OPEN MIND

Discoveries in Cognitive Science

an open access journal

Citation: Feldman, N. H., Goldwater, S., Dupoux, E., & Schatz, T. (2021). Do Infants Really Learn Phonetic Categories? *Open Mind: Discoveries in Cognitive Science*, *5*, 113–131. https://doi.org/10.1162/opmi_a_00046

DOI:
https://doi.org/10.1162/opmi_a_00046

Received: 31 January 2020
Accepted: 6 August 2021

Competing Interests:
The authors declare no conflict of interest.

Corresponding Authors:
Naomi H. Feldman
nhf@umd.edu
Sharon Goldwater
sgwater@inf.ed.ac.uk

The MIT Press

PERSPECTIVE

Do Infants Really Learn Phonetic Categories?

Naomi H. Feldman[1*], Sharon Goldwater[2*], Emmanuel Dupoux[3,4], and Thomas Schatz[1]

[1]Department of Linguistics and UMIACS, University of Maryland, College Park, MD, USA
[2]School of Informatics, University of Edinburgh, United Kingdom
[3]Cognitive Machine Learning (ENS - EHESS - PSL Research University - CNRS - INRIA), Paris, France
[4]Facebook A.I. Research, Paris, France
*These authors contributed equally to this work.

Keywords: language acquisition, speech perception, computational modeling, representation learning

ABSTRACT

Early changes in infants' ability to perceive native and nonnative speech sound contrasts are typically attributed to their developing knowledge of phonetic categories. We critically examine this hypothesis and argue that there is little direct evidence of category knowledge in infancy. We then propose an alternative account in which infants' perception changes because they are learning a perceptual space that is appropriate to represent speech, without yet carving up that space into phonetic categories. If correct, this new account has substantial implications for understanding early language development.

INTRODUCTION

Infants' perception of speech becomes specialized for the native language even before their first birthday. Discrimination of native contrasts improves, and discrimination of nonnative contrasts declines (Kuhl et al., 2006; Werker & Tees, 1984). These changes are often assumed to reflect the development of adultlike perceptual patterns, and more specifically of adultlike *phonetic category* representations: linguistically relevant categories that are phoneme-length and correspond roughly to the consonants and vowels of a language (Best, 1994; Kuhl et al., 1992; Werker et al., 2007; Zevin, 2012).[1] These assumptions have been motivated by the close ties observed in adults between native language phonetic categories and language-specific patterns of discrimination along phonetically relevant dimensions, as shown schematically in Figure 1 (Liberman et al., 1957).

If early changes in discrimination result from early knowledge of phonetic categories—discrete units, with or without explicit labels, that roughly correspond to linguistically relevant sounds like [ɹ] (as in *rock*) and [l] (as in *lock*)—then infants must learn these categories by their first birthday. The categories would then drive changes to their perceptual space (Figure 2a). However, phonetic categories are difficult to learn from the speech infants hear (Antetomaso et al., 2017; Bion et al., 2013), raising doubts about the feasibility of early

[1] Contextual variants of a phoneme are generally treated as different categories, with phonetic categories corresponding roughly to allophones (Dillon et al., 2013; Werker & Curtin, 2005; but see Pegg & Werker, 1997).

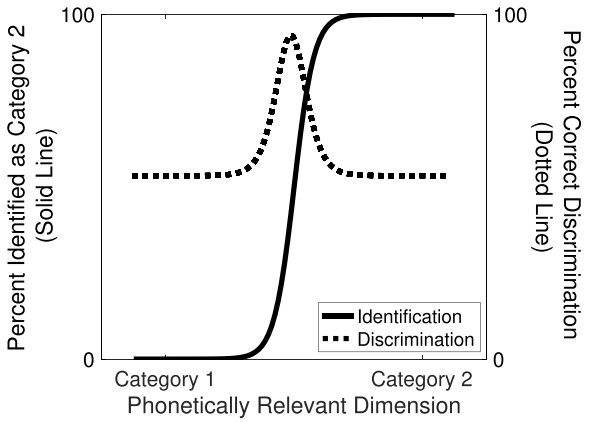

Figure 1. Hypothetical identification and discrimination functions in two-alternative forced choice tasks.

phonetic category learning. Early phonetic category learning has been questioned before (Jusczyk, 1992), yet only a few concrete alternative accounts of infants' changes in discrimination have been proposed (Guenther & Gjaja, 1996; Herrmann et al., 1995; Matusevych et al., 2020; Schatz et al., 2021).

Here we critically examine the evidence for phonetic category learning in infancy and highlight recent developments in speech technology which, we argue, can inspire an alternative account of early perceptual learning where phonetic categories are not involved. Under this account, early changes in discrimination are caused by a learning process that—without recourse to phonetic categories—transforms the acoustic similarity space, changing the perceptual distances between

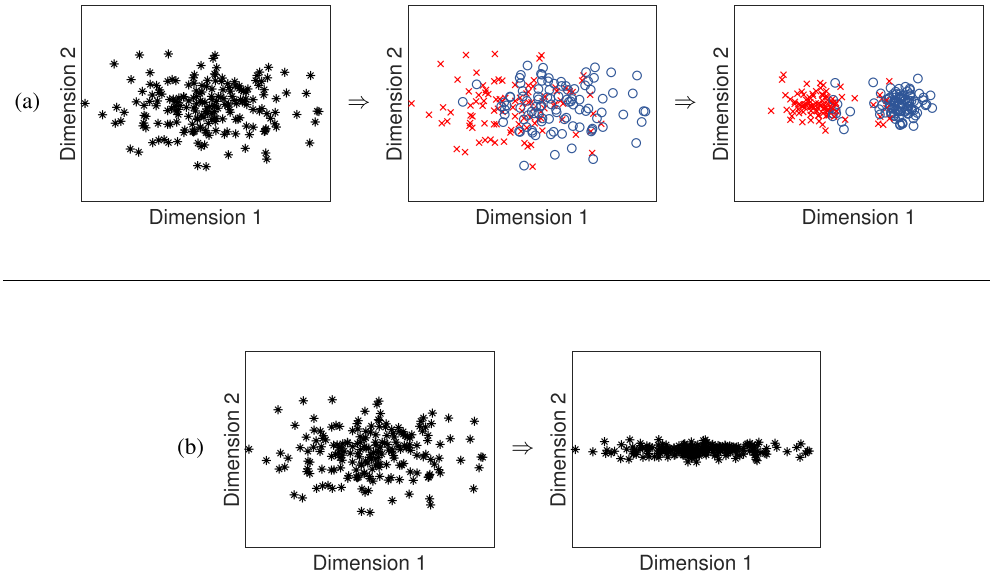

Figure 2. Phonetic category learning vs. perceptual space learning. (a) Under standard phonetic category learning theories, infants identify categories early. As a result, perception becomes warped along phonetically relevant dimensions (Dimension 1) and discrimination decreases along phonetically irrelevant dimensions (Dimension 2). (b) An alternative theory is that learners' perceptual space undergoes substantial changes before phonetic categories are learned. In this simplistic example, perceptual learning collapses the dimension of lower variance, decreasing discrimination along Dimension 2. As described later, we believe perceptual space learning actually involves more complex transformations.

sounds (Figure 2b). Phonetic categories are learned later, or more gradually, by carving up this learned space. We refer to the earlier phase of learning as *perceptual space learning*[2] and discuss several algorithms that might be used to implement such learning, including learning without any discrete units, or with units that do not correspond meaningfully to phones. Changes in discrimination driven by knowledge of phonetic categories could in principle also be considered a type of perceptual space learning, but here we restrict the term to mean learning *without* phonetic categories. We do not argue conclusively *against* the early phonetic category learning hypothesis; instead, we argue that perceptual space learning, which has thus far received little attention in the language acquisition literature, should be seriously considered as a plausible alternative theory of what causes infants' perceptual changes.

