrastive feature when establishing new lexical entries and in recognizing words. A recent study by Singh, White, and Morgan (2008) provides some evidence for development in infants' categorization of word forms varying in pitch. Singh et al. familiarized infants to words in isolation and tested their recognition of those words in sentences, using a procedure that evaluates infants' preference for familiarized versus novel materials. When the pitch realization matched between familiarization and test, both 7.5-month-olds and 9-month-olds preferred to listen to the sentences containing familiarized words. When the familiarized words were realized with different pitch, however, only the 9-month-olds preferred to listen to the familiarized words, suggesting that the younger infants failed to recognize them. In the second half of the first year, therefore, infants appear to become better able to recognize words despite changes in pitch. Still, this leaves open the phonological status of linguistic pitch in two ways. First, the pitch manipulation tested by Singh et al. (2008) involved an absolute change in the words' pitch levels, produced by raising or lowering all pitch samples by six semitones. Nine-month-old infants might still be thrown off by changes in intonation *contour* (e.g., Trehub & Hannon, 2006). Second, developmental changes in infants' matching of different realizations of a word form may bear more on a general property of infant memory (e.g., a decrease over development in the number of

perfectly matching features required for a new stimulus to be matched to a prior one) than on children's *interpretation* of how speech conveys meaning.

The distinction between interpretation and simple acoustic matching is also an issue for studies showing similar improvement in children's ability to recognize words despite changes in talker's voice or affect. At 10.5 months—but not at 7.5 months—infants successfully generalize familiarized words from male to female voices (Houston & Jusczyk, 2000),[3] or when the affect changes (from happy to neutral or vice-versa) across familiarization and test (Singh, Morgan, & White, 2004; see also Houston & Jusczyk, 2003). Studies of how infants match different tokens of a word form are informative about foundational mental capacities that underlie language acquisition, but they do not necessarily indicate how phonetic variation is interpreted referentially. Even adults are better at recognizing a word when it is spoken by the original voice (Palmieri, Goldinger, & Pisoni, 1993; Goldinger, 1996). Rather than tuning out irrelevant information completely, we apparently become more adept, over development, at focusing on essential properties of words, like phonemes and stress patterns. One way to view this process follows Jusczyk (1993) in proposing that exposure to the native language leads the system to weight relevant features more heavily and irrelevant features less heavily.

Learning how pitch is used in English requires separating pitch from the lexical level and learning intonational categories cued by pitch. Young children's speech does contain a range of intonational contours that often sound familiar enough to be

[3] Male versus female voices differ more in their fundamental frequency than two female or two male speakers (Houston & Jusczyk, 2000).

interpreted referentially by adults, but few studies have shown that young children analytically separate the intonational characteristics of words from their segmental characteristics (see Vihman, 1996, for a review). Galligan (1987) reports a case study of two children, who amid their second year each used single words with more than one intonational contour in ways that could be interpreted as being appropriate for the communicative context. This sort of evidence suggests that children attempt to interpret and produce sentence intonation, and may succeed in separating the pitch properties of an utterance from the utterance's lexical context. However, it does not necessarily follow that children command a linguistic system that rules out pitch contours as relevant for distinguishing words. Establishing this stronger claim requires an empirical test, like the current one, in which the child's experience with a word provides evidence for a (grammar-inconsistent) interpretation in which the word has intrinsic pitch, and the child must attribute that consistent pitch pattern to the intonational level rather than the lexical level. The apparent difficulty of this correct attribution depends upon whether one assumes that toddlers interpret speech in a holistic fashion, encoding words as a mass of relatively unanalyzed sensory properties, or in an analytical fashion, potentially attributing various phonetic properties of a word token to separate linguistic levels of interpretation.

The issue of interpreting pitch at the appropriate levels of structure has hardly been addressed in the developmental speech perception literature, in which discussion of holistic or analytic representations has focused on segmental phonology (consonants and vowels) rather than intonation. In that context, the analytic viewpoint holds that young children's lexical representations can be described using the conventional inventory of

consonants and vowels (e.g., Swingley, 2003), whereas the holistic viewpoint argues either that children's knowledge of the sounds of words is less clearly specified (many features are missing) or that children's lexical representations are not made up of a sequence of categories at all (e.g., Metsala & Walley, 1998; Storkel, 2002; for discussions, see Swingley, 2007; Vihman & Croft, 2007; Werker & Curtin, 2005).

More generally, the notion that children interpret speech analytically is at variance with simple exemplar models in which the lexicon provides the sole level of organization relevant to word recognition (see Goldinger, 1998, for what he describes as an "extreme" model of this sort, and for discussion of more richly structured alternatives). If the recognition of words depends entirely on the overall phonetic or acoustic match between the current token and the mass of previously experienced tokens of that word, prior experience with a particular word's realizations should trump phonological generalizations derived from analysis of the other known words of the language. Listeners do retain voice- or otherwise token-specific information about experienced words (e.g., Goldinger, 1996; Nygaard & Pisoni, 1998), which rules out models in which formal linguistic content *alone* guides behavior. But the existence of such effects does not imply that phonological analysis is unnecessary (Pierrehumbert, 2006). Studies supporting exemplar models rarely calibrate the effects of nonphonological information, like talker's-voice characteristics, against a phonological baseline. In Experiment 1, we test the hypothesis that adults will weigh much more heavily those phonetic changes that are *relevant* for distinguishing words in English, than changes that, though perceptually salient, are not lexically contrastive. If we find that adults are sensitive to changes in pitch contour, this will support the holistic, or exemplar, perspective; we will then be in a

position to assess the relative importance of lexically relevant and irrelevant phonetic variation within that perspective. If we find that adults show large effects of lexically relevant changes, but not changes in pitch contour, this will support analytic views of speech interpretation—or exemplar views in which the phonetic dimension of pitch is weighted extremely weakly.

2.1.2 Overview of the Two Experiments

We taught both adults and 2.5-year-olds a new word, always pronounced with a consistent, salient pitch contour, and then tested their interpretations of a nonphonemic change in the word's pitch contour versus a phonemic change in the word's vowel. We first tested adults, in Experiment 1, in order to establish the mature interpretation of these changes. In Experiment 2, we tested 2.5-year-olds in the same task. We selected an age at which children should treat the vowel change as relevant, since we wanted to compare interpretations of the pitch-contour change to this phonological baseline. Seventeen- to twenty-month-olds sometimes struggle to differentiate similar-sounding words in teaching contexts (Swingley & Aslin, 2007), so we wanted to ensure that processing constraints (e.g., failure to remember which version of the word was taught and which was the change) would not prevent children from interpreting a mispronunciation as a new word. Our selection of 2.5-year-olds for Experiment 2 was also motivated by evidence of developmental change in children's interpretation of pitch cues to emotion, over the ages of 3 to 4 years (Quam, Swingley, & Park, 2009). This suggests an especially protracted learning course for interpretation of pitch structure in English.

2.2 Experiment 1

Three questions led us to test adults as well as 2.5-year-olds. First, although adult native-English speakers are naturally expected to have acquired the phonology of English, in which pitch contours cannot be interpreted lexically, adults might still recognize words best when the test instances are most similar to the training instances. This result would be consistent with the episodic-lexicon model and with evidence that adults retain subsegmental and indexical information about words (e.g., Goldinger, 1996; Nygaard & Pisoni, 1998). A comparison between adults' and children's sensitivity to changes in pitch contour could also shed light on whether children's interpretation of nonphonemic dimensions becomes adult-like through the fine-tuning of attention weights to different acoustic dimensions (Jusczyk, 1993). Second, despite their knowledge of native phonology, adults could choose to interpret the highly salient pitch change as relevant, treating a word with altered, "mispronounced" pitch as a worse version of the newly learned word than the word with the original pitch contour. Third, adults could interpret the vowel change either as an entirely new word, referring to a different object, or as a mispronunciation of the taught word. We were interested in whether adults, who have reached the endpoint of phonological development, would be uniform in their responses, or whether we would still see individual variation in interpretations of the two changes.

2.2.1 Method

2.2.1.1 Participants

Twenty-four adults, nine male, and all native speakers of English were included in the analysis. (One of these participants was also a native speaker of Spanish; his

responses were typical.) All participants but one were undergraduates (the exception was a postdoctoral researcher), assumed to be between 17 and 23 years old. Ten more participated but were excluded: six for experimenter error / equipment failure, two for failure to follow instructions to fixate the pictures, and two for their language backgrounds (one was a nonnative speaker of English, the other was a native bilingual of English and Chinese).

2.2.1.2 Apparatus and Procedure

We used a language-guided looking procedure to investigate how adults would interpret a phonological (vowel-quality) versus nonphonological (pitch-contour) change in a newly learned word. Since adults participated in essentially the same experiment as the toddlers in Experiment 2, the stimuli were designed for children. To make this experience less odd, adult participants were told before the study that they would be helping to calibrate an experiment designed for two-year-olds.

Participants sat in front of a large display screen, on which they viewed pictures. Concealed speakers played recorded sentences that referred to the pictures, and a hidden video camera in the center of the display captured participants' eye movements, which were later coded by hand.

The experiment lasted twenty minutes and consisted of four phases (see **Figure 1** for the experimental design). The first two phases, the *animation* and *ostensive-labeling* phases, taught participants a novel word. In the animation phase, adults watched a five-minute, narrated, animated video in which a monkey presented his two toys to several potential playmates. One toy was labeled ten times as the "deebo" (IPA: [diboʊ]) in sentences like, "This is my deebo. Would you like to play with it?" The word was

pronounced with a highly consistent, distinctive intonation contour commonly found in speech to infants: either a rise-fall or a low fall (see **Figure 2** for spectrograms and pitch tracks of the two pitch contours). The other toy was present and talked about equally often, but never labeled, in sentences like, "This is my other toy. Would you like to play with it?" In the ostensive-labeling phase, each toy appeared independently on the screen. The *deebo* was labeled four times in each of three trials, for a total of twelve repetitions, in sentences like: "This is a deebo. Deebo. Look at the deebo. The deebo." The other toy was talked about, but not labeled, in sentences like: "Look at this toy. Isn't it pretty? Would you like to play with it?"

The third phase, the *test*, contained 18 critical trials. In these trials, the two toys appeared side by side. In eight *trained-pronunciation* trials, participants were asked to locate the "deebo," in sentences like, "Where's the deebo? Can you find it?" In the other ten trials, adults heard a word that differed from the taught pronunciation in one of two ways. In five *vowel-change* trials, participants heard "dahbo" (IPA: [dɑboʊ]) with the original pitch contour; in five *pitch-change* trials, they heard "deebo" with a different pitch contour. Half the participants were originally taught the word *deebo* with a rise-fall contour (which changed to the low fall on pitch-change trials), and the other half were taught *deebo* with a low fall contour (which changed to the rise-fall on pitch-change trials). In addition to these 18 critical trials, 69 familiar-word trials were interspersed throughout the ostensive-labeling and test phases. These familiar-word trials presented two familiar objects and asked adults to orient to one of them, in sentences like, "Look at the shoe. That's pretty." Target words in the familiar-word trials were produced with natural intonation and no segmental mispronunciations. These familiar-word trials, along

with 20 short, attention-getting animations, were intended to distract adults from the purpose of the experiment and also to prevent boredom and sleepiness.

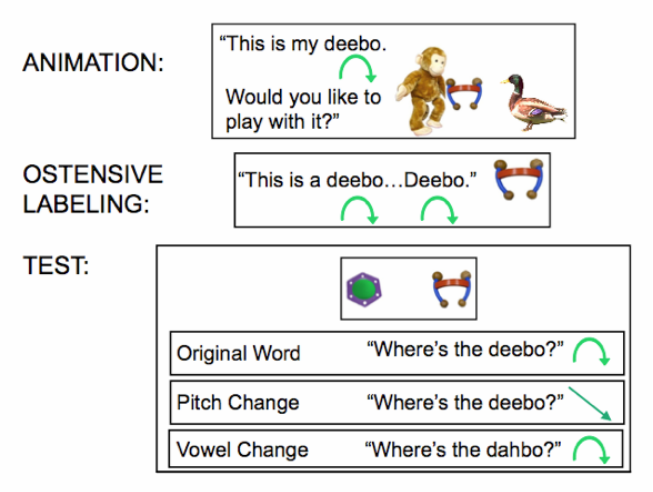


Figure 1: Experimental design

In the *animation* phase, participants heard the word "deebo" spoken with the same intonation contour (the rise-fall is used in this example) ten times in a story. Next, in the *ostensive-labeling* phase, the *deebo* was labeled directly twelve times. In *test* trials, the *deebo* and distracter objects were presented side-by-side. Adults heard eight trained-pronunciation (original-word) trials and five trials of each change type. Children heard the original word in eight trials and *either* the pitch change or the vowel change in the other eight trials. Finally, participants were asked to point to and name the objects (not pictured).

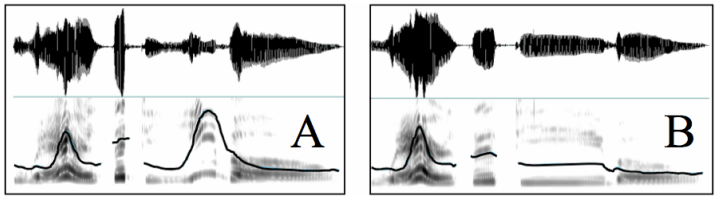


Figure 2: Intonation contours

Waveform, spectrogram, and pitch contour for the rise-fall contour (**A**) and low fall contour (**B**), in the sentence "Where's the deebo?"

Finally, in the *pointing-and-naming* phase, adults were asked to point to and name the objects. In *pointing* trials, both novel objects appeared on the screen and participants were asked to "Point to the [deebo]." The word was pronounced with the trained pronunciation and each of the changed pronunciations from the test phase, for a total of three trials. In *naming* trials, each object appeared separately on the screen next to a

picture of the Sesame Street character Elmo, and participants heard, “Elmo doesn’t know what that is. Tell Elmo what that is!”

After the experiment, adults filled out a questionnaire. The questions evaluated whether each participant had correctly learned the word-object pairing for *deebo*; whether she had noticed the pitch and vowel changes; and whether she had interpreted the word “dahbo” as a label for the distracter object, or merely as a mispronunciation of “deebo.”

2.2.1.3 Auditory Stimuli

A native English speaker (the first author) recorded auditory stimuli in clear child-directed speech, with exaggerated, infant-directed prosody and at a normal speaking volume. The animation sentences were embedded in a narration, similar to a storybook (e.g., “This is my deebo. Would you like to play with it?”). They accompanied an animated movie, meant to familiarize participants with the pairing of the word “deebo” and the object. “Deebo” was always spoken with a consistent intonation pattern: either a rising then falling contour (referred to as rise-fall) or a level, medium pitch followed by falling pitch (referred to as low fall; see **Figure 2** for spectrograms and pitch tracks of the two pitch contours). We chose pitch contours that could be interpreted either as lexical pitch or as phrasal intonation, because we wanted to avoid pushing participants into one interpretation or the other.

Ostensive-labeling sentences directly labeled the *deebo* object (e.g., “This is a deebo. That’s right. Look at the deebo. The deebo.”). In the animation and ostensive-labeling sentences, the word “deebo” was always spoken with the same intonation contour, though tokens were allowed to vary somewhat in length, absolute pitch, and amplitude (this variation helped them sound natural in context). In test sentences,

participants were asked either “Where’s the [deebo/dahbo]?” or “Which one is the [deebo/dahbo]?” The duration, pitch contour, and amplitude of test words were controlled carefully. Pointing sentences were comparable to test sentences, but asked participants to “Point to the [deebo/dahbo].” In all sentences, the word “deebo” (or “dahbo”) occurred at the end of the sentence, where the pitch contours and duration sounded most natural. Sentences were always naturally produced, but in some cases the length or amplitude of the word was modified slightly using Praat sound-editing software (Boersma & Weenink, 2008). See **Appendix 1** for duration, maximum pitch, and mean pitch of each word token.

2.2.1.4 Visual Stimuli

Visual stimuli were displayed on a rectangular plasma video screen measuring 37 by 21 inches. In the animation phase, these stimuli consisted of photographs of objects, moving around in front of a painted scene of a grassy hill. A plush toy monkey moved around the scene, manipulating two novel toys and playing with other animals. Visual stimuli in the ostensive-labeling and test phases consisted primarily of photographs of objects on gray backgrounds. In ostensive-labeling trials, each novel toy from the animation appeared on the screen alone, while in test trials, the two toys were displayed side by side. At the beginning of each ostensive-labeling and test trial, the deebo and/or distracter objects hopped or twisted on the screen (this was intended to get children’s attention in Experiment 2), after which they remained still. All photos were edited to balance their salience by roughly equating brightness and size. The two novel toys were a purple-and-green plastic disk (subsequently referred to as the *purple disk*) and a red-and-blue knobby wooden object (subsequently referred to as the *red knobs*; see **Figure 3**).

The particular object that was labeled the "deebo" varied across participants, and was crossed with which pitch pattern they heard during the teaching.

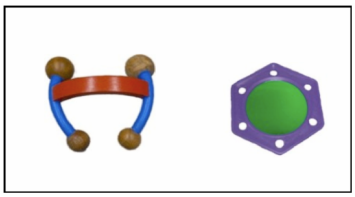


Figure 3: The two objects used in teaching and testing
On the left is the *red-knobs* object, and on the right is the *purple-disk* object. For each participant, one of these objects was labeled the "deebo" and the other was present equally often but never labeled.

2.2.1.5 Coding

After testing, trained coders, blind to target side, coded the direction and timing (beginning and end) of every eye movement a participant initiated during each trial. Eye movements were coded frame-by-frame with 33-millisecond resolution using the *SuperCoder* software (Hollich, 2005). Alignment of the timing of eye-movement events with auditory and visual stimulus events was ensured using a custom hardware unit that placed visible signals into the recorded video stream of the participant's face.

For each participant in each trial, we calculated the proportion of the time he or she fixated the *deebo* object (the amount of time spent looking at the *deebo* divided by the total time looking at either picture). We calculated this *deebo* fixation proportion over a specified time window after the onset of the target word: 200 to 2000 ms post–noun-onset. This time window is similar to the window commonly used with young children, 367 to 2000 ms (see **Experiment 2** for an explanation for the time window used with children), but begins earlier because adults are known to respond more quickly than toddlers in this procedure (e.g., Swingley, 2009).

2.2.2 Results and Discussion

Adults provided four types of responses: looking times to each picture, elicited pointing and naming of the pictures, and questionnaire responses. Looking times provide a gradient measure of interpretation of the auditory stimulus, while pointing and naming force participants to make a discrete and conscious choice. Naming responses also allow us to probe for encoding of pitch and segmental information. Finally, questionnaire responses allow us to determine participants' final interpretation of the stimuli.

The pronunciation of test words (trained pronunciation, pitch change, or vowel change) exerted a significant effect on adults' fixation of the *deebo* in an analysis of variance ($F(2,69) = 77.16$, $p < .001$). There were no main effects, or interactions with trial type, of which object was the *deebo*, which pitch contour was taught, or which type of change was presented first in the test phase. Planned comparisons thus further investigated only the effect of condition (pronunciation of the word) on *deebo* fixation.

When they heard the trained pronunciation of the word, adults fixated the *deebo* object significantly above chance, or 50% (mean, 91.8%; paired t(23) = 31.38; *p*(all tests 2-tailed) < .001). Participants also fixated the *deebo* above chance in response to the pitch change (mean, 89.3%; paired t(23) = 17.89; $p < .001$), and their accuracy did not differ significantly from their accuracy in response to the trained pronunciation.

In response to the vowel change, participants actually fixated the *deebo* below chance (this difference approached significance; mean, 39.7%; paired t(23) = -1.98; $p = 0.06$), and significantly less than in trained-pronunciation trials (paired t(23) = 10.13; $p < .001$) or pitch-change trials (paired t(23) = 9.29; $p < .001$). Every participant fixated the *deebo* less in vowel-change trials than in trained-pronunciation trials (see **Figure 4**).

Eighteen of the 24 participants (75%) fixated the *deebo* less than 50% of the time in response to the vowel change, suggesting they used a mutual-exclusivity strategy (Markman & Wachtel, 1988), interpreting "dahbo" as a label for the distracter object. In pitch-change trials, by contrast, no participants fixated the *deebo* less than 50% of the time, and exactly half of the participants (12/24) fixated the *deebo* less in pitch-change trials than in trained-pronunciation trials.

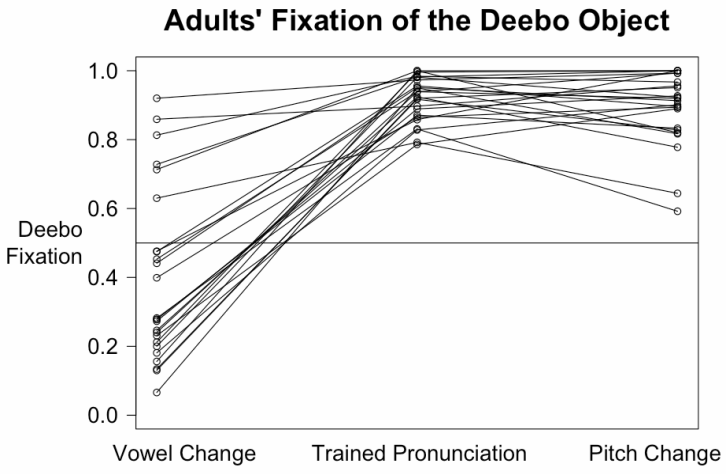


Figure 4: Adults' fixation of the *deebo* object in each trial type
The horizontal line indicates chance fixation, or 50%. Adults' fixation of the *deebo* object showed a large effect of the vowel change. All 24 participants fixated the *deebo* less in vowel-change trials than in trained-pronunciation trials, and 75% fixated the *deebo* less than 50% of the time in vowel-change trials. In contrast, adults showed no effect of the pitch change; only half of participants fixated the *deebo* less in pitch-change trials than in trained-pronunciation trials, and no participants fixated the *deebo* less than 50% of the time in pitch-change trials.

Adults' pointing, naming, and questionnaire responses provide additional insight into their interpretations of the pitch and vowel changes. **Tables 1** and **2** display adults' pointing and naming responses, respectively. When asked to "point to the deebo," regardless of the pitch contour used, all 24 adults pointed to the *deebo*. In contrast, responses to "point to the dahbo" were more varied: 19/24 participants pointed to the distracter object (though four of those showed uncertainty, assessed informally, either

through their facial expression, their words, or rising intonation), while the other five participants pointed to the *deebo*. When asked to label the *deebo*, 22/24 participants said "deebo," while the other two did not name it. When asked to label the distracter object, 15/24 participants said "dahbo," (five of whom showed uncertainty), seven did not label it, one said "deebo," and one said "doba." (The latter two participants wrote on the questionnaire that they interpreted "dahbo" as a label for the distracter, but they incorrectly reproduced "dahbo" as "dubbo" and "doba," respectively, suggesting they were having trouble remembering or reproducing the /a/ vowel.) The pitch characteristics of adults' labeling responses did not reflect the pitch contour used in teaching; analyses of variance predicting the f0 maximum and f0 mean of labeling responses from the interaction of taught pitch (rise-fall or low fall) and which object participants were labeling (*deebo* or distracter) showed no significant effects.

Table 1: Adults' pointing responses

Points to the *deebo* / number of adults pointing (percentage pointing to *deebo*), for each condition.

Trained pronunciation		Pitch change		Vowel change	
24 / 24	(100%)	24 / 24	(100%)	5 / 24	(21%)

Table 2: Adults' naming responses

Number of responses / number of adults (percentage giving the particular response), for each object.

	Viewing *deebo*		Viewing distracter	
Said "deebo"	22 / 24	(92%)	1 / 24	(4%)
Said "dahbo"	0 / 24	(0%)	15 / 24	(63%)
Did not name / Used different vowel	2 / 24	(8%)	8 / 24	(33%)

Though the acoustic measurements did not reveal differences between adults' productions depending on which pitch contour they were taught, it could be that human judges would be more sensitive to subtle differences not captured by the acoustic measurements we used. With this in mind, ten new adult judges were trained to identify rise-fall or low fall contours. They were first given training exemplars taken from the training and test phases of the original experiment, and then were tested on classification of twelve more exemplars. Only one adult made an error during this phase, on one of the twelve trials. The judges were then asked to categorize the experimental participants' productions as rise-fall or low fall contours. Adult productions were mixed in with child productions and presented in random order; classifications of the child productions are reported in the **Experiment 2 Results**. The judges' classifications of the adults' productions did not reflect the pitch contour participants were taught ($F(1,31) = .81$, $p = .38$), the object they were labeling ($F(1,31) = .09$, $p = .77$), or their interaction ($F(1,31) = 1.10$, $p = .30$) in an analysis of variance using the number of rise-fall classifications for each utterance (out of a possible ten) as the dependent variable. Judges assigned the "rise-fall" classification to participants' labels of the *deebo* object at similar rates regardless of which contour was taught (taught rise-fall, mean 4.56, SE 0.69; taught low fall, mean 4.55, SE 0.65). "Rise-fall" classification of participants' labels for the distracter object were also not significantly related to the taught contour (taught rise-fall, mean 3.57, SE 1.00; taught low fall, mean 5.00, SE 0.57). Participants' failure to imitate the taught pitch contour in their own productions suggests they did not consider the pitch pattern to be a relevant component of the word's sound.

In questionnaire responses, all 24 adults reported noticing the vowel change, and 17/24 reported having learned both "deebo" and "dahbo" as object labels. In contrast, only 12/24 participants reported noticing the pitch change. Eight of the twelve participants who did not report the change did remember it after prompting, either when the experimenter asked, "Did you notice any other changes in the word?" or when the experimenter reproduced the pitch contrast for them. No participants reported learning two words that contrasted in pitch.

Adults' responses across our measures of their learning were fairly consistent. Similar numbers of participants demonstrated learning of "dahbo" on each measure. Eighteen participants looked more at the distracter, and 19 pointed to the distracter, in response to "dahbo"; 15 labeled the distracter "dahbo"; and 17 reported learning the word "dahbo." Still, individual participants were not always wholly consistent. Thirteen participants showed all the behaviors consistent with learning the word "dahbo" (looking more to, pointing to, and labeling the distracter; and reporting having learned both words), and three participants showed *no* evidence of learning the word "dahbo." But eight participants exhibited some but not all behaviors associated with learning "dahbo," suggesting they did not commit to one single interpretation of the vowel change.

To summarize, adults universally showed no effect of the pitch change, fixating the *deebo* object equally in response to the trained pronunciation and the pitch change. They also universally showed sensitivity to the vowel change; all participants fixated the *deebo* less in response to the vowel change than in response to the trained pronunciation. Though we expected that adults might consistently interpret the large vowel change (from /i/ to /a/) as signaling a new word, we found instead that adults were fairly variable in

their interpretations. This was true both across participants and, sometimes, within individuals. All participants *noticed* the vowel change, as evidenced both by their questionnaire responses and their decreased fixation of the *deebo* in response to the vowel change. Detection and interpretation, however, are distinct.

2.3 Experiment 2

We next tested 2.5-year-olds in the same experiment, asking whether their interpretations of the pitch and vowel changes would be adult-like, reflecting their native phonology, or not yet fully developed. Children's responses could differ from the adult standard in two ways: children could treat the pitch change as lexically relevant, or they could fail to show sensitivity to the segmental change. Sensitivity to the pitch change would be consistent with evidence that young children are more open-minded than older listeners in interpreting new words (e.g., Namy, 2001; Namy & Waxman, 1998; and Woodward & Hoyne, 1999), and with evidence of a protracted developmental course for correct interpretation of pitch at other levels (e.g., pitch cues to emotions; Quam, Swingley, & Park, 2009). Lack of sensitivity to the vowel change is less likely, since 30-month-olds should be more sensitive to segmental changes than the younger children tested in previous experiments (e.g., 14-month-olds in Stager & Werker, 1997; 1.5-year-olds in Swingley & Aslin, 2007 and White & Morgan, 2008). We chose 30-month-olds for this reason, since we wanted the phonologically relevant change in the vowel to serve as a baseline for comparison with interpretations of the pitch-contour change. Still, children appear to be less sensitive to vowel changes than to consonant changes (Nazzi, 2005, testing 20-month-olds), and the pitch consistency in our teaching phase could also

dampen children's sensitivity to the vowel change, given that increased variability in talker's voice (Rost & McMurray, 2009) and in affect (Singh, 2008) improve children's sensitivity to subtle contrasts.

2.3.1 Method

The design, apparatus, and stimuli were comparable to Experiment 1. Children saw the same *animation* and *ostensive-labeling* phases as in Experiment 1. The other two phases differed slightly from the adult version. The *test* phase had three important modifications. First, because of children's more limited attention spans, each child heard *either* the vowel or the pitch change in the test trials, not both. The experiment contained eight *trained-pronunciation* trials, either eight *pitch-change* trials or eight *vowel-change* trials, and only ten *familiar-word* trials (instead of the 69 included in the adult experiment), so that the experiment was less than 10 minutes long. Finally, children also participated in the *pointing-and-naming* phase, but heard only two pointing trials, corresponding to the trained pronunciation and the pronunciation change the child heard in test. As in Experiment 1, there were two naming trials, one for each toy. In each pointing or naming trial, if the child did not point or speak, the trial was replayed and the parent and experimenter encouraged the child to respond without biasing her response. Parents kept their eyes closed in both the test and pointing-and-naming phases to avoid biasing the child's responses. Within a week of the test date, parents completed the MacArthur Communicative Development Inventory of Words and Sentences (Fenson et al., 1994), which measured their child's productive vocabulary.

2.3.1.1 Participants

Forty-eight children between the ages of 29 months, 3 days and 32 months, 8 days were included in the analysis. All participants were learning English as their dominant language and hearing it at least 2/3 of the time, as reported by their caregivers. Twenty-four children, 13 male, were included in the *vowel-change* condition (mean age 30 months, 19 days, *SD* = 24 days; mean productive vocabulary 512 words, *SD* = 154 words); and 24 children, 13 male, were included in the *pitch-change* condition (mean age 30 months, 17 days, *SD* = 30 days; mean productive vocabulary 468 words, *SD* = 181 words).

Fifteen more children participated but were excluded (four from the pitch condition, eleven from the vowel condition) for having fewer than six usable trials (including the point trial) in any of the trial types (familiar-word, trained-pronunciation, or changed-pronunciation trials). Trials were only included as usable if the child fixated the pictures for at least 10 frames during the analysis window, out of a possible 50.

