Psychon Bull Rev (2014) 21:884–906
DOI 10.3758/s13423-013-0569-y

THEORETICAL REVIEW

The development of voicing categories: A quantitative review of over 40 years of infant speech perception research

Marcus E. Galle · Bob McMurray

Published online: 19 February 2014

Abstract Most research on infant speech categories has relied on measures of discrimination. Such work often employs categorical perception as a linking hypothesis to enable inferences about categorization on the basis of discrimination measures. However, a large number of studies with adults challenge the utility of categorical perception in describing adult speech perception, and this in turn calls into question how to interpret measures of infant speech discrimination. We propose here a parallel channels model of discrimination (built on Pisoni and Tash Perception & Psychophysics, 15(2), 285–290, 1974), which posits that both a noncategorical or veridical encoding of speech cues and category representations can simultaneously contribute to discrimination. This can thus produce categorical perception effects without positing any warping of the acoustic signal, but it also reframes how we think about infant discrimination and development. We test this model by conducting a quantitative review of 20 studies examining infants' discrimination of voice onset time contrasts. This review suggests that within-category discrimination is surprisingly prevalent even in classic studies and that, averaging across studies, discrimination is related to continuous acoustic distance. It also identifies several methodological factors that may mask our ability to see this. Finally, it suggests that infant discrimination may improve over development, contrary to commonly held notion of perceptual narrowing. These results are discussed in terms of theories of speech development that may require such continuous sensitivity.

Keywords Infant · Speech perception · Categorization · Phonological development · Voicing · Meta-analysis · Gradiency · Categorical perception · Discrimination · Development

Introduction

For over 4 decades, an important area of inquiry in language acquisition has been infants' ability to discriminate speech sounds. Researchers have asked how young infants perceive acoustic differences relevant to language (Eimas, Siqueland, Jusczyk, & Vigorito, 1971), whether this ability has an auditory basis (Jusczyk, Rosner, Reed, & Kennedy, 1989), when these abilities are tuned to their native language (Werker & Tees, 1984), and what learning mechanisms are involved (Maye, Werker, & Gerken, 2003). These are significant issues, and the field has reached something of a consensus on many of them. Research has shown that even very young infants are sensitive to a range of speech contrasts, including contrasts not used in their native language, and that a shift in this sensitivity occurs between 6 and 12 months as perceptual ability narrows toward the categories of the infant's native language (Aslin, Werker, & Morgan, 2002; Werker & Curtin, 2005). These findings have led to the view that phonological categories gradually emerge during this period of development (see Gottlieb et al., 1977, for a discussion).

Electronic supplementary material The online version of this article (doi:10.3758/s13423-013-0569-y) contains supplementary material, which is available to authorized users.

M. E. Galle
Department of Psychology & Delta Center, University of Iowa, Iowa City, IA 52242, USA

B. McMurray
Department of Psychology, Department of Communication Sciences and Disorders, Department of Linguistics & Delta Center, University of Iowa, Iowa City, IA 52242, USA

B. McMurray (✉)
Department of Psychology, E11 SSH, University of Iowa, Iowa City, IA 52242, USA
e-mail: Bob-mcmurray@uiowa.edu

Out of necessity, most empirical work has relied on measures of discrimination, the ability to perceive the difference between sounds in an experimental task (not categorization, the ability to treat different sounds equivalently). Methods like habituation/dishabituation and conditioned head-turning form the bulk of the relevant data, and these are largely measures of how infants discriminate a target stimulus from a baseline stimulus. Yet, the broader interest of this work is phonological development—specifically, the development of speech categories.[1] Discrimination in these tasks is largely seen as a proxy for categorization. However, our growing understanding of how people (largely adults) discriminate speech sounds in experimental tasks raises concerns about whether the mapping between discrimination and categorization is isomorphic. The goal this article is to consider this link more carefully and, by doing so, begin to refine our understanding of the development of speech categories. We start by considering theoretical views on the relationship between discrimination and categorization (the debates around categorical perception) and their implication for infancy. We next consider a classic alternative framing of the relationship between discrimination and categorization in adult discrimination (Pisoni & Tash, 1974). Finally, we evaluate this model with a quantitative review of over 40 years of research on speech discrimination infancy.

Categorical perception as a linking function in infancy

Rarely do behaviors map onto cognitive constructs in a simple, one-to-one manner. As a result, researchers adopt linking functions (explicit or implied) that relate behavior to cognitive constructs. For example, consider an experiment in which a listener hears an ambiguous speech sound. While /b/, /d/, /p/, and /m/ may be partially considered to various degrees, the experiment may only offer a two-alternative /b-p/ decision. In this case, the Luce choice rule (Luce, 1959) can be used to describe how the probabilities across all the possibilities are mapped to the actual probabilities obtained in the experiment. That is, the Luce choice rule serves to link the underlying construct to the behavioral data. While such linking functions are adequate for paradigms like identification, studies of infant speech perception require a more subtle approach, since researchers must infer categorization without the aid of identification measures.

As in many developmental domains, research on infant speech perception has been heavily influenced by adult work. Early investigations of speech perception in adults found a robust relationship between categorization and discrimination (Liberman, Harris, Hoffman, & Griffith, 1957). When asked to label tokens on a continuum from one stop consonant to another, adult listeners showed high agreement for each token, with only small regions of uncertainty at the boundaries. More important, listeners were very good at discriminating speech sounds from different categories but poor at discriminating those from the same category, even for contrasts with the same physical distance. In fact, each participant's discrimination was predicted by their own labeling, suggesting that discrimination may be a useful proxy for identification. Thus, when two sounds are discriminated poorly, they are likely to be members of the same category, and when they are discriminated well, they are likely to be members of separate categories.

For example, consider one commonly studied acoustic cue: voice onset time (VOT). VOT is a continuous temporal cue that marks the time elapsed from the opening of the articulators (the lips or tongue) to the onset of laryngeal voicing (Lisker & Abramson, 1964). VOT distinguishes voiced sounds like /b, d, g/ from their voiceless counterparts /p, t, k/. Voiced sounds tend to have lower VOTs in the 0- to 10-ms range (in English), while voiceless sounds have higher VOTs in the 40- to 70-ms range, and there is a boundary around 25 ms of VOT. If VOT is perceived categorically, VOTs of 0 and 10 ms (both voiced sounds) would be harder to discriminate than VOTs of 20 and 30 ms (which cross the boundary), despite an equivalent acoustic distance in terms of VOT (Liberman, Harris, Kinney, & Lane, 1961).

This principle is described by categorical perception (CP), which refers to the specific relationship between discrimination and categorization. CP has played an important role in our understanding of development by serving as a linking hypothesis. Assessing identification in infancy is difficult (although see Husaim & Cohen, 1981; McMurray & Aslin, 2004), but assessing discrimination is easier. CP thus has offered a linking function to relate discrimination to categorization. Supported by categorical perception, habituation/discrimination measures have fueled a large number of studies on the development of speech categories (see Saffran, Werker, & Werner, 2006, for a review). Eimas et al. (1971) first showed that infants fail to discriminate consonants from the same voicing category but succeed at cross-category discriminations. Since assessing identification in infancy is difficult, Eimas et al. assumed adult category boundaries. These results and numerous replications (Miller & Eimas, 1983; Streeter, 1976; Werker, Gilbert, Humphrey, & Tees, 1981) have demonstrated a pattern of discrimination similar enough to the adult pattern to support CP and sustain the use of discrimination as a means of assessing categorization.

This paradigm enabled a number of important findings in infant speech perception. Werker and Tees (1984) reported that very young infants can discriminate a wide variety of speech contrasts, including some not present in their parental

[1] We use the more neutral term "speech category," rather than "phonological category," here and throughout this article, since phonological category implies an association with language and a particular form of representation that infants may not develop until much later.

language, but lose this sensitivity by 12 months (for reviews, see also Werker & Curtin, 2005; Werker & Lalonde, 1988; Werker & Polka, 1993). A handful of speech contrasts are not discriminable early but are acquired later (Eilers & Minifie, 1975; Eilers, Wilson, & Moore, 1977; Narayan, 2013; Narayan, Werker, & Beddor, 2010; Polka, Colantonio, & Sundara, 2001). In addition, the perceptual structure of adult categories relative to new stimuli predicts which nonnative contrasts are retained or lost past 12 months (Best, McRoberts, & Sithole, 1988). This sample of studies makes it clear that we have learned a great deal using CP as a linking function. Indeed, without it (or something like it), these are simply studies of discrimination—a skill that is, no doubt, important for language processing, but one that lacks the pivotal role of categorization.

However, CP was and is more than a linking function; it also is a theory of perceptual encoding. CP implies that low-level acoustic cues are perceived *in terms of* categories. That is, the categories of a language shape the precategorical representations of sound. Here, we define a cue as a measureable (typically, continuous) property of the acoustic signal that is potentially used by the perceptual system. Thus, CP would posit that continuous cues like formant frequencies are not perceived continuously but that speech categories, to some extent, warp perception at the sensory level, making acoustic differences that cross category boundaries easier to detect than those that fall within category boundaries. Therefore, if speech perception operates on two levels—the encoding of continuous cues and a representation of speech categories—CP suggests that this first level is not a veridical representation of the input but rather, is a nonveridical encoding shaped by the categories of the language. The use of categorical perception in infancy thus entails more than a convenient linking function; it also assumes a particular theoretical description of auditory representation.

Problems with categorical perception

While the use of categorical perception as a linking function has advanced our understanding of early speech perception, as a theoretical account it has faced significant challenges, and this debate has fundamentally shaped our understanding of discrimination. To be clear, none of this work refutes the notions of categories as an important component of speech perception; rather, it challenges whether these categories shape lower-level perception.

Work on adult speech perception has shown that vowels (Fry, Abramson, Eimas, & Liberman, 1962) and, to a lesser extent, fricatives (Healy & Repp, 1982) are perceived noncategorically, undercutting the isomorphy between discrimination and identification. More problematic for infant work, many of the non-English phoneme contrasts, like clicks, retroflexes, and ejectives (crucial to many infant studies; see, e.g., Best et al., 1988; Narayan et al., 2010; Werker & Tees, 1984), have never been tested for categorical perception in adults, leaving it unclear how to interpret discrimination results in infancy.

Early versions of categorical perception, in which listeners *only* discriminate differences that cross phonemic category boundaries, were really never supported. Even Liberman et al. (1957) showed significant within-category discrimination (albeit poorer than between-categories), and subsequent studies have shown measurable within-category discrimination for stop consonants (Carney, Widin, & Viemeister, 1977; Massaro & Cohen, 1983; Miller, 1997; Pisoni & Lazarus, 1974; Pisoni & Tash, 1974). This makes it challenging to directly infer categories from discrimination in infancy.

Theoretically, these findings are easy to rectify with a version of CP in which categories reduce, but do not eliminate, within-category sensitivity. More challenging for CP is a series of studies showing that the apparent warping of discrimination near the category boundary may be largely an artifact of discrimination tasks. Schouten, Gerrits, and Van Hessen (2003; Gerrits & Schouten, 2004) suggested that many discrimination tasks show various degrees of bias (for example, in an ABX task, participants are more likely to choose the A-response), and with unbiased tasks, discrimination is no longer predictable from categorization (see also Carney et al., 1977; Massaro & Cohen, 1983; Pisoni & Lazarus, 1974).

In congruence with these findings, cognitive neuroscience shows that when we measure sensory encoding more directly, there are clearly levels of sensory processing that show no warping, along with components that respond more invariantly to stimulus changes (Frye et al., 2007; Myers, Blumstein, Walsh, & Eliassen, 2009; Toscano, McMurray, Dennhardt, & Luck, 2010). Frye et al. showed that the M1 (an early auditory component detectable in MEG paradigms) reflects continuous changes in VOT. Similarly, Toscano et al. found two ERP components that responded to changes in VOT; the earlier N1 response appeared to represent a veridical encoding of VOT with no warping at the boundary, while the other, later component responded to the stimuli's prototypicality. This second component was not itself a measure of category-level representation but could not exist without the influence of speech categories. Finally, in an MRI study, Myers et al. (2009) found distinct brain regions that either show gradient representation of the speech signal or respond only to acoustic differences that span a category boundary. As a whole, these studies suggest that categories do not warp low-level cue encoding but that category membership and low-level acoustic cue-values are simultaneously available.

Given this emerging consensus, what are we to make about the classic evidence for CP? CP was bolstered, for a time, by findings that both humans and animals show evidence for

discontinuities in discrimination, particularly for cues like VOT. For example, macaques and chinchillas (Kuhl & Miller, 1975; Kuhl & Padden, 1982) show a boundary around 20 ms for English voicing contrasts. However, more recent work has demonstrated that many of these findings could be due to range effects (Ohlemiller, Jones, Heidbreder, Clark, & Miller, 1999); when animals are trained on a shifted range of stimuli (e.g., 10–50 ms, rather than 0–40), their boundary also shifts. Humans do not show this effect to the same degree. Thus, it is not clear that the animal work makes a strong case for discontinuities at these particular VOTs; this may be the result of the particular training ranges used.