Attributing infants' perceptual changes to perceptual space learning would have major implications for theories of language acquisition. Phonetic category learning has conventionally been thought to occur before (Werker et al., 2009) or alongside (Swingley, 2009) word learning, enabling word forms to be composed of sequences of phones from the earliest stages. This hypothesized trajectory makes phonetic category learning a difficult problem because it cannot draw on extensive knowledge of word meanings, which would provide information about which sounds in a language are meaningfully different (Trubetzkoy, 1939). However, if phonetic category learning occurs later in childhood, it could draw on a broad array of word meanings and minimal pairs, making it an easier problem (McMurray et al., 2018). Perceptual space learning would also have broad implications for other areas of language acquisition, such as understanding when and how infants notice that words are mispronounced (Curtin et al., 2009; Fennell & Werker, 2003; Rost & McMurray, 2009; Stager & Werker, 1997), studying whether infant-directed speech is optimized for phonetic learning (Cristia & Seidl, 2014; Eaves et al., 2016; Kuhl et al., 1997; McMurray et al., 2013), or understanding the challenges of adult second language learning (Flege & Hillenbrand, 1986; Francis & Nusbaum, 2002; Lipski et al., 2012; Underbakke et al., 1988; Ylinen et al., 2009). More generally, it would radically change our view of what children know at the beginning of their second year, a period when they rapidly acquire aspects of language related to grammar and meaning.

CHILDREN'S PERCEPTUAL LEARNING

The primary evidence for phonetic category learning in infancy comes from experiments that measure infants' discrimination of native and nonnative sound contrasts. The discrimination tasks do not inherently require category knowledge (Box 1), but they do reveal changes in discrimination that are suggestive of category learning (as articulated by Zevin, 2012). Discrimination of nonnative speech contrasts generally declines during the first year of life: by 10–12 months for consonants and by 6–8 months for vowels (Anderson et al., 2003; Best & McRoberts, 2003; Best et al., 1995; Bosch & Sebastián-Gallés, 2003; Burns et al., 2007; Kuhl et al., 1992; Segal et al., 2016; Tsuji & Cristia, 2014; Werker & Lalonde, 1988; Werker & Tees, 1984). During the same time period, discrimination of native contrasts generally improves (Burns et al., 2007; Kuhl et al., 2006; Narayan et al., 2010; Tsao et al., 2006). Although there are exceptions to this pattern (Best et al., 1988; L. Liu & Kager, 2014, 2016; Mattock & Burnham, 2006; Mazuka et al., 2014; Mugitani et al., 2009; Polka & Bohn, 1996; Polka et al., 2001; Sundara et al., 2006; Yeung et al., 2013), it is clear that infants' perception becomes more native-like as they are exposed to their native language.

[2] In machine learning the usual term is *unsupervised representation learning*, but we want to avoid confusion caused by the broader meaning of *representation* in cognitive science.

Box 1. DO INFANT DISCRIMINATION TASKS REQUIRE CATEGORY KNOWLEDGE?

Most tests of infant speech perception have used one of two paradigms. In a *habituation* experiment, infants experience repeated trials in which they hear a habituation stimulus—exemplars from one phonetic category—while viewing a visual display. Once their looking time to habituation trials falls below a threshold, discrimination is measured as the extent to which they look longer at *change* trials (with exemplars from another category) than at *same* trials (with exemplars from the habituated category). Infants need to be able to discriminate a contrast in order to show different looking behavior toward *change* trials and *same* trials. However, infants can succeed at this task without knowing phonetic categories, as long as they perceive the stimuli on *change* trials to be acoustically anomalous, relative to the habituation trials. Similar considerations hold for the oddball paradigm used by Hochmann and Papeo (2014).

The other paradigm that is frequently used to measure infant speech perception is the *conditioned head turn* (CHT) procedure, in which infants face an experimenter who is playing with toys and hear a background stimulus from a loudspeaker on the side of the room. On *change* trials, the stimulus changes to an exemplar from the other phonetic category, and they can look toward the loudspeaker and see toys light up and start to move. On *same* trials, when the category does not change, looking toward the loudspeaker does not yield any visual reward. After an initial conditioning phase, discrimination is assessed by measuring head turns on *change* trials, relative to *same* trials. As in habituation experiments, infants need to be able to discriminate a contrast in order to show different looking behavior toward *change* trials and *same* trials. However, because this paradigm involves a decision of whether to perform a head turn, it resembles identification tasks in some ways. Particularly striking are studies showing that when trained on a phonetic contrast, infants can generalize to novel speakers during test in a CHT paradigm (Kuhl, 1979, 1983). This seems to suggest that infants already know that phonetic differences, but not speaker differences, signal a category distinction.

However, it is possible that the categorical patterns of generalization reflect learning that has occurred during the experiment. The visual reinforcements that infants see during a CHT experiment provide a reward signal that could engage reinforcement learning mechanisms, which appear to be particularly successful in driving auditory perceptual learning in adults (Lim et al., 2019; Lim & Holt, 2011; Tricomi et al., 2006). In line with this, Kuhl (1979) notes that the infants initially make head turns toward stimuli that vary from the background stimulus along irrelevant dimensions, such as speaker or pitch, but that this tendency lessens over the course of the experiment. She hypothesizes that learning has occurred during the experiment and suggests that

> the infant demonstrates a proclivity to try to discover a criterial attribute which separates the two categories. The infant, in effect, displays a tendency to be a "natural sorter," and is attracted to a dimension which makes a set of multidimensional auditory stimuli fit into easily recognized perceptual groupings. (p. 1674)

In other words, Kuhl hypothesizes that it is the functional equivalence of different exemplars with respect to the visual reinforcement in the CHT paradigm that supports learning of new cue weights. Given that this learning could occur within the experiment itself, the categorical head-turn behavior that infants exhibit within this paradigm does not necessarily support the strong hypothesis that they come into the lab with well-formed phonetic categories (see Apfelbaum & McMurray, 2011, for a similar argument). Whether, and at what age, children use the same strategy to learn phonetic categories in more naturalistic settings remains an open question.

A category-based account of these perceptual changes would entail that learners group stimuli into discrete units that correspond roughly to the phones of a language. As shown in Figure 2a, the categories would then drive changes in the perceptual space (Bonnasse-Gahot & Nadal, 2008; Kuhl, 1979). However, there are reasons to question whether categories are the driving force behind infants' perceptual changes. Box 2 distinguishes three perceptual effects that are often associated with category knowledge. If all three are direct results of category knowledge, then they should develop in tandem, as categories are learned. Given the substantial evidence that discrimination of nonnative contrasts declines sharply relative to native contrasts during infants' first year (Effect 3), one might also expect to find sharpening category boundaries (Effect 1) or sharpening discrimination peaks along phonetically relevant dimensions (Effect 2) in young infants. Yet there is little evidence that these effects develop during the same time period.

Box 2. PERCEPTUAL EFFECTS ASSOCIATED WITH CATEGORIES

Three types of perceptual effects are typically assumed to arise from category knowledge. While there is substantial evidence that the first two are closely tied to knowledge of categories, or at least distinct clusters of sounds, we argue that the third effect is more general, and need not reflect such knowledge.