2.3.2 Results and Discussion

We calculated children's fixation of the *deebo* over a specified time window after the onset of the target word (beginning slightly later than the window used with adults): 367 to 2000 ms after noun onset. Before 367 ms, children are unlikely to be responding to the target word (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998; Swingley & Aslin, 2000). After 2000 ms, they are likely to have completed their response and moved their attention elsewhere.

Before asking whether children responded to changes in the word's pronunciation, we had to determine whether they learned the word at all, by comparing

children's *deebo* fixation to chance fixation, or 50%. In trials where "deebo" was spoken with the trained pronunciation, both groups' *deebo* fixation was significantly above chance (vowel-change group: mean, 67.4%; paired t(23) = 5.73; *p*(all t-tests 2-tailed) < .001; pitch-change group: mean, 66.3%; paired t(23) = 5.60; $p < .001$).

Next, we considered whether either the pitch change or the vowel change significantly affected children's fixation of the *deebo* object. **Figure 5** displays *deebo*-fixation proportions for each group in trained-pronunciation and change trials. Trial type (trained versus changed pronunciation) interacted significantly with condition (pitch versus vowel) in an analysis of variance ($F(1,92) = 11.57$, $p < .001$). The vowel change caused a significant decrease in *deebo* fixation compared with responses to the trained pronunciation (mean decrease, 15.0%; paired t(23) = -3.50; $p < .005$), exhibited by 20/24 participants (binomial $p < .001$). Additionally, 11/24 participants actually fixated the *deebo* less than 50% of the time in response to the vowel change (compared with only 2/24 children who did so in response to the pitch change), suggesting they may have used a mutual exclusivity strategy to map the word "dahbo" onto the distracter object (Markman and Wachtel, 1988). Overall, looking to the *deebo* in response to the vowel change did not differ from chance (mean, 52.8%; paired t(23) = 0.77; $p = 0.45$).

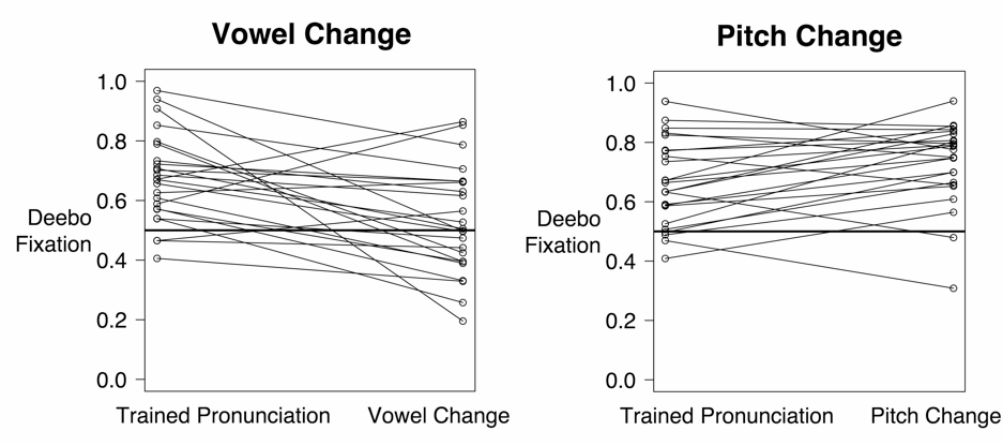


Figure 5: Thirty-month-old children's fixation of the *deebo* object in each trial type.
Left: Each vowel-change participant's fixation of the target object (the *deebo*) in response to the trained pronunciation and the vowel change; the horizontal line indicates chance fixation, or 50%. The vowel change caused a significant decrease in *deebo* fixation (15% on average) compared with responses to the trained pronunciation. *Right:* Each pitch-change participant's *deebo* fixation in response to the trained pronunciation and the pitch change. The pitch change actually caused a significant, though smaller, *increase* in target fixation (6.6% on average), perhaps because its novelty increased children's attentiveness.

Children exhibited a much different response to the pitch-contour change. Instead of a decrease in *deebo* fixation, we found a small *increase* compared with responses to the trained pronunciation (mean increase, 6.6%; paired $t(23) = 2.40$; $p < .05$). This effect of the pitch change was less than half the size of the vowel-change effect, and less consistent: 16/24 participants fixated the *deebo* more in response to the changed pitch (binomial $p > .05$). Still, this effect was unexpected. We speculate that the pitch change, after a long familiarization with one consistent pitch contour, may have made children more attentive and thus more successful at orienting to the target. Overall, looking to the *deebo* in response to the pitch change was significantly above chance (mean, 72.9%; paired $t(23) = 8.16$; $p < .001$).

When comparing children's fixation of the *deebo* object to chance, we face the risk that children might be biased to look at one picture or the other, making 50% an

inadequate baseline. To alleviate this concern, we conducted analogous tests in which we subtracted *deebo* fixation before the word's onset from *deebo* fixation in the 367-2000 ms window. We then compared this difference score to chance, or 0%. These tests yielded the same pattern of significance as the tests reported above. Children increased their *deebo* looking upon hearing the *trained pronunciation* in both the pitch-change group (mean increase, 13.1%; paired $t(23) = 3.32$; $p < .005$) and the vowel-change group (mean increase, 14.8%; paired $t(23) = 3.39$; $p < .005$). Children also increased their *deebo* looking in response to the pitch-change pronunciation (mean increase, 20.7%; paired $t(23) = 5.70$; $p < .001$). In response to the vowel-change pronunciation, in contrast, children's increase in *deebo* fixation did not differ from chance (mean increase, 2.2%; paired $t(23) = 0.73$; $p = 0.47$).

Next, we asked whether participants' age would affect their sensitivity to either change in pronunciation. We computed an analysis of covariance (ANCOVA) using each child's difference in *deebo* fixation between familiar and changed pronunciations as the dependent variable, and condition (pitch-contour change or vowel change), age in days, and their interaction as predictors. The effect of condition was significant ($t(44) = 2.11$, $p < .05$), as was the interaction of age and condition ($t(44) = 2.20$, $p < .05$). The effect of the interaction term arose because sensitivity to the vowel change was positively correlated with age ($r = 0.57$, $p < .005$), but there was essentially no correlation between age and children's sensitivity to the pitch change ($r = -0.17$, ns). Prior studies testing children's sensitivity to changes in the pronunciations of familiar words have, in most cases, failed to find a relationship between children's age and the magnitude of the effects of pronunciation changes (e.g., Swingley & Aslin, 2000; Bailey & Plunkett,

2002). In a shorter-term word-learning situation like the current one, however, older children may be better able to encode the vowel information than younger children, or may be more likely to consider the vowel change relevant to identification of the referent. An analogous ANCOVA testing effects of productive vocabulary size, rather than age, yielded no significant effects or interactions involving vocabulary size. However, a ceiling effect may have reduced the predictive power of the Communicative Development Inventory (the vocabulary checklist); over half of children (25/48) were reported to produce more than 80% of the words on the form. In analyses of variance, neither gender, the pitch contour used in teaching (rise-fall or low fall), nor the object used as the *deebo* (*red knobs* or *purple disk*) interacted with the effects of either mispronunciation.

Children's pointing and naming responses provided a useful supplement to the eyegaze data. Eyegaze, while a sensitive measure of word recognition, does not necessarily reliably index children's conscious interpretation of the utterance. For example, reduced looking to the *deebo* object upon hearing the vowel change could mean only that the changed pronunciation was not prototypical (and thus an inferior cue to the target), thus delaying or interfering with recognition. Pointing and naming responses involve discrete choices, and measure children's ultimate interpretation of the spoken words. Here, we found that children's pointing and naming responses were consistent with the results of the eye-movement analyses. **Table 3** shows pointing responses for children in each condition in response to the trained pronunciation and the changed pronunciation. Only pointing responses for children who responded in both trials are included (vowel change, n = 11; pitch change, n = 12). Children in the vowel-change

condition pointed much more often to the *deebo* (as opposed to the distracter object) when they heard "deebo" than when they heard "dahbo." Children in the pitch-change condition, by contrast, pointed more to the *deebo* than to the distracter in both trained-pronunciation trials and pitch-change trials. Pitch-change children pointed significantly more to the *deebo* object than would be expected by chance in response to the pitch change (binomial $p < .05$), and showed a trend in the same direction in response to the trained pronunciation (binomial $p = .146$). Vowel-change children pointed to the *deebo* above chance in response to the trained pronunciation (binomial $p < .001$), but not in response to the vowel change (binomial $p = 1$).

Table 3: Children's pointing responses

Points to target / number of children pointing (percentage of points to target), for each combination of condition and trial type.

	Trained pronunciation		Changed pronunciation	
Pitch-change children	9 / 12	(75%)	10 / 12	(83%)
Vowel-change children	11 / 11	(100%)	6 / 11	(55%)

Children's pointing responses to the trained pronunciation and the change could take four forms: pointing to the target for both pronunciations (abbreviated TT), pointing to the distracter for both (DD), pointing to the target for the trained pronunciation and to the distracter for the change (TD), or vice versa (DT). Children's distribution over these categories varied with mispronunciation type (X^2 (3, n = 23) = 8.57, $p < .05$), reflecting the fact that the vowel change caused children to point more to the distracter (TT = 6, **TD = 5**, DT = 0, DD = 0), while the pitch change did not (TT = 9, **TD = 0**, DT = 1, DD = 2). The pointing results indicate that children in the pitch-change condition considered both

pronunciations good matches to the *deebo* object, while for children in the vowel-change condition, "dahbo" was a worse match.

In naming trials, children were asked by a recorded voice to label both the *deebo* and distracter objects. We do not have responses from many children, either because they refused to respond, they said something other than a label for the object (e.g., "Elmo"), or they did not participate in the trials. Children were not always able to correctly pronounce all the sounds of the word (e.g., they sometimes said "teenbo" or "deedo" instead of "deebo"), so we scored productions for whether the first syllable contained the /i/ vowel (as in "deebo") or the /a/ vowel (as in "dahbo"). **Table 4** displays children's use of these vowels in their labeling of the objects. All children who produced either vowel are included, whether or not they responded in both naming trials. When asked to label the *deebo* object, both groups produced more /i/ vowels (vowel-change group: 15; pitch-change group: 14) than /a/ vowels (vowel-change group: 0; pitch-change group: 1). Children were more reluctant to label the distracter object, but the data we have are consistent with the looking and pointing responses: vowel-change participants labeled the distracter object with an /a/ vowel (5 responses) slightly more than with an /i/ vowel (1 response). In the pitch-change group, we expected children to have no name for the distracter object, and their responses are consistent with that: only two children produced /i/ vowels, and no children produced /a/ vowels. Like adults' productions, children's labeling of the *deebo* did not reflect the pitch contour they were taught; analyses of variance predicting f0 maximum and f0 mean, respectively, from taught pitch (rise-fall or low fall) showed no significant effects. (Since only seven children labeled the distracter

object, we did not include *object* as a predictor, instead excluding trials where the child was labeling the distracter object.)

Table 4: Children's naming responses

Responses with the /i/ vowel / responses with either vowel (percentage using /i/ vowel).

	Viewing *deebo* object		Viewing distracter object	
Pitch-change children	15 / 15	(100%)	2 / 2	(100%)
Vowel-change children	14 / 15	(93%)	1 / 6	(17%)

Recall from Experiment 1 that ten adult judges, trained to identify rise-fall and low fall contours in our stimulus materials, categorized participants' productions of our test words. Judges' classifications of children's productions as having rise-fall or low fall contours revealed no effect of taught pitch ($F(1,26) = .47$, $p = .50$) in an analysis of variance (again, there were too few instances of distracter-labeling to include *object* as a predictor). Judges assigned the "rise-fall" classification at similar rates for productions from children who were taught the rise-fall (and were labeling the *deebo* object; mean 6.33, SE 0.49); and those taught the low fall (mean 5.94, SE 0.37). Children's failure to imitate the taught pitch contour in their own productions suggests that they did not treat the pitch pattern as relevant for reproducing the word.

To summarize, our findings from the pointing and naming trials are consistent with our eye-movement result that children treated the vowel change—but not the pitch change—as relevant. Children pointed predominantly to the *deebo* when they heard both the trained pronunciation of the word and the pitch change, but pointed roughly equally to the *deebo* and the distracter object in response to the vowel change. In their naming of the objects, both groups of children used the /i/ vowel (as in "deebo") more often than the

/a/ vowel (as in “dahbo”) to label the *deebo* object. Children who had heard the word “dahbo” during the test trials were slightly more likely to use the /a/ vowel than the /i/ vowel to label the distracter object, while pitch-change children were not.

2.4 General Discussion

We addressed the development of interpretation of nonphonemic, but consistently realized, dimensions of the sounds of words by teaching 2.5-year-olds and adults a novel word, “deebo,” which was always produced with a consistent, salient pitch contour. In test, we changed either the pitch contour or the vowel (from /i/ to /a/). All of the 22 tokens participants heard in the teaching phase had the same vowel and the same pitch contour. If participants were storing each exemplar of this new word without selective emphasis on the native-language dimensions of contrast (as predicted by Goldinger’s 1998 “extreme” model), they would be expected to treat both changes as equally relevant in word recognition. We found instead that both children and adults interpreted these changes in accordance with English phonology, reacting to the segmental change but not to the pitch change. Even 2.5-year-olds were able to override the consistency of the teaching exemplars to assign the pitch variation to the appropriate level, possibly interpreting it as phrasal intonation rather than as part of the word.

At both ages, we saw individual variation in participants’ interpretations of the vowel change. Adults’ and children’s interpretations may have varied partly because of tension between their phonological knowledge and the pragmatics of the experiment. Participants’ phonological knowledge may tell them that a change from /i/ to /a/ signals a new word. Consistent with that knowledge, 18/24 adults and 11/24 children fixated the

deebo less than 50% percent of the time in response to "dahbo," suggesting they hypothesized that "dahbo" was a new word referring to the previously unlabeled distracter object. The pragmatics of the experiment, however, may support the alternative interpretation that "dahbo" is simply a mispronunciation of "deebo." In vowel-change trials, the *deebo* object was on the screen (with a distracter object), and participants heard a word that differed from "deebo" in only one segment. In the real world, interlocutors occasionally mispronounce words, requiring listeners to accommodate some variation. When an object is present and a speaker produces a word differing from that object's label in only one segment, this variant may well be a mispronunciation rather than a new word. Consistent with this interpretation, 6/24 adults and 13/24 children fixated the *deebo* more than 50% of the time in response to "dahbo," suggesting they hypothesized that "dahbo" was simply a mispronunciation of "deebo." The tension between English phonology and the pragmatics of the experiment may explain why many adults were inconsistent in their treatment of "dahbo" across different measures, apparently unable to settle on one interpretation or the other.

2.4.1 Pitting Children's Experience with a Word Against Their Phonology

Our finding that children do not treat all dimensions alike when representing and recognizing a new word is relevant to an ongoing debate over the abstractness of young children's—and even adults'—word representations. Psychological speech-recognition models have typically assumed that representations of words are composed of abstract phonemes (cf. Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986; and Norris, 1994), but experimental evidence suggests that adults' word representations are highly detailed. In word recognition, adults are sensitive to subphonemic information (Andruski,

Blumstein, & Burton, 1994; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; McMurray, Tanenhaus, & Aslin, 2002; Salverda, Dahan, & McQueen, 2003; Salverda et al., 2007) and to characteristics of the speaker's voice (Palmieri, Goldinger, & Pisoni, 1993; Goldinger, 1996; Luce & Lyons, 1998). And they are better at recalling a list of words spoken by one talker, at one speaking rate, than a list spoken by different talkers or at different speaking rates (Nygaard, Sommers, & Pisoni, 1995). Pronunciation of words in speech also reflects knowledge of word frequencies, information not available in abstract phonological representations. For example, speakers are more likely to reduce high-frequency words in production than low-frequency words (for reviews, see Pierrehumbert, 2001; Bybee, 2001a, 2007), and words that are used frequently together are more susceptible to liaison (Bybee, 2001b).

This evidence for nonphonemic information in word representations has led to the development of exemplar theories of speech-sound learning (Jusczyk, 1993), perception (Johnson, 1997), and production (Pierrehumbert, 2002). According to exemplar theories, word and speech-sound categories emerge from the storage of many detailed exemplars of the category. In word recognition, a word form activates the stored exemplars, and that pattern of activation is used to categorize the new token. Through the incorporation of attention weights (Johnson, 1997; Jusczyk, 1993), exemplar models can selectively emphasize certain acoustic or phonetic dimensions over others. Jusczyk (1993) proposed that phonological development proceeds by fine-tuning attention weights to emphasize dimensions relevant in the native phonology. Because less-relevant dimensions are not completely deweighted, even adults show sensitivity to variation on these dimensions in implicit tasks, but their word recognition is not impaired. In contrast, young children are

much more sensitive to episodic details, failing to recognize a word when the fundamental frequency (Singh, White, & Morgan, 2008), talker's voice (Houston & Jusczyk, 2000), or affect (Singh, Morgan, & White, 2004) has changed between familiarization and test. Presumably, infants are more sensitive to these dimensions in word recognition because they are still fine-tuning the weights of acoustic dimensions to match their native phonology.

Our results could be consistent with either the Jusczyk-style (1993) exemplar perspective or the abstraction view. The abstraction view is transparently consistent with our finding that English-learning children disregard lexical pitch. According to the abstraction perspective, children categorize new words as sequences of consonants and vowels, and do not store information like pitch in the lexical representation if it is not phonologically distinctive.

If viewed from the exemplar perspective, our results could be seen as evidence that 2.5-year-olds have already tuned their weights of acoustic dimensions to match the phonology of English, so that pitch information is downweighted sufficiently to not impact word recognition. This characterization is not typical of exemplar models, which were designed to account for listeners' retention of noncontrastive information (e.g., Goldinger, 1998). Still, this weaker version of exemplar models (e.g., Jusczyk, 1993) is consistent with our results.

Though simple exemplar models help account for effects of nonphonemic variation on word recognition, recall, and production, some questions remain. If people store nonphonemic detail about individual tokens of a word, how do we seem to make the phonologically normative interpretive decisions so consistently? Attention weights,

which emphasize those dimensions on which sounds contrast, begin to suggest an answer, but they are an incomplete solution in two ways. As Francis and Nusbaum (2002) point out, attention weights that operate at the level of the entire dimension (following in the vein of Nosofsky's 1986 generalized context model; Jusczyk, 1993, 1994; Johnson, 1997) are insufficient. Mature interpretation of speech requires more than attending just to an entire relevant dimension (e.g., Iverson & Kuhl, 1995). Instead, it appears to require *localized* variation in attention along a dimension, in which differences near the category center are compressed, and differences near the category boundary are expanded (Goldstone, 1994; Guenther, Husain, Cohen, & Shinn-Cunningham, 1999).

More fundamentally, the demands of ordinary conversation require listeners to attend to word-level *and* utterance-level phonetic information, both of which are given in the very same signal. Rather than supposing that listeners attend to one level at the expense of the other, we argue that listeners construct a *model* of the utterance, based on linguistic knowledge, to estimate the most probable interpretation (e.g., Dahan, Drucker, & Scarborough, 2008). For a given phonetic attribute (whether it be pitch, duration, glottalization, etc.), responsibility for the value of that attribute may need to be partitioned among several factors. In the case of duration, the length of a vowel results from word-level characteristics (e.g., vowel identity, syllable position, identity of the following consonant) and utterance-level characteristics (e.g., speaking rate, location relative to prosodic boundaries), as shown in numerous phonetic studies (e.g., Klatt, 1973; van Santen, 1992). The child's task is to discover the linguistic model that aligns best with that of her community.

We have shown that 2.5-year-olds have settled on the correct linguistic model for interpretation of pitch variation at the lexical level. An important extension of the present research will be to investigate the developmental trajectory of the interpretation of pitch. This trajectory could take two forms. Children could start out disregarding pitch variation, and then learn, through exposure to their native language, to attend to pitch at the relevant levels. Alternately, children could start out treating pitch as potentially relevant (e.g., at the lexical level), and then learn to ignore it if their native language doesn't provide evidence of structure at that level. We find the latter trajectory more likely, because of evidence that children start out more open-minded about what can be a word, constraining their hypotheses over development (Namy, 2001; Namy & Waxman, 1998; Woodward & Hoyne, 1999). However, further research is required to pinpoint the precise developmental trajectory. The present work provides an important starting point by demonstrating that by 2.5 years, interpretation of lexical pitch is similar to the adult interpretation, at least under the conditions tested here.

Studies like the current one shed light on outstanding questions about the nature of speech interpretation by providing evidence about the development of interpretation of perceptible, but nonphonemic, variation. We considered the interplay between the acoustic particulars of listeners' experience with a word and the constraints of their phonological system. From previous research, we know that adults show "echoes" of nonphonemic variation in word recognition, and infants often have even more trouble disregarding this variation. Young children often seem to struggle to interpret novel words through the lens of their native-language sound system. Yet we found that both children and adults could disregard consistency in the pitch contour of a novel word,

recognizing a newly learned word even when the consistency of its pitch contour was violated. This result tells us that by 2.5 years, children do not treat all dimensions of the sounds of words equally, but instead interpret a nonphonological change in pitch contour differently from a phonological vowel change.

Chapter 3: Development in Children's Sensitivity to Pitch as a Cue to Emotions

Carolyn Quam and Daniel Swingley[4]

Abstract

Even young infants respond to positive/negative prosody in parental speech (Fernald, 1993), yet 4-year-olds rely on lexical information when it conflicts with paralinguistic cues to approval/disapproval (Friend, 2003). The present research explores this surprising phenomenon, testing 2- to 5-year-olds' sensitivity to an isolated pitch cue to emotions in interactive tasks. Only by 4–5 years did children consistently use exaggerated, stereotypical pitch contours as evidence that a puppet had succeeded or failed to find his toy (Experiment 1) or was happy or sad (Experiment 2). Two- and three-year-olds exploited facial and body-language cues in the same task. The authors discuss the implications of this late development of use of pitch cues to emotions, relating them to other functions of pitch.

3.1 Introduction

Intonation plays a special role in young infants' linguistic interactions with their parents, well before infants know any words. Parents speak to their infants in a distinctive way, using higher mean fundamental frequency (f0) and a wider f0 range than when they speak to adults. These pitch characteristics are well suited to the infant's developing auditory system, making infant-directed speech more interesting to infants, and easier for

[4] This study has been presented by Quam, Swingley, & Park (2009; see also Quam & Swingley, under review); see **References** for details.

them to perceive, than adult-directed speech (Fernald, 1992). Infants prefer listening to infant-directed speech over adult-directed speech (Fernald, 1985), primarily because of its distinctive pitch features (Fernald & Kuhl, 1987; Katz, Cohn, & Moore, 1996). Adults express pragmatic functions (like comfort or prohibition) more clearly in the infant-directed register than in adult-adult speech (Fernald, 1989), and infants can group together infant-directed utterances by their pragmatic functions (Moore, Spence, & Katz, 1997). The intonation contours of infant-directed speech even help infants to segment words from the speech stream (Thiessen, Hill, & Saffran, 2005). In sum, there is abundant evidence that parents speak in a distinctive way to infants, shaping the pitch patterns of their speech to attract the infant's attention and communicate emotional and pragmatic information, and that infants are highly sensitive to these modifications.

In addition to its role in attracting infants' attention to speech and conveying emotional and pragmatic information, pitch may drive infants' sensitivities to several linguistic features, including phrase boundaries. Infants are sensitive to prosodic cues to phrase boundaries by 2 months; their memory for the phonetic properties of words is better when the words appear within versus between prosodically marked clauses (e.g., Mandel, Jusczyk, & Kemler Nelson, 1994). By 6 months, infants use pitch contours to parse utterances into clauses (e.g., Seidl, 2007), and by 10 months infants' recognition of a word is impaired when a prosodic break is inserted between its syllables (Gout, Christophe, & Morgan, 2004; see also Johnson & Jusczyk, 2001).

Infants' early sensitivity to intonation in parents' speech does not automatically provide them with an adult-like *understanding* of the many functions of pitch contours, however. Pitch is exploited in language at several levels of structure, and different

languages use pitch differently. For example, lexical-tone languages use pitch to contrast words. English does not have lexical tone, but uses higher pitch as one of several correlated cues to word stress (which sometimes differentiates words, as in noun-verb pairs like CONduct–conDUCT). Other languages mark stress with pitch lowering rather than raising, or do not use pitch to mark stress at all (Pierrehumbert, 2003). A child learning English must identify the particular role of pitch in marking lexical stress, while a child learning Mandarin must identify the role of pitch in the tonal system; and both children must also attend to pitch for demarcation of phrase boundaries, marking of yes/no questions, and emotional and pragmatic information, among other functions. Infants' early sensitivity to acoustic features of infant-directed speech—and even their apparent ability to respond appropriately to positive and negative prosody—do not necessarily entail the ability to *interpret* another person's vocal expressions of emotion.

3.1.1 Sensitivity to Vocal Cues to Emotions in Infancy

To understand the development of children's interpretation of emotional prosody, a distinction must be made between language-universal, direct effects of prosodic variations on infants' emotions, and a more reflective, interpretive capacity to integrate emotional prosody with the rest of the talker's linguistic message. It is in the former sense that young infants "understand" language before they know any words. Emotional prosody in parental speech induces infant emotion in appropriate and predictable ways, even if that speech was recorded in an unfamiliar language (Fernald, 1993). In one study, 12-month-olds showed reduced exploration of a toy and more negative affect upon hearing fearful-sounding speech, as opposed to more neutral speech, from a recorded actress (Mumme & Fernald, 2003). Similarly, Friend (2001) found that 15-month-olds'

exploration of a novel toy was affected by a combination of facial and vocal expressions of approval versus disapproval. Infants in such studies tend to respond more consistently to negative (e.g., fearful) messages than positive ones, compatible with the idea that parent-infant alignment about dangerous situations is evolutionarily more important than agreement about happy or contented states (Mumme, Fernald, & Herrera, 1996). Though infants display some sensitivity to vocal expressions of emotions, they appear even more sensitive to facial expressions. D'Entremont and Muir (1999) found that even 5-month-olds smiled more in response to happy than to sad facial expressions, but the addition of vocal paralanguage did not affect their responses, and they showed no differential responding to vocal paralanguage alone.

Infants have some knowledge that certain facial expressions go with certain vocal expressions. In intermodal-preference tasks, in which infants see two faces conveying different emotions and hear a voice that is more consistent with one face than the other, infants often gaze more at the matching face. In the youngest infants (3.5 months), this effect is found with the emotions *happy* and *sad* for the mother's face and voice, but not for unfamiliar women (Kahana-Kalman & Walker-Andrews, 2001); by 5 to 7 months, infants match unfamiliar faces and voices as well (e.g., Walker, 1982; see also Soken & Pick, 1992; 1999).

Infants can also detect certain *changes* in the pairings of faces and voices. In the multimodal-habituation task, infants are habituated to an affectively matching face and voice, and then some aspect of the stimuli, like voice affect, is changed; increased looking time indicates detection of the change. Five-month-olds detect changes in voice affect from happy to sad and vice-versa (Walker-Andrews & Grolnick, 1983, Walker-

Andrews & Lennon, 1991), but there is mixed evidence about their sensitivity to changes from happy to angry and vice-versa (Walker-Andrews & Lennon, 1991; Walker-Andrews, 1998). Walker-Andrews and Lennon (1991) found sensitivity to vocal-affect changes only when the same face was present during the entire experiment; they argue that the presence of a face is necessary during this intermediate stage in the progression from "featurally based discrimination" in early infancy to "meaningful discriminations among vocal-only and facial-only displays" later in development.

3.1.2 Surprising Difficulty Interpreting Paralinguistic Cues to Emotions in Early Childhood

Early-developing reactions to emotional prosody, and the capacity to link facial and vocal affective signals appropriately, do not appear to provide young children with a ready appreciation of how emotional prosody affects talkers' linguistic messages. This has been shown in a number of studies, most of which have presented children with discrepant linguistic and paralinguistic stimuli. When prosodic or facial cues conflict with lexical information, young children usually rely on the meaning of the words rather than facial expressions or prosodic contours. For example, Friend and Bryant (2000) asked children to place the emotion expressed by a disembodied voice on a 5-point scale ranging from "very happy" to "very mad." Four- and 7-year-olds relied more heavily on lexical information when it was in conflict with prosody, while 10-year-olds relied more on prosody (though in a similar experiment, even 10-year-olds relied more on lexical information than on paralanguage; Friend, 2000). Friend (2003) examined a more naturalistic behavioral response—interaction with a novel toy—to consistent versus discrepant lexical and paralinguistic (facial plus vocal) affective information from an

adult face on a video-screen. Four-year-olds approached the toy faster and played with it longer when the adult's affect was consistently approving than when it was consistently disapproving. When the cues were discrepant, words trumped facial-and-vocal paralanguage. Finally, Morton and Trehub (2001) examined 4- to 10-year-olds' and adults' ability to judge the speaker's happiness or sadness from vocal paralanguage versus lexical cues. Four-year-olds relied on lexical information when the cues conflicted, while adults relied exclusively on paralanguage. In between, there was a gradual increase in reliance on paralanguage; only half of 10-year-olds relied primarily on paralanguage. When 6-year-olds were primed to attend to paralanguage, however, they successfully relied more on paralanguage than on lexical information (Morton, Trehub, & Zelazo, 2003).

Like infants, preschoolers show some sensitivity to paralinguistic cues to emotion—when they are not pitted against lexical information. Friend (2000) found that 4-year-olds can identify the affect of *happy* versus *angry* reiterant speech, in which lexical content is replaced with repetitive syllables (e.g., "mama ma"; Friend and Bryant, 2000, found a similar result with 7- to 11-year-olds). They fail with low-pass-filtered speech, however, which preserves primarily f0, suggesting that f0 alone is not a sufficient cue. Still, this failure could be due to the unnaturalness of either low-pass-filtered speech or of the task, in which children listened to a sentence out of context. Morton and Trehub (2001) did find some success at age 4 with low-pass filtered speech,[5] as well as with paralanguage in Italian, though the Italian stimuli differed on many acoustic dimensions:

[5] Morton and Trehub (2001) excluded 22 4- and 5-year-olds—for exhibiting a response bias—from their sample of 40 children. A response bias might indicate that the child cannot access the relevant cue. If those children exhibiting a response bias had been retained, the overall success rate of 4-year-olds would be much lower.

happy sentences had higher pitch, a faster speaking rate, and greater pitch and loudness variability.