Similarly, speakers of languages with different boundaries than English are reported to show enhanced discrimination at the same psychophysical discontinuity. For example, Spanish speakers (Williams, 1977), whose VOT boundary is around −20 ms, also show enhanced discrimination at 20 ms. However, this may not necessitate an auditory discontinuity; Spanish speakers could be discriminating good exemplars of their categories from poor ones (cf. Miller & Volaitis, 1989, which studied extended VOT continua). This case was also classically bolstered by infant data showing that very young infants (who may not have robust categories) show heightened discrimination around 20 ms of VOT (1971). This is the topic of the present article.

While, again, none of these lines of work deny the importance of some form of category-level representation, they do not offer clear evidence for a perceptual warping. Thus, together, the adult and animal work, like the previously mentioned neuroscience work, suggest a model of speech perception in which both low-level acoustic variation and speech categories are present and can shape perception. Nonhumans (who we assume lack speech categories) discriminate VOTs on the basis of the presented range, while humans show additional effects of category membership. However, at the same time, these lines of work offer little clear evidence for a warped encoding at the perceptual level.

Moreover, theoretical views of adult speech perception suggest that the sensitivity to fine-grained detail may be fundamental to word recognition. In adult speech perception, such differences affect lexical processing (Andruski, Blumstein, & Burton, 1994; McMurray, Tanenhaus, & Aslin, 2002), help listeners anticipate future sounds (Gow, 2001, 2003; Gow & McMurray, 2007; Martin & Bunnell, 1981), and resolve ambiguous prior sounds (Gow, 2003; McMurray, Tanenhaus, & Aslin, 2009b). Such detail is also necessary for a variety of models to cope with factors like talker variability and coarticulation (Fowler & Smith, 1986; Goldinger, 1998; McMurray & Jongman, 2011; Smits, 2001), making it an integral part of word recognition. Thus, if cue representations were warped by categories, this would eliminate substantial information that listeners could use to cope with the variability in speech.

For infants, however, the picture is less clear. It may actually be beneficial for infants to use a more coarsely coded representation of the signal when their goal is not to utilize well-developed speech categories but to acquire them. However, as we describe next, this strategy does not fit well with many theoretical accounts of development.

Implications for development

The assumption that speech categories strongly influence sensory encoding in infancy is difficult to reconcile with our growing understanding of perceptual development. CP fundamentally implies a partial loss of fine-grained continuous information at the level of cue encoding. However, to account for much of what we know about the development of speech perception, there must be considerable plasticity in the system and considerable sensitivity to fine-grained detail. This can be seen in three important avenues of research.

First, a number of empirical studies (Cristia, 2011; Maye, Weiss, & Aslin, 2008b; Maye et al., 2003) and theoretical accounts (de Boer & Kuhl, 2003; McMurray, Aslin, & Toscano, 2009a; Toscano & McMurray, 2010; Vallabha, McClelland, Pons, Werker, & Amano, 2007; Werker et al., 2007) suggest that infants acquire the categories of their language, in part, through statistical learning based on the distribution of acoustic cues in the environment. That is, at some level, infants count the occurrence of individual cue values (e.g., individual VOTs) and use these to estimate categories. This requires infants to perceive subtle differences between tokens along a continuum. This is because it would simply not be possible to count individual VOTs if infants were unable to differentiate them.

Second, speech categories are tuned continually throughout late infancy and early childhood as children learn to combine and use multiple acoustic cues (Dietrich, Swingley, & Werker, 2007; Galle, Apfelbaum, & McMurray, in press; Nittrouer, 1992, 1996; Nittrouer & Miller, 1997; Rost & McMurray, 2009) and learn which are relevant for discriminating words. This is true even for acoustic cues like VOT that appear to be perceived "categorically" very early in infancy (Bernstein, 1983; Galle et al., in press; Rost & McMurray, 2009). This is difficult to rectify with CP, for two reasons. First, to integrate cues across dimensions, listeners must be sensitive to structure within each dimension. For example, VOTs of 10 ms show stronger effects of speaking rate than do VOTs of 0 ms, despite the fact that both are clearly voiced. By down-weighting within-category differences, CP hampers cue integration. Second, this perceptual weighting process may, in part, rely on statistical information (Apfelbaum & McMurray, 2011; Toscano & McMurray, 2010), which, as we described, may be less available if CP eliminates within-category detail.

Finally, and perhaps more important, given the evidence that adults appear to code cues like VOT veridically, if infants started with a "warped" representation of such cues, they would somehow have to unlearn this. No current theories of development provide a mechanism for this.

Together, these insights favor a theory of development in which infants maintain sensitivity to fine-grained detail for within-category contrasts and use this to construct multidimensional speech categories slowly. This is difficult to rectify with a model in which, during early infancy, infants have a warped perceptual encoding of speech cues.

Toward a "new" linking function

If categorical perception does not describe adult perception and is theoretically problematic for development, how then do we explain the decades of infant discrimination research that have concluded in its favor? And what sort of linking function might we adopt to help interpret infant discrimination data?

One possibility is that the manner in which CP is typically assessed in infancy, via measures of habituation/discrimination, is biased (in the psychophysical sense). If the much more sophisticated discrimination tasks that can be employed with adults show various forms of bias (cf. Massaro & Cohen, 1983; Pisoni & Lazarus, 1974; Schouten et al., 2003), it seems likely that the much coarser grained infant measures of discrimination may as well. Indeed, recent work using other variants of the typical infant paradigms has shown some evidence for a more gradient mode of phonetic perception (McMurray & Aslin, 2005; Miller & Eimas, 1996).

What is needed, then, is not necessarily a new theory of speech perception, but a model of performance in speech discrimination tasks. By understanding what infant discrimination tasks are really measuring, perhaps we can make some inferences about both perceptual encoding and categorization in infancy. Such a model must capture both sensitivity to continuous acoustic detail (within-category differences) and the apparent heightened discrimination at category boundaries. A critical question that such a model could help answer is whether both findings are possible with a veridical encoding at the level of auditory cues, without any sensory warping.

A complete model of infant discrimination must necessarily include the complex contributions of habituation and dishabituation. While there are examples of this in vision (Hunter & Ames, 1988; Perone & Spencer, 2013), this may be more difficult for speech perception given that, in audition, we must always use some form of operant conditioning to measure listening time. In audition, there is no direct measure of preference (like eye movements in vision), and as a result, infants must learn to suck a pacifier or turn their head to indicate their interest. Given the added difficulty of modeling habituation/dishabituation in the context of some kind of operant learning, a complete model of discrimination may thus be out of reach.

However, if we abstract away from the mechanics of habituation, there is still insight to be gained from considering the problem of discrimination in a new light. Pisoni and Tash (1974) provide a useful model of discrimination that can be applied to infant work. Their model assumes that speech cues are encoded continuously and veridically and are mapped onto discrete categories. That is, perception is not categorical, but there are categories that are identified from a veridical encoding of the signal. Both levels of representation exist concurrently. Under this view, categorization can occur on the basis of boundaries, prototypes, or even exemplars (the traditional categorization metrics of cognitive psychology) that are applied to this veridical representation at the cue level (without necessarily modifying it).

Crucially, discrimination judgments can be made on the basis of both levels of representation. That is, differences (or sameness) in both speech categories and continuous sensory detail are combined to decide whether two stimuli are equal. Congruent decisions, where the sounds both are acoustically different and lie in different categories (or where sounds are acoustically identical and, hence, lie in the same category) lead to better discrimination (short RTs in adults) than do incongruent decisions, since the decision is supported by both levels of representation. In contrast, if only one level of representation supports discrimination (e.g., sounds that are acoustically different but in the same category), discrimination is harder. For example, VOTs of 0 and 15 (which both lie in the same category) offer only weak evidence for discrimination (and hence, a longer RT or less accuracy), while VOTs of 15 and 30, which are the same distance but in two different categories, offer evidence for discrimination at both the cue and category level. Thus, this type of model can account for heightened discrimination at the boundary as an additive effect. However, it does so without positing any perceptual warping of any kind at the cue level—that is, without positing CP.

While the Pisoni and Tash (1974) model was initially posed primarily as a model of reaction time, this model can also explain the results of discrimination tasks in infancy if we consider dishabituation as a measure of discrimination. In this adapted model, which we term the *parallel channels model* as a convenience (although credit belongs to Pisoni and Tash, 1974), listeners are more likely to dishabituate when speech sounds are both acoustically distinct and members of separate categories and are less likely to dishabituate when they are only acoustically distinct.

A critical aspect of this approach is that different tasks may cause listeners to weigh the contributions of each channel differently. For example, discrimination requires that participants maintain at least one speech token in memory, and the parallel channels model suggests that this

may require multiple stores (sensory and categorical). However, task demands could make maintenance more difficult (for example, an ABX task requires two tokens) or make the more fragile sensory store decay faster. In this case, subjects may only be sensitive to phonemic (between-category) differences, since they are easier to maintain in memory than fine-grained differences. In fact, a number of studies that control these demands find that discrimination between and within categories is equivalent (Carney et al., 1977; Gerrits & Schouten, 2004; Massaro & Cohen, 1983; Pisoni & Tash, 1974). This suggests that while categories are present in the system, their effects lie not on perception, but on short-term memory and discrimination processes. This is clearly applicable to infancy, since the wide variants of tasks in use could shape both memory demands and bias toward one channel or the other.

Such a model fits the finding that between-category discrimination often exceeds within-category discrimination. It does so by allowing categories to participate in discrimination, but without positing any change in perception. However, this model also predicts that within-category differences can also lead to discrimination, albeit at a lesser rate. Moreover, since cue encoding is still fundamentally veridical in this model, it also predicts that the magnitude of the perceptual difference should correlate with the magnitude of discrimination (over and above the heightened between-category discrimination), something that has not been observed (or tested) in any infant study.

Of course, one could argue that evidence for within-category discrimination is consistent with any but the most extreme versions of CP. While this is true, the parallel channels model suggests that both heightened between-category sensitivity and some nonzero within-category sensitivity are not unique evidence for CP (as a theoretical account) but are equally consistent with a completely veridical encoding of the speech signal. Thus, while we cannot rule out CP as a mode of infant perception, if the parallel channels model holds, we also find ourselves lacking any strong evidence for CP (absent evidence for complete failure to discriminate tokens within category). This then provides a way to rectify the lack of evidence for CP in adults with the theoretical benefits of a more gradient mode of perception in infancy.

Finding evidence for within-category discrimination

To evaluate this model, we must address three questions. First, we must evaluate whether infants are sensitive to within-category differences at all. Clearly, both classic categorical framings of infant speech perception and a parallel channels approach predict good between-category discrimination, and the evidence for this is consistent. Where these accounts differ is in their sensitivity to within-category differences and the factors that can affect this. Most studies have not found such sensitivity, and only two have reported this as a primary finding (McMurray & Aslin, 2005; Miller & Eimas, 1996). However, even within these, McMurray and Aslin (2005) reported within-category discrimination in only two of three conditions, and Miller and Eimas (1996) in only half of theirs. The difficulty in observing this may be due to methods used with infants: When infants hear a baseline stimulus many times before the contrasting stimulus, the repetition may lead to adaptation or cause infants to become inattentive.

Second, we must ask whether this sensitivity is related to the magnitude of the difference between tokens. Normally, this would require a design similar to that in Carney et al. (1977), who tested multiple step sizes with adults. Here, if there is a veridical encoding at the level of acoustic cues, we should see increasing discrimination as the distance between the tokens increases. However, with infants, it is not easy to test more than a few tokens. Finally, we must examine whether task variants modulate the degree to which category- or cue-level differences drive discrimination. Again, this would be best done in the context of a study that uses multiple tasks simultaneously (as Gerrits & Schouten, 2004, did with adults), but this is difficult with infants.

A meta-analytic approach may offer a more sensitive way to address these issues. Meta-analyses of existing studies can often pull out small effects that are not significant or meaningful in individual studies but recur across many studies. They also allow us to pool conditions across studies, so that even if any individual study examined only one or two VOT contrasts, across studies we may examine more. Finally, given the variation in the tasks used to assess discrimination and the parallel channels model's sensitivity to task demands, we may be able to pull out information about the task dependence of these findings.

Such an analysis requires a speech contrast with a common unit of measurement across studies. This is difficult with dimensions like place of articulation, where there is no universally agreed-upon unit. However, this is not the case for the phonetic dimension of choice in infant speech discrimination, VOT. VOT is widely used in infant speech studies for several reasons. It can be manipulated easily in synthetic and natural speech. Voicing distinctions are found in the many languages, although the location and number of categories varies (Lisker & Abramson, 1964). Finally, VOT shows evidence of CP in adults (Liberman et al., 1961), infants (Eimas et al., 1971), and even animals (Kuhl & Miller, 1975; Kuhl & Padden, 1982). Thus, VOT has become a model speech cue for studying a range of issues in speech categorization and development.