Effect 1 is a sharp category boundary in identification tasks (Liberman et al., 1957; Figure 1). Performing an identification task requires category knowledge, given the use of category labels in the task. However, changes in steepness of the category boundary during learning could arise either from changes in category knowledge, or from children's improving ability to perform an identification task. These two possibilities can be disambiguated through a phenomenon known as *cue weighting*, which refers to the relative steepness of the identification curve across different dimensions. Changes in cue weighting have been tied to category learning across many studies (Francis et al., 2000; Francis et al., 2008; Francis & Nusbaum, 2002; Holt & Lotto, 2006; Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; R. Liu & Holt, 2015; Yang & Sundara, 2019; Ylinen et al., 2009), and cue weights are also key to many models of categorization (Kruschke, 1992; Love et al., 2004; Nosofsky, 1986; Toscano & McMurray, 2010), suggesting that this effect is closely tied to category knowledge.

Effect 2 is a discrimination peak near the category boundary (Liberman et al., 1957; Figure 1). While some models do attribute peaks in discrimination near the category boundary to category knowledge (Feldman et al., 2009; Kuhl, 1993; Lacerda, 1995), other models have suggested that this effect may only require distinct clusters in the distribution of sounds in the acoustic space (like the distributions in the third panel in Figure 2a) even if the clusters are not recognized as discrete units (Guenther & Gjaja, 1996; Herrmann et al., 1995; Shi et al., 2010). Moreover, categories with high variability (Figure 2a, second panel) may not yield a distinctive discrimination peak (Kronrod et al., 2016). Thus, we take the discrimination peak to index how tightly clustered the distribution of sounds is in listeners' perceptual space. Whether well-separated clusters of sounds constitute perceptual categories is a matter of some debate; to avoid overloading terminology, we simply refer to these as clusters of sounds in a perceptual space.

Effect 3 is listeners' differential ability to discriminate sounds along different dimensions. For example, English listeners discriminating instances of [ɹ] and [l] are more sensitive to differences in the third formant than to differences in the second formant, whereas Japanese listeners have roughly equal sensitivity to both dimensions (Iverson et al., 2003). Listeners can retain sensitivity to cues even when they stop using those cues to categorize sounds (Lehet & Holt, 2020), so changes in sensitivity

in discrimination tasks are not necessarily the same thing as changes in cue weighting. In theory, it is possible to lose or gain the ability to discriminate along certain dimensions even without representing well-separated clusters of sounds in a perceptual space (Figure 2b; Figure 3; Figure 4); that is the possibility we explore in this article.

The scope of this last effect merits consideration, because although discrimination is typically assumed to be better along phonetically relevant dimensions than along phonetically irrelevant dimensions (cf. Goldstone, 1994), there are exceptions to this generalization (Best et al., 1988). Moreover, even if there were no exceptions, predicting exactly which contrasts are difficult to discriminate requires knowing the dimensions of listeners' perceptual space. The second formant in tokens of [l] or [r] may be a different perceptual dimension than the second formant in vowels, for instance. For the purposes of this article, we take the primary signature of Effect 2 to be a peak in discrimination near a category boundary. Absent evidence of the development of such a peak, we tentatively assume that any changes in discrimination could instead be instances of Effect 3.

Identification tasks are challenging to carry out with infants, but the few studies that have directly measured English-learning infants' categorization have found extremely shallow identification boundaries (Burnham, 1986; Burnham et al., 1991). Boundaries become steeper—as measured through aggregated data and individual participants' identification functions—between 3 and 7 years, with differences even between 6- or 7-year-olds and adults in some cases (Burnham, 1986; Burnham et al., 1991; Chen et al., 2017; Hazan & Barrett, 2000; Krause, 1982; Kuijpers, 1996; McMurray et al., 2018; Ohde & Haley, 1997; Simon & Fourcin, 1978; Zlatin & Koenigsknecht, 1975). These changes could be partly due to children's improving ability to perform identification tasks, but task difficulty is not the only factor. Across much of the range between 3-year-olds and adults, the increase in category boundary steepness depends on the category being tested (Slawinski & Fitzgerald, 1998) and on the specific phonetic dimensions along which those categories are tested (Greenlee, 1980; Hazan & Barrett, 2000; Nittrouer, 1992; Nittrouer & Miller, 1997; Nittrouer & Studdert-Kennedy, 1987; Ohde et al., 1995; Ohde & Haley, 1997), indicating that children are reweighting different dimensions as cues to category membership. These differential changes in category boundary steepness strongly suggest that at least some category learning occurs later in childhood.

Discrimination peaks along phonetically relevant dimensions sharpen in tandem with the changes in category boundary steepness later in childhood (Chen et al., 2017; Medina et al., 2010), whereas in infants, evidence for the development of discrimination peaks is mixed.

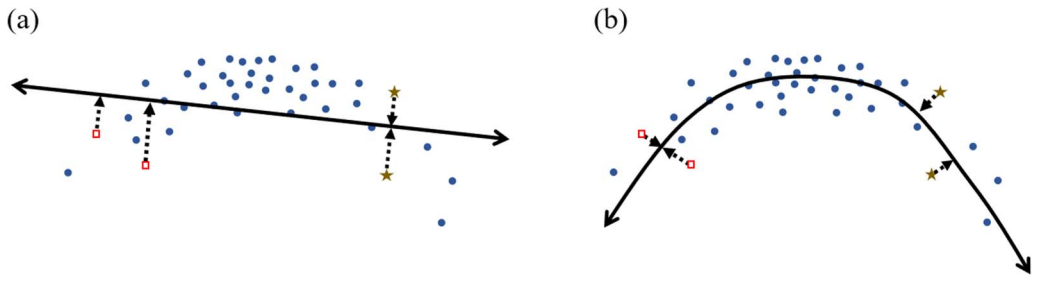

Figure 3. Example illustrating how different perceptual space learning methods could lead to different perceived distances between the same original points. Here, both methods map points from a two-dimensional space to a one-dimensional line. The mapping is shown explicitly for only four points; distances along the line correspond to perceptual distances in the learned space. (a) In the linear mapping, the brown stars are mapped to the same location, so the distinction between these points is lost, whereas the red squares remain distinct. (b) In the nonlinear mapping, the opposite holds.

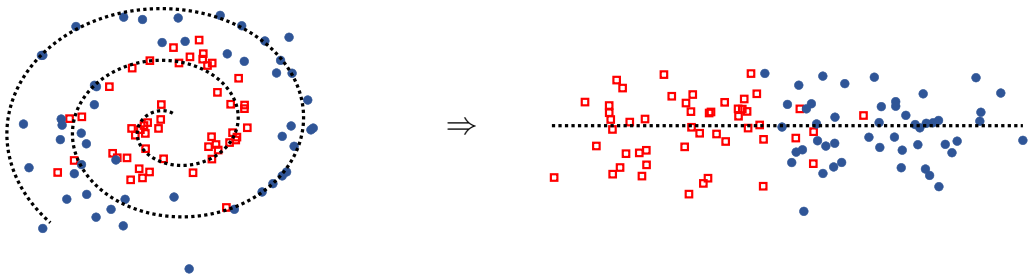

Figure 4. Perceptual space learning can make category learning easier. The marker shapes/colors represent ground truth category labels, which are unknown to the learner; the dotted line highlights the transformation. The decision boundary is simpler after transforming the space.