The present research reexamined children's sensitivity to vocal paralanguage in the absence of conflicting lexical information, using interactive and age-appropriate tasks. As discussed, in prior studies infants only demonstrated clear sensitivity to vocal expressions of fear, which is likely evidence of a low-level, evolved behavioral response rather than interpretation of another person's emotions. Evidence from preschoolers is mixed, and we know very little about children's sensitivity to *pitch* cues in particular. Some previous studies have combined facial and vocal paralanguage (e.g., Friend, 2003; Mumme, Fernald, & Herrera, 1996). Those studies that considered only vocal paralanguage rarely disentangled pitch, speaking rate, loudness, and breathiness (e.g., Friend, 2000; Friend & Bryant, 2000; Morton & Trehub, 2001; Mumme & Fernald, 2003), except when using low-pass filtered speech, which introduces additional naturalness issues. Considering the arguments and evidence that pitch plays a crucial role in children's early language processing, it would be useful to isolate the pitch cue to the speaker's emotions—by which we mean both pitch height and pitch contour—and identify when children can exploit it.

An important question concerns whether pitch cues to all levels of structure are equally accessible to the child, or whether cues to different levels of linguistic structure are acquired at different developmental points, despite being carried by the same acoustic dimension. In line with Fernald's (1992) qualitative model of the changing role of pitch over development, we argue that different levels of pitch structure should be available to the child at different points, depending on the child's ability to access the cue in the

signal, the reliability of the cue in the signal, and the developmental relevance of the cue (see also Werker & Curtin's 2005 PRIMIR model of infant speech processing, and Hollich et al.'s 2000 emergentist coalition model of word learning). *Access* to a particular cue in the signal may require certain linguistic preconditions. For example, an infant would be unable to access the pitch cue to a word's stress pattern if he/she did not know the word. The *reliability* of the realization of a pitch cue (e.g., to lexical stress) may be compromised because pitch is also being used to convey phrasal information and emotional content. These types of trade-offs might be even more striking in a lexical-tone language, in which pitch must convey syllable-level tones in addition to phrasal and emotional information. In fact, to preserve tone information, Mandarin-speaking mothers appear to expand their pitch range and raise their pitch mean *less* than speakers of nontone languages, though they still produce the same intonational meanings (M. Papousek, H. Papousek, & Symmes, 1991; but see Kitamura, Thanavishuth, Burnham, and Luksaneeyanawin, 2002).

Finally, the *developmental relevance* of the cue, in addition to modulating children's attention to the cue, may even impact the reliability of its realization in the signal. Mandarin-speaking mothers appear to reduce or neglect tone information in favor of producing simple intonation contours to 2-month-olds (Papousek & Hwang, 1991), but actually *exaggerate* tone categories in speech to 10- to 12-month-olds (Liu, Tsao, & Kuhl, 2007) similarly to how vowel categories are exaggerated in infant-directed speech (Burnham, Kitamura, & Vollmer-Conna, 2002). This difference could arise because intonational meaning is more relevant to younger infants, and tone and segmental

information is more relevant to older infants (Kitamura & Burnham, 2003; Stern, Spieker, Barnett, & MacKain, 1983).

Infants acquiring a tone language appear to learn tones at about the same time that they learn consonant and vowel categories (Mattock & Burnham, 2006; Harrison, 2000; Hua & Dodd, 2000), while tones that are less consistently realized appear to be learned more slowly (Demuth, 1995; see also Ota, 2003). *Interpretation* of highly discriminable pitch variation also appears to follow a slower time-course; English-learning children learn to disregard potentially lexical pitch sometime between 18 (Quam & Swingley, in progress) and 30 months (Quam & Swingley, in press), possibly by detecting the variability of pitch contours of words across tokens (Quam, Yuan, & Swingley, 2008). These different acquisition trajectories suggest that pitch cues to different levels of structure are indeed acquired at different time-points. We might therefore find a relatively late development of successful interpretation of pitch cues to emotions, despite the early importance of intonation in infancy.

The present work addresses children's understanding of intonational cues to the emotions *happy* and *sad*, comparing these cues to nonlinguistic, facial and body-language cues to the same emotions. We chose these two emotions to maximize the contrast between the two emotions children had to distinguish. Vocal expressions of emotions have been described as varying on two independent dimensions: valence (positive or negative) and activation/arousal (high or low; e.g., Russell, Bachorowski, & Fernández-Dolz, 2003). The emotional expressions used here contrasted on both these dimensions: *happy* had positive valence and high activation/arousal (reflected in high pitch means and large pitch excursions), while *sad* had negative valence and low activation/arousal

(reflected in low pitch means and small excursions). This particular contrast should therefore provide the best opportunity for children to demonstrate knowledge of how pitch indicates emotions.

Again, though young infants are sensitive to the prosodic characteristics of infant-directed speech (e.g., Fernald, 1993), preschool children have difficulty interpreting prosodic cues to the speaker's emotions (e.g., Friend, 2003). Two experiments tested preschoolers' interpretations of a pitch cue to emotions in the absence of conflicting lexical information, and using interactive tasks. Experiment 1 used a task inspired by Tomasello and Barton's (1994) nonostensive word-learning study (Experiment 4). Children had to interpret the emotions of a puppet, "Puppy," in order to infer which toy was the object of his search; Puppy was happy if he found his toy, and sad if he found a different toy. Children responded by giving the toy to Puppy if he was happy, and throwing it in a trashcan if he was sad. Experiment 2 used a simpler and more direct test of sensitivity to emotions. Puppy was again searching for toys, but this time children simply responded by pointing to a happy face (or saying "happy") if Puppy was happy, or pointing to a sad face (or saying "sad") if Puppy was sad.

3.2 Experiment 1

3.2.1 Method

3.2.1.1 Participants

Thirty-six children participated in Experiment 1 (20 female, 16 male): 13 3-year-olds, 15 4-year-olds, and 8 5-year-olds. Children were recruited by staff in preschools, via letters sent to parent addresses from a commercial database, and by word of mouth.

Of the 36 children included in the study, 2 were Asian, 2 were African-American, 5 were of mixed race or reported to be "Other," and 27 were Caucasian; 3 of the 36 children were Hispanic/Latino. These counts of racial groups are estimates based on voluntary parental report for some children and observation for others. Parental SES was not evaluated. Three more 3-year-olds participated, but were excluded: 2 for failure to participate in enough trials, and one for experimenter error. Since many 3-year-olds needed some help with both pretrials (and many of these children still succeeded in the body-language trials), failure in the pretrials was not used as grounds for exclusion in this experiment.

3.2.1.2 Apparatus and Procedure

Participants sat at a table across from the experimenter (the first author), either at the child's preschool or in a university developmental-laboratory suite. A red cylindrical container (the "trashcan") was placed to the child's right, with a cardboard box behind it, closer to the experimenter. In the preschool setting, one camcorder recorded the experimenter's face, while another camcorder, connected to an external microphone, recorded the child and the table. In the laboratory setting, a single camcorder, connected to the external microphone, recorded the child and the table and captured the experimenter's face in a mirror placed above the child. See **Figures 6** and **7** for a photograph and diagram of the testing setup, respectively.

Figure 6: Photographs of the experimental setup for Experiments 1 (left) and 2 (right)

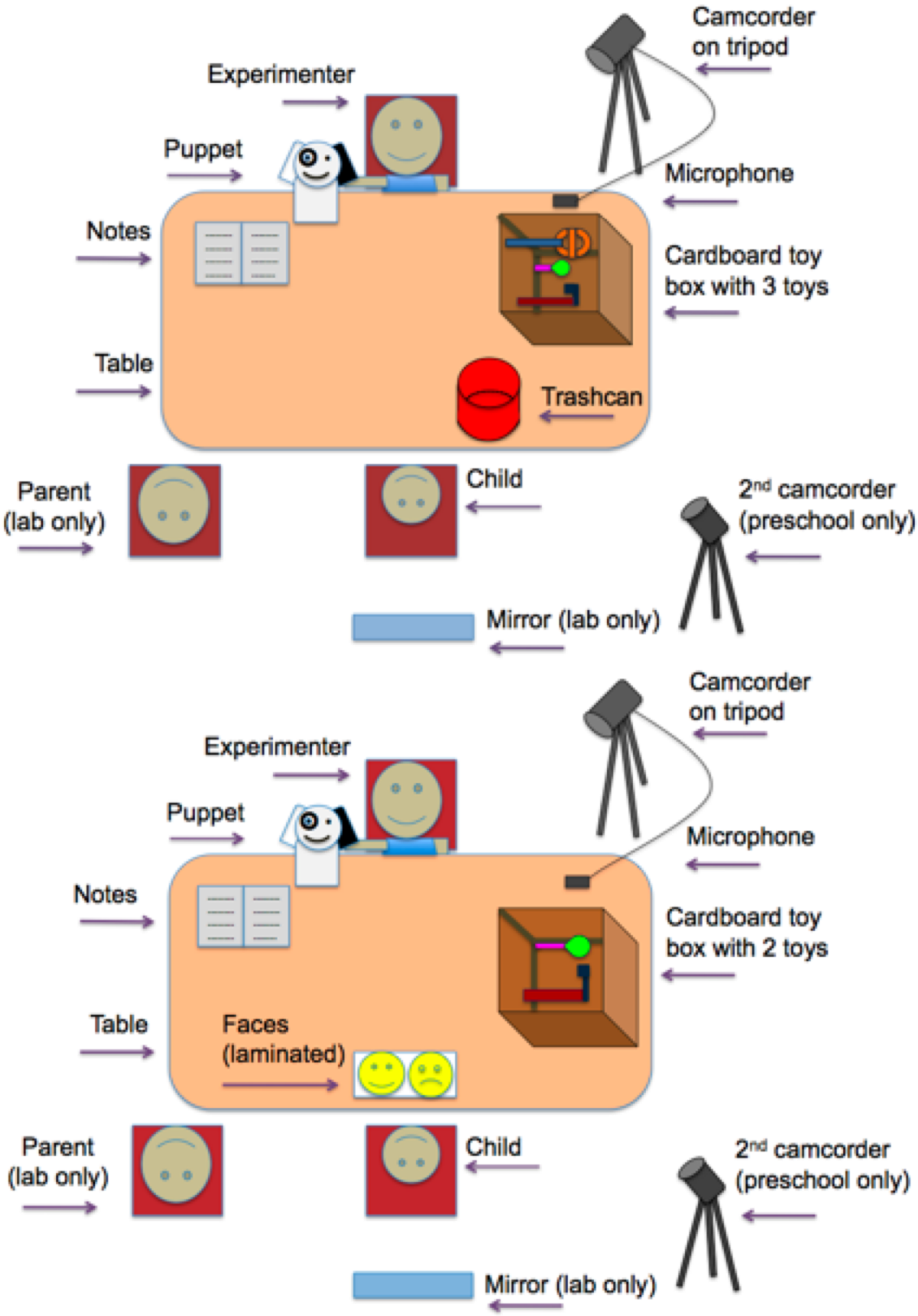


Figure 7: Diagrams of the setup for Experiments 1 (top) and 2 (bottom)

Before the experiment, children were told they would be playing a game in which they would see several toys and meet the experimenter's friend Puppy, a puppet. In each trial, the experimenter put three toys in the box (different toys for each trial). Children were first permitted to examine each of the three toys. Children were then told that Puppy was looking for a particular toy (e.g., the *toma*). Puppy would be happy when he found the *toma*, and sad if he found a different toy. Children were instructed to give the *toma* to Puppy and throw the other toys in the trash. The experiment began with one or two *pretest trials*, intended to teach children the task and let them practice the giving/throwing-away response (if a child failed the first pretest trial, a second one was included). In pretest trials, Puppy was "feeling shy," so he whispered in the experimenter's ear whether each toy was the target, and the experimenter told the child explicitly. In the next three *body-language trials*, Puppy produced a body-language cue to indicate whether each toy was the target. He expressed excitement by nodding and dancing side-to-side, and disappointment by shaking his head and slumping down. Finally, in the *pitch trials*, the experimenter, speaking for Puppy, produced excited pitch (high pitch with large excursions) or disappointed pitch (low pitch with small excursions). The first seven children participated in three pitch trials, but for the remainder we added an extra pitch trial; see **Appendix 2** for a sample trial order for the four-trial version. The extra trial was added to make it easier to distinguish between children who truly understood the *happy* versus *sad* pitches and children who happened to guess correctly part of the time. In particular, in the three-trial version children often responded correctly in 2/3 trials, making it difficult to tell whether they truly understood

the pitch cues or just happened to guess correctly on two trials. This was less of an issue in the four-trial version, since 3/4 correct is less likely to occur by chance.

3.2.1.3 Visual Stimuli

The toys used in the experiment were all intended to be novel. **Appendix 3** displays four of the roughly two-dozen toys used in the experiment. Most were handmade from parts of kitchen appliances, dog toys, and electronics, though some toys were unmodified from their original form (e.g., an unusual-looking potato masher—the first toy in **Appendix 3**). Children occasionally recognized parts of toys, saying, e.g., "That's a rolling pin!" The experimenter responded, "It looks kind of like a rolling pin, but it's just a silly toy." If the child asked, "What is that?" the experimenter responded, "I don't know—it's just a silly toy."

The puppet was a plush, black-and-white spotted dog measuring twelve inches high and six inches across (arm-span eleven inches). When the puppet was "talking," the experimenter moved Puppy's left hand once for each syllable so it was clear that Puppy was the one talking to the child. In all experimental trials, the puppet was placed between the experimenter's face and the child's face to prevent the experimenter from conveying any facial cues (see **Figure 8**).

Figure 8: Puppy blocks experimenter's face during all trials in Experiment 1 and in the pitch condition of Experiment 2.

3.2.1.4 Auditory Stimuli

Auditory stimuli were produced live by the experimenter. The experimenter mostly talked directly to the child, but during the crucial part of the test trials, she said, "Look what I found! Puppy, is this the [toma]?" and then, keeping the puppet between her face and the child's face, said "mmm, mm mm mmm" with stereotypical excited/happy pitch or disappointed/sad pitch. The sad pitch was characterized by low, falling pitch with small pitch excursions, while the happy pitch was high, with rise-fall patterns and large pitch excursions; see **Figure 9** for waveforms and pitch tracks of two example contours. The experimenter reminded the child to listen before producing each pitch contour.

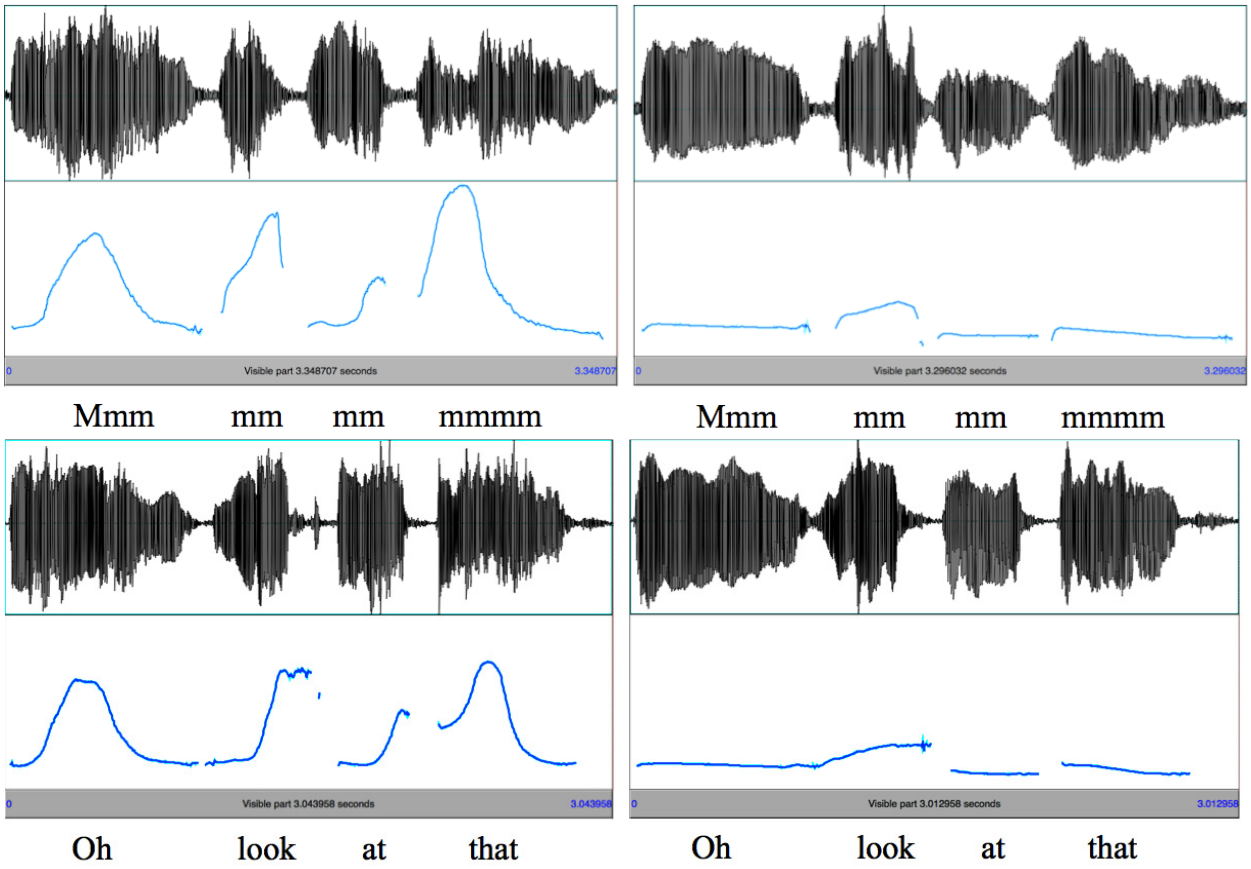


Figure 9: Waveforms and pitch contours for examples of the happy (left) and sad (right) pitch contours used in both Experiments 1 (top) and 2 (bottom)

In addition to the differences in pitch height and contour that we intentionally produced, happy/excited speech is stereotypically higher in amplitude and faster; spectral (or timbre) differences can also result from these differences in amplitude and pitch as well as from differences in mouth shape (e.g., smiling versus frowning). The experimenter was aware of these correlated cues, and attempted to equate duration, amplitude, and mouth shape when producing the stimuli. Nevertheless, we evaluated whether some differences in duration and amplitude may have persisted alongside the intentional differences in pitch characteristics. To numerically compare the acoustics of the experimenter's productions of *happy* versus *sad* pitches, the acoustic measurements from the two *sad* productions were averaged to produce a single value, which was then

compared to the *happy* value from that trial. This analysis was conducted on only the pitch trials from participants who had responded correctly on at least two of the first three pitch trials. In this and all following acoustic analyses, duration and intensity values were natural-log-normalized, and pitch measurements were converted from Hz to ERB (ERB = 11.17 * ln((Hz + 312)/(Hz + 14675)) + 43; Moore & Glasberg, 1983). Results were comparable without these conversions, however, and means are given here in Hz, seconds, and dB for ease of interpretation.

Happy and sad productions differed significantly on all acoustic dimensions measured. Happy productions had **higher pitch means** (happy, 416.74 Hz; sad, 255.72 Hz; paired t(23) = 57.74), **larger standard deviations of pitch samples** (happy, 129.60 Hz; sad, 51.78 Hz; paired t(23) = 80.59), **higher pitch maxima** (happy, 745.61 Hz; sad, 386.86 Hz; paired t(23) = 57.22), **higher pitch minima** (happy, 210.80 Hz; sad, 163.33 Hz; paired t(23) = 7.06), **greater intensities** (happy, 72.72 dB; sad, 71.26 dB; paired t(23) = 5.34), and **greater durations** (happy, 3.30 seconds; sad, 3.24 seconds; paired t(23) = 3.08, all $p < .01$, all tests 2-tailed). Though all of the tests indicated significant differences, the ratios of HappyValue / SadValue for intensity and duration were very close to one (see **Appendix 4**), suggesting that differences in intensity and duration, though consistent, were small in magnitude compared with the large differences in pitch that we intentionally produced, and they may not have been noticeable to participants.

3.2.2 Results and Discussion

Each participant gave responses in three body-language trials and three or four pitch trials.[6] **Table 5** reports the number of children at each age that succeeded with each cue. Success was defined as choosing the correct toy (the one to which Puppy responded with happy body-language or pitch) in at least two of the first three trials. Taken as a whole, the group performed significantly better with the body-language cue (89.8%) than with the pitch cue (61.1%, paired $t(35) = 4.55$; $p < .001$, 2-tailed). **Figure 10** plots accuracy in each condition against age. There was little change in success with the body-language cue across development, though there was a significant correlation between accuracy and age ($r = 0.36$, $p < .05$). At age 3, 11/13 children successfully exploited the body-language cue, while all of the 4- and 5-year-olds succeeded (15/15 4-year-olds and 8/8 5-year-olds). By contrast, the pitch cue showed marked improvement with age. There was a significant correlation between accuracy and age ($r = 0.46$, $p < .005$). At age 3, only 7/13 children succeeded with the pitch cue; at age 4, 10/15 succeeded; and by age 5, 7/8 children succeeded. While almost all of the 3-year-olds successfully exploited the body-language cue, the pitch cue was more difficult for children, and showed a more protracted developmental course.

Table 5: Success at each age with pitch versus body-language cues in Experiment 1

Age	***Body-language***	***Pitch***	***(At least 2 of first 3 trials correct)***
3	**11 / 13 (85%)**	**7 / 13**	**(54%)**
4	**15 / 15 (100%)**	**10 / 15**	**(67%)**
5	**8 / 8 (100%)**	**7 / 8**	**(88%)**

[6] Mean accuracy on pitch trials was not materially affected by inclusion of only the first three pitch trials rather than all four (63.9% with three trials, vs. 61.1% with all trials included).

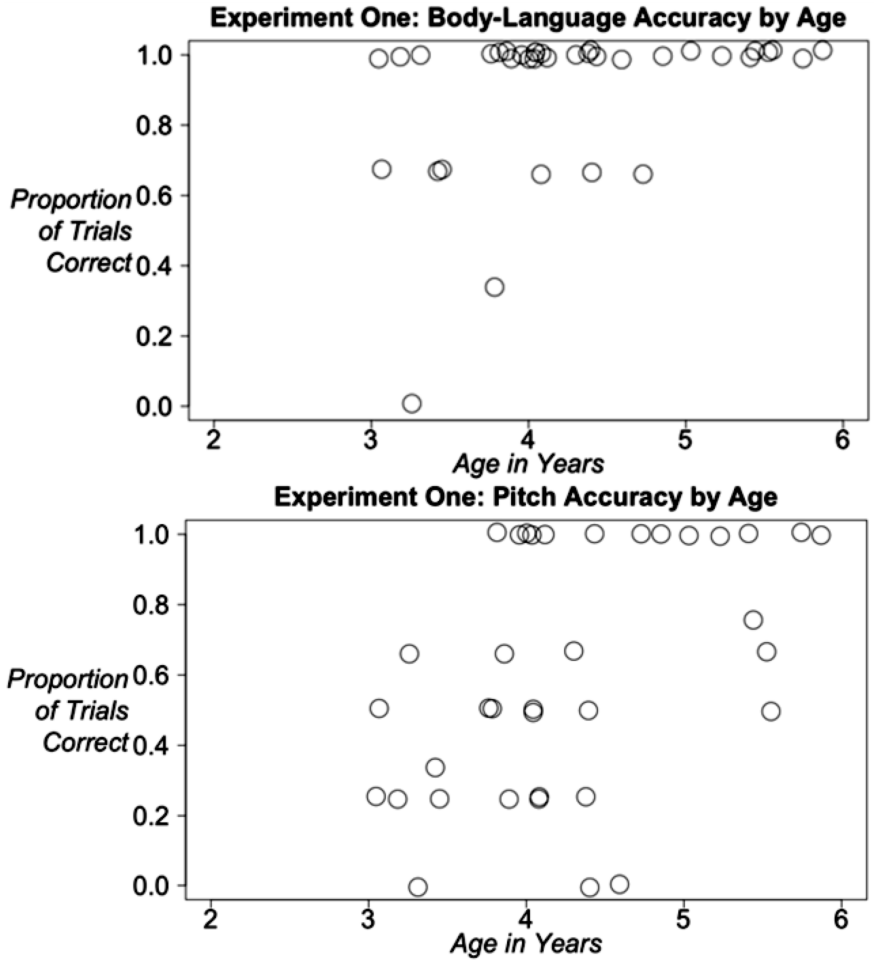


Figure 10: Scatterplots of accuracy with body-language (top) and pitch (bottom) cues across age in Experiment 1

3.3 Experiment 2

Children's success with the body-language cue in Experiment 1 suggests that the task itself was not responsible for children's difficulty with the pitch cue. Still, several concerns motivated us to make changes to the procedure, implemented in Experiment 2. First, the body-language cue in Experiment 1 involved the puppet nodding his head and then dancing, to express excitement, or shaking his head and slumping down, to express disappointment. We became concerned that these physical cues might be better mapped onto the meanings *yes* versus *no* than onto *excited* versus *disappointed.* If this were the

case, children might have an easy time with interpreting the nodding/shaking-head as yes/no, but then have a difficult time applying that mapping to the new cue, in which the pitch patterns better correspond with the meanings *excited* versus *disappointed* than with *yes* versus *no.* In other words, the transparent yes/no meanings of nodding/shaking-head might actually have *interfered* with children's interpretations of the pitch contours.

Another concern is that the inherent difficulty of the task might have blocked children's access to the pitch cue. Though children did succeed with the body-language cue, they still might have performed better with the pitch cue if the task had been simpler. Children, especially 3-year-olds, often struggled to remember to give the target toy to Puppy and throw the others in the trash. Finally, we were concerned that the task was not a direct test of interpretation of emotions. Children were not simply asked to tell the experimenter whether Puppy was happy or sad. Instead, they were required to make that judgment, and then make the further inference that if Puppy was happy, this was the toy he was searching for; and if Puppy was sad, this was *not* the toy he was searching for. Then, they had to perform the additional task of putting each toy in the correct location. As Baldwin and Moses (1996) point out, the ability to interpret emotions like happiness and sadness may be dissociable from the understanding that these emotions *refer* to things in the world; the latter understanding may take longer to develop. Removing the referential component of the task might therefore reveal children's understanding of the emotions themselves.

Experiment 2 implemented a simpler and more direct test of interpretation of emotions. Again, Puppy was presented with toys—this time only 2 toys per trial—and responded to each toy with excitement or disappointment. Children were simply asked to

tell the experimenter whether Puppy was happy or sad. Given the simplicity of this task relative to the task in Experiment 1, we included younger children—2-year-olds—in Experiment 2. The body-language cue was also better matched to the pitch cue; both of them mapped onto the meanings *excited/happy* versus *disappointed/sad.* The pitch contrast was identical to that tested in Experiment 1, but was produced on the words "Oh, look at that," which should be more naturalistic than hummed speech. The body-language and pitch cues were tested between subjects, unlike in Experiment 1, to avoid the possibility of transfer from one condition to the other.

3.3.1 Method

3.3.1.1 Participants

Sixty-two children participated in Experiment 2 (31 female, 31 male): 12 2-year-olds (6 in each condition), 26 3-year-olds (10 in the body-language condition, and 16 in the pitch condition), 12 4-year-olds (in the pitch condition), and 12 5-year-olds (in the pitch condition). Of the 62 children, one was Asian, 10 were African-American, 11 were of mixed race or reported to be "Other," and 40 were Caucasian; 5 of the 62 children were Hispanic/Latino. These counts of racial distribution were estimated as in Experiment 1. Participants were recruited as in Experiment 1, and parental SES was not evaluated. Seven more children participated but were excluded from the analysis: 3 2-year-olds (2 for failing both pretrials—i.e., not knowing the happy/sad faces—and one for having fewer than 6 usable trials), and 4 3-year-olds (2 for failing the pretrials, one for having fewer than 6 usable trials, and one because she was loudly singing along to the auditory stimuli, likely preventing her from adequately hearing them).

3.3.1.2 Apparatus and Procedure

The experimental setup of Experiment 2 was similar to that of Experiment 1. A cardboard box was again placed on the table to the child's right, but there was no trashcan, and a laminated piece of paper depicting a smiley-face (on the left) and a frowny-face (on the right) was placed directly in front of the child on the table. See **Figures 6** and **7** for a photograph and diagram of the testing setup, respectively.

At the beginning of the experiment, the child was told that Puppy was searching for his lost toys; that there were 2 toys in the box, one of which was Puppy's lost toy; and that Puppy would be happy if he found his lost toy, and sad if he found the other toy. The child was taught to point to the happy face when Puppy was happy, and to the sad face when Puppy was sad. Once the child was able to point correctly to each face, the experiment began. The experimenter pulled each toy out of the box one at a time and said, "Puppy, look what I found!" In the pretrials, Puppy was "feeling shy," so he whispered in the experimenter's ear, and the experimenter told the child directly how Puppy felt, and whether this was his lost toy. Then the experimenter asked the child, "Can you show me how Puppy feels?" If the child did not point immediately, the experimenter asked follow-up questions like "Can you point to the face?" or "Is Puppy happy or sad?" Verbal responses, e.g., "He's happy/sad" were also accepted. In response to the second toy, Puppy expressed the opposite emotion. After children had described Puppy's emotions in response to each toy, the experimenter placed both toys in front of the child and asked, "Which one is the toy Puppy lost?" If the child was unable to point to the correct faces, the experimenter ran a second pretrial. After the first 14 participants were tested, the experimenter ran both pretrials regardless of children's ability to point to

the faces, in order to reduce the possibility of response bias by presenting examples of *both* the first and second toys being the target. Children who were unable to point to the correct faces in the second pretrial were excluded from the analysis.

The 12 test trials had a similar structure to pretrials (see **Appendix 5** for an example trial-order), except that each child was given either pitch cues or facial and body-language cues to Puppy's emotions. Before the first test trial, children were again asked to show the experimenter that they could point to the happy and sad faces. In the first test trial, children in the *pitch condition* were told that Puppy was "not feeling shy anymore, so he's gonna talk this time!" Children were told they would have to listen carefully to tell if Puppy was happy or sad when he saw each toy. Then, the experimenter pulled each toy out of the box (using different toys in each trial) and again said, "Puppy, look what I found!" The experimenter reminded the child to listen, then kept the puppet between her face and the child's face (see **Figure 8**) while saying "Oh, look at that." The pitch contours were identical to those used in Experiment 1; see **Figure 9** for waveforms and pitch tracks of two example contours.