VOT is also an ideal platform for assessing discrimination across studies because it refers to an absolute measure. Many infant (and adult) studies make use of speech continua, a series of tokens that gradually change from one sound to another.

However, for continua like place of articulation, such changes are describable only relative to the endpoints (the unambiguous tokens), making it difficult to compare stimuli across experiments or even across different continua within an experiment. In contrast, VOT is meaningful over and above the particular set of tokens it is instantiated in. For example, the meaning of "step 3" on a b/d continuum is not universal, but there is no ambiguity about what a 15-ms VOT refers to across different vowels or places of articulation. While other cues like *f*0, *F*1, vowel length, and aspiration amplitude also contribute to voicing perception (Ohde, 1984; Summerfield & Haggard, 1977; Toscano & McMurray, 2012) for most languages, VOT is widely seen as the most important cue for voicing distinctions, and as a result, the majority of studies that have investigated voicing discrimination in infancy have used VOT exclusively.

Using VOT, we can attempt to look beyond any individual study to see infants' performance on an entire speech continuum. Thus, the present study collected the results of every study to date that investigated infant VOT discrimination (see also, Narayan, 2013, for a similar, although more limited, analysis). This affords a unique opportunity to answer several questions. First, we can investigate the effect of presumed category membership on perception by assessing the presence and location of peaks in discrimination that may correspond to boundaries. Second, and consistent with a continuous encoding of cues like VOT, we wanted to know whether infants' discrimination abilities are at least partially based on raw acoustic differences between VOTs. Third, there may be methodological factors that would validate the prediction that task differences can change the discrimination profile.

Finally, the development of speech perception (particularly consonants) is commonly discussed in what we term an overgeneration and pruning framework (Aslin & Pisoni, 1980; Aslin et al., 2002; Werker & Tees, 1984). This proposes that infants start life with the ability to discriminate a wide variety of speech contrasts and gradually lose sensitivity to contrasts that are not in their native language (although perhaps not completely). This has been assessed frequently with nonnative contrasts, but we can also examine this in within-category VOT contrasts.

General method

Twenty papers were identified that investigated VOT discrimination in infants under 12 months. This was compiled using a wide number of search terms in the PsycInfo, Linguistics and Language Behavior Abstracts, and PubMed databases. This set represents, to the best of our knowledge, every study examining VOT discrimination in infancy since Eimas et al.'s (1971) initial study (Table 1). A handful of studies were excluded because they used a training procedure to alter speech perception (Maye, Aslin, & Tanenhaus, 2008a; Maye et al., 2003) or did not include a control condition that would allow us to determine whether individual pairs of VOTs were discriminable (Hoonhorst et al., 2009).

For each study, a uniform description of the methods and results was coded. Our primary unit of analysis is the *condition*, which we define as any VOT contrast (pair of VOTs) for which discrimination was assessed over a group of infants. For example, Experiment 1 of Eimas et al. (1971) habituated infants to a 20-ms VOT and tested them on 0 and 40 ms. This experiment thus had two conditions (0 vs. 20 ms and 0 vs. 40 ms). From the 17 papers in our data set, we identified 220 conditions with approximately 10 babies per condition ($M = 9.99$; Table 1).

Eight attributes of each condition were identified, including experimental method (e.g., conditioned head-turn, visual/auditory habituation, etc.), mean subject age, and method of stimulus construction (see Table 2 for a complete list). Individual attributes will be discussed in depth in the relevant analyses. Our primary dependent measure was whether or not infants discriminated the contrast being tested at a statistically significant ($p < .05$) level. This was determined on the basis of group data and was coded as one if discrimination across the group of subjects (within a condition) was statistically significant and as zero otherwise.

From this data set, we conducted four analyses. The first two asked whether discrimination is related to both continuous distance in VOT and the presence of categories. The third focused on discrimination within categories to ask why within-category discrimination has been so hard to observe and to examine how differences in task and stimuli can change infants' apparent sensitivity to these differences. Finally, the fourth analysis examined the course of development.

While many meta-analyses pool values such as Cohen's *d* or correlation coefficients to arrive at inferential statistics, we did not use this approach for several reasons. First, there was not always a common dependent measure across experiments. This was largely due to methodological differences: Visual habituation, for example, is based on differences in looking time, while conditioned head-turn is based on the proportion of head-turns or, sometimes, the number of infants who learned to criterion. Since some of these dependent variables are not even on the same scales (e.g., linear looking time vs. proportional head-turning), it was not straightforward to construct a single measure of effect size. Thus, we simply assigned a 1 or 0 to each condition corresponding to whether the group of babies showed evidence for discrimination. This discards any within-condition variance, so while it is a good measure of group-level performance (particularly across many studies), it is not sufficient for a proper statistical analysis. Second, there were significant confounds for many of the

Table 1 Papers included in analysis (ordered chronologically)

Study	No. of conditions	Average infants/ condition
Eimas, Siqueland, Jusczyk, & Vigorito (1971)	6	8
Trehub & Rabinovitch (1972)	3	10
Lasky, Syrdal-Lasky, & Klein (1975)	5	10
Eilers, Wilson, & Moore (1977)	8	8
Streeter (1976)	3	21.3
Eilers, Gavin, & Wilson (1979)	4	8
Eilers, Wilson, & Moore (1979)	8	8
Aslin, Pisoni, Hennessy, & Perey (1981)	70	5.6
Werker, Gilbert, Humphrey, & Tees (1981)	1	15
Miller & Eimas (1983)	4	15
Jusczyk, Rosner, Reed, & Kennedy (1989)	8	12
Burnham, Earnshaw, & Clark (1991)	14	15
Miller & Eimas (1996)	6	16
Litwin (1998)	10	7.2
McMurray & Aslin (2005)	12	24
Rivera-Gaxiola, Silva-Pereyra, & Kuhl (2005)	4	14
Burns, Yoshida, Hill, & Werker (2007)	12	11.8

factors we studied. For example, the earliest studies focused on infants younger than 6 months and used synthetic speech with habituation. In contrast, newer studies used a mixture of methods, natural speech, and older infants. Given this unbalanced data set, we were not confident in the conclusions that could be obtained by a single multivariate analysis. Thus, we adopted a hypothesis-testing style of analysis, proposing specific questions in each section, presenting a descriptive (although quantitative) analysis, and then progressively burrowing through confounds. This, then, is more of a quantitative literature review than a true meta-analysis, but given the nature of this data set, this seemed more appropriate.

Analysis 1

Most infant studies assume adult boundaries when assessing discrimination, but few verify these boundaries in infancy, since this requires both a measure of identification and the ability to test many contrasts. While one could use the location of a peak in discrimination as a proxy (since it is predicted by both CP and the parallel channels account), few studies include enough contrasts to gauge this with any accuracy. However, because our data set consisted of dozens of VOT contrasts (across several studies), we had the unique opportunity of assessing infants' voicing boundary via discrimination. Crucially, we cannot assume that a boundary reflects differences in sensory encoding, since it could arise without it in the parallel channels model. However, it is important to look for such peaks in discrimination to help refine our understanding of other factors. Thus, Analysis 1 asked whether there is evidence for heightened discrimination near perceptual boundaries along the voicing continuum in infancy.

On the basis of phonetic and perceptual work with English-speaking adults, we expected heightened discrimination near 20–30 ms of VOT. We examined the rate of discrimination as a function of the mean VOT for the two test tokens of each condition (VOT location). A peak in discrimination at a VOT location indicates that when speech stimuli fall on either side of this VOT location they are easier to discriminate. If infant speech discrimination is affected by category membership alone, we should find a very strong and narrow peak of discrimination centered near the adult VOT boundary. However, if discrimination is affected by both low-level

Table 2 Attributes that were coded for each condition

Name	Description	Possible values
Method	Methodology used to assess discrimination	Head-turn preference procedure, cardiac deceleration, visual–auditory habituation, event related potential, conditioned head-turn, high-amplitude sucking
Language background	Native language of the subjects	English, Spanish, English and Spanish (bilinguals), Hindi, Kikuyu
Stimulus type	Either naturally produced or synthetically generated	Natural or synthetic
VOT distance	Distance in milliseconds between the VOT for familiarization and test stimuli	Between 0 and 250 ms
VOT location	Mean in milliseconds of the VOT for familiarization and test stimuli	Between −100 and 200 ms
Age	Average age in months for the subjects	Between 0 and 15 months
Place of articulation	Feature of speech determined by the position of the articulators during production	Bilabial, alveolar, velar, and dental
Boundary type	Which boundaries, if any, the VOT contrast spans	Within prevoiced, between prevoiced and voiced, within voiced, between prevoiced and voiceless, between voiced and voiceless, within voiceless

acoustic cues and category membership, we should observe a wide, graded peak centered near the adult VOT boundary.

Method

Of the 220 total conditions, 178 were included in this analysis. This included 22 conditions in which the infants' native language was not English. These conditions were included because we wanted to examine the effect of perceptual distance independently of categories. The remaining 42 control conditions were excluded because the two VOTs were identical. VOT location was calculated as the mean VOT between training and test stimuli in each condition, regardless of the distance between the two tokens. So, for example, a contrast between 25 and 45 ms would be assigned a location of 35 ms, as would a contrast between 5 and 65 ms. These VOT locations were binned into 10 ms increments centered at 5 ms (so that a +5-ms bin refers to 1–10 ms VOTs, a 15-ms bin to 11–20, and so on).

Results

Figure 1 shows the proportion of conditions reporting discrimination as a function of VOT location. Two locations of heightened discrimination can be seen: one centered at 25 ± 5 ms, and a second at 135 ± 5 ms. The peak at extremely high VOTs does not correspond to any known voicing boundary. However, it is based on only a handful of conditions in a single study (Miller & Eimas, 1996), whose authors suggest that it reflects discrimination across the right edge of the category (e.g., from a good /p/ to a poor one). Thus, this peak was not considered further.

The strongest evidence for a peak is seen at 25 ± 5 ms, which roughly corresponds to the voiced/voiceless boundary of English. However this may be an oversimplification. While discrimination was better at this location than at any other, there is a relatively large region of heightened discrimination between 0 and 50 ms, which seems more consistent with a graded category than with one discrete boundary. This broad region of heightened discrimination may be due, at least in part, to the manner in which we have determined VOT location. If the adult boundary for voicing truly lies at 25 ms of VOT (even though we know that it varies with regard to several factors), then all of the contrasts labeled as having a 25-ms VOT location will cross this boundary, many of the 15- and 35-ms VOT locations will include contrasts that cross this boundary, and relatively few of the more peripheral VOT locations will cross the boundary, resulting in what appears to be a graded discrimination function. This will be addressed in the next analysis, where we partial out conditions relative to this boundary.

An important concern is the possibility of variability in the boundary across participants and/or items in these studies.

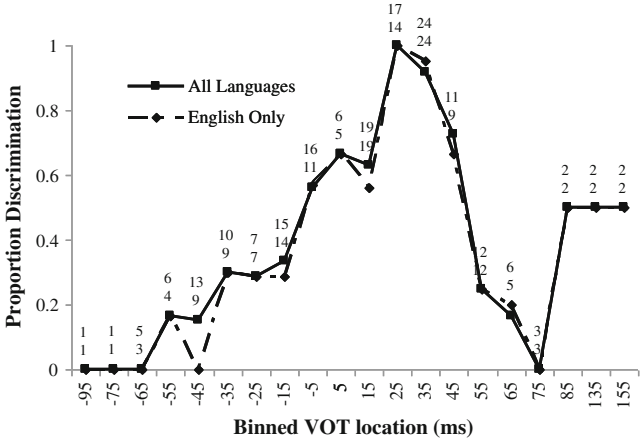

Fig. 1 Proportion of conditions showing discrimination as a function of binned voice onset time (VOT) location for all experimental conditions

Here, some studies may test participants or items with slightly different boundaries, which we have no way of documenting. This may be due to variable secondary cues across studies or to regional or other differences among the infants. In any case, if the boundary is variable across studies, this broader region of discriminability could derive from variation in the boundary across studies (e.g., the average of a lot of sharp peaks with different locations would be a relatively broad peak). We conducted extensive analyses to rule this out, in which we eliminated any condition with a token within 10 ms of this boundary, and found no substantive changes in this or the other analyses (see Online Supplement S1 for more details). This functionally permits variability of up to 20 ms in the boundary (more than is typically observed within any study). Thus, while we cannot conclusively rule out that the broad region of discriminability derives from an averaging artifact, it seems unlikely.