Newborn and 6-month-old English and Swedish learners show cross-linguistic differences in vowel perception for [i] and [y] (Kuhl et al., 1992; Moon et al., 2013), and English-learning 6-month-olds' discrimination is worse near a prototypical [i] than near a nonprototypical [i], similar to adults (Grieser & Kuhl, 1989; Kuhl, 1991). These studies are suggestive, but do not provide direct evidence that between-category discrimination peaks are developing in infancy. In consonants, there are cross-linguistic differences in infants' voice onset time (VOT) discrimination (Eilers et al., 1979; Streeter, 1976), with a clear peak in discrimination near the phonetic category boundary in English-learning 1- and 4-month-old infants (Eimas et al., 1971). However, a meta-analysis of infant studies with English learners did not find evidence that the VOT discrimination peak sharpens over the first year of life (Galle & McMurray, 2014). Moreover, the discrimination peak is also present in nonhuman animals (Kuhl, 1981; Kuhl & Miller, 1975; Kuhl & Padden, 1982), suggesting that it arises from an auditory discontinuity. Whether auditory discontinuities constitute knowledge of categories, and how they relate to subsequent perceptual learning, is less clear (Chládková & Paillereau, 2020). One study did find that French-learning infants' VOT discrimination changes between 4 and 8 months in the direction that would be expected if they were learning phonetic categories (Hoonhorst et al., 2009), providing some evidence of a developing discrimination peak. Overall, however, there is little convincing evidence that peaks in discrimination along phonetically relevant dimensions sharpen substantially during infants' first year.

The literature thus suggests that different perceptual changes occur at different ages. Infants' discrimination changes substantially during the first year (Effect 3), but changes that are diagnostic of category learning (Effect 1) and of increasing perceptual separation between clusters of sounds (Effect 2) are most clearly documented later in childhood. Existing accounts nevertheless attribute both infant and childhood perceptual changes to category learning (Burnham, 1986; Zevin, 2012). We question this interpretation for two reasons. First, as we argue in the next section, general changes in discrimination are compatible with various perceptual space learning algorithms that do not require phonetic categories at all. Second, for phonetic categories to be the cause of those drastic early perceptual changes, one must either posit well-developed categories (in which case the missing evidence of Effect 2 is puzzling), or suppose that noisy, poorly developed categories can drive a drastic reshaping of the perceptual space to yield Effect 3, even though those same category representations are too noisy to yield discrimination peaks along phonetically relevant dimensions (Effect 2).

For these reasons, we believe it is time for the field to consider the possibility that infants' perceptual changes primarily reflect a perceptual space learning process. Early perceptual development would look more like Figure 2b, or a more sophisticated variant (discussed following). Learning phonetic categories to carve up this perceptual space could then extend

well into childhood and even adolescence. Although there is, as yet, little empirical evidence to distinguish this hypothesis from the early phonetic category learning hypothesis, the latter makes stronger assumptions about the nature of early representations that have yet to be clearly validated.

COMPUTATIONAL APPROACHES TO PERCEPTUAL SPACE LEARNING

Although cognitive scientists have proposed a handful of perceptual space learning models for speech (Gauthier et al., 2007; Guenther & Gjaja, 1996; Herrmann et al., 1995; Nixon & Tomaschek, 2021; Westermann & Reck Miranda, 2004), perceptual space learning is more actively studied in the machine learning community, where it is well-known that modified representations of input features can be learned without access to, and without necessarily resulting in, categorical knowledge. This type of learning has been used in many domains, including vision and speech (Chung et al., 2019; Erhan et al., 2010; Kamper et al., 2015; Ranzato et al., 2007; Schneider et al., 2019; van den Oord et al., 2018; Yu et al., 2010), and there is even a series of recent speech technology challenge tasks devoted to the topic (Dunbar et al., 2017, 2019; Versteegh et al., 2015).

Perceptual space learning is popular in machine learning because it can improve a system's ability to learn from the signal: for example for speech, spectral information, or even waveforms. In contrast, cognitive models often use more abstract features (such as formants) as input. However, starting from abstract features skips over a critical part of the learning process, wherein infants must learn which of the many dimensions of raw speech are relevant to processing their native language. We argue that this aspect of learning, which most cognitive models do not consider at all, could explain many of the perceptual changes seen in young infants.

To illustrate, consider a well-known method for perceptual space learning: principal component analysis (PCA). PCA reduces the dimensionality of data in order to learn a more compact representation that still preserves the most important information. For example, in the speech domain each input data point might represent a short (10 ms) slice of speech using a vector where each dimension represents the value of some acoustic measure such as spectral energy. Some of these dimensions may vary independently, while others may be highly correlated or simply record random noise—thus, most of the information can be represented using a smaller number of dimensions. PCA identifies the orthogonal dimensions of greatest variation in the original data, rotates these to align with the axes of the vector space, and discards dimensions with low variation. That is, it learns a representation that is optimized to capture the greatest amount of variance in the data.

The transformation learned by PCA is *linear*, since it simply rotates the axes of the space before collapsing some dimensions. However, many perceptual space learning methods are more powerful, in that they learn a *nonlinear* transformation, warping the original space in potentially arbitrary ways (Figure 3).[3] The result is that points that were close together in the input space may end up far apart in the learned space or vice versa. Therefore, if discrimination depends on distance in some perceptual space (Shepard, 1987), perceptual space learning could lead to changes in discrimination.

Although perceptual space learning is not directly optimized for categorization, it could nevertheless help with later category learning by factoring out irrelevant features or warping

[3] Although both of the illustrated methods reduce dimensionality, perceptual space learning can also maintain or even increase dimensions; the key property is that it changes the shape of the input space.

the space in a way that makes the category structure more obvious (Figure 4). This effect has been demonstrated both in cognitive models of auditory learning (Gauthier et al., 2007; Roark et al., 2020) and in machine learning models, where "pretraining" a system's perceptual space on a generic unsupervised task (such as predicting the next input in a sequence) can improve performance on a variety of downstream tasks (such as question answering or phone classification) (Chung et al., 2019; Devlin et al., 2019; Erhan et al., 2010; Peters et al., 2018; Schneider et al., 2019). While it is theoretically possible that systems pretrained on speech could be implicitly learning phonetic categories, evidence from models that do learn quantized representations (latent categories) suggests otherwise: the learned units are typically far more granular than phonetic categories, and often cannot even be well-characterized as sub-phones or subsets of phonetic categories (Baevski et al., 2020; Baevski, Schneider, & Auli, 2019; Chorowski et al., 2019; Hsu et al., 2021; Schatz et al., 2021).

These recent successes in machine learning have led to a proliferation of new work on perceptual space learning algorithms. Thus, cognitive scientists should be considering not just whether perceptual space learning could explain infants' early perceptual development, but more specifically which algorithms might provide good models for infant learning. These algorithms differ in the source of the learning signal and the cognitive plausibility and domain-specificity of the mechanism. For example, self-organizing maps (Kohonen, 1989, 2001) are an early method for nonlinear dimensionality reduction, based on competitive learning. More popular in the speech community are autoencoder neural networks (Chorowski et al., 2019; van Niekerk et al., 2020), which can be viewed as a domain-general learning mechanism inspired by memory encoding: they learn to encode each input into an internal representation that allows the original input to be reconstructed as closely as possible. Other recent algorithms aim to predict missing or upcoming stretches of speech, with the learning signal coming from prediction errors—another cognitively plausible domain-general mechanism (Baevski, Auli, & Mohamed, 2019; Baevski et al., 2020; Baevski, Schneider, & Auli, 2019; Chung et al., 2019; Hsu et al., 2021).