In the *body-language condition*, children were told they would have to watch carefully to tell if Puppy was happy or sad when he saw each toy. After the experimenter pulled each toy out of the box, children were asked, "Are you ready to watch? Let's see what Puppy does!" For a happy response, the experimenter smiled and raised her eyebrows, and she and Puppy danced side-to-side. For a sad response, the experimenter frowned and brought her eyebrows down, and she and Puppy slumped down (see **Figure 11**).

Figure 11: Happy and sad facial expressions produced during Experiment 2 facial / body-language condition

As children became more familiar with the task, they sometimes participated in the repetitive story, by filling in the words "happy" and "sad":

Experimenter: "One of these toys is the one Puppy lost, so if he finds it he'll be…"
Child: "Happy!"
Experimenter: "That's right! But the other toy is *not* the toy Puppy lost, so if he finds it he'll be…"
Child: "Sad!"

Children's productions of "happy" and "sad," either during this repetitive story or as verbal responses during test trials, were recorded and analyzed to see whether children produced a pitch contrast in their own productions of "happy" versus "sad." At the end of the experiment, the experimenter sometimes asked the child to imitate how Puppy sounded when he was happy or sad. Twelve children tried to imitate Puppy's happy and sad pitch contours. An acoustic analysis of all three response-types is reported in the **Results and Discussion** section.

In a few trials, children changed from the wrong answer to the correct answer, apparently catching themselves making an error. In these cases, coders carefully analyzed

the videos of both the child's face and the experimenter's face to determine whether the experimenter might have inadvertently shown surprise at the incorrect response, leading the child to switch to the correct response. Three trials were excluded from the analysis for this reason.

3.3.1.3 Auditory Stimuli

As in Experiment 1, the experimenter attempted to equate duration and amplitude. Paired t-tests comparing happy versus sad productions in each trial on several acoustic dimensions (again converting Hz to ERB, natural-log-normalizing duration and intensity, and including only those children who succeeded with the pitch cue—defined here as 75% correct responses) revealed significant differences between happy and sad productions on all acoustic dimensions measured except for duration. Happy productions again had **higher pitch means** (happy, 379.65 Hz; sad, 230.27 Hz; paired $t(20) = 48.68$), **larger standard deviations of pitch samples** (happy, 142.82 Hz; sad, 50.34 Hz; paired $t(20) = 43.52$), **higher pitch maxima** (happy, 749.05 Hz; sad, 397.87 Hz; paired $t(20) = 44.94$), **higher pitch minima** (happy, 201.16 Hz; sad, 155.57 Hz; paired $t(20) = 12.45$), and **greater intensities** (happy, 71.69 dB; sad, 70.79 dB; paired $t(20) = 6.91$; all $p < .001$, all tests 2-tailed). As in Experiment 1, only the pitch measurements had ratios (HappyValue / SadValue) that appear to be meaningfully different from one (see **Appendix 4**), suggesting that the differences in intensity may not have been noticeable to participants.

3.3.2 Results and Discussion

Each participant gave responses in either body-language trials or pitch trials. Participants were included if they were able to complete at least 6 of the 12 trials. **Table**

6 reports the number of children at each age that succeeded with each cue; success is defined as responding with the correct emotion ("happy" or "sad") at least 75% of the time. In the youngest group, 2-year-olds, children given the body-language cues performed significantly better (mean, 93.8%) than children given the pitch cues (mean, 48.6%, t(6.76) = 7.61; $p < .001$); this pattern held for 3-year-olds (body-language, 86.0%; pitch, 67.7%, t(20.59) = 2.53; $p < .05$; all tests 2-tailed and assumed unequal variances; results are comparable when equal variances are assumed). **Figure 12** plots accuracy in each condition against age. Again, there was little change in success with the body-language cue across development. If anything, older children performed slightly worse than younger children (7/10 three-year-olds succeeded, versus 5/6 two-year-olds)—likely because of boredom—but the correlation with age was not significant (r = -0.27, p = 0.31). For the pitch cue, there was a statistically significant correlation between accuracy and age (r = 0.50, $p < .001$). None of the 6 2-year-olds succeeded with the pitch cue, and only 5/16 3-year-olds succeeded. By age 4, over half of children (7/12) succeeded (mean, 79.0%), and by age 5, 75% of children (9/12) succeeded (mean, 84.8%).

Table 6: Success at each age with pitch versus facial/body-language cues in Experiment 2

Age	***Facial/Body-language***	***Pitch (At least 75% of responses correct)***
2	**5 / 6 (83%)**	**0 / 6 (0%)**
3	**7 / 10 (70%)**	**5 / 16 (31%)**
4	NA	**7 / 12 (58%)**
5	NA	**9 / 12 (75%)**

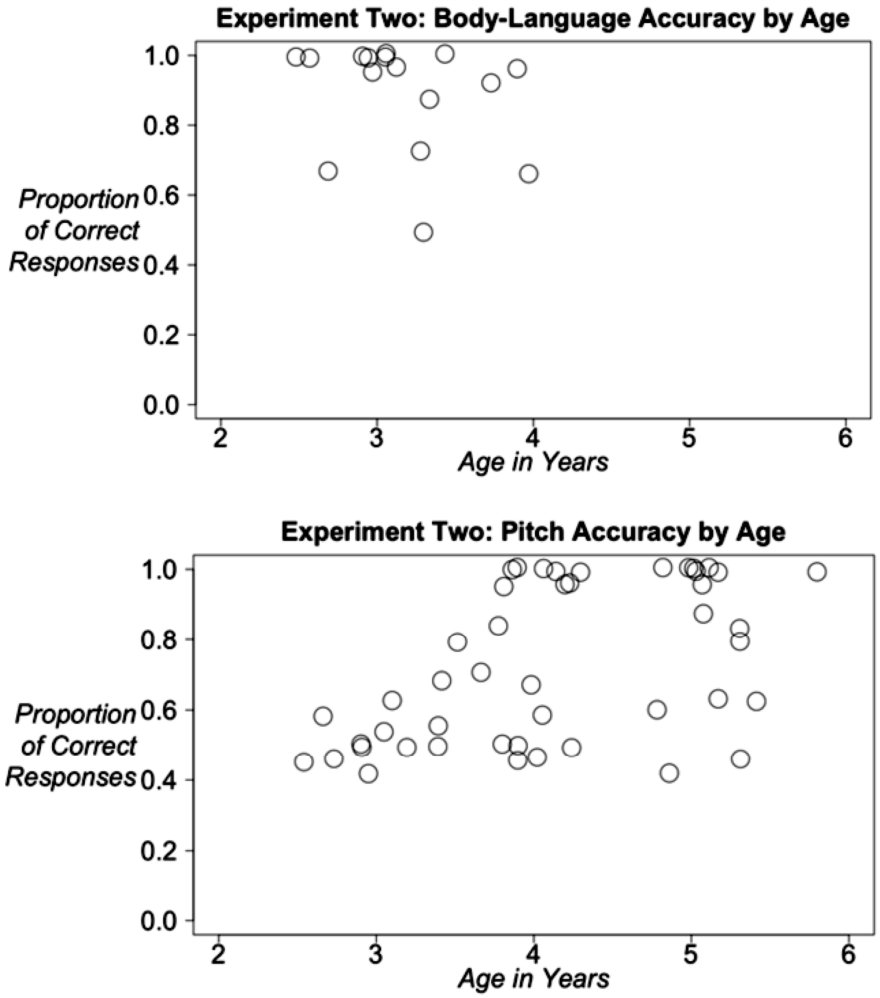


Figure 12: Scatterplots of accuracy with body-language / facial (top) and pitch (bottom) cues across age in Experiment 2

The greater acoustic salience (higher pitch mean and larger excursions) of the happy pitch, and greater visual salience of the happy body-language, could have made the happy stimuli easier for children to identify, but children in fact showed slightly better accuracy in trials in which the *sad* stimuli came first; this difference was significant for the pitch group (paired $t(45) = -2.23$, p(2-tailed) $< .05$).

Because of the length and repetitiveness of the experiment, children (especially the younger ones) sometimes became fatigued. This fatigue often resulted in children reverting from responding correctly to responding with "happy" for both toys or "sad" for

both. To reduce the impact of fatigue, we conducted a second analysis, excluding trials in which the child responded with the same emotion for both toys: this analysis is summarized in **Table 7**. After excluding "happy"/ "happy" and "sad"/ "sad" trials, 6/6 2-year-olds (mean, 97.2%) and 8/10 3-year-olds (mean, 91.1%) succeeded with the body-language cue. In the pitch-cue group, 3 2-year-olds and one 3-year-old are necessarily excluded from this analysis because they responded with "happy"/ "happy" or "sad"/ "sad" on *every* trial. Of the remaining children, 1/3 2-year-olds (mean, 50.4%), 8/15 3-year-olds (mean, 76.9%), 7/12 4-year-olds (mean, 76.7%), and 9/12 5-year-olds (mean, 88.9%) succeeded with the pitch cue.

Table 7: Success at each age in Experiment 2 after excluding *happy*/*happy* and *sad*/*sad* trials

Age	***Facial/Body-language***	***Pitch***	***(At least 75% of responses correct)***
2	**6 / 6 (100%)**	**1 / 3**	**(33%)**
3	**8 / 10 (80%)**	**8 / 15**	**(53%)**
4	NA	**7 / 12**	**(58%)**
5	NA	**9 / 12**	**(75%)**

Children often produced their own pitch contours, either on the words "happy"/ "sad" during the experiment, or when asked to imitate Puppy at the end of the experiment using the words, "Oh, look at that." To determine whether children could produce the happy/sad pitch contrast themselves, we performed an acoustic analysis of children's productions, comparable to the one performed on the experimenter's speech (reported in **Auditory Stimuli**). Children said the words "happy" and "sad" either as their response on each trial, or during the routine at the start of each trial (see **Method** section for more details); we combined these two response-types in the analysis, since their acoustical properties were very similar. This analysis included only children from the *pitch condition*, since we could relate their own productions to their interpretations of the pitch

contours in the experiment. In response to the experimenter's query, "how did Puppy sound when he was happy/sad?", 12 children imitated the experimenter's pitch contours at the end of the experiment using the words, "Oh, look at that." We also analyzed these imitations. For both analyses, only cases in which the child produced both words/intonations were included, and t-tests were computed on data grouped by child.

Children's productions of "happy" and "sad" during the trials differed on several acoustic dimensions. Productions of "happy" had **higher pitch means** (325.5 Hz) than productions of "sad" (280.8 Hz; $p < .001$), **larger standard deviations of pitch samples** (happy, 60.6 Hz; sad, 42.1 Hz; $p < .05$), **higher pitch maxima** (happy, 420.3 Hz; sad, 363.3 Hz; $p < .001$), and **higher intensities** (happy, 62.00 dB; sad, 60.47 dB; $p < .05$; all paired $t(35) > 2.0$, all tests 2-tailed). The 2 words did not differ significantly in pitch minima or durations.

Impressionistically, children's imitations of Puppy's *happy* and *sad* productions usually matched the experimenter's pitch contours very well, and the acoustic measurements reflect that. Children's imitations of the *happy* pitch had **higher pitch means** (happy, 423.1 Hz; sad, 283.2 Hz), **higher pitch maxima** (happy, 622.4 Hz; sad, 435.5 Hz), **higher pitch minima** (happy, 235.5 Hz; sad, 147.3 Hz), and **higher intensities** (happy, 67.8 dB; sad, 62.7 dB, all paired $t(11) > 3.75$, all $p < .005$, all tests 2-tailed). The 2 pitch contours did not differ significantly in standard deviations of pitch samples or in durations.

Children's ability to produce the *happy* versus *sad* pitch contrast (operationalized as the happy – sad subtraction for each acoustic measurement in turn, averaged across all productions made during the trials, but excluding imitations) was not predicted by age,

success in interpreting pitch contours during the experiment, or their interaction in an analyses of covariance (ANCOVA). Since children who responded verbally in the task tended to be older children, who were also more likely to succeed in the task, there may not have been enough variance in either predictor to find an effect.

The results from Experiment 2 showed improvement in use of the pitch cue with age, similar to what we found in Experiment 1. We again found that even the youngest children succeeded with the body-language cues. Children produced the happy/sad pitch contrast themselves, both in their "happy"/"sad" responses during the experiment and when imitating Puppy with the words, "Oh, look at that."

3.3.3 General Discussion

Children did not consistently interpret *happy*- or *sad*-sounding pitch contours in accordance with the emotions they cue until about age 4. By 5 years, children's interpretations of our stereotyped pitch contours accorded with our own. This late development contrasts with young infants' sensitivity to prosodic cues to stress (Nazzi, Bertoncini, & Mehler, 1998; Jusczyk, Cutler, & Redanz, 1993; Jusczyk & Houston, 1999) and phrase boundaries (Hirsh-Pasek et al., 1987; Mandel, Jusczyk, & Kemler Nelson, 1994), and with early acquisition of lexical-tone categories (Mattock & Burnham, 2006; Harrison, 2000; Hua & Dodd, 2000), which appear to be acquired synchronously with consonants and vowels (at least in some languages; Demuth, 1995). But these developments before the child's first birthday all concern perceptual categorization and generalization within the speech domain, before meaningful (semantic) interpretation of phrases or stress patterns is central—so they do not provide crucial connections between diverse types of phonetic variation and the *meanings* they

convey. For example, infants and young children excel at distinguishing the consonants of their language, but do not reliably infer that a consonantal change in a familiar word yields another, different word, even well into the second year (e.g., Stager & Werker, 1997; Swingley & Aslin, 2007; White & Morgan, 2008). The phonetic categories that compose the language's phonology do not come supplied with rules for their interpretation.

One might expect earlier sensitivity to pitch cues to emotional states, since in some cases they appear to be universal (Bryant & Barrett, 2007) and to evoke an innate response (e.g., Fernald, 1993; Mumme & Fernald, 2003; Kahana-Kalman & Walker-Andrews, 2001). However, we find that, despite early sensitivity to pragmatic functions and emotions cued by prosody in maternal speech, preschoolers have trouble detecting the pitch contours that convey *happy* versus *sad*. There are two explanations for this apparent discrepancy. One is that the *happy* and *sad* contours we tested are not among the set of universal mood-inducing contours (used to attract attention, express approval, prohibit behaviors, and comfort the infant; Fernald, 1992). The contours tested by Fernald (1992) are also produced with the goal of shaping the infant's behavior, rather than expressing the mother's emotional state, so they may be qualitatively different from *happy* and *sad*. Our *sad* contours could theoretically be interpreted by young infants as comforting contours, which also have fairly low mean f0, a narrow f0 range, and an often falling shape (Fernald, 1989). The connection between the stereotypical intonational patterns we used and the emotions *happy* and *sad* may therefore need to be learned from linguistic experience.

Another possible explanation is that *happy* and *sad* contours may be accessible as such in infancy (perhaps by inducing these emotions in infants, rather than actually signaling the talker's internal state), but may lose their iconicity through reinterpretation during language acquisition. This would explain why infants are sensitive to *happy* versus *sad* prosody (e.g., Kahana-Kalman & Walker-Andrews, 2001), but younger preschoolers do not show sensitivity in our task. Loss of iconicity through reinterpretation has been documented in other cases. For example, deaf children initially use pointing gestures much the same way hearing infants do. As they acquire American Sign Language, however—in which pointing is used both pronominally and for other functions—they stop using pointing for first- or second-person reference altogether for several months, then for several weeks actually make reversal errors, such as pointing to their interlocutor to mean "me" (Petitto, 1987). In general, sign-language words are not markedly easier for children to learn when they are more iconic (Orlansky & Bonvillian, 1984; see also Namy, 2008). As children acquire language, they seem to accept the possibility that their earliest (and sometimes the most intuitive) hypotheses may be wrong. These types of reinterpretations, which lead to a U-shaped developmental function in children's performance, do not imply regression or loss of ability, but instead reflect a fundamental change in the way children are processing their input (Werker, Hall, & Fais, 2004).

Regardless of whether meaningful interpretation of *happy* and *sad* contours occurs in early infancy, we still need to account for the consistently late understanding of these contours in our task and in previous conflict tasks (Friend, 2000, 2003; Friend & Bryant, 2000; Morton & Trehub, 2001). Children in our task were not puzzled by the

semantic categories of *happy* and *sad*, readily linking them to nonlinguistic behavioral manifestations like joyful dancing or distraught slumping. Furthermore, intuition suggests that children are not deprived of real-world experience with joyful and sad emotions and their vocal expressions, which seem to be on abundant display in daycares and playgrounds. Most likely, children's late learning of connections between the modeled intonational types and their associated emotions is due to the complexity of pitch-contour patterning in the language as a whole.

Intonation functions at both the paralinguistic and phonological levels in English (Ladd, 2008, section 1.4; Scherer, Ladd, & Silverman, 1984; Ladd, Silverman, Tolkmitt, Bergmann, & Scherer, 1985), which may make discovering the connections between specific intonational patterns and conveyed emotions more difficult for children. In addition, the prototypical intonational patterns for *happy* and *sad* are not produced every time someone feels happiness or sadness. Elated joy and quiet happiness have very different vocal signatures, for example (Scherer, Johnstone, & Klasmeyer, 2003), though both could be described as expressions of happiness. The converse can also be true: emotions that are very distinct semantically can have similar pitch characteristics. Happiness, anger, and fear, for example, are all often characterized by elevated pitch (though other pitch characteristics like pitch range and pitch contour may help differentiate these emotional expressions). These factors likely reduce the cue validity of these pitch patterns in speech, making them harder to learn. Pitch cues to emotions also typically occur in combination with facial cues, so children may not be used to interpreting pitch cues in the absence of facial information (though this is less of an issue for vocal than for facial cues, since children do frequently hear voices when they cannot

see the person's face; Baldwin & Moses, 1996). As Walker-Andrews and Lennon (1991) point out, during the intermediate state in the progression from "featurally based discrimination" to "meaningful discrimination," children may require the presence of a face in order to interpret vocal expressions of emotions. Our findings suggest that meaningful discrimination of vocal-only displays may not fully develop until age 4 or 5.

We have found that children have surprising difficulty interpreting a pitch cue to the speaker's emotions, despite the well-attested early accessibility of pitch cues at other levels of structure. This is consistent with Fernald's (1992) suggestion that different functions of pitch in language are accessed by the child at different points depending on their developmental relevance, and—we would add—the cue validity in the signal. The present findings emphasize the importance of considering not just the perceptual availability of a particular acoustic dimension (like pitch, or vowel duration; see Dietrich, Swingley, & Werker, 2007), but the cue validity and developmental relevance of each *particular cue* being conveyed by that dimension.

Chapter 4: Bunny? Banana? Late Development of Sensitivity to the Pitch Cue to Lexical Stress

Carolyn Quam[7]

Abstract

We tested adults' and preschoolers' sensitivity to an isolated pitch cue to the lexical-stress contrast between "bunny" and "banana." Though infants respond to pitch characteristics of infant-directed speech (Fernald, 1993), children lack explicit access to how pitch conveys emotion in speech until age four (Quam, Swingley, & Park, 2009). The multiple linguistic functions of pitch and the complex, interacting nature of their phonetic realization may impede acquisition of pitch cues to emotions and to other linguistic structure. Here, we isolated the pitch cue to lexical stress using Praat *Pitch Resynthesis.* Adults and 2.5–5-year-olds saw pictures of a bunny and a banana, and heard versions of "bunny" and "banana" in which stress was marked using only pitch. Adults (N=32) fixated the target picture longer in response to correctly versus incorrectly stressed items, but children (N = 96) showed less robust sensitivity to the stress changes. The pitch cue to stress may be acquired late because of relatively low cue validity; pitch conveys stress in combination with vowel-reduction, duration, and relative-amplitude cues, and pitch also helps to mark vowel identity, prosodic boundaries, and pragmatic aspects of speaker intention.

[7] This study has been presented by Quam and Swingley (2010b); see **References** for details.

4.1 Introduction

All of the world's languages use pitch to organize the speech stream, but the particular ways in which languages exploit pitch for suprasegmental and contrastive functions differ. At the suprasegmental level, most languages use pitch to mark phrase boundaries, differentiate yes/no questions from statements, and convey emotions and intentions, among other functions (Ladd, 2008; Gussenhoven, 2004). Languages are classified into three types depending on how they use pitch at the word level. Lexical-tone languages use pitch to contrast words; for example, in Mandarin the syllable *ma* can mean 'horse,' 'mother,' 'hemp,' or 'to scold' depending on which of the four tones it is paired with (McCawley, 1978). In pitch-accent (or "accentual") languages, pitch is also part of the representation and realization of words, but these languages have a small inventory of tone melodies (one or two), usually only one tone per morpheme or word is allowed (Yip, 2002, Chapter 9; Cutler & Otake, 1999), and the domain of a pitch-accent pattern (e.g., HL) is usually multisyllabic (Cutler, Dahan, & Donselaar, 1997), whereas tone languages can specify a distinct tone for each syllable (though some argue that accentual languages are a subset of tone languages, e.g., Yip, 2002, Chapter 9). The inventory of tones in a lexical-tone language can be much more complex and can include contour tones, in which the pitch changes within a syllable. Cantonese, a fairly extreme example, has seven tones, at least three of which are contour tones (Yip, 2002, p. 175). Though the same acoustic dimension, fundamental frequency (perceived as pitch), is used to mark both contrastive and suprasegmental categories, these pitch categories appear to be processed very differently in the brain. Mandarin speakers process Mandarin tones in their left hemisphere—which typically processes language—but process intonational

pitch contours in both hemispheres, with a right-hemisphere bias (Gandour et al., 2003). English speakers process the tones bilaterally, suggesting that Mandarin speakers' left-hemisphere bias emerges from experience with speech.

Finally, in lexical-stress languages like English and German, prominence is indicated on one syllable (and secondary stress can be indicated on additional syllables) through a combination of pitch, duration, and amplitude, rather than through pitch alone (Fry, 1958; Lieberman, 1960). The precise use and weighting of these cues differs across languages; for example, Italian relies more heavily on duration than on the other cues (Bertinetto, 1980), while the Mayan language K'etchi appears not to use duration at all (Berinstein, 1979). English has been argued to rely more heavily on pitch than on intensity or duration (e.g., Morton & Jassem, 1965).

4.1.1 Lexical Stress in English

The lexical-stress system of English differs from those of other languages in two primary ways. First, English has variable stress placement, meaning that primary stress can occur on any syllable of a word (Peperkamp, 2004). However, words in English, especially nouns, usually have first-syllable (trochaic) stress. Variable stress placement allows for minimal stress pairs, which, though fairly rare in English, contrast noun/verb pairs like 'record/re'cord. Minimal stress pairs are more frequent in Spanish, largely because of verb conjugations like 'hablo (I speak) and ha'blo (he spoke; Hochberg, 1988). Other languages have fixed stress or accent; accent in French, for example, always falls on the final syllable of the word or phrase (con'cept; concep'tuel; conceptuali'ser, etc.; Peperkamp, 2004). Compared with fixed-stress languages, the English stress system, in which stress can occur on any syllable, is relatively complex (Peperkamp, 2004).

The second distinctive characteristic of stress in English is its correlation with vowel reduction. English is fairly typical in its use of longer durations, higher amplitudes, and pitch targets (high or low, depending on the intonational context; Hayes, 1995) to indicate the stressed syllable, though languages do vary in the weighting of these cues (e.g., in the Mayan language K'etchi, the presence of contrastive vowel duration apparently blocked the use of duration as a cue to lexical stress; Berinstein, 1979). English stress is also highly correlated with the amount of vowel reduction (Cutler & Clifton, 1984; Fear, Cutler, & Butterfield, 1995). In the 'record/re'cord example, the vowel in the unstressed second syllable of 'record is reduced to schwa, while the vowel in the unstressed first syllable of re'cord is similarly reduced. This means that the two words actually differ in their segmental content as a result of their differing stress. The differences across languages in stress placement (fixed or variable) and cues to stress (varying combinations and relative weighting of duration, amplitude, pitch, and vowel reduction) mean that children must learn the particulars of their language's stress system from experience with speech.

4.1.2 Phonological Acquisition: Evidence of Early Sensitivity to Rhythmic and Lexical Stress

Research on children's learning of sound categories from speech has focused primarily on acquisition of consonant and vowel categories. Infants begin life able to discriminate many of the sound contrasts that occur in the world's languages; through exposure to the native language, they fine-tune and reshape those original categories to match the categories they are hearing (e.g., Jusczyk, 1993; Nittrauer, 2002). This process leads to a loss of discrimination for some nonnative contrasts (by 4-6 months for vowels,

Polka & Werker, 1994; Bosch & Sebastián-Gallés, 2003; and by 10-12 months for consonants, Werker & Tees, 1984) and, in some cases, an improvement in discrimination of acoustically difficult native contrasts (Kuhl et al., 2006; Kuhl, Conboy, Padden, Nelson, & Pruitt, 2005; Narayan, Werker, & Beddor, 2010). Acquisition of lexical-tone contrasts appears to parallel that of consonants and vowels (Mattock & Burnham, 2006).

Children also acquire a basic knowledge of the rhythmic properties of their native language in the first year, but this knowledge must be supplemented over the subsequent months and years by more sophisticated phonological learning. Newborns can discriminate disyllabic stress patterns (Sansavini, Bertoncini, & Giovanelli, 1997, with Italian newborns; see also Jusczyk & Thompson, 1978, with English-learning 2-month-olds) and can discriminate foreign languages based on their rhythmic class (e.g., mora-timed Japanese versus stress-timed English versus syllable-timed Italian; Nazzi, Bertoncini, & Mehler, 1998). By 4 to 5 months, infants already display a processing advantage for their native-language stress patterns (initial stress for German and final stress for French; Friederici, Friedrich & Christophe, 2007, using event-related potentials), and can discriminate their native language from foreign languages from the same rhythmic class (Nazzi, Jusczyk, & Johnson, 2000). By 7 months, infants learning English prefer to listen to trochaic words (Curtin, Mintz, & Christiansen, 2005; see also Jusczyk, Cutler, & Redanz, 1993, who found a similar preference at 9 months but not at 6 months).

English-learning infants' trochaic bias allows them to segment strong-weak (SW) words from the speech stream, and even to segment SWS words, as long as the primary stress is on the first syllable (Houston, Santelmann, & Jusczyk, 2004). Children initially

overweight their trochaic preference in word segmentation, however; 7-month-olds cannot segment words with primary stress on the third syllable (Houston et al., 2004), and, when familiarized to a speech stream that includes "guitar is," they segment the trochaic sequence "taris" rather than the iambic word "guitar." At 8 months, children's trochaic bias outweighs their attention to transitional probabilities between syllables (Johnson & Jusczyk, 2001; but not at 7 months; Thiessen & Saffran, 2003) and phonotactic probabilities (Mattys, Jusczyk, Luce, & Morgan) in word segmentation. By contrast, 10.5-month-olds can integrate their expectations for stress with other information, like phonotactics and conditional probabilities, to correctly segment iambic words (Jusczyk, Houston, & Newsome, 1999). Still, even 11-month-olds still rely on stress information over transitional probabilities for word segmentation (Johnson & Seidl, 2009). At 9 months, infants will even treat a single cue to stress—spectral tilt, or the distribution of energy across frequencies—as more important for word segmentation than statistical information, though slightly older infants and adults do not (Thiessen & Saffran, 2004).

Infants' word segmentation relies on other prosodic cues as well, which get integrated with the trochaic bias over development. The intonation contours of infant-directed speech appear to help infants segment words (Thiessen, Hill, & Saffran, 2005). Four-month-olds identify clause boundaries by relying equally on pause-, pitch-, and vowel-duration cues, perhaps because they are processing clause boundaries holistically rather than attending to individual cues (Seidl & Cristià, 2008). By 6 months, infants rely primarily on pitch, though they still require convergence from the other two cues (Seidl, 2007). Six-month-olds prefer to listen to word sequences that are prosodically marked as

a noun or verb phrase than to the same words presented as a syntactic nonunit (Soderstrom, Seidl, Kemler Nelson, & Jusczyk, 2003), and prefer word sequences that are prosodically cohesive to those that straddle a clause boundary (Soderstrom, Kemler Nelson, & Jusczyk, 2005). Nine-month-olds prefer to listen to speech in which pauses coincide with pitch and durational cues to linguistic boundaries (e.g., between the lexical noun phrase and the verb; Gerken, Jusczyk, & Mandel, 1994), and they prefer that pauses occur before strong-weak units rather than in between the strong and weak syllables (Echols, Crowhurst, & Childers, 1997). By 10 months, infants' word recognition is disrupted if a phrase boundary divides the word in half, even when the syllables form a trochee, and they improve at integrating phrase-boundary information in word segmentation between 10 and 13 months (Gout, Christophe, & Morgan, 2004).

Whether infants develop a preference for iambic or trochaic words depends on the role of stress in the particular language they are learning. Infants learning English (Curtin, Mintz, & Christiansen, 2005) and German (Höhle, Bijeljac-Babic, Herold, Weissenborn, & Nazzi, 2009) develop a preference for trochaic words by 7–9 months, because trochaic words are highly frequent in both languages. By contrast, infants learning French (Höhle et al., 2009), Catalan, and Spanish (Pons & Bosch, 2007) can discriminate trochaic versus iambic stress patterns, but show no listening preference. French, Catalan, and Spanish are all syllable timed rather than stress timed, making stress groupings less important for word segmentation (Pons & Bosch, 2007; Höhle et al., 2009). Additionally, Spanish and Catalan have only a weak iambic tendency, compared with the strong trochaic tendencies of English and German (Pons & Bosch, 2007), and, though French words always end in an accented syllable, French accent is acoustically

weaker than stress is in German and English (Höhle et al., 2009). The strength of the iambic or trochaic tendency in the native language, and the importance of stress groupings for word segmentation, therefore appear to impact whether infants will show a listening preference. Even short-term experience can affect infants' segmentation strategies; English-learning infants can be trained to segment iambic words (Thiessen & Saffran, 2007), and 7-month-olds, who typically weight statistical information over stress information in word segmentation (Thiessen & Saffran, 2003) can be trained to rely on stress (Thiessen & Saffran, 2007).