This 25-ms peak also ignores the possible effects of context and secondary cues. It is well known that the voicing boundary is both context dependent and built on multiple cues. It can differ as a result of factors like speaking rate, speaker differences, pitch and formant frequency, and place of articulation. Evidence of sensitivity to such factors is scant in infancy, although two studies suggest that young infants can use multiple cues and integrate context (Fowler, Best, & McRoberts, 1990; Miller & Eimas, 1983). While of these, only Miller and Eimas examined voicing, we cannot rule out such effects. Many of the factors that may have influenced the VOT boundary in these studies are unavailable in our data set (e.g., speaking rate, formant frequencies). And we were unable to assess the one factor that is available to us (place of articulation), due to a lack of nonbilabial conditions within our data set ($N_{\text{bilabials}} = 158$, $N_{\text{alveolars}} = 15$, $N_{\text{velars}} = 4$, $N_{\text{dentals}} = 1$). Therefore, while this analysis clearly shows heightened discrimination in the region of 25 ± 5 ms, we do not know whether this is a fixed boundary or something more context sensitive. However, it is important to point out that in adult work, such factors rarely shift the boundary by more than

10 ms; thus, our analysis excluding tokens that approached the boundary (Supplement S1) in some ways accounts for this.

Thus, while we cannot rule out a discrete boundary from this analysis alone, there is little evidence to support it, and the presence of a graded discrimination function around the boundary is a better fit with the existing literature on adult phoneme categorization (Andruski et al., 1994; Carney et al., 1977; McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008; McMurray et al., 2002; Miller, 1997). This is quite consistent with the veridical encoding of VOT posited by the parallel channels model. But more important, the strong peak centered at 25 ms (the voicing boundary for the bulk of the infants studied here) also supports the influence of categories in the parallel channels model.

Given this discrimination peak, we next considered the role of category membership. Figure 2a provides a summary of the effect of discrimination as a function of category membership (ignoring other factors), or *contrast type*. We used a +25-ms VOT boundary as the boundary between voiced (short-lag) and voiceless; we also considered a 0-ms VOT boundary[2] as potentially dividing prevoiced and voiced (as in languages like French or Hungarian). This yielded six contrast types: *within-prevoiced*, *between prevoiced/voiced*, *within-voiced*, *between voiced/voiceless*, *between prevoiced/voiceless*, and *within-voiceless*. We further subdivided conditions by language background of the infants.

The best discrimination was observed for the between voiced/voiceless contrast (93.5 %; English-only: 92.3 %). Of the remaining contrast types, within prevoiced had the lowest rate of discrimination (9.4 %; English-only: 3.3 %), and prevoiced/voiced (38.5 %; English-only: 36.1 %), within-voiced (28.5 %; English-only: 28.5 %), and within-voiceless (35.9 %; English-only: 34.3 %) contrasts had intermediate discrimination. As the bulk (85.9 %) of our conditions tested monolingual English-listening infants, infants are responding in a way that is largely appropriate to their language. Moreover, since discrimination across the prevoiced/voiced boundary did not differ from other within-category contrasts, there is not robust evidence favoring heightened discrimination at this boundary. This is quite consistent with the role of categories in discrimination posited by the parallel channels model.

In addition to the clear evidence of between-category discrimination, this analysis offers substantial evidence for discrimination of tokens that fall *within* categories boundaries, with over a third of both within-voiceless and within-voiced contrast conditions showing discrimination (despite only two studies highlighting this finding). Of course, one possibility is that, as was discussed above, variability in the location of the boundary across participants and items is driving this within-category effect. To control for that, we again excluded any condition in which either member of the contrast was within 10 ms of the boundary. These results are presented in Fig. 2b and show an almost identical pattern.

To summarize, there is robust evidence for heightened discrimination around +25 ms of VOT (although the region of heightened discrimination is broad) and little evidence for any other peaks in the discrimination function. There is also uncertainty as to the exact form of this boundary, since at least in adults, it is quite context dependent, but we could not assess this in our data set. When a boundary of +25 ms is used to classify conditions, we see a robust effect of category membership, with higher rates of discrimination for between-category contrasts. Yet there is still evidence for the effect of acoustic differences for within-category contrasts, and this holds using even a very conservative criterion for what constitutes within-category.

Analysis 2

Analysis 1 indicated a relatively high rate of discrimination for within-category contrasts. However, it is yet unclear whether the rate of within-category discrimination represents a uniform baseline (all within-category discriminations are roughly equal at around 30 %) or whether as is predicted by the parallel channels account, discrimination is a function of the magnitude of acoustic difference. Thus, we investigated both *within-* and *between-category* discrimination (again assuming the 25-ms boundary) as a function of the distance between tokens.

Method

As in previous analyses, 42 control conditions were excluded. Because infants in non-English environments may have different boundaries and the majority ($N = 162$ of 178) of the conditions used English-learning infants, we considered these condition to be more confident only in the 25-ms boundary. Our independent variables were *VOT distance* and *contrast type*. VOT distance was the difference in VOT between the tokens presented during the training phase and those presented in the test phase. These were binned into increments of 10 ms.[3] Contrast type was coded as two values: between or within category based on the 25-ms boundary of Analysis 1. This led to 108 within-category contrasts and 54 between-category contrasts.

[2] Of course, a 0-ms boundary is a purely perceptual construct, since it is nearly impossible for the articulatory system to actually produce a VOT of 0 ms.

[3] No attempt was made to distinguish the direction of comparison. So, for example, a condition in which infants were habituated to 20 ms and tested on 40 ms had an equivalent 20-ms distance to one in which infants were habituated to 40 and tested on 20. While this may be important (Miller & Eimas, 1996), most studies used only a single direction, and several methodologies could not be mapped clearly onto this construct.

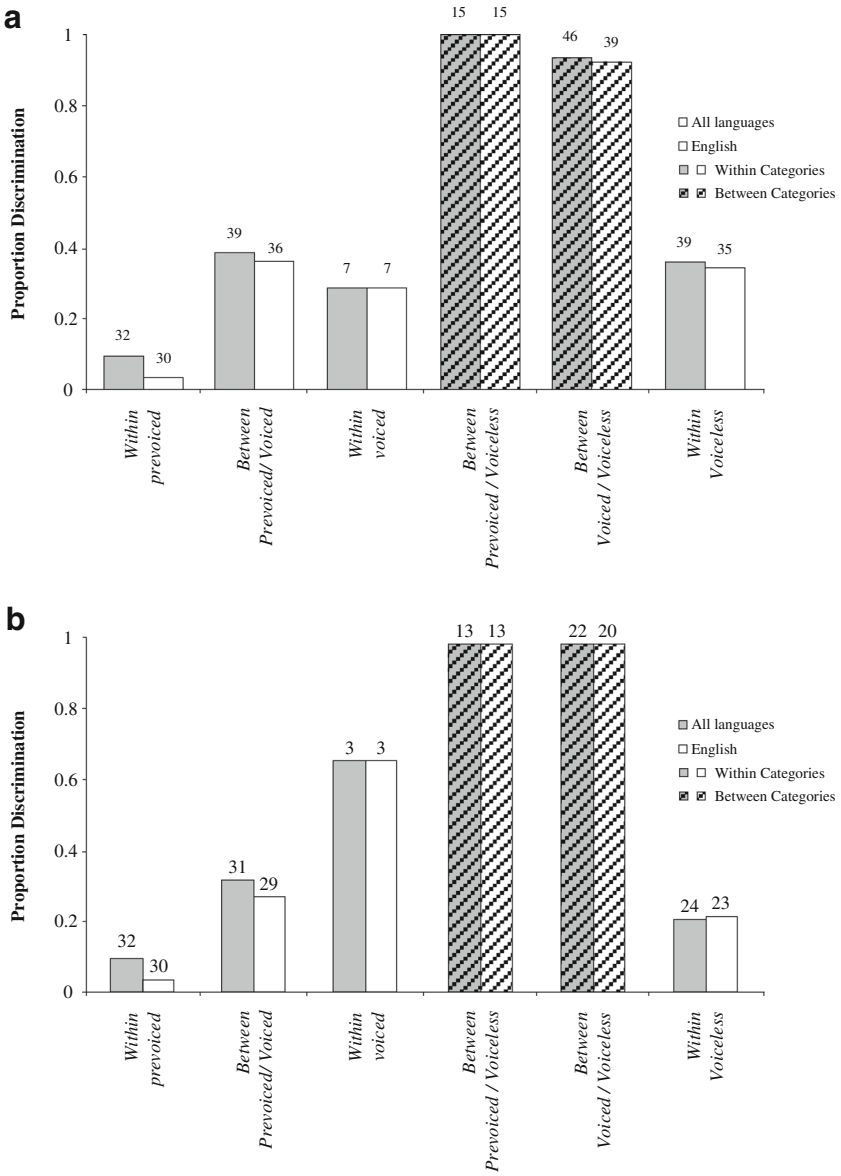

Fig. 2 Percentage of conditions showing discrimination as a function of boundary type. Numbers indicate the total number of conditions in each category. **a** Category membership determined with 25-ms voicing boundary. **b** Category membership determined with conservative boundary

Results

Figure 3 shows the proportion of conditions showing discrimination as a function of VOT distance and contrast type. As before, we found a prominent effect of contrast type, with between-category contrasts having a much higher rate of discrimination across all VOT distances than the within-category contrasts. The 20- to 40-ms distances, in particular, showed large differences as a function of contrast type; not surprisingly, these distances were most commonly used during the heyday of CP. Nonetheless, there is a trend for within-category discrimination to increase with VOT distance, and at the smallest distances (<50 ms), both within- and between-category discrimination suffers. Most important, for within-category conditions, there is a steady, albeit more shallow, rise in discrimination.

Two points violated this trend. At a VOT distance of 80 ms, we found unexpectedly poor discrimination; but this was the result of only one condition. At a VOT distance of 10 ms, we found unexpectedly good discrimination. However, this included seven conditions using natural speech (more than any other data point), which, as Analysis 3 demonstrates, results in more sensitive discrimination. When all of the conditions utilizing natural speech are excluded, the within-category rate of discrimination shows a much more clearly linear effect of VOT distance (Fig. 3b). Use of the more conservative approach (in which tokens within 10 ms of the boundary were excluded) did not change the results (Supplement S2). This

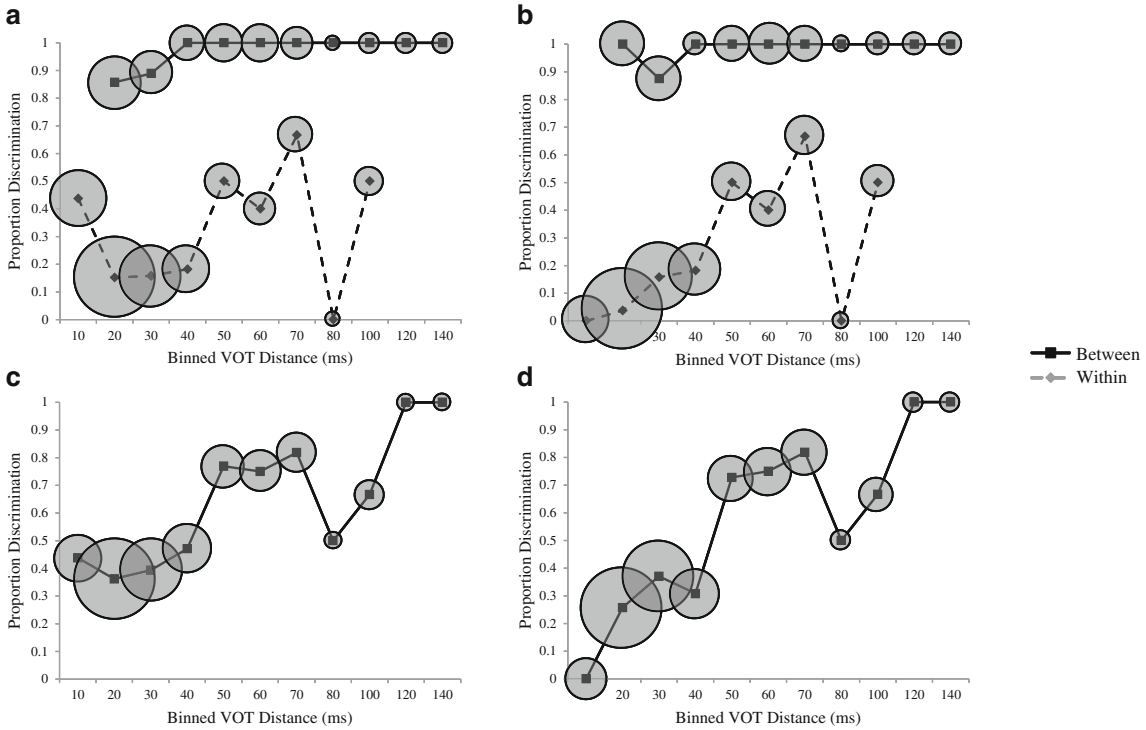

Fig. 3 Proportion of conditions showing discrimination as a function of voice onset time (VOT) distance between speech tokens. The size of the bubble around each data point indicates the number of contributing conditions. **a** All conditions as a function of contrast type. **b** Only conditions using synthetic speech tokens as a function of contrast type. **c** All conditions collapsed across contrast type. **d** Only conditions using synthetic speech tokens collapsed across contrast type

cautions against the assumption that discrimination is driven solely by categorization: Although infants discriminate between-category differences better than within-category ones, low-level acoustic/perceptual factors (VOT distance) also affect it.