There have also been recent proposals for more domain-specific perceptual space learning methods that rely on a noisy top-down signal provided by knowledge of some word-like units (Kamper et al., 2015; Renshaw et al., 2015; Riad et al., 2018; Thiollière et al., 2015). These units can be found by searching for stretches of speech that form similar pairs or clusters, without any knowledge of phones (Jansen & Van Durme, 2011; McInnes & Goldwater, 2011; Park & Glass, 2008; Räsänen & Blandon, 2020). Assuming that the clusters represent different instances of the same word, the learner can then adjust its current representation of the low-level speech features to make these instances even closer together in perceptual space. Preliminary evidence suggests that models using this mechanism can learn representations that demonstrate some of the effects seen in infants (Matusevych et al., 2020). At a high level, this is essentially the mechanism proposed by Jusczyk (1992), and—unlike the other methods described above—it does use a form of categorical knowledge (word categories) to guide learning. Whereas we argue in the next section that phonetic categories are difficult to learn due to high acoustic overlap, word-like units are likely to have fewer near acoustic neighbors than phones (Swingley, 2009), which could make them easier for infants to discover in naturalistic speech (cf. Jusczyk & Aslin, 1995; Jusczyk et al., 1999).

REVISITING PHONETIC CATEGORY LEARNING

Learners eventually develop sharp identification boundaries and discrimination peaks, providing evidence of well-separated categories (Box 2). Under a phonetic category learning account

of infants' perceptual changes, much of the category learning process happens in infancy. Under a perceptual space learning account, category learning might occur later or more gradually, and even if it begins in infancy, it is not the primary driver of infants' perceptual changes. Either way, there must be a mechanism for learning phonetic categories.

Distributional learning (Maye et al., 2002) has emerged as a leading hypothesis for a mechanism that could operate in infancy. Infants discriminate stimuli better after hearing a bimodal distribution—with two distinct clusters of sounds—along the relevant phonetic dimension than after hearing a unimodal distribution (Cristia, 2011; Maye et al., 2002; Maye et al., 2008; Wanrooij et al., 2014; Yoshida et al., 2010; see Cristia, 2018, for a meta-analysis). This ability to track acoustic distributions of sounds could support category learning if phonetic categories corresponded to well-separated clusters of sounds.

However, while some contrasts in laboratory speech are well-separated acoustically (Lisker & Abramson, 1964), categories overlap substantially in naturalistic speech, as in the second panel of Figure 2a (Antetomaso et al., 2017; Bard & Anderson, 1982; Bion et al., 2013; Hitczenko et al., 2020; Pollack & Pickett, 1963; Swingley, 2019).[4] Most models that have tested the feasibility of distributional learning for identifying phonetic categories have simplified the learning problem, for example, by using artificial data with low variability (McMurray et al., 2009; Pajak et al., 2013; Vallabha et al., 2007), focusing only on subsets of the categories infants would need to acquire (Adriaans & Swingley, 2017; de Boer & Kuhl, 2003; Gauthier et al., 2007), or limiting the training data to a single speaker (Miyazawa et al., 2010; Miyazawa et al., 2011). Similar models that were tested on more realistic datasets showed much worse performance at learning phonetic categories (Adriaans & Swingley, 2012; Jones et al., 2012; Schatz et al., 2021). Therefore, the distributional sensitivity that infants exhibit in simplified laboratory settings may not be sufficient to learn phonetic categories in naturalistic settings. This may still be true even after perceptual space learning (as in the second panels of Figure 2b and Figure 4).

Aside from distributional information, phonetic category learners can draw on additional sources of information, such as word forms or meanings (Swingley, 2009). Infants recognize word forms in fluent speech (Bortfeld et al., 2005; Jusczyk & Aslin, 1995; Jusczyk et al., 1999) and know some word meanings (Bergelson & Swingley, 2012); both can affect infants' discrimination in laboratory settings (Feldman, Myers, et al., 2013; Yeung & Werker, 2009). However, unsupervised phonetic category learning models that use contextual information have again done better when trained in idealized settings than in more naturalistic settings (Antetomaso et al., 2017; Feldman et al., 2013; Frank et al., 2014; C.-Y. Lee et al., 2015).

These differences between naturalistic and idealized settings make category-based accounts of infants' perceptual changes less parsimonious than previously believed. When categories are heavily overlapping along some dimensions, as in the second panel of Figure 2a, separating them—even imperfectly, as in the third panel of Figure 2a—requires finding better dimensions for representing the sounds in the underlying perceptual space. Such a transformation is similar to perceptual space learning, but is driven by category knowledge. Thus, both the category-based account and the perceptual space learning account require the same two learning processes. What is at stake is the interdependence and relative timing of those processes. If phonetic category learning is as difficult as the above evidence suggests, it might be more feasible for older children, who can draw on more knowledge of higher level

[4] Although the degree of overlap depends on the specific dimensions measured, we know of no language-universal set of dimensions that reliably yields well-separated phonetic categories (see also Chládková & Paillereau, 2020).

linguistic structure (McMurray et al., 2018) and benefit from using a learned perceptual space with fewer irrelevant dimensions.

EMPIRICAL EVIDENCE FOR PERCEPTUAL SPACE LEARNING

There is not yet any direct evidence for a perceptual space learning process in infants. However, evidence from adults lends plausibility to such an account. After hearing nonspeech stimuli in which two auditory dimensions are perfectly correlated, listeners can discriminate between stimuli that follow the same correlation as in training, but not those that violate the correlation (Stilp et al., 2010; Stilp & Kluender, 2012), suggesting that correlations among dimensions can drive auditory perceptual space learning. The integration of perceptual dimensions for perceiving speech is not always determined by experience (Kingston et al., 2008; S. Lee & Katz, 2016), but several studies have suggested that an experience-based perceptual space learning process could play a role (Holt et al., 2001; Nearey, 1997; Schertz et al., 2020) and could interact in nontrivial ways with subsequent learning of cue weights (Roark et al., 2020; Roark & Holt, 2019; Scharinger et al., 2013).

Adults are additionally sensitive to temporal structure *within* perceptual dimensions. Their attention to dimensions in visual perception, such as color or shape, is affected by the temporal statistics within each dimension (Zhao et al., 2013)—that is, conditional probabilities, which infants are sensitive to in auditory perception (Saffran et al., 1996). This attentional benefit may well have an analogue in the auditory domain, given that auditory exposure to temporal regularities elicits increased MEG amplitude in auditory cortex relative to random sequences (Barascud et al., 2016). Although there is not yet evidence linking this attentional benefit of temporal structure to infants' early perceptual changes, such a strategy could potentially be effective at identifying informative perceptual dimensions, because language has considerable internal structure.

THE WAY FORWARD

To begin testing which type of theory best accounts for early perceptual development in speech, it is important to take seriously the complexity of speech produced in naturalistic environments. Naturalistic speech varies along many more acoustic dimensions than are typically manipulated in stimuli for speech perception experiments, or represented in phonetic learning models, and several studies have already shown that considering the variability of naturalistic speech can change our understanding of perceptual development (Antetomaso et al., 2017; Bion et al., 2013; Hitczenko et al., 2020). Methods for working with speech in naturalistic settings have been developed in the context of engineering applications, and naturalistic speech corpora now exist in numerous languages. By adapting these tools (e.g., Räsänen, 2011; Räsänen & Rasilo, 2015; Schatz, 2016; Schatz et al., 2013, 2021; Schatz et al., 2018; Schatz & Feldman, 2018), cognitive scientists can begin investigating the role of perceptual space learning in explaining how infants' perception of speech becomes specialized for their native language.[5]

Thus far, we know of only a handful of models that have been evaluated against infant behavioral data after training on natural continuous speech. Schatz et al. (2021) trained a bottom-up distributional learner—specifically, a Dirichlet process Gaussian mixture model—on low-level spectral representations of speech from Japanese or English. The model reproduced infants' discrimination of [ɹ] and [l], but the units it learned did not resemble phonetic categories.