Though the work of Pons & Bosch (2007) and Höhle et al. (2009) suggests that infants who show no listening preference for trochaic or iambic words can still discriminate stress contrasts, other work suggests that the role and complexity of stress in the native language can impact even discrimination. Skoruppa et al. (2009) found that French- and Spanish-learning 9-month-olds could discriminate iambic versus trochaic pronunciations at the acoustic level—after familiarization with a single word, either with an iambic or a trochaic stress pattern—but only Spanish learners succeeded after familiarization with *eight* different words with the same stress pattern. Multiple words required infants to abstract the stress pattern over phonetic variability, putatively requiring phonological knowledge of stress. Dupoux and colleagues found comparable effects with French- (Dupoux, Pallier, Sebastián-Gallés, & Mehler, 1997; Dupoux, Peperkamp, & Sebastián-Gallés, 2001), Finnish-, and Hungarian-speaking adults (Peperkamp & Dupoux, 2002), suggesting that the "stress deafness" effect caused by learning a language without contrastive stress persists into adulthood.

4.1.3 Production Evidence for the Trochaic Bias

English-learning children's production of lexical stress does not become adult-like until age 3; like their perception, younger children's productions reflect a strong influence of the trochaic bias, which must be tempered over development. Both children and adults produce more segmental errors and more motor variability for weak syllables than for strong syllables (Goffman, Gerken, & Lucchesi, 2007). Children learning English often omit unstressed function words (Gerken, Landau, & Remez, 1990) and unstressed initial syllables of nouns (Carter & Gerken, 2003, 2004; Goffman et al., 2007) in imitation tasks. Children's omissions often leave acoustic traces of the unstressed syllable, however. Two-year-olds produce a longer duration between the previous syllable and the onset of the stressed syllable when imitating "Lucinda" as "Cinda" than when imitating "Cindy." This suggests that omissions do not occur because of a failure to hear unstressed syllables, nor because children's phonology deletes the syllables completely (Carter & Gerken 2003, 2004).

Children are more successful at producing "irregular" stress (iambic words, for English speakers) in spontaneous speech (Hochberg, 1988) than in imitations. During natural interactions with their mothers, English learning 13- to 20-month-olds produce roughly equal numbers of iambic and trochaic phrases (Vihman, DePaolis, & Davis, 1998). Though English words are typically trochaic, English *phrases* are typically iambic (e.g., *with light*; Vihman et al., 1998). The presence of iambic phrases in English learners' speech matched the prevalence of iambic phrases in their mothers' speech (e.g., *a ball*). Vihman et al. (1998) argue that these iambic phrases in toddlers' speech (e.g., [ʔə'βɪ] for *a bead*) are unanalyzed (into article and noun) in early productions, and that

children begin omitting the unstressed article once they analyze it separately from the noun.

Children's productions also reflect development in the weighting of acoustic cues to stress. Kehoe, Stoel-Gammon, & Buder (1995) found that 18- to 30-month-olds produced correct stress placement about 70% of the time, using pitch, duration, and intensity to mark stress. The pitch cue was produced most reliably across age, though Kehoe et al. (1995) found large individual variability in which cues were produced. Pollock, Brammer, & Hagerman (1993) found that while 3- and 4-year-olds reliably produced all 3 acoustic cues to stress, 2-year-olds produced only duration, and were judged to produce correct stress placement only 55% of the time. However, these findings might underestimate 2-year-olds' abilities because the task used, imitation of nonwords, is less sensitive than spontaneous speech tasks (Hochberg, 1988; Klein, 1984). Finally, the complexity of structure in the input language can determine the age of acquisition of rhythmic properties of language. In French, lengthening of the final syllable of a phrase complements phrase-final accent placement, whereas in English, final lengthening goes against the trochaic-stress bias. Vihman, DePaolis, and Davis (1998) found that French-learning 13- to 20-month-olds produced clear final lengthening, while English learners did not consistently lengthen phrase-final syllables.

4.1.4 Lexical Stress in Infants' Word Representations

Peperkamp and Dupoux (2002; Peperkamp, 2004) argue that the transparency of the stress system *before* infants have access to word boundaries determines whether stress will be encoded in word representations, and therefore whether infants and adults will be able to discriminate stress contrasts. In other words, infants will encode stress, even if it

is noncontrastive, if they *fail to observe* its predictability before they begin learning words (by contrast, Mehler, Dupoux, & Segui, 1990, claim that infants store words using *only* the dimensions that are contrastive in their native language). This would explain why adult speakers of Polish, which lacks contrastive stress, are nonetheless able to discriminate stress patterns (Peperkamp and Dupoux, 2002), unlike speakers of French, Finnish, and Hungarian.

Peperkamp (2004) argues that before the onset of word segmentation (around 7.5 months; Jusczyk & Aslin, 1995), infants have access to stress patterns only at *utterance* boundaries. In Polish, stress is penultimate, but this structure is not detectable from utterance boundaries; because utterances ending in a monosyllabic content word have final stress, stress assignment appears irregular or nonphonological. This lack of transparency may lead Polish infants to store stress information in word representations, which would explain why Polish adults successfully discriminate stress (Peperkamp 2004; Peperkamp & Dupoux, 2002).

Since English has variable stress placement—i.e., stress is not assigned in a purely phonological way—stress in English needs to be stored in word representations under Peperkamp's (2004) model. This would allow the disambiguation of minimal stress pairs in English, and would probably also improve word segmentation. In a corpus analysis of child-directed speech, Curtin, Mintz, and Christiansen (2005) found that representing stressed and unstressed syllables as distinct types makes cues to some word boundaries clearer.

There is some evidence that English-learning infants store stress information in their word representations. Curtin (in press) found that 12-month-olds could learn two

novel words differing only in their stress pattern, suggesting that English-learning infants specify stress in their representations of newly learned words. Curtin, Mintz, and Christiansen (2005) familiarized 7-month-old English learners to a speech string in which every third syllable was stressed, so that infants might posit trisyllabic words with first-syllable stress (e.g., *'dobita*). In test, infants were presented with neutrally stressed words that matched strings that had contained first syllable stress during familiarization; words that matched strings with medial stress; words that matched strings with final stress; or control words that did not match the familiarization. Infants looked equally in response to the control words, the medially stressed words, and the finally stressed words, but looked significantly less to initially stressed words, suggesting they recognized the SWW words as familiar and exhibited a novelty preference for the control stimuli.

Vihman, Nakai, DePaolis, & Halle (2004) found only weak evidence that stress information is included in infants' word representations of familiar words. When familiar words were misstressed, 11-month-olds listened just as long as they did to correctly stressed familiar words, and they also listened longer to misstressed familiar words than they did to rare words; both these findings suggest that misstressings did not impair infants' word recognition. Infants' word identification was delayed as a result of the misstressing. Infants also responded to segmental mispronunciations more reliably when they occurred at the onset of stressed syllables versus unstressed syllables, but Vihman et al. (2004) point out that this could be because the unstressed syllable provides a "weaker acoustic signal," not because it is segmentally underrepresented compared with the stressed syllable.

4.1.5 Lexical Stress in Adults' Word Representations

There is some evidence that stress information is stored in adults' word representations and exploited during word recognition and retrieval. Brown and McNeill (1966) found that adults in the "tip-of-the-tongue" state, just before recalling a word, often remembered the location of primary stress. They argued that stress was one of the more easily retrieved features of a word, and might be "one of the features to which we chiefly attend in word-perception." Connine, Clifton, and Cutler (1987) found that stress affected adults' judgments of an ambiguous phoneme. Connine et al. (1987) selected word pairs that differed both in their stress pattern and initial consonant (e.g., /di'gress/-/'tigress/), and then synthesized the initial consonant to make its voicing ambiguous. The stress pattern participants heard affected their judgments of the initial consonant; they were more likely to report the consonant as 'd,' for example, if that would create a real-word given the stress pattern (as in di'gress). Finally, in a gating task, Lindfield, Wingfield, & Goodglass (1999) demonstrated that adults use prosodic information to constrain word identification. Participants heard increasing portions of target words starting with the first 50 milliseconds and then increasing by 50-millisecond intervals until they identified the word; in some conditions, partial information about the rest of the word was also included. Adults identified words more quickly when the rest of the word was low-pass filtered, preserving syllable and stress information, than when duration alone was signaled using white noise.

Though the above findings suggest an important role for stress in word recall and recognition, other work has found very mixed evidence for use of stress information in lexical access. Cutler and colleagues have found no priming between minimal stress pairs

in Dutch (Cutler & Donselaar, 2001), but have found priming in English (Cutler, 1986), suggesting that minimal stress pairs are treated as more distinct by Dutch listeners than by English listeners, who treat them as homophones. Priming tasks have also failed to demonstrate effects of lexical stress in English (Cutler, 1986; Slowiaczek, Soltano, & Bernstein, 2006; but see Cooper, Cutler, & Wales, 2002, with word-fragment primes), leading Slowiaczek, Soltano, and Bernstein (2006) to conclude that "the influence of metrical stress lies in pre-lexical segmentation and early accessing of information from lexical memory based on the phonetics of the stimulus items. Evidence does not exist to support a lexical architecture based on stress information."

More recently, however, Cooper, Cutler, and Wales (2002) demonstrated sensitivity to stress during English word recognition, though English-speakers showed weaker effects than speakers of Spanish and Dutch. For Spanish (Soto-Faraco, Sebastián-Gallés, & Cutler, 2001) and Dutch (Donselaar, Koster, & Cutler, 2005), stress-mismatched prime words actually *slow* identification of words relative to unrelated primes. Similar effects have been found for accentual mismatches in Japanese (Cutler & Otake, 1999; Sekiguchi & Nakajima, 1999). Cooper et al. (2002) did not find a comparable inhibition effect for English, but they did find some sensitivity to stress: bisyllabic prime words with mismatched stress did not facilitate identification of target words. They still found facilitation for monosyllabic mismatched primes, however. Donselaar et al. (2005) failed to find inhibition effects for monosyllabic primes, so in both English and Dutch, stress appears to be better encoded or exploited for bisyllables (Soto-Faraco et al., 2001, did not include monosyllables, so it is unclear whether this is also true of Spanish).

Cutler and Clifton (1985) found that incorrectly stressing words interfered with word recognition in English, but only when the stress change was accompanied by vowel reduction (see also Fear, Cutler, & Butterfield, 1995). Accordingly, Cutler & Norris (1988; Cutler, 1990) proposed that adults use a metrical segmentation strategy to detect word boundaries in speech, using metrically stressed syllables—those with unreduced vowels—to infer the beginnings of words. They incorporated the metrical segmentation strategy into Shortlist, a connectionist model of word recognition (Norris, 1994; McQueen, Norris, & Cutler, 1994; Norris, McQueen, & Cutler, 1995). In its original form, Shortlist incorporated suprasegmental information only at the prelexical level of processing (e.g., for segmentation), not at the lexical level. Given findings that stress constrains lexical access in Spanish (Soto-Faraco, Sebastián-Gallés, & Cutler, 2001) and Dutch (Cutler & Donselaar, 2001), however, McQueen, Cutler, and Norris (2003) argued that suprasegmental information should also be represented at the lexical level, at least for these languages. Since stress information also constrains lexical access in English (Cooper, Cutler, & Wales, 2002), albeit more weakly than in Dutch and Spanish, Cutler and colleagues would probably now advocate incorporating stress information at the lexical level for English as well.

4.1.6 Learning to Interpret Phonological Variation

The fact that infants are sensitive to the rhythmic properties of their native language does not tell us about their interpretations either of prosodic structure at the lexical level or of interactions between different levels of prosodic structure, e.g., between lexical stress and sentence intonation. The early trochaic preference exhibited by English learners does not represent adult-like knowledge; children must overcome their

overreliance on the trochaic bias both to produce and to recognize iambic words. Though the evidence from Curtin and colleagues suggests that 7- and 12-month-olds store stress information in their representations of new words, we do not know how detailed this stress information is, whether it incorporates all 4 cues (duration, amplitude, pitch, and vowel quality), and whether children can exploit it rapidly during word recognition.

Despite the early sensitivity to rhythmic structure discussed above, and despite infants' early sensitivity to intonation (Fernald & Kuhl, 1987; Katz, Cohn, & Moore, 1996), children struggle to exploit more complex prosodic structure until the preschool years. Children do not exploit sentence prominence before age 5 or 6, failing to show a facilitation effect for accented words in a word-monitoring task the way adults do (Cutler & Swinney, 1987). In exploiting contrastive focus to identify the referent of sentences, Russian-learning 5- to 6-year-olds rely less on prosodic cues to contrastive focus than on syntactic cues, likely because the syntactic construction is a deterministic cue, while the prosodic cues are probabilistic (Sekerina & Trueswell, in press). Cutler and Swinney (1987) conclude that children under 6 are "poor at exploiting prosodic information in language comprehension."

Understanding prosodic cues to the speaker's emotions follows a similarly late trajectory. Though young infants respond to pitch characteristics of infant-directed speech (Fernald, 1993), children lack explicit access to how pitch conveys emotion in speech until age 4 (Friend, 2000; Morton & Trehub, 2001; Quam, Swingley, & Park, 2009). Finally, children appear to learn that English words cannot have tones sometime between 18 (Quam & Swingley, in progress) and 30 months (Quam & Swingley, 2010), as their interpretations become constrained by their linguistic input. These late

trajectories for interpretation of complex prosodic structure suggest that, despite infants' early sensitivity to rhythmic properties of speech, a complete understanding of lexical stress in English may take several years to develop.

4.1.7 Challenges for Acquiring the Pitch Cue to English Stress

Several aspects of stress in English may make the pitch cue to stress more difficult to learn. First, unlike in a lexical-tone system, pitch does not contrast words on its own in English, but combines with three other probabilistic cues to indicate the stressed syllable (Fry, 1958). The precise reliance on each of these cues differs somewhat across languages, so children must learn which cues to exploit, weight them properly, and then integrate them rapidly to recognize stressed syllables in speech.

Second, the pitch pattern of a stressed syllable is potentially ambiguous between indicating lexical stress versus other pitch categories. Pitch serves many other functions in English, including cuing phrase boundaries, yes/no questions, and the speaker's emotions (Ladd, 2008; Gussenhoven, 2004). Because of the many functions of pitch, a pitch peak is ambiguous between indicating a stressed syllable, the speaker's excitement, contrastive stress (e.g., "not the red one, the BLUE one") or a lexical tone (if the child has not yet ruled out tone as a possibility), which presents an interpretive challenge. In addition to sorting out variability from linguistically relevant sources, the child must also learn to disregard changes in pitch due to the talker's voice, perturbations from consonants, intrinsic vowel height, etc., and must disregard pitch-contour changes to recognize words across tokens (Quam & Swingley, 2010).

Third, the pitch cue to lexical stress interacts with other pitch structure, affecting the phonetic realization of the pitch cue to stress and likely complicating children's

ability to store stress information in word representations. Most notably, the underlying stress value of a syllable (stressed or unstressed) interacts with sentence context to determine the phonetic realization of the pitch target (Hayes, 1995). Typically, a stressed syllable in a declarative sentence context has a high pitch target, while one in a yes/no question context has a low target. Because of this variability, Yip (2002; p. 256) argues that pitch "is not lexically specified" in stress languages, compared with tone languages, in which "tones are crucially part of the lexical representation."

Thus, the need to combine the pitch cue with three other probabilistic cues, ambiguity in how a pitch target should be attributed, and variability in the realization of the pitch cue likely make the stress system more difficult for English-learners to acquire (see Peperkamp, 2004, Demuth, 1995, and Vihman, DePaolis, & Davis, 1998, for examples of phonetic variability slowing acquisition of phonological structure; see also Cohn, submitted, for discussion). Given the complexity and cross-linguistic differences in cue combination, interactions of each cue with the context (Hayes, 1995), and the variable-stress system of English (Peperkamp, 2004), we might expect relatively late acquisition of the lexical-stress system of English (especially compared with simply detecting the trochaic tendency of words in English).

A final reason why lexical stress acquisition might be delayed in English concerns its functional load. English words have relatively complex segmental structure; English allows consonant clusters in both syllable onsets and codas, and words are often polysyllabic. By contrast, Mandarin syllables are more constrained; consonants are optional, and consonant clusters are not allowed (Huang, 1992), so "homophonous morphemes are quite numerous in Mandarin because the constraints on the combination

of segments in the syllable admit only about 400 different segmental syllables" (Howie, 1976). As a result, lexical tone in Mandarin bears a much higher functional load than does lexical stress in English; tones in Mandarin frequently differentiate words, while there are few minimal stress pairs in English (Cooper, Cutler, & Wales, 2002; Cutler, Dahan, & Donselaar, 1997). Accent in Japanese also contrasts many more words than does stress in English (Shibata & Shibata, 1990; in Sekiguchi & Nakajima, 1999). The lower importance of lexical stress for contrast in English compared with the crucial importance of tone and accent in many languages (or even compared with the functional load of stress in Spanish and Dutch, which have more minimal stress pairs; Hochberg, 1988, Cooper et al., 2002) might also slow the acquisition of English lexical stress compared with other pitch categories. Still, even lexical-tone contrasts are processed less efficiently by adults relative to segmental contrasts (Cutler & Chen, 1997), possibly because they are dependent on vowels for their realization, so listeners appear to first identify the vowel and then the tone (Cutler, Dahan, & Donselaar, 1997).

4.1.8 Goals of the Present Research

After the early perceptual reorganization in which children learn the vowel and consonant categories of their language, children must still learn to interpret perceptible variation in speech. We propose (as does Cohn, submitted) that much of phonological development occurs after the first year, when children must learn to cope with multiple sources of acoustic variation and attribute them to the appropriate levels of linguistic structure. For this reason, we consider children's interpretation of correct and incorrect realizations of the pitch cue to stress for words that they already know ("bunny" and "banana"), and we investigate this question with preschoolers, aged 2 to 5.

Given the complexity of pitch structure in English and of pitch assignment to stressed syllables, we ask whether children and adults know that a pitch peak indicates a stressed syllable. We also ask whether children and adults weigh the pitch cue similarly. The pitch cue to stress may be realized more consistently in child-directed speech, the way segmental (Burnham, Kitamura, & Vollmer-Conna, 2002) and tone (Liu, Tsao, & Kuhl, 2007) categories are exaggerated in speech to infants. If so, children might rely more heavily on the pitch cue than adults do, the way young infants overrely on the trochaic bias (Jusczyk, Houston, & Newsome, 1999). On the other hand, the lack of minimal stress pairs (like 'record/re'cord) in children's vocabularies might make children less reliant on stress in word recognition, because it is not necessary for distinguishing words (Charles-Luce & Luce, 1990; but see Swingley & Aslin, 2002). The timeline by which children begin to integrate lexical stress information in word recognition will inform our understanding of whether minimal pairs are needed for children to exploit a phonetic cue.

We predict that preschoolers can exploit a probabilistic cue like the pitch peak, and that they can differentiate the pitch cue to stress from other functions of pitch and bind it with other cues to stress without having learned minimal stress pairs. However, the complexity of pitch in English and of lexical-stress cuing might slow children's acquisition of the pitch cue, so we predict that, as with the pitch cue to emotions (Quam, Swingley, & Park, 2009), children will not exploit the pitch cue to stress until after infancy. We predict that by age 5, children will exploit the pitch cue to stress in an adult-like way, but we test children beginning at age 2.5 to investigate the developmental trajectory of this ability.

To begin to investigate these questions, we considered preschoolers' and adults' interpretations of an isolated pitch cue to stress. This does not allow us to consider differential weighting of the pitch cue compared with other cues like duration or amplitude; that question must await future research. Because we tested 2-year-olds, we were restricted to the word pair "bunny"/"banana," since these words contrast in the stress of their first syllable and are contained in 2-year-olds' vocabularies. Our experimental paradigm, in which photos of a bunny and a banana appeared and participants heard either "bunny" or "banana," is admittedly a very simplified context, but it crucially allows us to determine whether children know that a pitch peak indicates a stressed syllable.

4.2 Experiment 1

In order to calibrate children's sensitivity to mispronunciations of the pitch cue to stress, we tested adults in roughly the same procedure. Because of the complexity of pitch target assignment to stressed syllables (Hayes, 1995), it is possible that even adults will not exploit the presence or absence of a pitch peak in predicting which word they are hearing. Llisterri, Machuca, de la Mota, Riera, and Rios (2003) found that Spanish-speaking adults did not use the pitch pattern alone to identify stressed syllables. Spanish is also a variable-stress language, and stress arguably has a higher functional load in Spanish than in English because of the greater prevalence of minimal stress pairs (Hochberg, 1988; though Spanish is a syllable-timed rather than a stress-timed language, which might make stress less important; Pons & Bosch, 2007; Höhle et al., 2009). Llisterri et al.'s (2003) null result thus justifies asking whether English-speaking adults

will exploit the pitch cue to stress during word recognition. If adults are sensitive to the mispronunciations, we can compare the magnitude and timing of their decrease in fixation of the target picture to children's responses at each age. This will allow us to determine whether children are more or less sensitive to the pitch cue than adults are, which may give us some insight into how sensitivity to the pitch cue (and to stress itself) changes across development.

4.2.1 Method

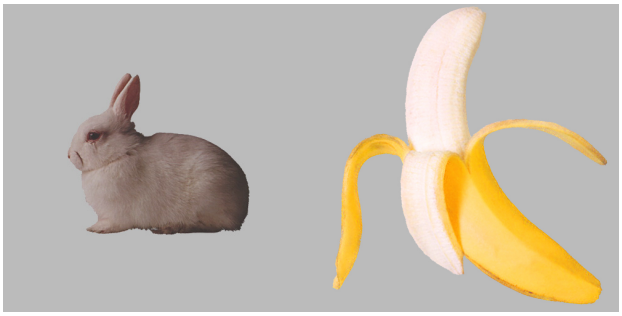
4.2.1.1 Participants

Thirty-two adults, fourteen female, and all native monolingual speakers of English, were included in the analysis. All but two participants were undergraduates or very recent graduates (the two exceptions were affiliated with the Psychology department), assumed to be between 17 and 23 years of age. Seven more participated but were excluded: five for their language backgrounds (they were native bilinguals of Hindi, Chinese, Korean, Russian, and Spanish), one for equipment failure, and one because his glasses interfered with the eyetracking.

4.2.1.2 Apparatus and Procedure

We used a language-guided looking procedure to investigate how adults would interpret a mispronunciation of the pitch cue to lexical stress during recognition of familiar words. Since 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 an infant-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.

For both adults and children, the experiment alternated between experimental (bunny/banana) trials and filler (other familiar-word) trials. There were 16 trials of each of these types, making 32 total trials. In each trial, two pictures appeared on the screen; two seconds later, recorded sentences, referring to one of the two pictures, played from speakers on either side of the screen (see **Figure 13**). 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 "banNAna" trials, and misstressed "BAnana" trials; these four 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. 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 prevent adults from detecting the purpose of the experiment. Because of these extra trials, the adult experiment was about 25 minutes long; much longer than the child version.



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

Figure 13: Example photographs used in all three experiments, with example sentences

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 at the bottom of the computer screen rather than on the participant's head, making the procedure more comfortable than mounted eyetracking, especially for children. According to the manufacturer, the system can accommodate a fair amount of movement by participants (8.7" x 7.1" x 7.9" of head movement, and a range of 15.7" to 27.6" in distance from the screen) without losing accuracy. When the eye-tracker does lose the pupil, the two speakers (embedded on either side of the screen; dimensions of 2" x 10.5") play a low-pitched pulsing noise that is intended to draw the infant's attention back to the screen; no eye-tracking data is recorded while this sound is playing.

Before the experiment, we conducted a procedure to calibrate and validate the eyetracking. First, a round sticker with a black-and-white target symbol printed on it was placed on participants' foreheads 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 divergence between the pupil/CR and the target symbol to compute the location of fixation. Once the target and pupil/CR were identified, the experimenter began the calibration procedure. An expanding and contracting target symbol appeared in the middle of the screen paired with a sound-effect (a boing), then moved to each of the four corners of the screen. The participant was

instructed to "look at the circle." As the participant fixated the target image in each location on the screen, the eye-tracking program calibrated its eye-gaze calculations to the individual participant. These calibrated settings were then validated as the participant fixated the target image again. 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 in between trials.

4.2.1.3 Auditory Stimuli

For experimental trials, we selected the words "bunny" and "banana" for three 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" is wedge (IPA: /ʌ/), which is acoustically very similar to schwa (IPA: /ə/), making it easier for us to neutralize the vowel-reduction contrast between stressed and unstressed first syllables. Finally, "bunny" and "banana" are the only word pair in most 2-year-olds' vocabularies that fit both these criteria.[8]

In English, four different acoustic cues—amplitude, duration, amount of vowel reduction, and location of the pitch target—jointly indicate the location of the stressed syllable. In order to test English-speakers' knowledge of the pitch cue in particular, we isolated the pitch cue to the stress contrast between "bunny" and "banana", controlling the other three cues. 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

[8] In pilot testing, we tried the pair "button"/"balloon", but the L-coloring of the first vowel in "balloon" eliminated any ambiguity between the first syllables of the words.

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 a much earlier pitch peak than a word with a stressed second syllable.

We created two versions of each word that differed only in the location of their pitch peaks. We first recorded tokens of each word with stress on the first or second syllable (attempting to neutralize duration differences between the first and second syllables) and a "neutrally stressed" version of each word (in which we attempted to neutralize amplitude, duration, and vowel quality). We then superimposed the pitch contour from each of the stressed versions onto the neutrally stressed token, using Praat *Pitch Resynthesis* (Boersma & Weenick, 2008).

In pilot testing, we noticed that for both "bunny" and "banana," despite our efforts to equalize amplitude between the first and second syllables, the first syllable was higher in amplitude. This was a concern because it might make the misstressing of "banana" more noticeable than the misstressing of "bunny" (since for "bunny" the higher amplitude of the first syllable would be consistent with its trochaic stress pattern). Using the Goldwave program's *Change Volume* feature, we manipulated the amplitude of particular regions in each word to make the first and second syllables more comparable. We then normalized the mean amplitude of each word to 70 decibels in Praat. Waveforms and pitch tracks for the resulting stimuli are shown in **Figure 14**.

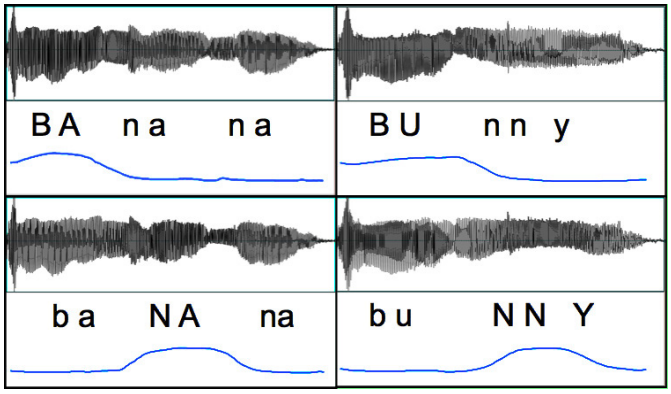


Figure 14: Waveforms of each of the words used in Experiments 1 and 2

4.2.1.4 Visual Stimuli

Visual stimuli were color photographs placed on gray backgrounds. There were two different *banana* photos and two *bunny* photos (see examples in **Figure 13**), as well as two versions of each of the filler pictures. The pictures were equated for size and salience. 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 ameliorated the baseline (before target-word) preference for the *bunny* object, but children still preferred the *bunny* photos, as discussed in the **Experiment 2 Results**.

4.2.1.5 Analysis

The output files of the EyeLink system (EDF format) were converted to ASC format, processed with custom JAVA-based software and Python scripts, and then imported to Excel. The result of this processing was information about which picture each participant was fixating (target, distracter, or "other") at each time point (with 20-millisecond resolution) in each trial.

4.2.2 Results and Discussion

Adults provided two types of responses: looking times to each picture over time and questionnaire responses after the experiment. Looking times provide a gradient measure of interpretation of the auditory stimulus, while questionnaire responses allow us to determine participants' post-hoc impressions of the stimuli. We first consider participants' looking responses, and then relate them to their questionnaire responses.

4.2.2.1 Looking-time measures

Figure 15 plots participants' fixation of the target picture over time in each of the four experimental conditions. The onset of the target word is 0 milliseconds (ms) on the x-axis, and the ambiguous region, "bun," ends at 610 ms in all conditions (the first syllable ends at roughly 450 ms; see **Table 8** for details). For both words, target fixation is higher in correctly stressed trials than in misstressed trials. In **Figure 16,** we further split the data by which object participants happened to be fixating at target onset (or within 60 ms of target onset). The graphs were generated by plotting distracter-initial fixations straightforwardly, and subtracting target-initial fixations from 1. In this type of plot, responses to correct pronunciations typically show the following pattern: distracter-initial fixations begin at 0% and rise to 100%, while target-initial fixations begin at 0% and remain close to 0%. These plots show effects of the misstressings for both target- and distracter-initial trials, for both "bunny" and "banana" trials. Compared with correctly stressed trials, participants in misstressed trials who began on the distracter object were slower to move to the target picture, while participants who began on the target were more likely to move their eyes away to the distracter picture.

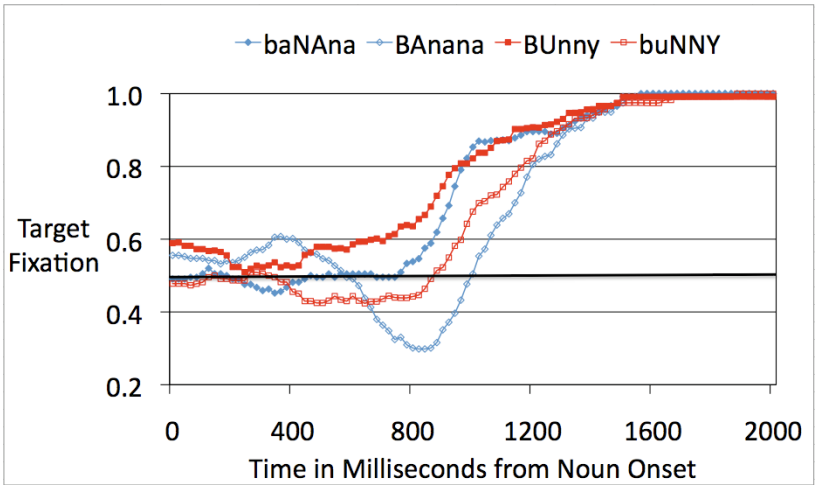


Figure 15: Adults' fixation of the target picture over time in each condition in Experiment 1
The x-axis is time in ms; 0 is target-word onset, and the ambiguous region "bun" ends at 610 ms in all conditions. The y-axis is the proportion of trials in which participants are fixating the target object.