The striking gradiency in within-category responding leads us to reconsider these data from a new perspective. In particular, we asked what these results would look like if we did not assume that infants had speech categories at all (perhaps the most basic assumption, even as it seems counterintuitive). Figure 3c and d show discrimination rate as a function of VOT distance collapsed across contrast type and suggest a striking linearity. While this should be interpreted in light of decades of work supporting some form of categories in infancy, they helpfully illustrate what the data would have looked like without such assumptions.

Finally, in assessing the relationship between VOT distance and contrast type, it is important to consider the number of studies examining different VOT distances and contrast types to determine whether there are any confounds. Across studies, there was a decided skew in favor of CP-like effects. The average VOT distances typically used to test for between-category contrasts (M = 54 ms, SD = 41 ms) are much larger than those used for within-category contrasts (M = 33 ms, SD = 22 ms). This methodological difference may derive from researchers' uncertainty as to the exact boundary. Because the VOT boundary is variable, many researchers use large between-category contrasts to ensure they cross the boundary. For example, assuming a 25-ms boundary, rather than assessing discrimination for 20 and 30 ms of VOT (a 10-ms VOT distance), researchers might compare 10 and 40 ms to account for uncertainty in the exact boundary. Conversely, researchers testing within-category discrimination may test smaller differences to be sure that they are truly within the category (as in our conservative approach) and to ensure the maximum impact of their findings (discriminating small differences is more noteworthy than discriminating large differences). This is reasonable, but in the context of the literature as a whole, it stacks the deck in favor of between-category discrimination in this data set. Thus, our estimate of roughly 30 %–40 % of conditions showing within-category discrimination likely underestimates infants' abilities to discriminate within-category contrasts, especially given evidence like this suggesting the importance of VOT distance in predicting discrimination performance.

Together, Analyses 1 and 2 indicate that the majority of studies investigating contrasts spanning the presumed 25-ms boundary show discrimination (N_{cond} = 51/54, 94.4 %). In contrast, within-category discrimination was less robust (N_{cond} = 28/108, 26 %; Fig. 2a), but still quite substantial,

and grew as a function of VOT distance. However, these studies spanned a variety of stimulus types and experimental methods, raising the possibility that some approaches may be more sensitive to small differences.

Analysis 3

The broad discrimination peak and within-category discrimination of Analysis 1, coupled with the effect of VOT distance in Analysis 2, suggest that within-category discrimination is more widespread than previously thought. Moreover, work on adult speech discrimination highlights the role that task and stimulus variables play in predicting within-category discrimination (Carney et al., 1977; Gerrits & Schouten, 2004; Massaro & Cohen, 1983; Pisoni & Lazarus, 1974; Pisoni & Tash, 1974; Schouten et al., 2003). Thus, Analysis 3 investigated the effects of several factors on within-category discrimination.

Method

As in Analysis 2, conditions were classified on the basis of category boundaries, so we only examined conditions that tested English-learning infants. Since we were interested only in within-category discrimination, we limited our analysis to the 149 within-category conditions, assuming a 25-ms boundary. Three factors were examined: *stimulus type*, *experimental method*, and *VOT distance*. With respect to stimulus type, we had hoped to investigate whether the particular method of VOT synthesis (e.g., the acoustic parameters that were manipulated) had any effect on rate of discrimination. However, among the 137 conditions using synthesis, all but one of the studies included in this analysis used the *F*1 cutback method (Aslin, Pisoni, Hennessy, & Perey, 1981, replaced pitch pulses with aspiration using the Klatt synthesizer). Therefore, stimulus type simply was coded as either manipulated natural recordings (N_{cond} = 12) or synthetically produced ones (N_{cond} = 137). With respect to *experimental method,* the studies included here used one of five methods: conditioned head-turn (CHT, N_{cond} = 84), high-amplitude sucking (HAS, N_{cond} = 42), visual/auditory habituation (VAH, N_{cond} = 15), event related potentials (ERPs, N_{cond} = 2), or the head-turn preference procedure (HTP, N_{cond} = 6). Finally, VOT distance was added here to control for confounds (e.g., if synthetic stimuli used smaller distances than did natural ones).

Factors affecting within-category discrimination

We found large differences in within-category discrimination as a function of both stimulus type and experimental method (Fig. 4). With respect to stimulus type, 83.3 % of conditions (N_{cond} = 10/12) using natural stimuli showed within-category discrimination, as compared with only 18.8 % of conditions using synthetic speech (N_{cond} = 18/96). Similarly, 55.5 % of the visual auditory habituation conditions (N_{cond} = 5/9), 50 % of the ERP conditions (N_{cond} = 1/2), and 100 % of the HTP conditions (N_{cond} = 6/6) showed within-category discrimination, while HAS (10.5 %, N_{cond} = 2/19) and CHT (19.4 %, N_{cond} = 14/72) showed the worst. We next investigated possible dependencies between these factors and between these factors and other factors in this data set.

Experimental method was somewhat confounded with stimulus type. The three methods with the highest percentages of within-category discrimination (HTP, VAH, and ERP) also tended to use more natural stimuli. HTP and ERP conditions used natural stimuli exclusively (HTP, N_{cond} = 6/6; ERP, N_{cond} = 2/2), while half of the VAH conditions used natural stimuli (N_{cond} = 3/6). In contrast, all of the CHT experiments and 18 of 19 HAS experiments used synthetic stimuli and had the lowest within-category discrimination. Given these unbalanced distributions, it may seem difficult to determine which factor is driving the discrepancy between experimental methods and stimulus types. However, while all of the conditions that used HTP and ERP used only one stimulus type, two methodologies have been used with both synthetic and natural stimuli—HAS and VAH—allowing for a cleaner comparison.

There was only one condition that used HAS with natural stimuli, and it did show within-category discrimination; in contrast, only 1 of the 17 HAS conditions (5.6 %) using synthetic stimuli showed evidence of discrimination (Fig. 5). We observed a similar, if less pronounced, trend with VAH: 66.7 % of VAH conditions with natural stimuli showed within-category discrimination (N_{cond} = 2/3), as compared with only 50 % of those using synthetic stimuli (N_{cond} = 3/6;

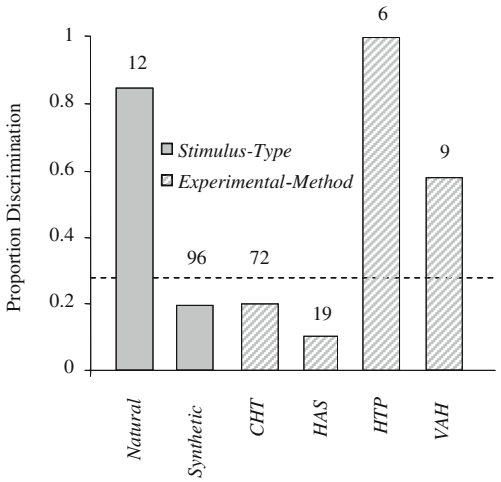

Fig. 4 Proportion of conditions showing discrimination grouped by stimulus type and experimental type. Natural: stimuli constructed from natural speech; synthetic: stimuli constructed from synthetic speech. CHT, conditioned head-turn procedure; HAS, high-amplitude sucking procedure; HTP, head-turn preference procedure; VAH, visual/auditory habituation procedure

Fig. 5). This suggests that the use of natural stimuli may be the more important predictor of within-category discrimination. This could be due to the fact that synthetic speech lacks some cues to voicing that are present in natural speech. Alternatively, infants may not recognize synthetic speech as fully linguistic. This is not to say that methodology has no effect. When VAH and HTP are used with synthesized stimuli, they still show higher rates of within-category discrimination than do other methods. However, additional experiments are needed to test this.

It is also important to ask whether VOT distance is confounded with stimulus type or experimental method (as in Analysis 2). To examine this, we collapsed VOT distance into two categories: less than 25 ms and greater than 25 ms (25 ms was the median VOT distance). If VOT distance is driving discrimination rates across methodologies, we would expect to see a higher proportion of conditions in which VOT distance exceeds 25 ms for HTP, ERP, and VAH conditions, as compared with CHT and HAS conditions. This was not the case: 57 % of the conditions using the methodologies showing the worst within-category performance (CHT and HAS) used large VOT distances greater than 25 ms (CHT, N_{cond} = 50/72, used large VOT distances; HAS, N_{cond} = 2/19), while only 47 % of the conditions using the three best methodologies used large VOT distances (HTP, N_{cond} = 0/6, used large VOT distances; ERP, N_{cond} = 2/2; VAH, N_{cond} = 6/9). Thus, these variables are not particularly confounded, and if anything, the relationship is slightly inverted. Interestingly, all of the HTP conditions used test stimuli less than 25 ms apart. Despite this disadvantage, HTP conditions have the highest rate of within-category discrimination, suggesting a robust effect of either methodology or stimulus type or their interaction.

The case for stimulus type is even more convincing when we examine the proportion of natural stimuli conditions by VOT distance. Of the natural stimuli conditions 83 % used small VOT differences (N_{cond} = 10/12), as compared with only 39.6 % of the synthetic conditions (N_{cond} = 38/96). Thus, the vast majority of conditions using natural stimuli showed within-category discrimination (83.4 %; Fig. 5), despite using more acoustically similar VOT contrasts.

The foregoing analysis indicates that stimulus type plays a large role in within-category discrimination. However, more research is required to validate this. The number of within-category conditions using natural speech is relatively small (N_{cond} = 12), but it does span four different studies (Burns, Yoshida, Hill, & Werker, 2007; McMurray & Aslin, 2005; Rivera-Gaxiola, Silva-Pereyra, & Kuhl, 2005; Trehub & Rabinovitch, 1972), and three of these are the most recent studies on VOT discrimination. These studies differ from older studies in both their methodology and subject age and come from theoretical perspectives that may also have shaped design choices. However, these results raise the possibility that the peak observed in Analysis 2 might have disappeared if these studies had been conducted using natural speech, since infants may be uniformly able to discriminate at all VOT locations. Indeed, we found five conditions showing discrimination for VOT distances less than or equal to 10 ms; all of them used natural stimuli.

Finally, we also attempted to ask whether the quality of synthesized speech affected discrimination. Since natural speech showed such a large advantage over synthetic speech and earlier synthesizers produced less natural speech, it may be that synthetic stimuli from newer synthesizers show a greater discrimination rate. Unfortunately, it is difficult to determine which synthesizer was used for each experiment. Thus, we classified studies as either older than 1988 or newer than 1988 (the year of the last version of the Klatt synthesizer, on which a number of widely available synthesizers are based). Interestingly, this rough classification technique yielded an increase in discrimination rate. Only 9.64 % of

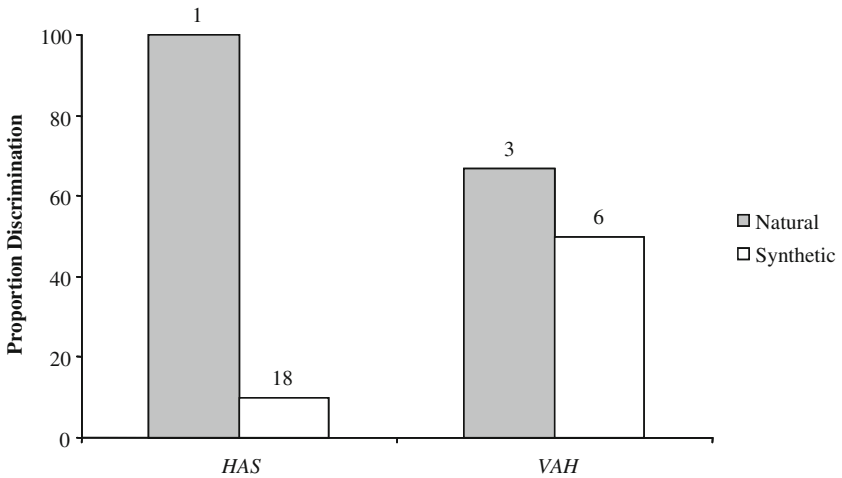

Fig. 5 Proportion of conditions showing discrimination as a function of stimulus type. HAS, high-amplitude sucking procedure; VAH, visual/auditory habituation procedure

synthetic conditions before 1988 ($N = 52$) found within category discrimination, as compared with 22.22 % of synthetic conditions after 1988 ($N = 27$).

Analysis 4

Previous analyses collapsed across age to maximize the amount of data. Thus, our final analysis examined the development of discrimination as a function of age.

Method

This analysis used the same conditions as Analysis 2. Conditions in which the native language was not English were excluded. Age was binned into four groups: less than 3 months ($N_{cond} = 24$); 3–6 months ($N_{cond} = 29$), 6–9 months ($N_{cond} = 87$), and older than 9 months ($N_{cond} = 22$), an age that has just lost the ability to discriminate nonnative phoneme contrasts (Werker & Tees, 1984). We also examined contrast type, using the simpler within- versus between-category levels (assuming a 25-ms VOT boundary).