[5] A complete model would also need to incorporate social factors (Conboy et al., 2015; Kuhl et al., 2003; Lytle et al., 2018; Tripp et al., 2021).

Matusevych et al. (2020) found that a recurrent neural network that optimized its hidden representations to represent correspondences between tokens of the same word achieved performance comparable to the model from Schatz et al. (2021). The success of these models suggests that alternatives to the phonetic category learning hypothesis, including perceptual space learning models that have no sub-word categories at all, are well worth exploring. In contrast, we are not aware of a phonetic category–based model that has been trained on continuous, unsegmented speech and used to predict cross-linguistic patterns of infants' discrimination (see Schatz et al., 2021, supplementary discussion 1, for further discussion of this gap in the literature).

Parallels between the phonetic learning and machine learning literatures provide other reasons to be optimistic about perceptual space learning theories. Perceptual space learning algorithms that rely on word-like units (Kamper et al., 2015; Renshaw et al., 2015; Riad et al., 2018; Thiollière et al., 2015) are reminiscent of proposals that the words infants segment from fluent speech can constrain phonetic category learning (Feldman, Griffiths, et al., 2013; Swingley, 2009). The distributional learning strategy that Schatz et al. (2021) used is similar to that proposed by Maye et al. (2002) to learn phonetic categories. Both of these strategies have struggled to scale to more realistic data under a phonetic category learning account (Antetomaso et al., 2017; Bion et al., 2013; Taniguchi et al., 2016), but perform well once the constraint that phonetic categories need to be learned is dropped.

Jusczyk (1992) proposed over 25 years ago that phonetic learning might not rely on phonetic categories, but this idea has largely been disregarded in the literature on phonetic learning. Here we have argued that this idea is consistent with a large body of empirical literature on infant phonetic learning and have connected the proposal to recent trends in speech technology that provide paths toward a formal theory. The time course of phonetic category learning has major implications for our understanding of language acquisition as a whole, and as such we hope this article will inspire serious consideration of the perceptual space learning hypothesis and encourage the kind of rigorous empirical and computational tests that can ultimately distinguish it from the currently popular alternative.

ACKNOWLEDGMENTS

We thank Adam Albright, Richard Aslin, Yevgen Matusevych, Bob McMurray, and two anonymous reviewers for insightful comments.

FUNDING INFORMATION

NHF, National Science Foundation (https://dx.doi.org/10.13039/100000001), Award ID: BCS-1734245. SG, Economic and Social Research Council (https://dx.doi.org/10.13039/501100000269), Award ID: ES/R006660/1. SG, James S. McDonnell Foundation (https://dx.doi.org/10.13039/100000913), Award ID: Scholar Award 220020374. ED, Agence Nationale pour la Recherche, Award ID: ANR-17-EURE-0017 Frontcog. ED, Agence Nationale pour la Recherche, Award ID: ANR-10-IDEX-0001-02 PSL*. ED, Agence Nationale pour la Recherche, Award ID: ANR-19-P3IA-0001 PRAIRIE 31A Institute. ED, Facebook AI Research, Award ID: Research Grant.

AUTHOR CONTRIBUTIONS

NHF: Conceptualization: Lead; Funding acquisition: Lead; Investigation: Lead; Writing – original draft: Lead; Writing – review & editing: Lead. SG: Conceptualization: Lead; Funding acquisition: Lead; Investigation: Lead; Writing – original draft: Lead; Writing – review & editing: Lead.

ED: Conceptualization: Supporting; Funding acquisition: Supporting; Writing – review & editing: Supporting. TS: Conceptualization: Supporting; Writing – review & editing: Supporting.

REFERENCES

Adriaans, F., & Swingley, D. (2012). Distributional learning of vowel categories is supported by prosody in infant-directed speech. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 72–77). Cognitive Science Society.

Adriaans, F., & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *Journal of the Acoustical Society of America*, *141*(3070), 3070–3078. https://doi.org/10.1121/1.4982246, PubMed: 28599541

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and Speech*, *46*(2–3), 155–182. https://doi.org/10.1177/00238309030460020601, PubMed: 14748443

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling phonetic category learning from natural acoustic data. In M. LaMendola & J. Scott (Eds.), *Proceedings of the 41st Boston University Conference on Language Development* (pp. 32–35). Cascadilla Press.

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138. https://doi.org/10.1111/j.1551-6709.2011.01181.x, PubMed: 21609356

Baevski, A., Auli, M., & Mohamed, A. (2019). *Effectiveness of self-supervised pre-training for speech recognition*. ArXiv. https://arxiv.org/abs/1911.03912

Baevski, A., Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*. OpenReview.net.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33* (pp. 12449–12460). Curran Associates.

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., & Chait, M. (2016). Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences*, *113*(5), E616–E625. https://doi.org/10.1073/pnas.1508523113, PubMed: 26787854

Bard, E. G., & Anderson, A. H. (1982). The unintelligibility of speech to children. *Journal of Child Language*, *10*(2), 265–292. https://doi.org/10.1017/S0305000900007777, PubMed: 6874768

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109, PubMed: 22331874

Best, C. T. (1994). Emergence of native-language influences. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). MIT Press.

Best, C. T., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*, *46*(2–3), 183–216. https://doi.org/10.1177/00238309030460020701, PubMed: 14748444

Best, C. T., McRoberts, G. W., LaFleur, R., & Silver-Isenstadt, J. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant Behavior and Development*, *18*(3), 339–350. https://doi.org/10.1016/0163-6383(95)90022-5

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. https://doi.org/10.1037/0096-1523.14.3.345, PubMed: 2971765

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), Article e51594. https://doi.org/10.1371/journal.pone.0051594, PubMed: 23437036

Bonnasse-Gahot, L., & Nadal, J.-P. (2008). Neural coding of categories: information efficiency and optimal population codes. *Journal of Computational Neuroscience*, *25*(1), 169–187. https://doi.org/10.1007/s10827-007-0071-5, PubMed: 18236147

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x, PubMed: 15828977

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech*, *46*(2–3), 217–243. https://doi.org/10.1177/00238309030460020801, PubMed: 14748445

Burnham, D. K. (1986). Developmental loss of speech perception: Exposure to and experience with a first language. *Applied Psycholinguistics*, *7*(3), 207–240. https://doi.org/10.1017/S0142716400007542

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*, 231–260. https://doi.org/10.1017/S0305000900011041, PubMed: 1874826

Burns, T. C., Yoshida, K. A., Hill, K., & Werker, J. F. (2007). The development of phonetic representation in bilingual and monolingual infants. *Applied Psycholinguistics*, *28*(3), 455–474. https://doi.org/10.1017/S0142716407070257

Chen, F., Peng, G., Yan, N., & Wang, L. (2017). The development of categorical perception of Mandarin tones in four- to seven-year-old children. *Journal of Child Language*, *44*(6), 1413–1434. https://doi.org/10.1017/S0305000916000581, PubMed: 27916015