Table 8: Acoustic measurements for the *first syllable* of each target word used in Experiments 1 and 2

	Pitch mean (SD)	Pitch max	Intensity	Duration	F1/F2
baNAna	200.3 Hz (2.3 Hz)	207.4 Hz	70.6 dB	0.49 sec	805.4/1452.4 Hz
BAnana	390.0 Hz (27.0 Hz)	420.5 Hz	72.1 dB	0.49 sec	803.3/1593.7 Hz
buNNY	200.6 Hz (4.4 Hz)	214.9 Hz	70.0 dB	0.42 sec	739.2/1655.6 Hz
BUnny	367.37 Hz (20.4 Hz)	392.6 Hz	69.9 dB	0.42 sec	882.6/1650.9 Hz

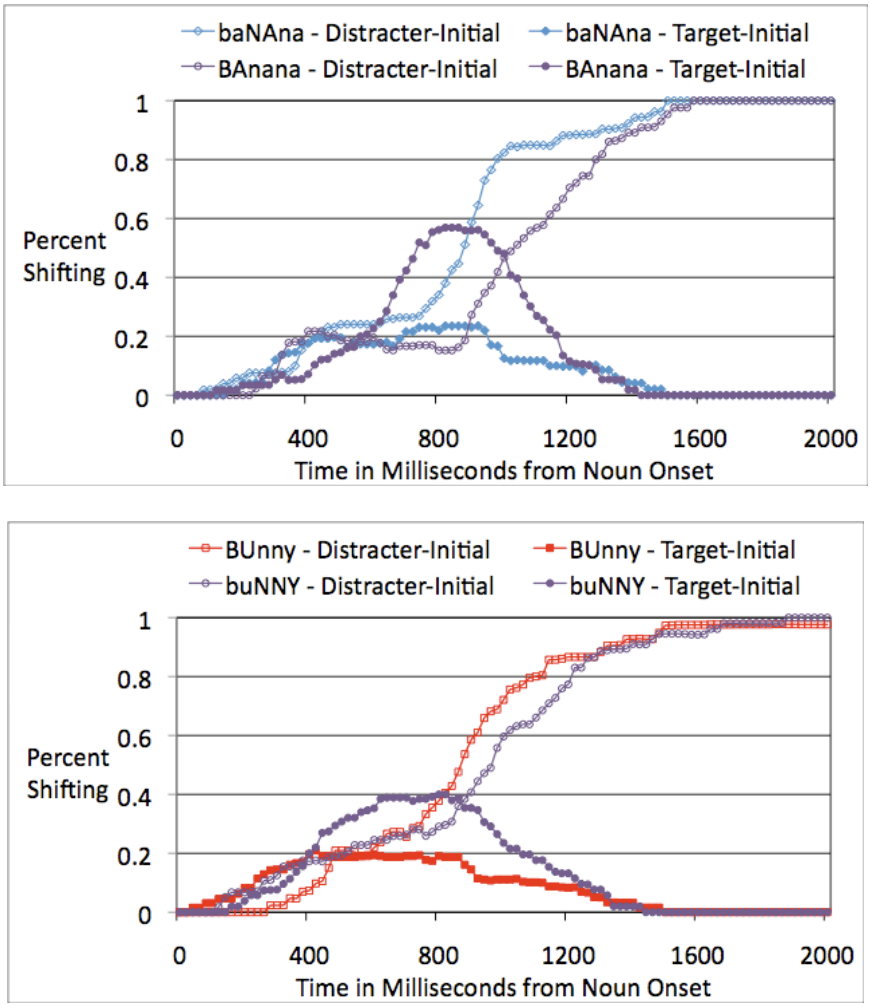


Figure 16: Target-fixation split by object (target vs. distracter) adults were fixating at target-word onset in Experiment 1, for "banana" trials (top) and "bunny" trials (bottom)
These plots are generated by plotting distracter-initial fixations straight-forwardly, and plotting target-initial fixations subtracted from 1. This means that in correct-pronunciation trials, distracter-initial fixations should begin at 0% and rise to 100%, while target-initial fixations should remain close to 0%.

To compute statistics on the looking-time data, we first averaged target-fixation proportions across the time window 360–2000 ms post–noun onset, the time-window typically used in eye-tracking studies with children (we use the same time-window in **Experiment 2**).[9] **Figure 17** and **Table 9** summarize adults' mean target fixations for this

[9] The time window 200–2000 milliseconds post–noun onset is typically used with adults. We use this slightly later time window here because the nature of the stimuli and the degree of stimulus overlap (over the entire 'bun' region of the words) delayed adults' responses.

time window. We then conducted an analysis of variance in which the dependent variable was target-fixation proportion, and the predictors were the word ("bunny" or "banana") and the pronunciation (correct or misstressed). Pronunciation exerted a significant effect on target fixation ($F(1,124) = 13.99$, $p < .001$), which was higher in response to correctly stressed words than misstressed words; there was no effect of the word. T-tests confirmed that participants fixated the target much more in response to correctly stressed versions of "bunny" (mean, 80.99%) than misstressings (mean, 73.15%; paired $t(31) = 2.77$, p(2-tailed) < .01). This was also true for "banana" (mean for correct pronunciations, 83.21%; mean for misstressings, 74.69%; paired $t(31) = 2.78$, p(2-tailed) < .01).

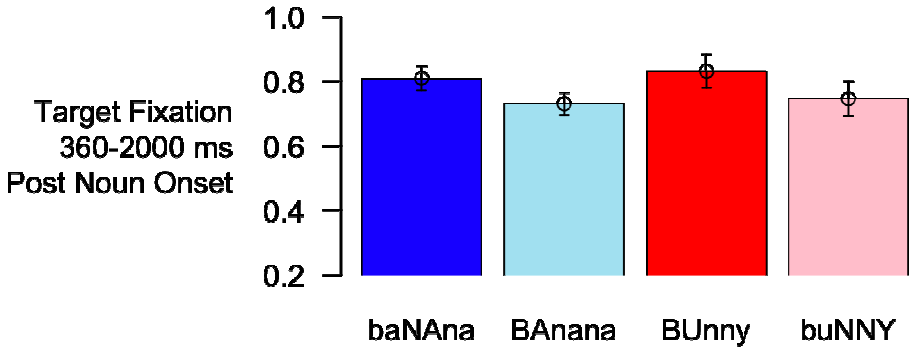


Figure 17: Adults' mean target fixation proportions in Experiment 1, averaged over the time window 360–2000 milliseconds after noun onset

Table 9: Target-fixation difference for correct – mispronounced versions of *bunny* and *banana* in Experiment 1, averaged across the time window 360–2000 ms after noun onset

	Mispronunciation effect	Proportion of participants with MP-effect > 0
Banana	7.84%	21/32 (66%)
Bunny	8.52%	20/32 (63%)

4.2.2.2 Questionnaire Responses

In the questionnaire, participants responded to questions intended to determine whether they noticed the misstressings of "bunny" and "banana" ('Did you notice

anything strange about the words you heard? If so, which word(s)? What was strange about them?' / 'Did you ever notice changes in any words? If so, which word(s)? How did the word(s) change?') and whether they guessed the purpose of the experiment ('Based on your experience in the whole experiment, what do you think we were testing? When did you first think this was what the experiment was about?').

Participants varied in how much they noticed about the pronunciations of "bunny" and "banana." All but one participant reported noticing something strange about the words; this was often described as similar-sounding first syllables (13 participants), elongated first syllables (8), strange pronunciations, emphasis, accent, or inflections (9), and/or mispronunciations (6). All but three participants noticed that the first syllables of "bunny" and "banana" overlapped. Though 21 participants reported that the words were made to sound more similar, only 12 participants believed that this was related to the purpose of the experiment; 5 more participants believed that the strange pronunciations of "bunny" and "banana" were related to the purpose of the experiment, and only 2 participants realized that we were testing stress perception.

Of the 32 participants, 15 mentioned that one or both of the words changed across trials (14/32 reported that "banana" changed, and 11/32 reported "bunny"), though only 3 mentioned stress changes specifically. Since roughly half of participants reported noticing changes in the words, we evaluated whether reporting or not reporting changes was related to moment-by-moment sensitivity to the misstressings during the experiment. To evaluate this, we split participants into "reporters" (N = 15) and "nonreporters" (N = 17). We then conducted an analysis of variance in which the dependent variable was target-fixation proportion, and the predictors were the word ("bunny" or "banana"), the

pronunciation (correct or misstressed), and whether the participant reported either mispronunciation (yes or no).

This analysis revealed a significant overall effect of pronunciation; target fixation was higher in response to correctly stressed words (mean, 81.88%) than misstressed words (mean, 73.92%; $F(1,120) = 12.93$, $p < .001$). There was also an effect of reporting either word; nonreporters had higher overall target fixation (mean, 80.00%) than reporters (mean, 75.53%; $F(1,120) = 4.07$, $p < .05$). This effect appears to be largely driven by reporters' reduced target fixation in response to misstressed words (means for reporters: correctly stressed words, 81.22%; misstressed words, 69.83%; see **Table 10**; means for nonreporters: correctly stressed words, 82.46%; misstressed words, 77.53%; see **Table 11**), but the interaction between reporting status and pronunciation was not significant ($p = .15$).

Table 10: Target-fixation differences for participants who reported noticing mispronunciation effects in Experiment 1

	Mispronunciation effect	Proportion of participants with MP-effect > 0
Banana	11.32%	10/15 (66.67%)
Bunny	12.40%	10/15 (66.67%)

Table 11: Target-fixation differences for participants who did *not* report noticing mispronunciation effects in Experiment 1

	Mispronunciation effect	Proportion of participants with MP-effect > 0
Banana	4.77%	11/17 (64.71%)
Bunny	5.10%	10/17 (58.82%)

In t-tests, only reporters had significant mispronunciation effects for the words "bunny" (correctly stressed mean, 82.46%; misstressed mean, 70.06%, $t(14) = 2.59$, p(2-

tailed) < .05) and "banana" (correctly stressed mean, 80.91%, misstressed mean, 69.60%, t(14) = 2.32, *p*(2-tailed) < .05). Nonreporters showed no significant mispronunciation effects, though they had a trend in the right direction for both "bunny" (correctly stressed, 83.87%, misstressed, 78.77%, t(16) = 1.31, *p*(2-tailed) = .21) and "banana" (correctly stressed, 81.06%, misstressed, 76.29%, t(16) = 1.55, *p*(2-tailed) = .14). Time-course plots also look similar for the two groups (see **Figure 18**), suggesting that nonreporters were sensitive to the mispronunciations.

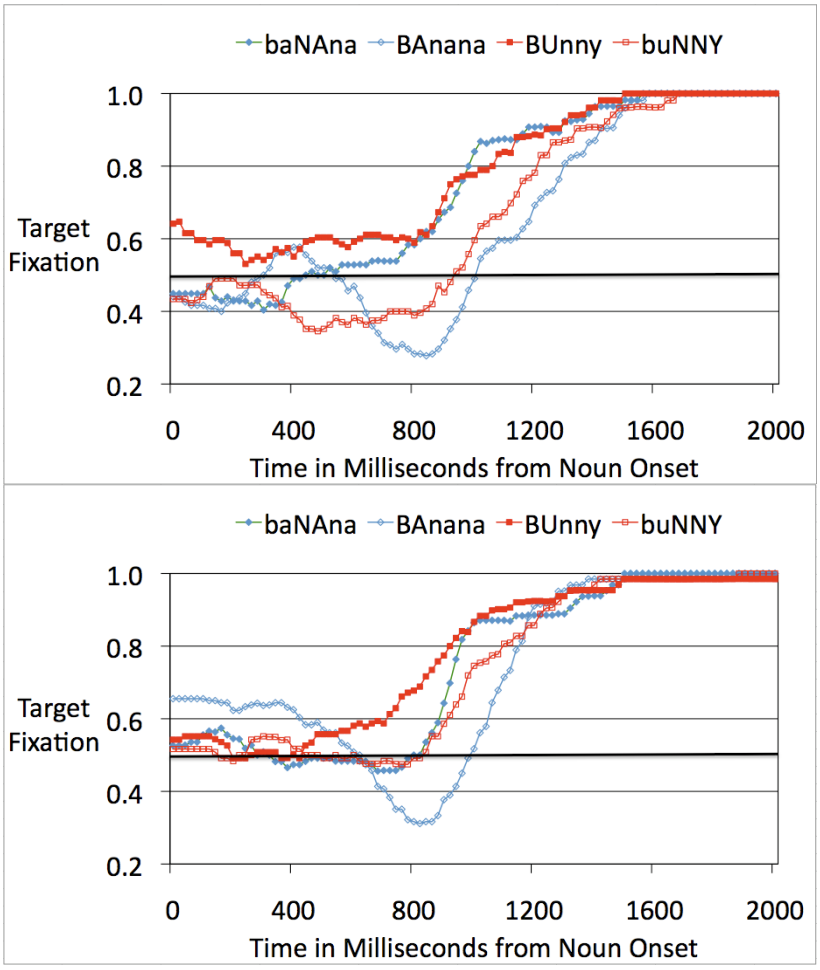


Figure 18: Target-fixation over time for Experiment 1 participants who *reported* noticing either mispronunciation (N = 15; top) and those who did not (N = 17; bottom)

We found overall that adults showed sensitivity to mispronunciations of the pitch cue to stress for both "bunny" and "banana." Testing adults allowed us to analyze their looking patterns over time in combination with their questionnaire responses. We found that participants who reported noticing the mispronunciations were also somewhat more responsive to the mispronunciations, as reflected by their eye movements. We next tested preschoolers in roughly the same task, to see whether they would also be sensitive to mispronunciations of the pitch cue to stress, and to see whether this sensitivity develops during the preschool years.

4.3 Experiment 2

We tested children from 2.5 to 5 years of age in roughly the same procedure used with adults. We collected information from parents about whether their children understood and/or said the words "bunny" and "banana," so that we could evaluate whether knowing the words made children more likely to respond to mispronunciations. Of the 96 parents, only 1 said the child did not understand one of the words, "bunny," but the child's looking patterns indicated comprehension of "bunny." Five more parents said their children understood but did not say one or both of the words. We also conducted a simplified, verbal version of the questionnaire with children, asking them whether they noticed "something funny" about the words they heard. Five children indicated (either during this verbal questionnaire or spontaneously during or after the experiment) that they had noticed either mispronunciations of the words or something funny or weird about them; three 5-year-olds, one 4-year-old, and one 3-year-old.

4.3.1 Method

4.3.1.1 Participants

We included 96 children between the ages of 2.5 and 5 years in the analysis; 48 2 ½- to 3-year-olds, 26 female (mean age 3 years, 2 months, and 2 days), and 48 4- to 5-year-olds, 24 female (mean age 5 years and 9 days). Ten more participants were tested but excluded from the analysis: two because they were hearing a language other than English more than 30% of the time, and eight because they were inattentive. Children were deemed inattentive if, in more than half (2) of the trials in each trial type (e.g., "BUnny" or "buNNY" trials), they failed to fixate the pictures for at least 300 milliseconds during the time window typically used for analysis with children, 360–2000 milliseconds after noun onset (out of 1660 possible milliseconds).

4.3.1.2 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 plus the attention-getting videos (this was much shorter than the adult experiment, which contained 4 extra filler trials inserted between each of the 32 trials that children saw). Instead of answering a questionnaire, children were simply asked if they had noticed anything weird about the words they heard.

4.3.2 Results and Discussion

For analysis, we divided children into two groups: 2 ½- to 3-year-olds (N=48) and 4- to 5-year-olds (N=48). We chose this age division because it created two groups of equal size, and also reflected similarities in looking patterns between 2- to 3-year-olds and between 4- to 5-year-olds. **Figure 19** plots, for each of these groups, fixation of the

target picture over time in each of the four experimental conditions. Three aspects of these plots are most salient. First, both age groups fixated the *bunny* picture more than the *banana* picture for most of the time window. Second, the younger children appeared to respond to the "bunny" mispronunciation late in the time window; target fixation in response to the correctly stressed version of "bunny" exceeded target fixation in response to the misstressing. Finally, the older children appeared most responsive to the "banana" mispronunciation, with only very small, if any, effects for "bunny."

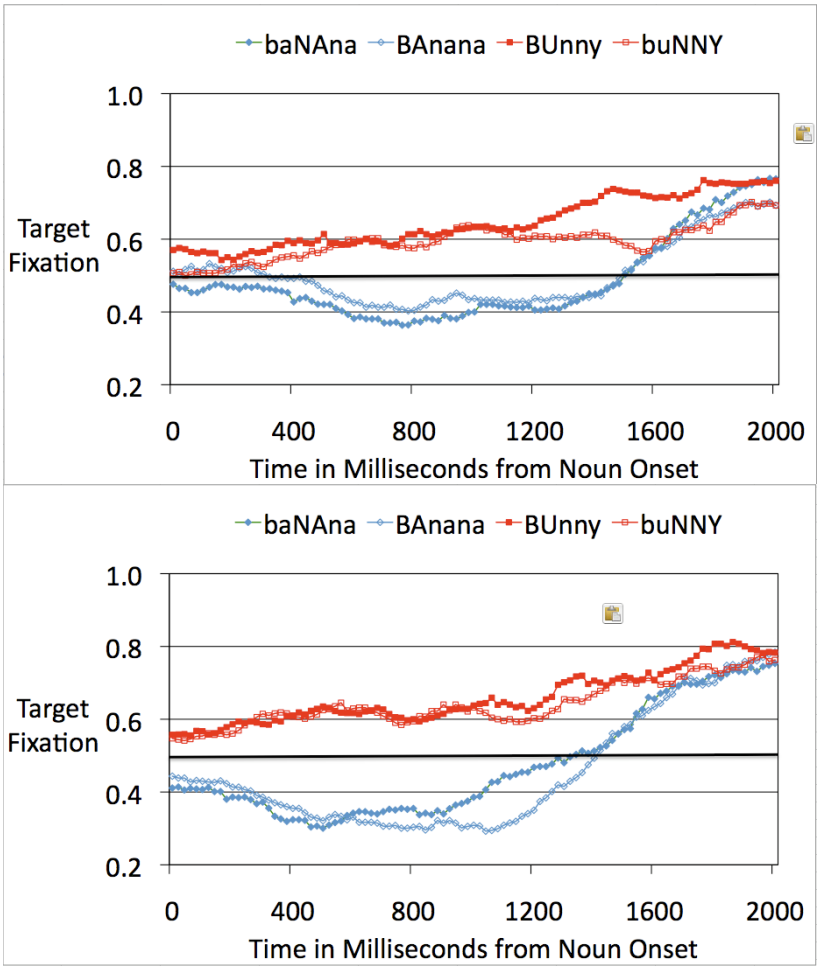


Figure 19: Children's fixation of the target picture over time in each condition in Experiment 2, for 2- to 3-year-olds (top) and 4- to 5-year-olds (bottom).
The x-axis is time in ms; 0 is target-word onset, and the ambiguous region "bun" ends at 610 ms in all conditions. The y-axis is the proportion of trials in which participants are fixating the target object. Younger children show late sensitivity to misstressings of "bunny," while older children show earlier sensitivity to misstressings of "banana."

Both groups showed mispronunciation effects later in the time window than adults did, with the correct versus mispronounced lines diverging after roughly 800 ms. It would not be surprising for children to show sensitivity to the misstressings later in time than adults do. Considering these apparently late mispronunciation effects, we averaged children's target-fixation proportions across a later time window than the one used with adults: 800–2000 ms post–noun onset; **Figure 20** and **Tables 12** and **13** summarize these

means. We then conducted analyses of variance for each age group in which the dependent variable was target-fixation proportion, and the predictors were the word ("bunny" or "banana"), and the pronunciation (correct or misstressed).

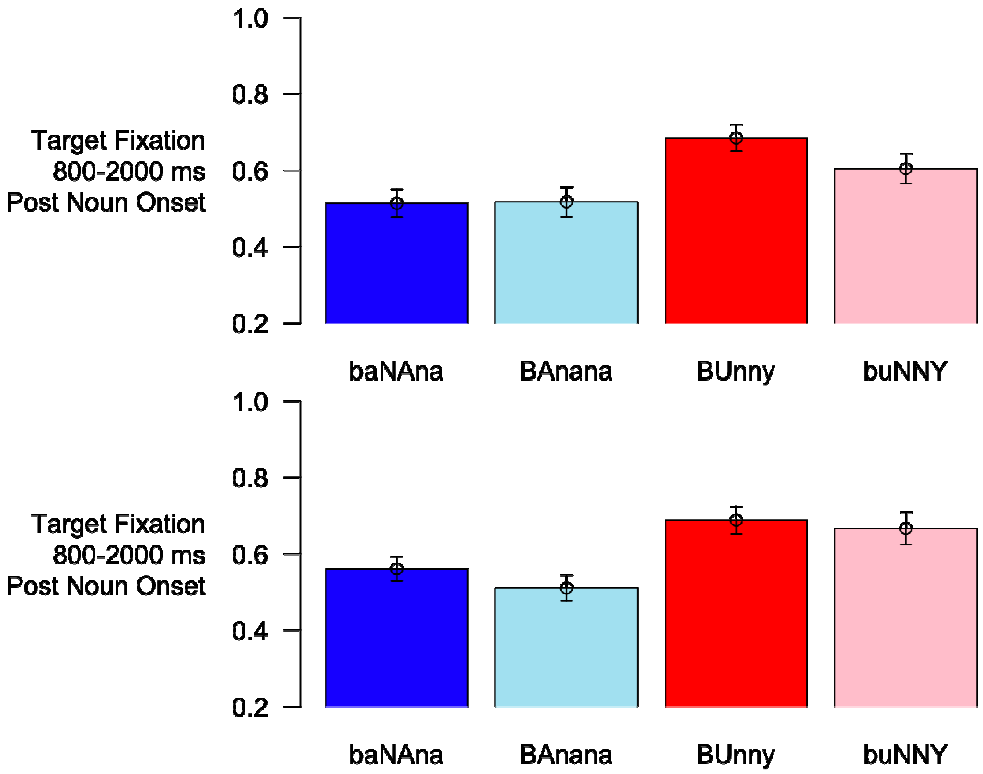


Figure 20: Mean target-fixation proportions in Experiment 2, averaged over the time window 800–2000 milliseconds after noun onset, for 2- and 3-year-olds (top), and 4- and 5-year-olds (bottom)
These means and the time-course plots in **Figure 19** suggest sensitivity to "bunny" mispronunciations at 2–3 years and sensitivity to "banana" mispronunciations at 4–5 years.

Table 12: Target-fixation differences for 2- and 3-year-olds in Experiment 2, averaged across the time-window 800–2000 milliseconds post–noun onset

	Mispronunciation effect	Proportion of participants with MP-effect > 0
Banana	-0.41%	26/48 (54.17%)
Bunny	8.01%	26/48 (54.17%)

Table 13: Target-fixation differences for 4- and 5-year-olds in Experiment 2, averaged across the time-window 800–2000 milliseconds post–noun onset

	Mispronunciation effect	Proportion of participants with MP-effect > 0
Banana	5.00%	29/48 (60.42%)
Bunny	2.20%	24/48 (50.00%)

For the younger age group, the word ("bunny" or "banana") exerted a significant effect on target fixation ($F(1,188) = 11.81$, $p < .001$); children fixated the target more in "bunny" trials (mean, 64.38%) than in "banana" trials (mean, 51.48%). This difference reflects a preference for the *bunny* picture, but it is not simply a baseline preference. There is an advantage for *bunny* at target onset (see **Figure 19**), but it becomes stronger as children process the target word. This may reflect children's bias to match "bun" with the *bunny* picture, which they prefer to look at. It may also reflect a bias in the stimuli. Though we tried to record "bunny" and "banana" with neutral duration and amplitude cues to stress (before superimposing the pitch contours onto them), these cues seem to actually better match stressed than unstressed syllables (see **Figure 14**). Children may weight these other cues more heavily than adults do, leading them to treat "bun" as a better match for *bunny* than for *banana* regardless of its pitch pattern.

While adults showed significant effects of pronunciation (correctly stressed or misstressed) on target fixation in an analysis of variance (ANOVA), 2- to 3-year-old children did not. We conducted post-hoc t-tests of the effects of misstressing for each word separately, using Bonferroni adjusted alpha levels of .025 (.05/2). These tests revealed significantly greater target fixation in response to correct versions of "bunny"

(mean, 68.39%) compared with misstressings (mean, 60.38%, paired t(47) = 2.45, p(2-tailed) = .018); there was no such effect for "banana" trials.

We conducted the same statistical tests with 4- to 5-year-olds. As with younger children, the ANOVA revealed a significant effect of the word ("bunny" or "banana"; ($F(1,188) = 15.35$, $p < .001$), but no effect of pronunciation. Unlike with younger children, however, Bonferroni-adjusted post-hoc t-tests revealed no significant effects, though there was a trend toward greater target fixation in response to correct versions of "banana" (mean, 56.05%) compared with misstressings (mean, 51.05%, paired t(47) = 1.38, p(2-tailed) = .18). There were no significant correlations between age (in days) and mispronunciation effects (target fixation in response to correct stress – misstressings) for "banana" or "bunny" (though there was a trend for more sensitivity to misstressings of "banana" with increasing age, $r = 0.13$, $p = 0.22$; and a trend for *less* sensitivity to "bunny" misstressings with age, $r = -.17$, $p = 0.10$).

To get a better sense of each age group's sensitivity to the misstressings, we considered two types of trials separately: trials in which, at the onset of the target word, children happened to be fixating the target picture (e.g., the *banana* in a "baNAna" or "BAnana" trial), or "target-initial" trials, versus trials where the child was initially fixating the distracter picture (e.g., the *bunny* in a "banana" trial). **Figures 21** and **22** plot these two types of trials for each of the four target-word types at each age. The plots suggest that both the "bunny" mispronunciation effect at the younger age and the "banana" mispronunciation effect at the older age were largely driven by target-initial trials. In other words, children who happened to be fixating the *target picture* when the target word began were more likely to *shift away* to the distracter picture if the first

syllable was misstressed, whereas misstressing had little effect if children were initially fixating the distracter. This is in contrast to adults, who showed mispronunciation effects in both target-initial and distracter-initial trials (see **Figure 16**).

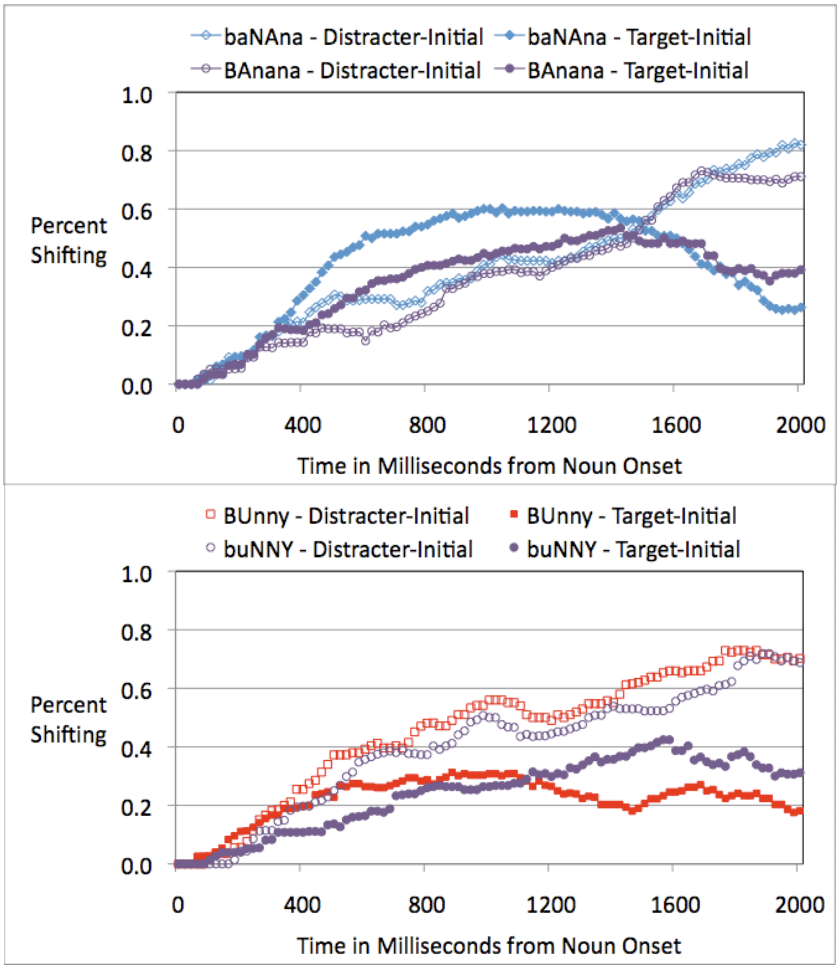


Figure 21: Target fixation split by object (target vs. distracter) children were fixating at target-word onset in Experiment 2, for 2- to 3-year-olds in "banana" trials (top) and "bunny" trials (bottom) These plots are generated by plotting distracter-initial fixations straight-forwardly, and plotting target-initial fixations subtracted from 1. This means that in correct-pronunciation trials, distracter-initial fixations should begin at 0% and rise to 100%, while target-initial fixations should remain close to 0%. These two plots support the impression from the bar graphs in **Figure 20** that younger children were sensitive to mispronunciations of "bunny" but not "banana" (though there is a hint of possible sensitivity to "banana" mispronunciations late in the time window).