Results

Not surprisingly, this analysis revealed that between-category discrimination begins and remains high throughout development. This supports a consistent presence of categories in shaping behavior throughout development, even at young ages. This is predicted by the parallel channels account, but the early presence of categories may be more difficult to account for in that framework, and we return to this issue in the General Discussion. We did not observe any evidence that between-category discrimination gets better over development (as was seen by Kuhl, Stevens, Hayashi, Deguchi, Kiritani, & Iverson, 2006, for r/l; and by Tsuji & Cristia, 2013, in a meta-analysis of vowels). However, this is likely due to our binary outcome variable; if infants are already discriminating fairly well, there is no room for improvement in our measure, but a continuous outcome measure may reveal such sensitivity.

When we turn to the discrimination of within-category contrasts, this ability appears to *improve* over the first year (Fig. 6). This is in opposition to the standard framing of speech development, which suggests that infants lose the ability to discriminate nonnative contrasts (Werker & Curtin, 2005), but in line with adult work demonstrating within-category discrimination for VOT (Carney et al., 1977; Kong & Edwards, 2011; Massaro & Cohen, 1983; Toscano et al., 2010). VOT, despite its status in the literature as a canonical speech dimension, does not appear to show this canonical pattern. Use of the conservative approach did not change the data significantly (Supplemental S3).

However, it is important to consider experimental factors from Analysis 3 that may affect sensitivity to within-category detail. When we look at the rate of within-category discrimination as a function of age and grouped by stimulus type (Fig. 7), an interesting pattern emerges. For conditions using natural stimuli, the rate of discrimination remains high throughout development but reduces slightly with age. For the synthetic stimuli, there is an increase in discrimination with age, although discrimination is still low throughout. Although there are relatively few natural conditions ($N_{cond} = 12$) and, thus, there is less confidence in this decline in discrimination with age, there are many synthetic conditions available for analysis ($N_{cond} = 96$), and therefore, this increase in discrimination rate is harder to dismiss. Use of the conservative approach did not change the data significantly (Supplemental S3).

Looking only at synthetic speech, this suggests either that infants' sensitivity to raw cue values increases or that some capacity for better analysis of the signal does. It is also possible that infants' emerging awareness of VOT categories (Andruski et al., 1994; McMurray et al., 2002; Miller, 1997) is what supports their increased sensitivity to within-category detail. That is, infants' sensitivity to the prototype structure of the category is what supports their ability to discriminate within-category differences (e.g., Miller & Eimas, 1996). This is consistent with recent work by Clayards, Tanenhaus, Aslin, and Jacobs (2008) suggesting that adult listeners' sensitivity to gradations within categories may stem from their sensitivity to the graded statistics of categories, and it supports a prediction made by the McMurray et al. (2009a) computational model. Thus, distributional learning may, in fact, give rise to within-category sensitivity.

At the same time, the reduction in discrimination seen in natural speech is more consistent with the standard approach. However, crucially, within-category discrimination only gets to about 40 %, suggesting a fairly incomplete pruning of these abilities. Taken together, it may then seem that infants are gradually gaining access to the more restricted information found in synthetic speech, while simultaneously losing some sensitivity when tested with natural stimuli, and that both ultimately converge at around 40 %. However, much more work with natural stimuli and younger infants will be needed to confirm this. At a broader level, however, this confirms the parallel channel model by demonstrating that not just phonological categories (and their development) contribute to discrimination but, rather, that both categories and emerging abilities to deal with the acoustic signal predict this developmental trend.

General discussion

The strong empirical evidence against CP in adults, along with the necessity of a flexible perceptual system during

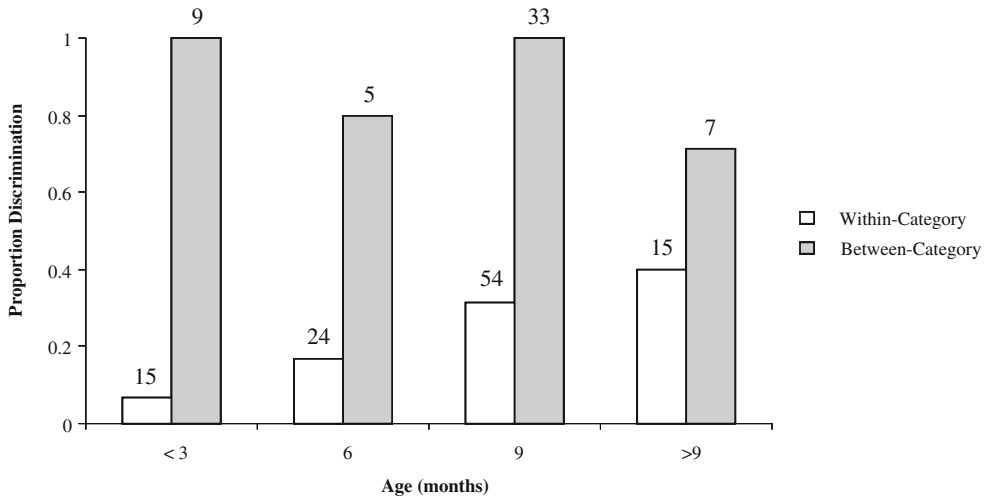

Fig. 6 Proportion of conditions showing discrimination as a function of relative subject age and grouped by contrast type

development, requires a new linking function for infant discrimination, one that reconciles both the decades of research demonstrating CP in infancy and the recent studies of both infants and adults arguing against it. To this end, we extended Pisoni and Tash's (1974) model of speech discrimination to infancy to suggest that the presence both of high-level categories and of low-level acoustics influences discrimination. To assess the validity of this parallel channels model, we conducted a quantitative analysis of over 40 years of infant VOT discrimination. Results go beyond the work on which they were based to support the parallel channels model by documenting a robust peak in discrimination, evidence of the contribution of speech categories, coupled with substantial within-category discrimination, sensitivity to the continuous metric distance in VOT, and sensitivity to task and stimulus factors. Importantly, our review suggests that even our surprisingly high 30 %–40 % of conditions showing within-category discrimination may be too low; natural stimuli (which are rarely used) show much higher rates, and the VOT distances tested for within-category discrimination are generally smaller than those tested for between.

As we now discuss, this body of findings supports a general auditory basis for discrimination, reveal useful insights into experimental design, and shed new light on development. We discuss each of these in turn, but first it is important to address two caveats.

Caveats

First, it is possible that evidence for graded speech discrimination is an artifact of our collapsing across studies that used different stimuli and, thus, may have had different boundaries.

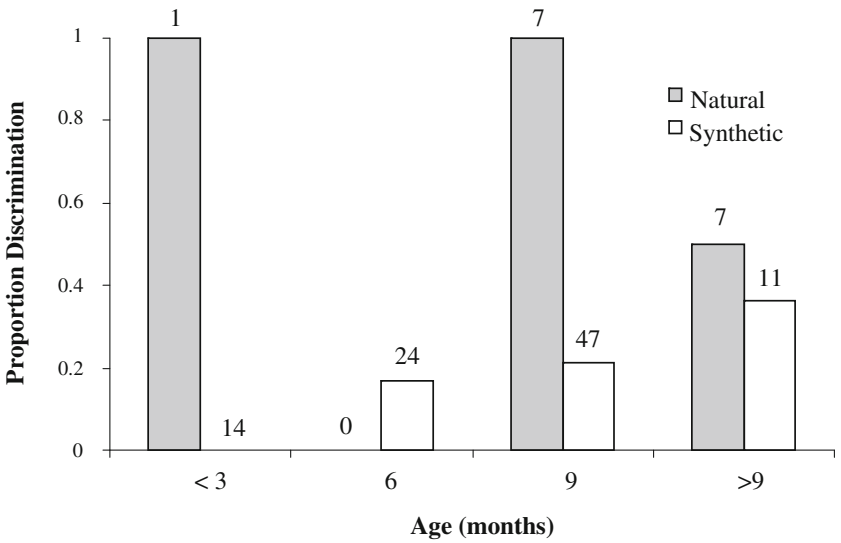

Fig. 7 Proportion of conditions showing discrimination as a function of subject age and stimulus type

We have accounted for this by adopting a more conservative analysis in which we excluded contrasts that were potentially ambiguous (see Fig. 2b and Online Supplement), and our results were largely unchanged by this. VOT boundaries may differ (across subjects or studies) by more than the 20 ms that our analytic approach accounts for. We cannot fully dismiss this alternate explanation, although two important facts minimize this concern. First, VOT boundaries in adult studies rarely shift beyond 20 ms (the extent of our more conservative analysis), and there is no reason to believe that they would in infants. Second and more important, findings of within-category sensitivity in infancy imply strong developmental continuity with adults, where the same finding is uncontroversial. The alternative is that infants start out categorical and somehow acquire sensitivity to within-category detail. This seems less plausible.

It is also important to point out that the averaging argument may be a double-edged sword. As we described, the parallel channels model suggests that phenomena that look like categorical perception are basically additive; when both category and sensory differences exist, researchers should naturally find improved discrimination. If this is true, it is not possible to make strong claims about perceptual levels of processing from discrimination data alone, particularly if there is any evidence for within-category sensitivity. Work that finds evidence for heightened discrimination at category boundaries cannot dismiss the possibility that discrimination measures are masking perception that is, at its core, graded.

Second, our meta-analysis is limited since it looked only at the "condition" as a unit of analysis (rather than individual subjects). Due to the large differences in methodologies between studies, we were not able derive standard summary statistics to complete a more standard meta-analysis. This oversimplification may mask some important effects (for example, an increase in between-category discrimination, as was observed in a recent meta-analysis of vowels: Tsuji & Cristia, 2013). However, the evidence for within-category sensitivity may be even stronger, if such an analysis were possible. Eimas et al. (1971) and Miller and Eimas (1983), for example, found no significant difference in sucking rate between within-category and control conditions. However, the *quantitative* difference was in the right direction, suggesting that their paradigm may be sensitive to such differences with more power. If it were possible to pool these small, nonsignificant quantitative differences across studies, we might find even more robust evidence.

Categorical perception versus parallel channels?

CP in all its forms invokes warping of low-level perception by high-level categories. The parallel channels model, however, proposes that category membership represents one channel of information that listeners can use during discrimination decisions but that it does not necessarily have an effect on perceptual encoding (the other channel). This model predicts an effect of both category membership and acoustic similarity on discrimination. While many studies demonstrate an effect of category membership, only recently have researchers found significant effects of acoustic similarity in infancy.

Analyses 1 and 2 found evidence of low-level acoustics on discrimination, along with an effect of category membership. Analysis 1 showed a *graded* peak of heightened discrimination centered at +25 ms along the VOT continua, indicative of within-category discrimination. When this boundary was used to classify contrasts as between or withincategories, we found a strong effect of category membership, with better discrimination for between-category contrasts, but also evidence for an effect of acoustic similarity, with fairly substantial rates of within-category discrimination. Analysis 2 demonstrated an effect of VOT distance on within-category discrimination, with a somewhat linear relationship between VOT distance and discrimination. This has never been observed before in infancy. Finally, Analysis 4 showed that development of both categories and the perceptual abilities to encode VOT are necessary to explain the data on development, suggesting an important confirmation of the model, since both channels may undergo development. As a whole, these findings support the parallel channels model by demonstrating an effect of both category membership and acoustic similarity on discrimination. Moreover, both Analysis 1 and 2 were able to demonstrate the effect of acoustic similarity within a data set made up primarily of data drawn from studies that concluded in favor of CP. Finally, while this review has emphasized the contribution of within-category sensitivity (which is where the two models differ), we find continued strong support for a peak in discrimination near the category boundary, consistent with the category channel in the parallel channels model. Thus, our quantitative review was able to provide evidence for the parallel channels model of speech discrimination and account for data that originally concluded both for and against CP.

Models of development

While our goal was not to develop a new model of speech perception, the parallel channels model and the associated findings of this meta-analysis are consistent with a number of recent proposals concerning the development of speech perception. The veridical encoding enabled by this model is an important prerequisite of some statistical learning theories and models (de Boer & Kuhl, 2003; Maye et al., 2003; McMurray, Aslin, & Toscano, 2009a; Pierrehumbert, 2003; Toscano & McMurray, 2010; Vallabha et al., 2007; Werker et al., 2007), as well as exemplar accounts (Goldinger, 1998), in which cues are coded veridically and mapped onto

categories at a second level of representation (interestingly, they may not be compatible with approaches such as that in Guenther & Gjaja, 1996, in which a single representation is "warped" by the statistical distribution of the input). Indeed, recent computational accounts suggest that such models can account for the loss of nonnative contrasts, the enhancement of native ones, the enhancement of within-category detail, and the differential weightings of acoustic cues (McMurray, Aslin, & Toscano 2009a; Toscano & McMurray, 2010).