Chládková, K., & Paillereau, N. (2020). The what and when of universal perception: A review of early speech sound acquisition. *Language Learning*, *70*(4), 1136–1182. https://doi.org/10.1111/lang.12422

Chorowski, J., Weiss, R. J., Bengio, S., & van den Oord, A. (2019). Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 2041–2053. https://doi.org/10.1109/TASLP.2019.2938863

Chung, Y.-A., Hsu, W.-N., Tang, H., & Glass, J. (2019). An unsupervised autoregressive model for speech representation learning. In *Proceedings of Interspeech* (pp. 146–150). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2019-1473

Conboy, B. T., Brooks, R., Meltzoff, A. N., & Kuhl, P. K. (2015). Social interaction in infants' learning of second-language phonetics: An exploration of brain-behavior relations. *Developmental Neuropsychology*, *40*(4), 216–229. https://doi.org/10.1080/87565641.2015.1014487, PubMed: 26179488

Cristia, A. (2011). Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category. *Journal of the Acoustical Society of America*, *129*(5), 3271–3280. https://doi.org/10.1121/1.3562562, PubMed: 21568428

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition*, *170*, 312–327. https://doi.org/10.1016/j.cognition.2017.09.016, PubMed: 29102857

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, *41*(4), 913–934. https://doi.org/10.1017/S0305000912000669, PubMed: 23406830

Curtin, S., Fennell, C., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word-object associative learning. *Developmental Science*, *12*(5), 725–731. https://doi.org/10.1111/j.1467-7687.2009.00814.x, PubMed: 19702765

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134. https://doi.org/10.1121/1.1613311

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science*, *37*(4), 344–377. https://doi.org/10.1111/cogs.12008, PubMed: 23137418

Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.-N., Miskic, L., Dugrain, C., Ondel, L., Black, A. W., Besacier, L., Sakti, S., & Dupoux, E. (2019). The zero resource speech challenge 2019: TTS without T. In *Interspeech 2019: 20th Annual Congress of the International Speech Communication Association*. https://doi.org/10.21437/Interspeech.2019-2904

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., & Dupoux, E. (2017). The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop* (pp. 323–330). IEEE. https://doi.org/10.1109/ASRU.2017.8268953

Eaves, B. S., Jr., Feldman, N. H., Griffiths, T. L., & Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological Review*, *123*(6), 758–771. https://doi.org/10.1037/rev0000031, PubMed: 27088361

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. https://doi.org/10.2307/1129035, PubMed: 446199

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303, PubMed: 5538846

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, *11*(19), 625–660.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778. https://doi.org/10.1037/a0034245, PubMed: 24219848

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782. https://doi.org/10.1037/a0017196, PubMed: 19839683

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438. https://doi.org/10.1016/j.cognition.2013.02.007, PubMed: 23562941

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, *46*(2), 245–264. https://doi.org/10.1177/00238309030460020901, PubMed: 14748446

Flege, J. E., & Hillenbrand, J. (1986). Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. *Journal of the Acoustical Society of America*, *79*(2), 508–517. https://doi.org/10.1121/1.393538, PubMed: 3950204

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception and Psychophysics*, *62*(8), 1668–1680. https://doi.org/10.3758/BF03212164, PubMed: 11140187

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *Journal of the Acoustical Society of America*, *124*(2), 1234–1251. https://doi.org/10.1121/1.2945161, PubMed: 18681610

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349–366. https://doi.org/10.1037/0096-1523.28.2.349, PubMed: 11999859

Frank, S., Feldman, N. H., & Goldwater, S. (2014). Weak semantic context helps phonetic learning in a model of infant language acquisition. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 1073–1083). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1101

Galle, M. E., & McMurray, B. (2014). The development of voicing categories: A quantitative review of over 40 years of infant speech perception research. *Psychonomic Bulletin and Review*, *21*(4), 884–906. https://doi.org/10.3758/s13423-013-0569-y, PubMed: 24550074

Gauthier, B., Shi, R., & Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition*, *103*(1), 80–106. https://doi.org/10.1016/j.cognition.2006.03.002, PubMed: 16650399

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037/0096-3445.123.2.178, PubMed: 8014612

Greenlee, M. (1980). Learning the phonetic cues to the voiced-voiceless distinction: A comparison of child and adult speech. *Journal of Child Language*, *7*(3), 459–468. https://doi.org/10.1017/S0305000900002786, PubMed: 7440672

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, *25*(4), 577–588. https://doi.org/10.1037/0012-1649.25.4.577

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*(2), 1111–1121. https://doi.org/10.1121/1.416296, PubMed: 8759964

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, *28*(4), 377–396. https://doi.org/10.1006/jpho.2000.0121

Herrmann, M., Bauer, H.-U., & Der, R. (1995). The "perceptual magnet" effect: A model based on self-organizing feature maps. In L. S. Smith & P. J. B. Hancock (Eds.), *Proceedings of the 3rd Neural Computation and Psychology Workshop* (pp. 107–116). Springer. https://doi.org/10.1007/978-1-4471-3579-1_9

Hitczenko, K., Mazuka, R., Elsner, M., & Feldman, N. H. (2020). When context is and isn't helpful: A corpus study of naturalistic speech. *Psychonomic Bulletin and Review*, *27*(4), 640–676. https://doi.org/10.3758/s13423-019-01687-6, PubMed: 32166605

Hochmann, J.-R., & Papeo, L. (2014). The invariance problem in infancy: A pupillometry study. *Psychological Science*, *25*(11), 2038–2046. https://doi.org/10.1177/0956797614547918, PubMed: 25269621

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, *119*(5), 3059–3071. https://doi.org/10.1121/1.2188377, PubMed: 16708961

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2001). Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement? *Journal of the Acoustical Society of America*, *109*(2), 764–774. https://doi.org/10.1121/1.1339825, PubMed: 11248980

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. https://doi.org/10.1016/j.jecp.2009.07.005, PubMed: 19709671

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). *HuBERT: Self-supervised speech representation learning by masked prediction of hidden units*. ArXiv. https://arxiv.org/abs/2106.07447

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956. https://doi.org/10.1037/a0025641, PubMed: 22004192

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009–1021. https://doi.org/10.1037/a0035269, PubMed: 24364708

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. https://doi.org/10.1016/S0010-0277(02)00198-1, PubMed: 12499111

Jansen, A., & Van Durme, B. (2011). Efficient spoken term discovery using randomized algorithms. In *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 401–406). IEEE. https://doi.org/10.1109/ASRU.2011.6163965

Jones, C., Meakins, F., & Muawiyath, S. (2012). Learning vowel categories from maternal speech in Gurindji Kriol. *Language Learning*, *62*(4), 1052–1078. https://doi.org/10.1111/j.1467-9922.2012.00725.x

Jusczyk, P. W. (1992). Developing phonological categories from the speech signal. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 17–64). York.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010, PubMed: 7641524

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*(3–4), 159–207. https://doi.org/10.1006/cogp.1999.0716, PubMed: 10631011

Kamper, H., Elsner, M., Jansen, A., & Goldwater, S. (2015). Unsupervised neural network based feature extraction using weak top-down constraints. In *Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5818–5822). IEEE. https://doi.org/10.1109/ICASSP.2015.7179087

Kingston, J., Diehl, R. L., Kirk, C. J., & Castleman, W. A. (2008). On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics*, *36*(1), 28–54. https://doi.org/10.1016/j.wocn.2007.02.001, PubMed: 19657466

Kohonen, T. (1989). *Self-organization and associative memory*. Springer. https://doi.org/10.1007/978-3-642-88163-3