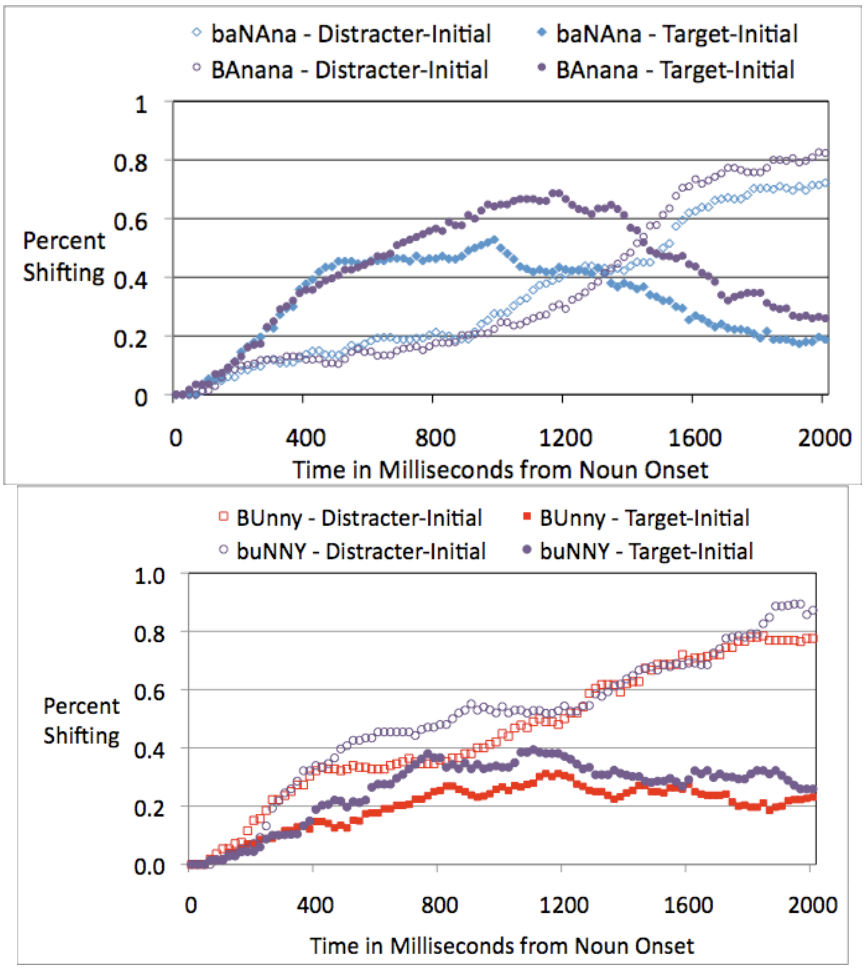


Figure 22: Target fixation split by object (target vs. distracter) children were fixating at target-word onset in Experiment 2, for 4- to 5-year-olds in "banana" trials (top) and "bunny" trials (bottom)
These plots confirm the impression from **Figure 19** that older children responded more to mispronunciations of "banana" than to "bunny." "Banana"-mispronunciation effects were mostly driven by target-initial trials. However, the "bunny" plot shows more subtle sensitivity to "bunny" mispronunciations, also in target-initial trials.

Though we found only fairly subtle effects of the mispronunciations on children's looking overall, we predicted that we might find more sensitivity to mispronunciations for the five children who reported noticing something strange about the words. One of these five children had been excluded from previous analyses for having only one usable correctly stressed "banana" trial. We included all her data in the time-course plots in **Figure 23**, and but included her target-fixation means only for "bunny" in **Table 14**.

Both the time-course plots (**Figure 23**) and mean target-fixation proportions (**Table 14**) indicate that "responders" did indeed show substantially more sensitivity to mispronunciations of both words than did children overall (**Figure 19** and **Tables 12** and **13**).

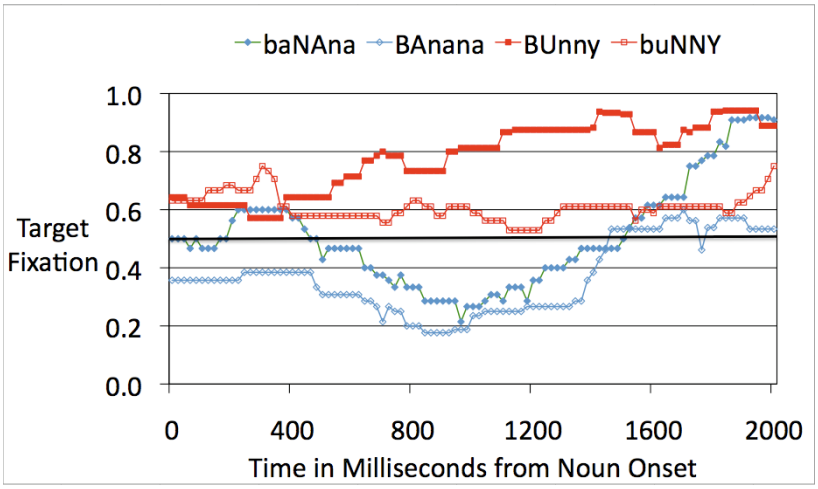


Figure 23: Target fixation over time in each condition in Experiment 2, for the five children who *reported* noticing misstressings of "bunny" and/or "banana"
Three 5-year-olds, one 4-year-old, and one 3-year-old reported noticing changes in the words (or something strange about them). As a group, these children show substantially larger mispronunciation effects for both words (see **Table 14** for means).

Table 14: Target-fixation differences for *reporters* in Experiment 2, averaged across the time-window 800–2000 milliseconds post–noun onset
One reporter was included only for "bunny" trials, since she only had one correctly stressed "banana" trial (for that reason, she was excluded from the previous analyses).

	Mispronunciation effect	Proportion of participants with MP-effect > 0
Banana	21.13%	4/4 (100%)
Bunny	21.98%	5/5 (100%)

To summarize, children were less sensitive to mispronunciations of the pitch cue to stress than adults were. Younger children (2- and 3-year-olds) responded to misstressings of "bunny," while older children (4- and 5-year-olds) showed some sensitivity to misstressings of both words (as evidenced by **Figures 19–22**), though these

effects were not significant in ANOVA and t-tests.[10] Children showed a strong preference to fixate the *bunny* picture, which increased over the course of the trial. This preference likely reflected a combination of greater interest in animate objects, combined with greater weighting of other cues to stress like duration, amplitude, and vowel quality. The "stress-neutral" stimuli, upon which we superimposed the pitch contours, appear to actually be a better match to stressed syllables than to unstressed syllables on these other acoustic dimensions. This might explain why children fixated the *bunny* more than the *banana*, regardless of the pitch pattern they heard.

While nearly half (15/32) of adults reported noticing the mispronunciations, a much small proportion of children (5/97, or roughly 5%) reported noticing something strange about the words. We did find that children who reported noticing something strange about the words showed greater sensitivity to the mispronunciations during the experiment (see **Figure 23** and **Table 14**); this parallels similar effects with adults. To some degree, the lower rate of reporting mispronunciations likely reflects children's more limited ability to reflect upon and verbalize their experiences; especially for something as subtle as detecting pitch mispronunciations. Nevertheless, it probably also reflects children's lower sensitivity to the pitch cue relative to adults.

Children's difficulty exploiting the pitch cue to lexical stress led us to wonder whether children's difficulty was related to the pitch cue in particular, or to a larger problem with exploiting lexical stress in word recognition. To tease this apart, we are currently testing whether children also struggle to exploit *all four* cues to stressed

[10] A more sensitive statistical method like Growth Curve Analysis (GCA; Mirman, Dixon, & Magnuson, 2008) might better represent the time-course of children's looking patterns. It could be that such an analysis would reveal statistically significant mispronunciation effects for 4- and 5-year-olds. Ultimately, we plan to analyze this data set using a method like GCA.

syllables. We predict that children will be more sensitive to mispronunciations of all four cues to stress than they were to just the pitch cue. However, it is possible that adults will show a similar pattern. Though adults successfully exploited the isolated pitch cue, they might be *even better* at exploiting all four convergent cues to stress. Such a finding would help us interpret children's responses, by giving us a sense of the adult cue weights for pitch versus the other three cues to stress.

4.4 Experiment 3

We tested 16 adults in roughly the same procedure used in Experiment 1. In this case, however, all four cues to stress varied naturally. The question of interest was whether adults would be more sensitive to misstressings when all 4 cues were mispronounced, versus when just the pitch cue was mispronounced in Experiment 1.

4.4.1 Method

4.4.1.1 Participants

Sixteen adults, six female, and all native speakers of English, were included in the analysis.[11] All participants were undergraduates assumed to be between 17 and 23 years of age. No participants were excluded.

4.4.1.2 Auditory Stimuli

The visual stimuli, trial sequence, procedure, and data analysis were almost identical to Experiment 1. The primary difference was the nature of the auditory stimuli. In English, four different acoustic cues—amplitude, duration, amount of vowel reduction,

[11] For Experiment 3 we were less concerned about the effects of language-specific cue weightings, so we did not require participants to be monolingual English-speakers. One participant was a native bilingual in English and Gujarati; her responses were comparable to those of monolinguals.

and location of the pitch target—jointly indicate the location of the stressed syllable. In contrast to Experiment 1, where we isolated the pitch cue to the stress contrast between "bunny" and "banana" and controlled the other three cues, in Experiment 3 we allowed all four cues to covary, by presenting naturally recorded words. This meant that the first syllable of "BAnana" and "BUnny," in addition to containing a pitch peak, was longer, higher in amplitude (the mean amplitude of the whole word was normalized to 70 dB, but relative differences between the syllables were maintained), and less reduced vocalically than the first syllables of "baNAna" and "buNNY." **Table 15** summarizes the acoustic measurements for stimuli in Experiment 3. We used two tokens of each word (e.g., two "BUnny" tokens) compared with one in Experiment 1; this was partly to reduce boredom in the child version of this experiment that is currently in progress, 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. Waveforms and pitch tracks for Experiment 3 stimuli are shown in **Figure 24**; for each word, one of the two tokens is depicted.

Table 15: Acoustic measurements for the *first syllable* of each target word used in Experiment 3, averaged over the two tokens of each word

	Pitch mean (SD)	Pitch max	Intensity	Duration	F1/F2
baNAna	201.4 Hz (7.8 Hz)	221.0 Hz	68.4 dB	0.23 sec	623.1/1690.83 Hz
BAnana	364.4 Hz (47.4 Hz)	422.2 Hz	75.7 dB	0.36 sec	861.07/1587.08 Hz
buNNY	195.75 Hz (4.9 Hz)	204.6 Hz	69.3 dB	0.17 sec	575.9/1739.7 Hz
BUnny	370.7 Hz (48.3 Hz)	425.9 Hz	74.7 dB	0.38 sec	827.7/1609.3 Hz

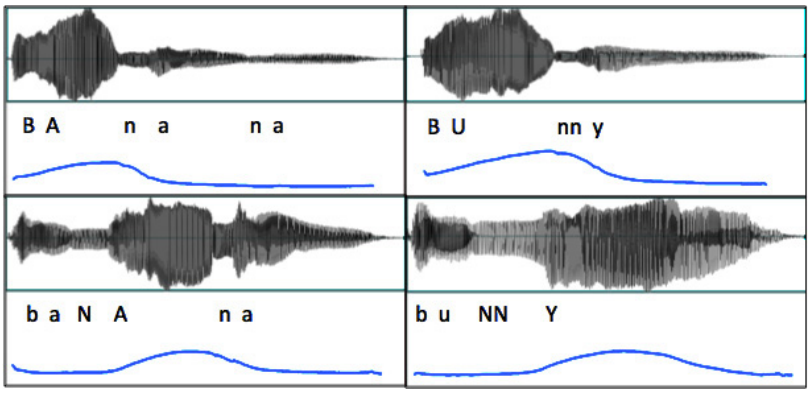


Figure 24: Waveforms and pitch tracks of one of the two tokens of each of the words used in Experiment 3

Note that duration and amplitude vary substantially between stressed and unstressed syllables in these stimuli (as does vowel quality).

4.4.2 Results and Discussion

We predicted that adults would be more sensitive to misstressings of "bunny" and "banana" when all four cues were allowed to covary than when only the pitch cue indicated stress. Plots of adults' target fixation over time in **Experiments 1** and **3** (**Figure 25**) support this hypothesis; mispronunciation effects for both words appear to be larger in **Experiment 3** than in **Experiment 1**. To determine whether looking patterns were statistically different in the two cases, we averaged target fixation over the time window from 360–2000 milliseconds after noun onset; see **Figure 26** and **Table 16** for these means. We then conducted an analysis of variance in which the dependent variable was target-fixation proportion, and the predictors were the word ("bunny" or "banana"), the pronunciation (correct or misstressed), and the experiment (1 or 3).

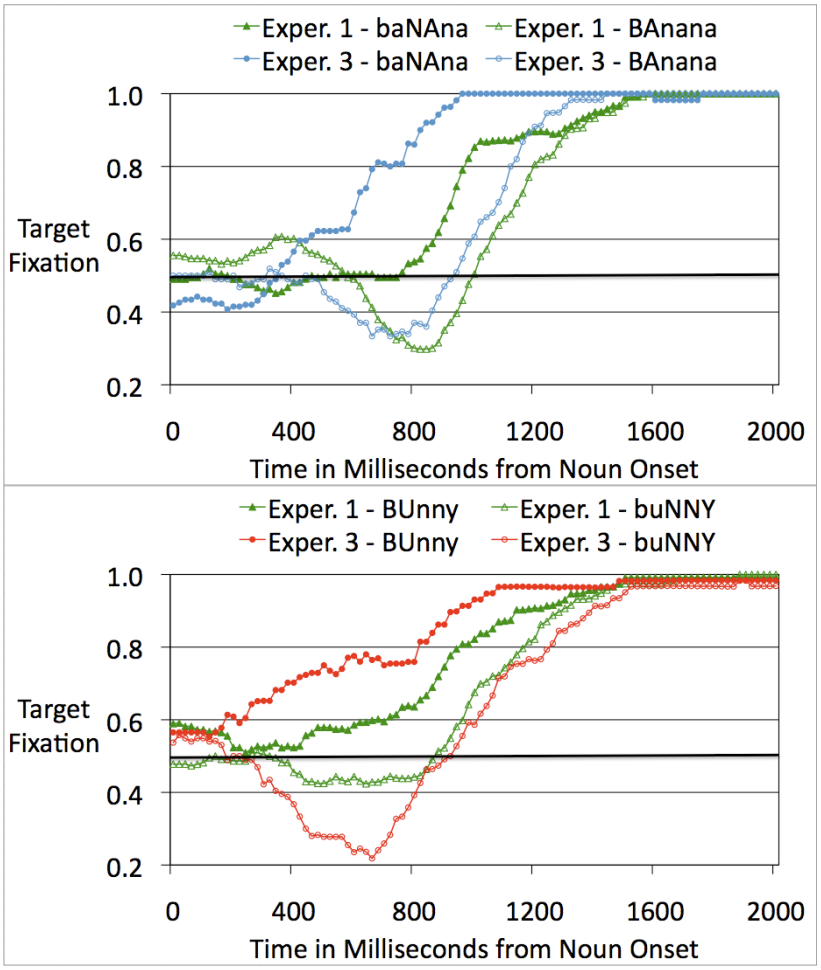


Figure 25: Target Fixation Over Time in "Banana" Trials (Top) and "Bunny" Trials (Bottom) Mispronunciation effects are greater in **Experiment 3** (blue circles for "banana" and red circles for "bunny") than in **Experiment 1** (green triangles).

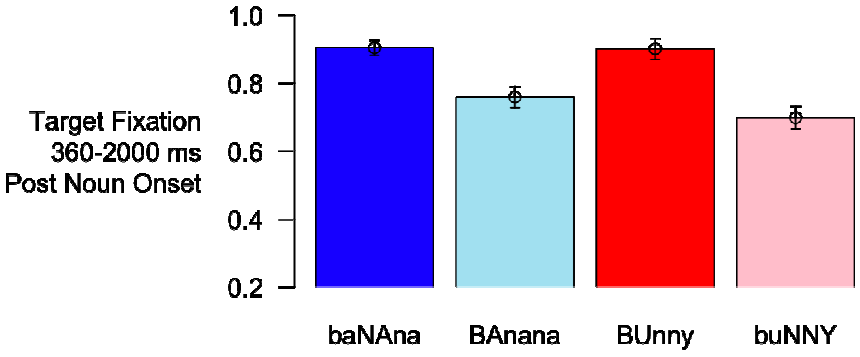


Figure 26: Adults' mean target fixation proportions in Experiment 3, averaged over the time window 360–2000 milliseconds after noun onset

Table 16: Target-fixation difference for correct – mispronounced *bunny* and *banana* in Experiment 3, averaged across the time window 360–2000 ms after noun onset

	Mispronunciation effect	Proportion of participants with MP-effect > 0
Banana	14.50%	13/16 (81%)
Bunny	20.23%	15/16 (94%)

The ANOVA revealed a significant overall effect of pronunciation; target fixation was higher in response to correctly stressed words (mean, 84.67%) than misstressed words (mean, 73.57%; $F(1,184) = 39.55$, $p < .001$). Importantly, it also revealed a significant interaction between pronunciation and experiment; the target-fixation advantage for correctly stressed words over misstressed words was greater in Experiment 3 (a difference of 17.37%) than in Experiment 1 (a difference of 7.96%; $F(1,184) = 6.32$; $p < .05$). This suggests that adults were more sensitive to naturalistic misstressings, in which all four cues jointly indicated stress, than to misstressings that relied on the pitch cue alone.

4.5 General Discussion

The three experiments described here begin to paint a picture of adults' and children's use of different acoustic cues to lexical stress. In **Experiment 1**, we found that adults exploited isolated pitch cues to lexical stress in word recognition. Participants had more trouble identifying the target picture, *bunny* or *banana*, when the pitch of the first syllable mismatched the stress of the word (as in "buNNY" and "BAnana") than when the pitch matched adults' expectations (as in "BUnny" and "baNAna"). In **Experiment 2**, we found that 2- to 5-year-old children were less skilled at exploiting pitch cues to stress than adults were. Time-course plots and t-tests computed on looking-time averages

revealed that younger children (2- to 3-year-olds) showed sensitivity to misstressings of "bunny" but not "banana." Time-course plots indicated that older children (4- to 5-year-olds) were sensitive to misstressings of both words (though effects were more noticeable for "banana" mispronunciations), but these effects were not significant in t-tests. For both adults and children, we found that participants who reported noticing mispronunciations showed larger mispronunciation effects during the experiment. However, the rate of reporting mispronunciations was much greater for adults (47%) than for children (roughly 5%).

Several aspects of lexical stress in English, and the pitch cue in particular, may help explain the protracted acquisition course for this cue. Pitch conveys stress in combination with spectral, duration, and amplitude cues, and the pitch cue interacts with sentence intonation, leading to variability in its realization. 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 interpreted. All of these factors likely reduce the reliability of the pitch cue. More generally, children may acquire the lexical-stress system of English slowly relative to other phonological systems because of its low functional load relative to categories like lexical tones.

The results of **Experiment 2** do not tell us whether children are struggling to exploit the pitch cue in particular in word recognition, or lexical stress more generally. **Experiment 3** represents the first step in teasing apart these two explanations for children's difficulty with the pitch cue to stress. In that experiment, we presented adults simultaneously with all four cues to lexical stress: duration, intensity, vowel quality, and pitch. Adults' target fixation was dramatically affected when all four cues were

mispronounced; they were significantly better at exploiting stress in word recognition when all four cues converged than when only the pitch cue was manipulated.

The greater sensitivity to mispronunciations of all four cues compared with the pitch cue alone suggests that even adults struggle to exploit isolated pitch cues to stress, relative to all four converging cues. This result may help shed light on children's difficulty exploiting the isolated pitch cue. It could be that children initially rely either on a holistic integration of all four cues, or on cues other than pitch (e.g., duration and intensity). Over developmental time, they learn to flexibly shift the weights of different cues to adapt to the particular context (e.g., a noisy environment). By adulthood, listeners can ratchet up their weighting of the pitch cue, as adults presumably did in **Experiment 1**, when it is the most reliable cue in the local context. Unsurprisingly, however, even adults perform best when given all sources of information about the location of stress.

Chapter 5: Conclusion

This dissertation examined how and when children learn the various roles of pitch in English, focusing in particular on how they rule out lexical pitch and attend to pitch when it cues the speaker's emotions and the location of word stress. Despite evidence that young infants are highly sensitive to pitch, and despite the early acquisition of consonant and vowel categories—which might suggest that phonological acquisition is completed in the first year—we have found that correct *interpretation* of discriminable pitch exhibits a more protracted learning course. Children learn to rule out pitch as lexically contrastive in English between 18 (Quam & Swingley, in progress) and 30 months (**Chapter 2**). We also found protracted time-courses for learning to exploit pitch when it *is* relevant in English; children did not correctly interpret pitch cues to emotions until around age 4 (**Chapter 3**), and were still struggling to exploit the pitch cue to lexical stress at age 5 (**Chapter 4**).

In these concluding paragraphs, we discuss the particular challenge of interpreting pitch variation and offer some ways that children might learn to properly and fluently interpret it. To offer another perspective on the task of the learner, we also discuss ongoing phonetic corpus work, which asks how the input to children might help them identify pitch categories. Integrating phonetic corpus analysis methods with the experimental methods used in this dissertation offers the best hope of fully describing how children converge on adult-like processing of phonetic variation.

5.1 The Learning Problem: Interpreting Ambiguous Pitch Patterns

As discussed throughout this dissertation, pitch peaks in English are ambiguous between multiple linguistic categories. For example, a pitch peak on the first syllable of the word "bunny" could indicate a lexical tone (if a child has not learned that English is not a tone language), a stressed syllable, sentence focus, or excitement; it could even indicate several of these categories at once (e.g., a stressed syllable produced by an excited speaker). The child must learn how to resolve this ambiguity.

Distributional learning over instances may help children correctly interpret pitch peaks by allowing them to identify correlations between levels of structure. For instance, it seems likely that children's knowledge of global stress patterns (e.g., English learners' trochaic bias) and their lexical knowledge interact during early word learning. Discovering that English words tend to be trochaic helps infants to segment word-forms that they can later map to meanings, though it leads them to initially missegment iambic words, which require integration of other cues (Jusczyk, Houston, & Newsome, 1999).

Once they have established a rudimentary inventory of words, infants can begin tracking lexical-stress properties within words. Knowing a word provides children with a *domain* within which they can integrate multiple cues to stress and detect properties of the lexical-stress system. Tracking cues to stress within the word domain, for example, would allow children to notice that many words are trochaic but that some are iambic. Tracking stress within the word and the syllable also likely helps children detect the four covariant cues to stress.

For instance, once children know the word "bunny," they can start tracking its properties. They will notice that the first syllable of "bunny" is relatively long and loud,

often contains a pitch peak, and has an unreduced vowel instead of a schwa (IPA: /ə/). With experience, they can observe that the pitch pattern of "bunny" varies as a function of context; in a yes/no question context, "bunny" usually has a rising pitch, with a low pitch target in the first syllable. Many of the words in children's developing vocabularies—the trochaic words—will exhibit a similar pattern, with a pitch peak in the first syllable that changes to a low target in yes/no-question contexts. Factors like speaker excitement, sentence focus, utterance position, and speech rate will affect the pitch mean and range of this pattern across tokens, but the essential pattern will be reasonably invariant. Comparable structure will emerge for iambic words, which exhibit a pitch peak in or near the second syllable that also changes to a low target in yes/no-question contexts.

Within the domain of the word "bunny," children can also track distributions of amplitude, duration, and vowel quality, and learn that these dimensions are correlated with each other and with the pitch cue. Once all four cues to stress have been identified in this way, the child must learn the optimal weight of each cue, and must learn to flexibly modify these weights in different linguistic contexts and in different conditions of noise. Children's difficulty learning to flexibly adapt phonetic weights to particular contexts (Cohn, submitted; Nittrouer, Miller, Crowther, & Manhart, 2000; Hazan & Barrett, 2000) means that this last stage of phonological development takes much longer than the acquisition of the native-language sound categories in the first year (Polka & Werker, 1994; Bosch & Sebastián-Gallés, 2003; Werker & Tees, 1984; Kuhl et al., 2006; Kuhl, Conboy, Padden, Nelson, & Pruitt, 2005).

Since children struggle to adapt cue weights to the context, they may initially rely more heavily on cues that are relatively context-invariant. This could explain why children relied less heavily on the pitch cue to lexical stress (which is very different in statements versus yes/no questions) than adults did in **Chapter 4**. Over time, children's ability to track distributions of pitch patterns (and of other cues) across tokens of particular words may allow them to cope with variability in the realization of the pitch cue.

Children's relative inflexibility in interpreting phonetic cues may also explain why they sometimes have more trouble exploiting a single, isolated cue than adults do (e.g., Seidl, 2007; Seidl & Cristià, 2008). By presenting children with mispronunciations of all four cues to stress, we are now exploring whether children have difficulty exploiting stress in general, or flexibly adjusting their cue weights to selectively attend to the pitch cue. Adults can flexibly shift their cue weights in response to changes in the reliability of different cues; hence their success at interpreting the pitch cues in **Chapters 3** and **4**. This flexibility is crucial for identifying linguistic categories in noise and across different contexts.

5.2 Corpus Phonetics Provide Another Perspective on the Learning Problem

The work in this dissertation has focused on children's interpretations of speech. Another perspective from which to examine the child's learning task is to ask how the input to children might help them eventually converge on the adult interpretations of phonetic variation. How might it explain the late trajectories we have found in the present

experiments? In ongoing work (Quam, Yuan, & Swingley, 2008; Quam, Yuan, Swingley, & Wang, in progress), we are analyzing the phonetic patterns of mothers' speech to their infants to ask how these patterns might convey pitch categories. Because languages differ in their pitch categories, children must learn language-specific categories from their parents' speech. We have argued that distributional learning goes a long way in enabling the child to learn acoustic cues to different linguistic categories and correctly attribute pitch variation. Phonetic analyses of mothers' speech allow us to investigate how distributions of pitch patterns actually occur in speech, to help inform our investigations of how children detect and exploit those distributions. Quam, Yuan, & Swingley (2008) analyzed the pitch patterns of English infant-directed speech (IDS); ongoing work (Quam, Yuan, Swingley, & Wang, in progress) is now comparing the pitch patterns of English versus Mandarin IDS.

Quam, Yuan, & Swingley (2008) investigated pitch patterns of infant-directed English speech. IDS is characterized by exaggerated intonation patterns and short, simple phrases. Because these exaggerated intonation patterns frequently convey a small, stereotyped range of emotional signals, one might expect particular words, like "good" or "no," to be realized with consistent pitch contours. This consistency in a word's pitch realization might facilitate word recognition; however, in an intonation language like English, it could falsely suggest lexical tones. Using corpus phonetics methods, Quam et al. (2008) analyzed the speech input to English-learning children, identifying the amount, nature, and sources of pitch variation across about 3,300 tokens of 8 highly frequent words in the Brent corpus (Brent & Siskind, 2001) from the CHILDES database (MacWhinney, 2000).

Quam et al. (2008) found two basic results. First, although intonation in IDS is prototypically exaggerated, about half the instances of frequently occurring, utterance-final words were flat in contour. Second, although each frequent word varied substantially in its intonation contours (e.g., rises versus rise-falls), words like "good" and "no" differed in ways that seemed to reflect the pragmatic categories typical of each word's use. For instance, "no" was generally flat or falling, and consistently low in pitch, reflecting its occurrence in prohibitive utterances, while "good" occurred more often with a rise-fall contour, reflecting its use in approving utterances. Even the word "good," however, still had more flat contours than rise-fall contours. Quam et al. (2008) proposed that this within-word variability in pitch realization could help the child rule out lexical tone in English.

In order to test the hypothesis that within-word variability in pitch patterns helps children rule out lexical tone, word-level pitch patterns in English must be compared to those in a tone language. To this end, Quam, Yuan, Swingley, & Wang (in progress) are currently collecting a corpus of Mandarin-speaking mothers' speech to their infants. These recordings and transcriptions will allow us to compare the pitch patterns of similar words in English versus Mandarin. We predict that, because words in Mandarin are constrained by their underlying tones, they will exhibit more consistency in pitch (contour, mean, range, etc.) than Quam, Yuan, & Swingley (2008) found with English words. This would suggest that the amount of pitch variability of words across tokens may help children rule out—or focus in on—lexical tone.

As discussed in the previous section (with respect to lexical stress), we propose an interactive model of learning in which knowledge of prosody and knowledge of words

support one another. Infants track prosodic patterns before they know any words; they pick up on their language's dominant stress pattern early on (e.g., Friederici, Friedrich & Christophe, 2007), and they use prosodic cues to segment words from fluent speech (e.g., Jusczyk, Houston, & Newsome, 1999). Tracking the frequency of pitch peaks across syllables may also help infants infer whether or not they are hearing a tone language (M. Liberman, personal communication, March, 2008; C. Phillips, personal communication, November, 2008).

Infants are clearly not waiting until they have a rudimentary lexicon before beginning to learn about lexical stress and other prosodic systems. At the same time, once children know some words, they can begin tracking prosodic properties within the word domain. This is likely important for learning that English words can be *either* trochaic or iambic, for example, or that Mandarin syllables can have one of four tone patterns. Having the word as an anchor may also help children learn to cope with lexical pitch variability, so that they can recognize the word "bunny" whether it is produced with a low or a high pitch target in the first syllable, for example. This in turn might help children learn to cope with variability in the pitch cue to stress more generally, so that knowing words makes children's knowledge of the lexical-stress system more flexible and sophisticated. Similar interactions between the lexicon and children's prosodic knowledge are also likely for children learning pitch-accent and lexical-tone systems.

The research presented in this dissertation finds a late trajectory for correctly interpreting lexical pitch; English-learning children seem to learn to disregard lexical pitch between 18 months (Quam & Swingley, in progress) and 30 months (**Chapter 2**). We find an even later time-course for correctly interpreting pitch cues to emotions, with

which children succeed by about age 4 (**Chapter 3**), and for exploiting pitch cues to lexical stress, with which 5-year-olds still struggle (**Chapter 4**). This late time-course is surprising given the evidence that young infants are highly sensitive to pitch, and that they learn consonant and vowel categories by 12 months. It emphasizes, however, that detecting patterns of sounds in language (consonants, vowels, and prosodic structure) is just the first step of phonological development. The next step is interpreting acoustic variation to recognize words and infer the meanings of utterances (by incorporating lexical and paralinguistic information). This interpretive process 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) and variability in the realization of cues, introduced by linguistic context, environmental noise, and other factors. Evidence from our studies and others (e.g., Hazan & Barrett, 2000; Nittrouer, Miller, Crowthers, & Manhart, 2000) suggests that this learning process continues well into childhood.

Appendix 1: Acoustics of the teaching and test words from Chapter 2

Mean and standard deviation of duration in seconds, pitch maximum in Hz, and pitch mean in Hz for each teaching and test word.