At the same time, there may be limits to distributional learning, since recent analyses of infant directed speech suggest that some categories may not be distinguishable on the basis of statistics alone, due to the differential frequencies of the phonemes (Bion, Miyazawa, Kikuchi, & Mazuka, 2013), and that caregivers do not always modify these statistics in a way that maximizes such contrasts (Cristia & Seidl, 2013; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013). More likely, such learning may only help generate a set of proto-categories along candidate cue values, proto-categories that are then refined by the developing lexicon. This fits with recent empirical data suggesting that during the second year, infants are determining what cues are relevant for making lexical/phonological distinctions (Apfelbaum & McMurray, 2011; Dietrich et al., 2007; Rost & McMurray, 2009) and is ultimately consistent with PRIMIR (Werker & Curtin, 2005).

General auditory principles and discontinuities

One question that is not well addressed by statistical learning is the origin of the early biases in discrimination. In this sense, a classic question has been whether these early between-category discrimination abilities (particularly for cues like VOT) are based on general auditory principles or discontinuities (which also apply to nonspeech; Gottlieb et al., 1977). That is, is there an auditory basis for categorical perception (Jusczyk, Pisoni, Walley, & Murray, 1980; Jusczyk et al., 1989)? The early thinking on this was that CP derives from specific psycho-acoustic properties of cues like VOT or formant transitions. In this light, the peaks of discrimination seen in infancy around 20–30 ms of VOT were used to argue for a sensory discontinuity in VOT encoding, prior to linguistic experience. This conclusion was supported by animal work showing a similar boundary (Kuhl & Miller, 1975) and the neurophysiological evidence of a discontinuity (Sharma, Marsh, & Dorman, 2000; Sinex, McDonald, & Mott, 1991).

However, as we have described, the animal work is problematic given evidence that their boundaries are sensitive to the range of the stimuli being tested (Ohlemiller et al., 1999), and the neuroscience is problematic given the more recent ERP, MEG, and MRI studies (Frye et al., 2007; Myers et al., 2009; Toscano et al., 2010) showing veridical encoding. Thus, there is no clear evidence for a discontinuity. In light of this evidence, if we take a standard view of an auditory discontinuity, this requires us to posit a developmental pathway in which infants start with a warped representation of auditory cues and then somehow lose it by adulthood. In contrast, taken in light of the parallel channels model and the evidence here, infant perception may be described by conventional perceptual principles without necessitating discontinuities.

Where we do observe heightened discrimination, around 20–30 ms of VOT in infants, the parallel channels account suggests that this can arise out of the contributions to category-level information to discrimination without any warping of the perceptual encoding. However, this leads to perhaps the most difficult finding to account for if VOT is encoded veridically: the fact that very young infants (1 month olds) show heightened discrimination at 20 ms of VOT (Eimas et al., 1971). This fact was one of the findings that, along with the animal work, led to the consensus favoring discontinuities. While we note that only a single study examining infants younger than 6 months used natural stimuli, which may offer a different profile (Trehub & Rabinovitch, 1972, which showed discrimination for both between and within categories), taking the other studies at face value (25 conditions at <3 months), there is substantial evidence from synthetic speech for heightened discrimination at this boundary that cannot be dismissed.

In the parallel channels formulation, the easiest way to account for such results are if infants actually have something like voicing categories by this age. Indeed, VOT categories may develop much earlier than expected. This would be supported by the fairly robust statistics of VOT in infant directed speech (Englund, 2005; Kovack-Lesh & Oakes, 2007) and the high frequency of stop consonants, which could enable very rapid statistical learning. It is possible that infants have some access to the relevant envelope cues prenatally or could simply learn them quickly postnatally.

Even if we don't want to posit that infants have fully formed categories by a month, they may possess several weak categories centered near eventual adult categories that contribute as a group to discrimination. This is basically a form of a distributed representation adapted to this problem (Hinton, McClelland, & Rumelhart, 1986). This idea is well illustrated by statistical learning models (de Boer & Kuhl, 2003; McMurray, Aslin, & Toscano, 2009a; Vallabha et al., 2007), which posit that infants represent dimensions like VOT using several Gaussian distributions. These distributions compete with each other for activation during learning. Each time a VOT is presented, the distribution that best fits that VOT is strengthened, and those nearest the winner are weakened. Over time, this competition eliminates some distributions, while others align to the input. Thus, a region of VOT that will correspond to a robust category in adulthood (e.g., Fig. 8a), may be represented by multiple primitive categories early in development (see McMurray, Aslin, & Toscano,

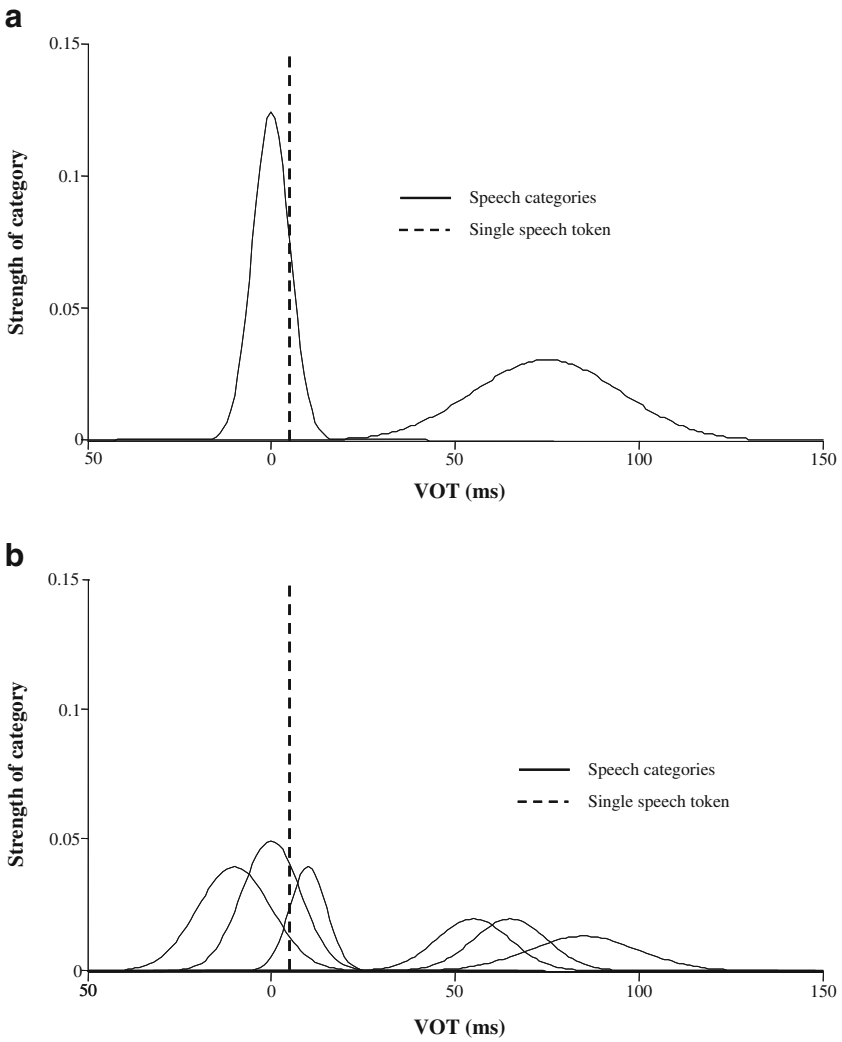

Fig. 8 Hypothetical voicing categories for adults (top) and young infants (bottom), plotted as Gaussians by strength and voice onset time (VOT)

2009a, for a discussion). Although each category is only weakly associated with a given speech sound, together they could provide enough information to heighten discrimination near the boundary. For example, the line in Fig. 8b represents an utterance with a VOT of 5 ms. In this situation, the infant may not know what particular category this VOT corresponds to, but the infant may know that it corresponds to one of the three low VOT categories and not to any of the high VOT categories. This could enable a discrimination peak in the middle, even as individual categories are poorly represented. Obtaining empirical evidence for this assertion is difficult, however, since assessing even robust categories during infancy is notoriously difficult.

Thus, the parallel channels account suggests that there must be something at the category level to give rise to such discontinuities. While this may be counterintuitive at such a young age, the alternative (that such abilities come from an auditory discontinuity) makes the prediction that such discontinuities must be lost over development to account for the adult data, something which is not described or predicted by any developmental models. It is also important to point out that the most robust evidence against this account—and for so-called innate auditory discontinuity—would have to come from infants learning a prevoicing language (which does not use the 20-ms boundary). However, in our data set, there were only two such studies testing 5.25 and 7.25 months—far too late to call anything "innate."

Implications for experimental methodologies

Analysis 3 revealed that methodological choices are not trivial and may account for some of the discrepancies between studies. Discrimination of within-category contrasts is greatly dependent on several factors. Although the average rate of discrimination for within-category contrasts is approximately 31 %, it ranges from under 20 % to over 90 % across specific condition types. In general, stimulus type (natural vs. synthetic) and experimental method appear to have the greatest effect.

However, as was noted previously, these factors are often confounded with one another.

Importantly, many of the studies carried out in recent years have used different methodologies than those in past decades and have tended to use naturally produced stimuli rather than synthetically produced stimuli. However, these methodological trends along with new theoretical goals (demonstrating within-category discrimination) are also confounded. In short, recent work on infant speech perception makes conclusions that contradict previous research (continuous vs. categorical perception) while using different methodologies (natural vs. synthetic speech). Our quantitative review indicates that this trend may be the result of more sophisticated and sensitive behavioral measures, experimental factors (including stimulus type), or both. Subsequent empirical work is necessary to disentangle these factors and assign causation. It is equally important to correct the imbalance with regard to stimulus type, methodology, and age, so that we can determine a single developmental time course for this ability, which despite dozens of studies, actually appears somewhat more uncertain than one would have expected.

More broadly perhaps, our analysis suggests an unbalanced literature. Older studies examined younger babies with habituation and synthetic speech, whereas more recent studies have examined older babies with other techniques (e.g., the head-turn preference procedure, visual/auditory habituation, and ERPs) and stimuli created from natural speech. Few have examined consonants other than bilabials, and there is little data on non-English-learning babies. This makes it challenging to draw strong conclusions about any single factor. Crucially, these methodological factors raise the question of what would be found with younger infants and natural speech; it is possible that the robust peak of discrimination near 25 ms of VOT may be seen only with synthetic speech. Would our picture of infant speech development be different if this methodological innovation had been employed from the start? Moving beyond bilabial stop consonants and English-learning infants is important for generalizing recent findings and improving our understanding of the development of speech perception.

Development

The canonical view of speech perception characterizes early development as narrowing of discrimination abilities toward those contrasts used in native language. (Aslin et al., 2002; Werker & Curtin, 2005; Werker & Tees, 1984). This trend should manifest as a reduction of sensitivity to nonnative and within-category contrasts over time. However, looking only at the synthetic stimuli, it appears that infants actually become more sensitive to within-category variation as they approach 12 months of age. This may be the result of infants simply gaining better access to the reduced information in synthetic stimuli, since natural stimuli show the more expected reduction in sensitivity to within-category contrasts (although this represents only 12 conditions), or it could derive from more robust encoding of the prototype structure of speech categories. Both synthetic and natural discrimination converge on a similar rate of within-category discrimination, however, suggesting that this 40 % discrimination rate may be the more or less "developed" level of within-category discrimination.

This challenges the narrowing or overgeneration/pruning account. It is possible that the narrowing applies only to so-called "nonnative" contrasts, whereas here, VOT is a native contrast (even as within-category ranges of it are not used in the language). However, this would somehow require the perceptual learning system to know which stimulus ranges are nonnative (e.g., potentially phonological in another language) and which are merely within-category.

Interestingly, a similar developmental trend of enhancing discrimination has been shown for other phonetic contrasts. In a cross-linguistic study of /r-l/ discrimination, Kuhl et al. (2006) showed that infants' ability to discriminate the /r-l/ contrast improved when it was a native one but decreased when it was nonnative. This trend has also been observed with Mandarin infants for an alveolo-palatal affricate fricative contrast and English infants for a native palatal-alveolar affricate fricative contrast (Tsao, Liu, & Kuhl, 2006), and in fricatives (Eilers & Minifie, 1975), and for the prevoiced contrast (Eilers et al., 1977). It has also been seen in a recent meta-analysis of vowels. This analysis of several dozen studies reported much more robust evidence for enhancement of native contrasts than for pruning of nonnative contrasts (Tsuji & Cristia, 2013). Thus, enhancement may be a much more common trend than the canonical view would admit.

Most studies of this pruning and enhancement have investigated complete phonological feature changes, not within-category sensitivity. It is intriguing, however, that the canonical view of speech perception does not fully describe the canonical phonetic cue used in infancy research (VOT) and that this pattern may be moderated by the type of stimuli (suggesting an emerging ability to extract single features from noncanonical input). Regardless of how the debates over the developmental time course turn out, the parallel channels model suggests that neither type of change need be seen as solely a perceptual change (e.g., perceptual narrowing). Rather, as long as our measure is predominantly discrimination, such changes could occur simply by adding, strengthening, or modifying categories at a more cognitive level, categories which then contribute along with lower level perceptual representations to yield discrimination.