Word	Pitch	Phase	Duration (SD)	Pitch Max (SD)	Pitch Mean (SD)
Deebo	Rise-fall	Teaching	1.245 (0.076)	587.7 (56.2)	284.8 (15.5)
Deebo	Low fall	Teaching	1.370 (0.121)	264.1 (11.7)	215.1 (6.8)
Deebo	Rise-fall	Test	1.321 (0.038)	673.4 (26.3)	300.1 (2.7)
Deebo	Low fall	Test	1.292 (0.077)	283.9 (2.9)	232.7 (9.1)
Dahbo	Rise-fall	Test	1.326 (0.044)	757.4 (1.2)	295.1 (7.1)
Dahbo	Low fall	Test	1.283 (0.007)	274.3 (16.1)	221.3 (5.3)

Appendix 2: Example trial order for Chapter 3, Experiment 1

	Cue	***Word***	***Target***
Pretrial 1	Words	*Gazzer*	Toy 1
Pretrial 2	Words	*Blicket*	Toy 2
Trial 1	Body-language	*Toma*	Toy 3
Trial 2	Body-language	*Zeemo*	Toy 1
Trial 3	Body-language	*Pumbie*	Toy 2
Trial 4	Pitch	*Noopa*	Toy 3
Trial 5	Pitch	*Dawnoo*	Toy 2
Trial 6	Pitch	*Tizzle*	Toy 1
Trial 7	Pitch	*Tawny*	Toy 2

Appendix 3: Example toys used in the Chapter 3 experiments

Appendix 4: Ratios of happy versus sad acoustic measurements of the experimenter's speech for each acoustic dimension in each experiment in Chapter 3

Cue	Ratios of Happy / Sad	
	Experiment 1	**Experiment 2**
Pitch Mean	1.41	1.44
Standard Deviation of Pitch Samples	3.12	2.39
Pitch Maximum	1.50	1.50
Pitch Minimum	1.22	1.23
Intensity	1.00	1.01
Duration	1.02	1.00 (n.s. difference)

Appendix 5: Example trial order for Chapter 3, Experiment 2

All twelve trials after pretrials are either pitch trials or facial/body-language trials, depending on the child.

	Target
Pretrial 1:	Toy 2
Pretrial 2:	Toy 1
Trial 1:	Toy 2
Trial 2:	Toy 1
Trial 3:	Toy 1
…	
Trial 12:	Toy 2

REFERENCES

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52,* 163–187.

Aravamudhan, R., Lotto, A. J., & Hawks, J. W. (2008). Perceptual context effects of speech and nonspeech sounds: The role of auditory categories. *Journal of the Acoustical Society of America, 124,* 1695–1703.

Ashby, F. G., Queller, S., and Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception and Psychophysics, 61,* 1178–1199.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9,* 321–324.

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-Stop clusters. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 536–563.

Bailey, T. M., & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development, 17,* 1265–1282.

Baldwin, D. A., & Moses, L. J. (1996). The ontogeny of social information gathering. *Child Development, 67,* 1915–1939.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 9–54). Oxford: Oxford University Press.

Berinstein, A. E. (1979). A cross-linguistic study on the perception and production of stress. *University of California Working Papers in Phonetics*, *47,* 1–59.

Bertinetto, P. M. (1980). The perception of stress by Italian speakers. *Journal of Phonetics, 8,* 385–395.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In Goodman, Judith C.; Nusbaum, Howard C. (Eds.). The development of speech perception: The transition from speech sounds to spoken words. (pp. 167–224). Cambridge, MA, US: The MIT Press.

Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptuareorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14(3),* 345–360.

Best, C. T., Morongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics, 29,* 191–211.

Best, C. T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in acoustic patterns. *Perception & Psychophysics, 45,* 237–250.

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

Boersma, P., Escudero, P., & Hayes, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-

specific sound categories. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1013–1016.

Bolinger, D. (1989). Intonation and its uses. *Stanford, CA: Stanford University Press.*

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast. *Language and Speech, 46*, 217–244.

Braine, M. D. S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In B. MacWhinney (Ed.). *Mechanisms of language formation* (pp. 65–87). Hillsdale, NJ: Erlbaum.

Brent, M. R. & Siskind, J. M (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81,* 31–44.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5,* 325–337.

Bryant, G. A., & Barrett, H. C. (2007). Recognizing intentions in infant-directed speech: Evidence for universals. *Psychological Science, 18,* 746–751.

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new pussycat? On talking to babies and animals. *Science, 296,* 1435.

Bybee, J. (2001a). *Phonology and language use.* Cambridge: Cambridge University Press.

Bybee, J. (2001b). Frequency effects on French liaison. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 337–359). Amsterdam, Netherlands: John Benjamins Publishing Company.

Bybee, J. (2007). *Frequency of use and the organization of language.* Oxford: Oxford University Press.

Carter, A. K., & Gerken, L.A. (2003). Similarities in weak syllable omissions between children with specific language impairment and normally developing language: A preliminary report. *Journal of Communication Disorders, 36,* 165–179.

Carter, A., & Gerken, L.A. (2004). Do children's omissions leave traces? *Journal of Child Language, 31,* 561–586.

Chen, Y., Robb, M., Gilbert, H., & Lerman, J. (2001). *Clinical Linguistics & Phonetics, 15*, 427–440.

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics, 31,* 585–598.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access. I. Adult data. *Journal of Memory and Language, 51,* 523–547.

Cohn (submitted). The source of universals: Features, segments, and the nature of phonological primitives. Clements & Ridouane (Eds.), *Where Do Features Come From?*

Connine, C. M., Clifton, C., & Cutler, A. (1987). Effects of lexical stress on phonetic categorization. *Phonetica, 44,* 133–146.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech, 45,* 207–228.

Curtin, S. (2010). Young infants encode lexical stress in newly encountered words. *Journal of Experimental Child Psychology, 105,* 376–385.

Curtin, S., Fennell, C., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word object associative learning. *Developmental Science, 12,* 725–731.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition, 96,* 233–262.

Cutler, A. (1986). Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech, 29,* 201–220.

Cutler, A. (1989). Auditory lexical access: Where do we start? In W. D. Marslen-Wilson (Ed.), Lexical representation and process. Cambridge, MA: MIT Press.

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing.* Cambridge, MA: The MIT Press; p. 105–121.

Cutler, A., & Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics, 59,* 165–179.

Cutler, A., & Clifton, C. (1984). The use of prosodic information in word recognition. In H. Bouma & D.G. Bouwhuis (Eds.), *Attention and Performance X: Control of Language Processes.* Hillsdale, N.J.: Erlbaum; 183–196.

Cutler, A., & Donselaar, W. van (2001). Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech, 44,* 171–195.

Cutler, A., & Norris, D. (1988). The role of strong syllables for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 113–121.

Cutler, A., & Otake, T. (1999). Pitch accent in spoken-word recognition in Japanese. *Journal of the Acoustical Society of America, 105,* 1877–1888.

Cutler, A., & Swinney, D. A. (1987). Prosody and the development of comprehension. Journal of *Child Language, 14,* 145–167.

Cutler, A., Dahan, D., & Donselaar, W. van (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40,* 141–201.

Cutler, E. A., & Clifton, C. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Proceedings of the Tenth International Symposium on Attention and Performance* (pp. 183–196). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

D'Entremont, B., & Muir, D. (1999). Infant responses to adult happy and sad vocal and facial expressions during face-to-face interactions. *Infant Behavior and Development, 22,* 527–539.

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

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16,* 507–534.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science, 208,* 1174–1176.

Demuth, K. (1995). The acquisition of tonal systems. In J. Archibald (Ed.), *Phonological Acquisition and Phonological Theory* (pp. 111–134). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dietrich, C., Swingley, D., & Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of*

the National Academy of Sciences of the USA, 104, 16027–16031.

Donselaar, W. van, Koster, M., & Cutler, A. (2005). Exploring the role of lexical stress in lexical recognition. *The Quarterly Journal of Experimental Psychology Section A, 58,* 251–273.

Dupoux, E., Pallier, C., Sebastián-Gallés, N., & Mehler, J. (1997). A destressing 'deafness' in French? *Journal of Memory and Language, 36*, 406–421.

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress "deafness." *Journal of the Acoustical Society of America, 110,* 1606–1618.

Dupoux, E., Sebastián-Gallés, N., Navarrete, E., & Peperkamp, S. (2008). Persistent stress 'deafness': The case of French learners of Spanish. *Cognition, 106,* 682–706.

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language, 36,* 202–225.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, 26, 551–585.

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America,* 97, 1893–1904.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-Shot learning of object categories. *IEEE Trans. Pattern Analysis and Machine Intelligence, 28,* 594–611.

Fennell, C. T. (2006). Infants of 14 months use phonetic detail in novel words embedded in naming phrases. In *Proceedings of the 30th Annual Boston University*

Conference on Language Development (pp. 178-189). Somerville, Massachusetts: Cascadilla Press.

Fennell, C. T., Waxman, S. R., & Weisleder, A. (2007). With referential cues, infants successfully use phonetic detail in word learning. In *Proceedings of the 31st Annual Boston University Conference on Language Development* (pp. 206–217). Somerville, Massachusetts: Cascadilla Press.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development, 59,* i–185.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development, 8,* 181–195.

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

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Development, 64*, 6571497–510674.

Fernald, A. (1992). Meaningful melodies in mothers' speech to infants. In H. Papousek, U. Jurgens, & M. Papousek (Eds.), *Nonverbal vocal communication: Comparative and developmental approaches* (pp. 262–282). Cambridge: Cambridge University Press.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development, 10*, 279–293.

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the second year. *Psychological Science, 9,* 72–75.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics, 27,* 343–350.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 241–256.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 349–366.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics, 62,* 1668–1680.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 234–257.

Friederici, A. D., Friedrich, M., & Christophe, A. (2007). Brain responses in 4-month-old infants are already language specific. *Current Biology, 17,* 1208–1211.

Friend, M. (2000). Developmental changes in sensitivity to vocal paralanguage. *Developmental Science, 3,* 148–162.

Friend, M. (2001). The transition from affective to linguistic meaning. *First Language, 21,* 219–243.

Friend, M. (2003). What should I do? Behavior regulation by language and paralanguage in early childhood. *Journal of Cognition and Development, 4,* 161–183.

Friend, M., & Bryant, J. B. (2000). A developmental lexical bias in the interpretation of discrepant messages. *Merrill-Palmer Quarterly, 46,* 342–369.

Fry, D. (1958). Experiments in the perception of stress. *Language and Speech, 1,* 205–213.

Fulkerson, A. L., & Haaf, R. A. (2003). The influence of labels, non-labeling sounds, and source of auditory input on 9- and 15-month-olds' object categorization. *Infancy, 4,* 349–369.

Galligan, R. R. (1987). Intonation with single words: Purposive and grammatical use. *Journal of Child Language, 14*, 1–21.

Gandour, J. T. (1978). The perception of tone. In V. A. Fromkin (Ed.), *Tone: A Linguistic Survey* (pp. 41–76). New York: Academic Press.

Gandour, J., Dzemidzic, M., Wong, D., Lowe, M., Tong, Y., Hsieh, L., Satthamnuwong, N., & Lurito, J. (2003). Temporal integration of speech prosody is shaped by language experience: An fMRI study. *Brain and Language, 84,* 318–336.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613–656.

Gauthier, B., Shi, R., & Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition, 103,* 80–106.

Gerken, L.A., Jusczyk, P. W., & Mandel, D. R. (1994). When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition, 51,* 237–265.

Gerken, L.A., Landau, B., & Remez, R. E. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology, 26,* 204–216.

Goffman, L., Gerken, L.A., & Lucchesi, J. (2007). Relations between segmental and motor variability in prosodically complex nonword sequences. *Journal of Speech, Language, and Hearing Research, 50,* 444–458.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22,* 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*, 178–200.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585–612.

Goldstone, R. L. & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130,* 116–139.

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language, 51,* 548–567.

Grieser, DA., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology, 25,* 577–588.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100,* 1111–1121.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America, 106*, 2900–2912.

Gussenhoven, C. (2004). The phonology of tone and intonation. Cambridge: Cambridge University Press.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology, 30,* 304–337.

Halle, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics, 32,* 395–421.

Harrison, P. (2000). Acquiring the phonology of lexical tone in infancy. *Lingua, 110,* 581–616.

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

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics, 28,* 377–396.

Hillenbrand, J.M., Clark, M.J., & Houde, R.A. (2000). Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America, 108,* 3013–3022.

Hirschberg, J., & Ward, G. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics, 20,* 241–251.

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

Hochberg, J. G. (1988). Learning Spanish stress: Developmental and theoretical perspectives. *Language, 64,* 683–706.

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

Hollich, G. (2005). Supercoder: A program for coding preferential looking (Version 1.5). [Computer Software]. West Lafayette: Purdue University.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., Hennon, E., & Rocroi, C. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development, 65,* v–123.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America, 119,* 3059–3071.

Holt, L. L., Lotto, A. J., & Diehl, R. L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *Journal of the Acoustical Society of America, 116,* 1763–1773.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2001). Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement? *Journal of the Acoustical Society of America, 109,* 764–774.

Houston, D. M., & Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 1143–1154.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology, 26,* 1570–1582.

Houston, D. M., Santelmann, L. M., & Jusczyk, P. W. (2004). English-learning infants' segmentation of trisyllabic words from fluent speech. *Language and Cognitive Processes, 19,* 97–136.

Howie, J. M. (1976). Acoustical Studies of Mandarin Vowels and Tones. Cambridge: Cambridge University Press.

Hua, Z., & Dodd, B. (2000). The phonological acquisition of Putonghua (Modern Standard Chinese). *Journal of Child Language, 27,* 3–42.

Huang, L. M. (1992). Remarks on the phonological structure of Mandarin Chinese. *Bulletin of the National Taiwan Normal University, 37,* 363–383.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America, 97,* 553–562.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Ketterman, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87,* B47–B57.

James, W. (1890). The Principles of Psychology, Vol. 1. New York, NY, US: Henry Holt and Co.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44,* 548–567.

Johnson, E. K., & Seidl, A. (2009). At 11 months, prosody still outranks statistics. *Developmental Science, 12,* 131–141.

Johnson, K. (1997) Speech perception without speaker normalization: An exemplar model. In Johnson & Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 145–165). San Diego: Academic Press.

Jusczyk, P. W., & Houston, D. M. (1999). The beginnings of words segmentation in English-learning infants. *Cognitive Psychology, 39,* 159–207.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64,* 675–687.

Jusczyk, P.W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics, 61,* 1465–1476.

Jusczyk, P. W., Luce, P. A. & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33**,** 630–645.

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA Model of how speech perception develops. *Journal of Phonetics, 21,* 3–28.

Jusczyk, P. W. (1994). Infant speech perception and the development of the mental lexicon. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech*

perception: The transition from speech sounds to spoken words (pp. 227–270). Cambridge, MA: The MIT Press.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29,* 1–23.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64,* 675–687.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39,* 159–207.

Jusczyk, P. W., Pisoni, D. B., Walley, A. C., and Murray, J. (1980). Discrimination of relative onset time of two-component tones by infants. *Journal of the Acoustical Society of America, 67,* 262–270.

Kahana-Kalman, R., & Walker-Andrews, A. S. (2001). The role of person familiarity in young infants' perception of emotional expressions. *Child Development, 72,* 352–369.

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*, 205–217.

Kehoe, M., Stoel-Gammon, C., & Buder, E. H. (1995). Acoustic correlates of stress in young children's speech. *Journal of Speech and Hearing Research, 38,* 338–350.

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mothers' speech: Adjustments for age and sex in the first year. *Infancy, 4,* 85–110.

Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2002). Universality and specificity in infant-directed speech: Pitch modifications as a

function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development, 24,* 372–392.

Klatt, D.H. (1973). Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America, 54,* 1102–1104.

Klein, H. B. (1984). Learning to stress: A case study. *Journal of Child Language, 11,* 375–390.

Kluender, K. R., Lotto, A. J., Holt, L. L., & Bloedel, S. L. (1998). Role of experience for language-specific functional mappings of vowel sounds. *Journal of the Acoustical Society of America, 104,* 3568–3582.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *Journal of the Acoustical Society of America, 70,* 340–349.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences, 97,* 11850–11857.

Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics, 32,* 542–550.

Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America, 23,* 1003–1010.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M. & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B, 363,* 979–1000.

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: Implications for the "critical period." *Language Learning and Development, 1,* 237–264.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S, & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9,* F13–F21.

Lacerda, F. (1993). Sonority contrasts dominate young infants' vowel perception. *Journal of the Acoustical Society of America, 93*, 2372–2372.

Lacerda, F. (1994). The asymmetric structure of the infant's perceptual vowel space. *Journal of the Acoustical Society of America, 95,* 3016–3016.

Ladd, D. R. (2008). *Intonational Phonology (Second Edition).* Cambridge: Cambridge University Press.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *Journal of the Acoustical Society of America, 78,* 435–444.

Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition, 38,* 245–294.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America, 32*, 451–454.

Lindfield, K. C., Wingfield, A., & Goodglass, H. (1999). The contribution of prosody to spoken word recognition. *Applied Psycholinguistics, 20,* 395–405.

Liu, H.-M., Tsao, F.-M., & Kuhl, P. K. (2007). Acoustic analysis of lexical tone in Mandarin infant-directed speech. *Developmental Psychology, 43,* 912–917.

Llisterri, J., Machuca, M. J., de la Mota, C., Riera, M., & Rios, A. (2003). The perception of lexical stress in Spanish. In Sole, M. J., Recasens, D., & Romero, J. (Eds.) *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona*, pp. 2023–2026.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1998). Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America, 103,* 3648–3655.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka, S. Manuel, & M. Matthies (Eds.), From Sound to Sense: 50+ Years of Discoveries in Speech Communication: MIT.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition, 26,* 708–715.

Luksaneeyanawin, S., Burnham, D., Francis, E., & Pansottee, S. (1997). The role of L1 background and L2 instruction in the perception of fricative contrasts: Thai and English children and adults. *Asia Pacific Journal of Speech, Language, and Hearing, 2,* 25–42.

MacWhinney, B. (2000). *The CHILDES database: Tools for analyzing talk, 3rd Edition.*

Vol 2: The database. Mahway, NJ: Lawrence Erlbaum Associates.

Mandel, D. R., Jusczyk, P. W., & Kemler Nelson, D. G. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition, 53,* 155–180.

Mareschal, D., & Quinn, P. C. (2001). Categorization in infancy. *TRENDS in Cognitive Sciences, 5,* 443–450.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Mattock, K., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy*, *10*, 241–265.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38,* 465–494.

Maye, J., Werker, J. F., & Gerken, LA. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–B111.

Mayo, C., & Turk, A. (2004). Adult-child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions. *Journal of the Acoustical Society of America, 115,* 3184–3194.

Mayo, C., & Turk, A. (2005). The influence of spectral distinctiveness on acoustic cue weighting in children's and adults' speech perception. *Journal of the Acoustical Society of America, 118,* 1730–1741.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese adults: Tests of a

Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience, 2,* 89–108.

McCawley, J. D. (1978). What is a tone language? In Fromkin, V. A. (Ed.), *Tone: A Linguistic Survey,* New York: Academic Press, pp. 113–131.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/–/l/ discrimination to Japanese adults: Behavioral and neural aspects. *Physiology & Behavior, 77,* 657–662.

McMurray, B., Tanenhaus, M., & Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*, B33–B42.

McQueen, J. M., Cutler, A., & Norris, D. (2003). Flow of information in the spoken word recognition system. *Speech Communication, 41,* 257–270.

McQueen, J. M., Norris, D, & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 621–638.

Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: The onset of word recognition. In G. T. M. Altmann (Ed.), Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives. Cambridge, MA: The MIT Press, pp. 236–62.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In Metsala, J. L. & Ehri, L. C. (Eds.), *Word recognition in*

beginning literacy (pp. 89–120). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics, 18,* 331–340.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America, 74,* 750–753.

Moore, D. S., Spence, M. J., & Katz, G. S. (1997). Six-month-olds' categorization of natural infant-directed utterances. *Developmental Psychology, 33(6),* 980–989.

Morton, J. & Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech, 8*, 159–181.

Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Development, 72,* 834–843.

Morton, J. B., Trehub, S. E., & Zelazo, P. D. (2003). Sources of inflexibility in 6-year-olds' understanding of emotion in speech. *Child Development, 74,* 1857–1868.

Morton, J., & Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech, 8,* 159–181.

Mumme, D. L., & Fernald, A. (2003). The infant as onlooker: Learning from emotional reactions observed in a television scenario. *Child Development, 74,* 221–237.

Mumme, D. L., Fernald, A., & Herrera, C. (1996). Infants' responses to facial and vocal emotional signals in a social referencing paradigm. *Child Development, 67,* 3219–3237.

Namy, L. L. (2008). Recognition of iconicity doesn't come for free. *Developmental Science, 11*, 841–846.

Namy, L. L. (2001). What's in a name when it isn't a word? 17-month-olds' mapping of nonverbal symbols to object categories. *Infancy, 2,* 73–86.

Namy, L. L., & Waxman, S. R. (1998). Words and gestures: Infants' interpretations of different forms of symbolic reference. *Child Development*, *69,* 295–308.

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science, 13,* 407–420.

Nazzi, T. (2005). Use of phonetic specificity during the acquisition of new words: Differences between consonants and vowels. *Cognition, 98,* 13–30.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology. Human Perception and Performance, 24,* 756–766.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language, 43,* 1–19.

Nittrauer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *Journal of the Acoustical Society of America, 112,* 711–719.

Nittrouer, S. (1996). Discriminability and perceptual weighting of some acoustic cues to speech perception by 3-year-olds. *Journal of Speech & Hearing Research, 39,* 278–297.

Nittrouer, S., & Lowenstein, J. H. (2007). Children's weighting strategies for word-final stop voicing are not explained by auditory sensitivities. *Journal of Speech, Language, and Hearing Research, 50,* 58–73.

Nittrouer, S., Miller, M. E., Crowther, C. S., & Manhart, M. J. (2000). The effect of segmental order on fricative labeling by children and adults. *Perception & Psychophysics, 62,* 266–284.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1209–1228.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Expermental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43,* 25–53.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60,* 355–376

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics, 57,* 989–1001.

Orlansky, M. D., & Bonvillian, J. D. (1984). The role of iconicity in early sign language acquisition. *Journal of Speech and Hearing Disorders, 49*, 287–292.

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

Palmieri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309–328.

Papousek, M., & Hwang, S. C. (1991). Tone and intonation in Mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction. *Applied Psycholinguistics, 12,* 481–504.

Papousek, M., Papousek, H., & Symmes, D. (1991). The meanings of melodies in motherese in tone and stress languages. *Infant Behavior and Development, 14,* 415–440.

Pater, J., Stager, C., & Werker, J. (2004). The perceptual acquisition of phonological contrasts. *Language, 80,* 384–402.

Pegg, J. E., & Werker, J. F. (1997). Adult and infant perception of two English phones. *Journal of the Acoustical Society of America, 102,* 3742–3753.

Peperkamp, S. A. (2004). Lexical exceptions in stress systems: Arguments from early language acquisition and adult speech perception. *Language, 80,* 98–126.

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress 'deafness'. In C. Gussenhoven & N. Warner (Eds.), Laboratory phonology VII. Berlin: Mouton de Gruyter, pp. 203–240.

Peperkamp, S., Le Calvez, R. L., Nadal, J.-P., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition, 101,* B31–B41.

Petitto, L. A. (1987). On the autonomy of language and gesture: Evidence from the acquisition of personal pronouns in American Sign Language. *Cognition, 27*, 1–52.

Pierrehumbert, J. (1980). The Phonology and Phonetics of English Intonation. Dissertation, MIT.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 337–359). Amsterdam, Netherlands: John Benjamins Publishing Company.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In *Laboratory Phonology VII*, (pp. 101–139). Berlin: Mouton de Gruyter.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46,* 115–154.

Pierrehumbert, J.B. (2006). The next toolkit. *Journal of Phonetics, 34*, 516–530.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 421–435.

Pollock, K. E., Brammer, D. M., Hageman, C. F. (1993). An acoustic analysis of young children's productions of word stress. *Journal of Phonetics, 21,* 183–203.

Pons, F., and Bosch, L. (2007). The perception of lexical stress patterns by Spanish and Catalan infants. In Prieto, P., Mascaro, J., & Sole, M.-J. (Eds.), *Segmental and Prosodic Issues in Romance Phonology.* Amsterdam, NE: John Benjamins, pp. 199–218.

Quam, C., & Swingley, D. (2010b). Bunny? Banana? Late development of sensitivity to the pitch cue to lexical stress. *International Conference on Infant Studies 2010 Biennial Meeting.*

Quam, C., & Swingley, D. (2010a). Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. *Journal of Memory and Language, 62,* 135–150.

Quam, C., & Swingley, D. (under review). Development in children's sensitivity to pitch as a cue to emotions. *Child Development.*

Quam, C., Swingley, D., & Park, J. (2009). Developmental change in preschoolers' sensitivity to pitch as a cue to the speaker's emotions. *Society for Research in Child Development 2009 Biennial Meeting*, Denver, CO.

Quam, C., Yuan, J., & Swingley, D. (2008). Relating intonational pragmatics to the pitch realizations of highly frequent words in English speech to infants. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 217–222). Austin, TX: Cognitive Science Society.

Roberts, K. (1995). Categorical responding in 15-month-olds: Influence of the noun-category bias and the covariation between visual fixation and auditory input. *Cognitive Development, 10,* 21–41.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*, 339–349.

Russell, J. A., Bachorowski, J.-A., & Fernández-Dolz, J.-M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54,* 329–349.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90,* 51–89.

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically-modulated sub-phonemic variations on lexical competition. *Cognition, 105,* 466–476.

Sansavini, A., Bertoncini, J., & Giovanelli, G. (1997). Newborns discriminate the rhythm of multisyllabic stressed words. *Developmental Psychology, 33,* 3–11.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 433–456). New York, New York: Oxford University Press.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America, 76,* 1346–1356.

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language, 57*, 24–48.

Seidl, A., & Cristià, A. (2008). Developmental changes in the weighting of prosodic cues. *Develomental Science, 11,* 596–606.

Sekerina, I. A., & Trueswell, J. C. (in press). Interactive processing of contrastive expressions by Russian children. *First Language.*

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

Sibata, T., & Shibata, R. (1990). 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, 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. (2008). Influences of high and low variability on infant word recognition. *Cognition, 106*, 833–870.

Singh, L., Morgan, J. L, & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language, 51,* 173–189.

Singh, L., White, K. S., & Morgan, J. L. (2008). 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, 4*, 157–178.

Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., Limissuri, R. A., & Peperkamp, S. Language-specific stress perception by 9-month-old French and Spanish infants. *Developmental Science, 12,* 914–919.

Slowiaczek, L. M., Soltano, E. G., & Bernstein, H. L. (2006). Lexical and metrical stress in word recognition: Lexical or pre-lexical influences? *Journal of Psycholinguistic Research, 35,* 491–512.

Soderstrom, M., Kemler Nelson, D. G., & Jusczyk, P. W. (2005). Six-month-olds recognize clauses embedded in different passages of fluent speech. *Infant Behavior and Development, 28,* 87–94.

Soderstrom, M., Seidl, A., Kemler Nelson, D. G., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49,* 249–267.

Soken, N. H., & Pick, A. D. (1992). Intermodal perception of happy and angry expressive behaviors by seven-month-old infants. *Child Development, 63,* 787–795.

Soken, N. H., & Pick, A. D. (1999). Infants' perception of dynamic affective expressions: Do infants distinguish specific expressions? *Child Development, 70,* 1275–1282.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language, 45,* 412–432.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*, 381–382.

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

Storkel, H. L. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language, 29*, 251–274.

Swingley, D. (2003). Phonetic detail in the developing lexicon. *Language and Speech, 46*, 265–294.

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology, 43*, 454–464.

Swingley, D. (2009). Onsets and codas in 1.5-year-olds' word recognition. *Journal of Memory and Language, 60,* 252–269.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition, 76,* 147–166.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13,* 480-484.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology, 54*, 99–132.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39,* 706–716.

Thiessen, E. D., & Saffran, J. R. (2004). Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception and Psychophysics, 66,* 779–791.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3,* 73–100.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7,* 53–71.

Thiessen, E.D. (2007). The effect of distributional information on children's use of phonemic contrasts. Journal of Memory and Language, 56, 16–34.

Tomasello, M., & Barton, M. E. (1994). Learning words in nonostensive contexts. *Developmental Psychology, 30,* 639–650.

Trehub, S.E., & Hannon, E.E. (2006). Infant music perception: Domain-general or domain-specific mechanisms? *Cognition, 100*, 73–99.

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35,* 445–472.

Turk, A.E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics, 28*, 397–440.

van Santen, J.P.H. (1992). Contextual effects on vowel duration. *Speech Communication, 11*, 513–546.

Vihman, M. (1996). *Phonological development: The origins of language in the child.* Malden, MA: Blackwell Publishing.

Vihman, M. M., DePaolis, R. A., & Davis, B. L. (1998). Is there a "trochaic bias" in early word learning? Evidence from infant production in English and French. *Child Development, 69,* 935–949.

Vihman, M. M., Nakai, S., DePaolis, R. A., & Hallé, P. (2004). The role of accentual pattern in early lexical representation. *Journal of Memory and Language, 50,* 336–353.

Vihman, M., & Croft, W. (2007). Phonological development: Toward a "radical" templatic phonology. *Linguistics, 45*, 683–725.

Walker-Andrews, A. S. (1998). Emotions and social development: Infants' recognition of emotions in others. *Pediatrics, 102,* 1268–1271.

Walker-Andrews, A. S., & Grolnick, W. (1983). Discrimination of vocal expressions by young infants. *Infant Behavior and Development, 6,* 491–498.

Walker-Andrews, A. S., & Lennon, E. (1991). Infants' discrimination of vocal expressions: Contributions of auditory and visual information. *Infant Behavior and Development, 14,* 131–142.

Walker, A. S. (1982). Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology, 33,* 514–535.

Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language, 61,* 747–776.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1,* 197–234.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7,* 49–63.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy, 3,* 1–30.

Werker, J. F., Hall, G., & Fais, L. (2004). Reconstructing U-shaped functions. *Journal of Cognition and Development, 5*, 147–151.

Werker, J.F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1,* 197–234.

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, *59*, 114–132.

White, K. S., Peperkamp, S., Kirk, C., & Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition, 107,* 238–265.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science, 30,* 945–982.

Woodward, A. L., & Hoyne, K. L. (1999). Infants' learning about words and sounds in relation to objects. *Child Development*, *70,* 65–77.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America, 95,* 2240–2253.

Yip, M. (2002). Tone. Cambridge: Cambridge University Press.

Yoshida, K., Fennell, C., Swingley, D., & Werker, J.F. (2009). 14-month-old infants learn similar-sounding words. *Developmental Science*, *12*, 412–418.

Younger, B. A. (1985). The segregation of items into categories by ten-month-old infants. *Child Development, 56,* 1574–1583.

Younger, B. A., & Cohen, L. B. (1983). Infant perception of correlations among attributes. *Child Development, 54,* 858–867.

Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *Journal of the Acoustical Society of America, 123,* 4498–4513.