A second developmental issue is the nature of these categories. Although this review has demonstrated that infants are able to discriminate speech sounds on the basis of acoustic and, presumably, categorical differences, recent work calls into question the stability of speech categories early in development. Maye et al. (2003) found that 9-month-old infants

(who have had voicing categories for quite some time) can learn to ignore categorical information after only a brief exposure to sounds with unimodal cue distributions. Similarly, building on seminal work on early word learning (Stager & Werker, 1997), Rost and McMurray (2009, 2010) showed that 14-month-old infants require significant variation in irrelevant cues (such as those attributed to speaker variability) in order to use supposedly well-learned voicing categories for learning word/object pairings, implying that they do not know what dimensions are relevant for word learning (for a theoretical discussion, see also Dietrich et al., 2007; Werker & Curtin, 2005). Thus, it may be that our relatively coarse metric of discrimination is overstating what may be a much slower-to-develop process. Indeed, it is tempting to presume that speech categories stabilize at some later point in development, and the trends found in Analysis 4 suggests that such a milestone may be later rather than sooner. Furthermore, work with adults demonstrating measurable within-category sensitivity suggests that such sensitivity is not an indication of immature development.

Acknowledgments The authors would like to thank Joel Dennhardt, Julianne Spoo, and William McEchron for assistance in coding the studies and Matt Goldrick and Chandan Narayan for helpful comments on earlier drafts. This research was supported by NIH Grant DC008089 awarded to B.M.

References

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

Apfelbaum, K., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science, 35*(6), 1105–1138.

Aslin, R. N., & Pisoni, D. B. (1980). Effects of early linguistic experience on speech discrimination by infants: A critique of Eilers, Gavin, and Wilson (1979). *Child Development, 51*(1), 107–112. doi:10.2307/1129596

Aslin, R. N., Pisoni, D. B., Hennessy, B. L., & Perey, A. J. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child Development, 52,* 1135–1145.

Aslin, R. N., Werker, J. F., & Morgan, J. (2002). Innate phonetic boundaries revisited. *Journal of the Acoustical Society of America, 112*(4), 1257–1260.

Bernstein, L. E. (1983). Perceptual development for labeling words varying in voice onset time and fundamental-frequency. *Journal of Phonetics, 11*(4), 383–393.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization 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.

Bion, R. A. H., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLoS ONE, 8*(2), e51594. doi:10.1371/journal.pone.0051594

Burnham, D. K., Earnshaw, L. J., & Clark, J. E. (1991). Development of categorical identification of native and non-native bilabial stops: Infants, children, and adults. *Journal of Child Language, 18*(2), 231–260.

Burns, T., Yoshida, K., Hill, K., & Werker, J. (2007). The development of phonetic representation in bilingual and monolingual infants. *Applied Psycholinguistics, 28,* 455–474.

Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Non categorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America, 62,* 961–970.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804–809.

Cristia, A. (2011). Fine-grained variation in caregivers'/s/predicts their infants'/s/category. *The Journal of the Acoustical Society of America, 129,* 3271.

Cristia, A., & Seidl, A. (2013). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language, FirstView*, 1–22. doi:10.1017/S0305000912000669

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Auditory Research Letters On-Line (ARLO), 4,* 129–134.

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, 104*(41), 16027–16031. doi:10.1073/pnas.0705270104

Eilers, R. E., Gavin, W., & Wilson, W. R. (1979). Linguistic experience and phonemic perception in infancy: A crosslinguistic study. *Child Development, 50*(1), 14–18.

Eilers, R. E., & Minifie, F. (1975). Fricative discrimination in early infancy. *Journal of Speech and Hearing Research, 18*(1), 158–167.

Eilers, R. E., Wilson, W., & Moore, J. (1977). Developmental changes in speech discrimination in infants. *Perception & Psychophysics, 16,* 513–521.

Eilers, R. E., Wilson, W. R., & Moore, J. M. (1979). Speech discrimination in the language-innocent and the language-wise: A study in the perception of voice onset time. *Journal of Child Language, 6*(01), 1–18. doi:10.1017/S0305000900007583

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171,* 303–306.

Englund, K. T. (2005). Voice onset time in infant directed speech over the first six months. *First Language, 25*(2), 219–234. doi:10.1177/0142723705050286

Fowler, C. A., Best, C. T., & McRoberts, G. W. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & Psychophysics, 48*(6), 559–570.

Fowler, C. A., & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. S. Perkell & D. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 123–136). Hillsdale: Erlbaum.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech, 5,* 171–189.

Frye, R. E., McGraw Fisher, J., Cody, A., Zarella, M., Liederman, J., & Halgren, E. (2007). Linear coding of voice onset time. *Journal of Cognitive Neuroscience, 19,* 1476–1487.

Galle, M., Apfelbaum, K., & McMurray, B. (2013). Within-speaker variability benefits phonological word learning. *Language Learning and Development* (in press).

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics, 66*(3), 363–376.

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

Gottlieb, G., Darwin, R. C., Eimas, P., Konishi, M., Liberman, A., Marler, P., . . . Pisoni, D. (1977). *Evidence for early species-typical biases in auditory sensory and perceptual mechanisms.* Paper presented at the

Recognition of complex acoustic signals: Report of the Dahlem Workshop on Recognition of Complex Acoustic Signals, Berlin 1976, September 27 to October 2.

Gow, D. W. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language, 45,* 133–139.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics, 65*(4), 575–590.

Gow, D. W., & McMurray, B. (2007). Word recognition and phonology: The case of English coronal place assimilation. In J. S. Cole & J. Hualdo (Eds.), *Papers in laboratory phonology 9* (pp. 173–200). New York: Mouton de Gruyter.

Guenther, F., & Gjaja, M. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100,* 1111–1112.

Healy, A. F., & Repp, B. H. (1982). Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception and Performance, 8*(1), 68–80.

Hinton, G., McClelland, J. L., & Rumelhart, D. (1986). Distributed representations. In D. Rumelhart, J. L. McClelland, & T. P. R. Group (Eds.), *Parallel distributed processing, vol. 1: Foundations* (Vol. 1, pp. 77–109). Cambridge, MA: The MIT Press.

Hoonhorst, I., Colin, C., Markessis, E., Radeau, M., Deltenre, P., & Serniclaes, W. (2009). French native speakers in the making: From language-general to language-specific voicing boundaries. *Journal of Experimental Child Psychology, 104*(4), 353–366.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research, 5,* 69–95.

Husaim, J., & Cohen, L. (1981). Infant learning of ill-defined categories. *Merrill-Palmer Quarterly, 27,* 443–456.

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

Jusczyk, P. W., Rosner, B., Reed, M., & Kennedy, L. (1989). Could temporal order differences underlie 2-month-olds' discrimination of English voicing contrasts? *Journal of the Acoustical Society of America, 85*(4), 1741–1749.

Kong, E. J., & Edwards, J. (2011). *Individual differences in speech perception: Evidence from visual analogue scaling and eye-tracking.* Paper presented at the Proceedings of the XVIIth International Congress of Phonetic Sciences, Hong Kong.

Kovack-Lesh, K. A., & Oakes, L. (2007). Hold your horses: How exposure to different items influences infant categorization. *Journal of Experimental Child Psychology, 98,* 69–93.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190*(4209), 69–72.

Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics, 32*(6), 542–550.

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.

Lasky, R. E., Syrdal-Lasky, A., & Klein, R. E. (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology, 20*(2), 215–225.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*(5), 358–368.

Liberman, A. M., Harris, K. S., Kinney, J., & Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and non-speech patterns. *Journal of Experimental Psychology, 61,* 379–388.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384–422.

Litwin, E. (1998). *Discrimination of native and non-native speech-sounds by newborns.* Ph.d., The City University of New York, New York.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Martin, J. G., & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America, 69*(2), 559–567.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication, 2,* 15–35.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008a). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science: A Multidisciplinary Journal, 32*(3), 543–562. doi:10.1080/03640210802035357

Maye, J., Weiss, D. J., & Aslin, R. N. (2008b). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11*(1), 122–134. doi:10.1111/j.1467-7687.2007.00653.x

Maye, J., Werker, J. F., & Gerken, L. (2003). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* 101–111.

McMurray, B., & Aslin, R. N. (2004). Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy, 6*(2), 203–229. doi:10.1207/s15327078in0602_4

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition, 95*(2), B15–B26.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance, 34*(6), 1609–1631.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009a). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science, 12*(3), 369–379.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*(2), 219–246.

McMurray, B., Kovack-Lesh, K., Goodwin, D., & McEchron, W. D. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition, 129*(2), 362–378.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*(2), B33–B42.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009b). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language, 60*(1), 65–91.

Miller, J. L. (1997). Internal structure of phonetic categories. *Language and Cognitive Processes, 12,* 865–869.

Miller, J. L., & Eimas, P. (1983). Studies on categorization of speech by infants. *Cognition, 13*(2), 135–165.

Miller, J. L., & Eimas, P. D. (1996). Internal structure of voicing categories in early infancy. *Perception & Psychophysics, 58*(8), 1157–1167.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46*(6), 505–512.

Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science, 20*(7), 895–903. doi:10.1111/j.1467-9280.2009.02380.x

Narayan, C. R. (2013). Developmental perspectives on phonological typology and sound change. In A. C. L. Yu (Ed.), *Origins of sound change: Approaches to phonologization* (pp. 128–146). Oxford: Oxford University Press.

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*(3), 407–420. doi:10.1111/j.1467-7687.2009.00898.x

Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics, 20,* 351–382.

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

Nittrouer, S., & Miller, M. E. (1997). Predicting developmental shifts in perceptual weighting schemes. *Journal of the Acoustical Society of America, 101,* 2253–2266.

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America, 75*(1), 224–230.

Ohlemiller, K., Jones, L., Heidbreder, A., Clark, W., & Miller, J. (1999). Voicing judgements by chinchillas trained with a reward paradigm. *Behavioural Brain Research, 100,* 185–195.

Perone, S., & Spencer, J. P. (2013). The co-development of looking dynamics and discrimination performance. *Developmental Psychology*. doi:10.1037/a0034137

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

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America, 55,* 328–333.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15*(2), 285–290.

Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-language comparison of /d/-/ð/ perception: Evidence for a new developmental pattern. *Journal of the Acoustical Society of America, 109*(5), 2190–2200.

Rivera-Gaxiola, M., Silva-Pereyra, J., & Kuhl, P. K. (2005). Brain potentials to native and non-native speech contrasts in 7- and 11-month-old American infants. *Developmental Science, 8*(2), 162–172.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*(2), 339–349.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of non-contrastive phonetic variability in early word learning. *Infancy, 15*(6), 608.

Saffran, J. R., Werker, J. F., & Werner, L. A. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In W. Damon, R. Lerner, D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology* (6th ed., pp. 58–108). Hoboken, NJ: John Wiley and Sons.

Schouten, M. E. H., Gerrits, E., & Van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication, 41,* 71–80.

Sharma, A., Marsh, C. M., & Dorman, M. F. (2000). Relationship between N1 evoked potential morphology and the perception of voicing. *Journal of the Acoustical Society of America, 108*(6), 3030–3035.

Sinex, D., McDonald, L., & Mott, J. (1991). Neural correlates of nonmonotonic temporal acuity for voice onset time. *Journal of the Acoustical Society of America, 90*(5), 2441–2449.

Smits, R. (2001). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception & Psychophysics, 63,* 1109–1139.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 338,* 381–382.

Streeter, L. A. (1976). Language perception of 2-mo-old infants shows effects of both innate mechanisms and experience. *Nature, 259*(5538), 39–41.

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America, 62*(2), 435–448.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: A statistical approach to cue weighting and combination in speech perception. *Cognitive Science, 34*(3), 436–464.

Toscano, J. C., & McMurray, B. (2012). Online integration of acoustic cues to voicing: Natural vs. synthetic speech. *Attention, Perception, & Psychophysics, 74*(6), 1284–1301.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. (2010). Continuous perception and graded categorization electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science, 21*(10), 1532–1540.

Trehub, S., & Rabinovitch, M. S. (1972). Auditory-linguistic sensitivity in early infancy. *Developmental Psychology, 6*(1), 74–77.

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *The Journal of the Acoustical Society of America, 120,* 2285.

Tsuji, S., & Cristia, A. (2013). Perceptual attunement in vowels: A meta-analysis. Manuscript submitted for publication.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104,* 13273–13278.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1*(2), 197–234. doi:10.1080/15475441.2005.9684216

Werker, J. F., Gilbert, J., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development, 52*(1), 349–355.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology, 24*(5), 672–683. doi:10.1037/0012-1649.24.5.672

Werker, J. F., & Polka, L. (1993). Developmental changes in speech perception: New challenges and new directions. *Journal of Phonetics, 21,* 83–101.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition, 103*(1), 147–162. doi:10.1016/j.cognition.2006.03.006

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.

Williams, L. (1977). The perception of stop consonant voicing by Spanish-English bilinguals. *Perception & Psychophysics, 21*(4), 289–297.