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Springer. https://doi.org/10.1007/978-3-642-56927-2

Krause, S. E. (1982). Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults. *Journal of the Acoustical Society of America*, *71*(4), 990–995. https://doi.org/10.1121/1.387580, PubMed: 7085987

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin and Review*, *23*(6), 1681–1712. https://doi.org/10.3758/s13423-016-1049-y, PubMed: 27220996

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22, PubMed: 1546117

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, *66*(6), 1668–1679. https://doi.org/10.1121/1.383639, PubMed: 521551

Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *Journal of the Acoustical Society of America*, *70*(2), 340–349. https://doi.org/10.1121/1.386782

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*(2–3), 263–285. https://doi.org/10.1016/S0163-6383(83)80036-8

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, *50*(2), 93–107. https://doi.org/10.3758/BF03212211, PubMed: 1945741

Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics*, *21*(1–2), 125–139. https://doi.org/10.1016/S0095-4470(19)31326-9

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686. https://doi.org/10.1126/science.277.5326.684, PubMed: 9235890

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, *190*(4209), 69–72. https://doi.org/10.1126/science.1166301, PubMed: 1166301

Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception and Psychophysics*, *32*(6), 542–550. https://doi.org/10.3758/BF03204208, PubMed: 7167352

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*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x, PubMed: 16472309

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, *100*(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100, PubMed: 12861072

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608. https://doi.org/10.1126/science.1736364, PubMed: 1736364

Kuijpers, C. T. L. (1996). Perception of the voicing contrast by Dutch children and adults. *Journal of Phonetics*, *24*(3), 367–382. https://doi.org/10.1006/jpho.1996.0020

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Vol. 2, pp. 140–147). KTH and Stockholm University.

Lee, C.-Y., O'Donnell, T. J., & Glass, J. R. (2015). Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, *3*, 389–403. https://doi.org/10.1162/tacl_a_00146

Lee, S., & Katz, J. (2016). Perceptual integration of acoustic cues to laryngeal contrasts in Korean fricatives. *Journal of the Acoustical Society of America*, *139*(2), 605–611. https://doi.org/10.1121/1.4926435, PubMed: 26936544

Lehet, M., & Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cognitive Science*, *41*(S1), 885–912. https://doi.org/10.1111/cogs.12413, PubMed: 27666146

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*, Article 104328. https://doi.org/10.1016/j.cognition.2020.104328, PubMed: 32502867

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. https://doi.org/10.1037/h0044417, PubMed: 13481283

Lim, S.-J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences*, *116*(10), 4671–4680. https://doi.org/10.1073/pnas.1811992116, PubMed: 30782817

Lim, S.-J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x, PubMed: 21827533

Lipski, S. C., Escudero, P., & Benders, T. (2012). Language experience modulates weighting of acoustic cues for vowel perception: An event-related potential study. *Psychophysiology*, *49*(5), 638–650. https://doi.org/10.1111/j.1469-8986.2011.01347.x, PubMed: 22335401

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Liu, L., & Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition*, *133*(2), 385–394. https://doi.org/10.1016/j.cognition.2014.06.004, PubMed: 25128796

Liu, L., & Kager, R. (2016). Perception of a native vowel contrast by Dutch monolingual and bilingual infants: A bilingual perceptual lead. *International Journal of Bilingualism*, *20*(3), 335–345. https://doi.org/10.1177/1367006914566082

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783–1798. https://doi.org/10.1037/xhp0000092, PubMed: 26280268

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309, PubMed: 15065912

Lytle, S. R., Garcia-Sierra, A., & Kuhl, P. K. (2018). Two are better than one: Infant language learning from video improves in the presence of peers. *Proceedings of the National Academy of Sciences*, *115*(40), 9859–9866. https://doi.org/10.1073/pnas.1611621115, PubMed: 30275298

Mattock, K., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy*, *10*(3), 241–265. https://doi.org/10.1207/s15327078in1003_3

Matusevych, Y., Schatz, T., Kamper, H., Feldman, N. H., & Goldwater, S. (2020). Evaluating computational models of infant phonetic learning across languages. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 571–577). Cognitive Science Society.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134. https://doi.org/10.1111/j.1467-7687.2007.00653.x, PubMed: 18171374

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. https://doi.org/10.1016/S0010-0277(01)00157-3, PubMed: 11747867

Mazuka, R., Hasegawa, M., & Tsuji, S. (2014). Development of non-native vowel discrimination: Improvement without exposure. *Developmental Psychobiology*, *56*(2), 192–209. https://doi.org/10.1002/dev.21193, PubMed: 24374789

McInnes, F. R., & Goldwater, S. (2011). Unsupervised extraction of recurring words from infant-directed speech. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2006–2011). Cognitive Science Society.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378. https://doi.org/10.1111/j.1467-7687.2009.00822.x, PubMed: 19371359

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental Psychobiology*, *54*(8), 1472–1491. https://doi.org/10.1037/dev0000542, PubMed: 29952600

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–378. https://doi.org/10.1016/j.cognition.2013.07.015, PubMed: 23973465

Medina, V., Hoonhorst, I., Bogliotti, C., & Serniclaes, W. (2010). Development of voicing perception in French: Comparing

adults, adolescents, and children. *Journal of Phonetics*, *38*(4), 493–503. https://doi.org/10.1016/j.wocn.2010.06.002

Miyazawa, K., Kikuchi, H., & Mazuka, R. (2010). Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model. In *Proceedings of Interspeech* (pp. 2914–2917). https://doi.org/10.21437/Interspeech.2010-757

Miyazawa, K., Miura, H., Kikuchi, H., & Mazuka, R. (2011). The multi timescale phoneme acquisition model of the self-organizing based on the dynamic features. In *Proceedings of Interspeech* (pp. 749–752). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2011-286

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experience in utero affects vowel perception after birth: A two-country study. *Acta Pediatrica*, *102*(2), 156–160. https://doi.org/10.1111/apa.12098, PubMed: 23173548

Mugitani, R., Pons, F., Fais, L., Dietrich, C., Werker, J. F., & Amano, S. (2009). Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology*, *45*(1), 236–247. https://doi.org/10.1037/a0014043, PubMed: 19210005

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. https://doi.org/10.1111/j.1467-7687.2009.00898.x, PubMed: 20443962

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*(6), 3241–3254. https://doi.org/10.1121/1.418290, PubMed: 9193041

Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics*, *20*(3), 351–382. https://doi.org/10.1016/S0095-4470(19)30639-4

Nittrouer, S., & Miller, M. E. (1997). Predicting developmental shifts in perceptual weighting schemes. *Journal of the Acoustical Society of America*, *101*(4), 2253–2266. https://doi.org/10.1121/1.418207, PubMed: 9104027

Nittrouer, S., & Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research*, *30*(3), 319–329. https://doi.org/10.1044/jshr.3003.319, PubMed: 3669639

Nixon, J. S., & Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition*, *212*, Article 104697. https://doi.org/10.1016/j.cognition.2021.104697, PubMed: 33798952

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39, PubMed: 2937873

Ohde, R. N., & Haley, K. L. (1997). Stop-consonant and vowel perception in 3- and 4-year-old children. *Journal of the Acoustical Society of America*, *102*(6), 3711–3722. https://doi.org/10.1121/1.420135, PubMed: 9407663

Ohde, R. N., Haley, K. L., Vorperian, H. K., & McMahon, C. W. (1995). A developmental study of the perception of onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, *97*(6), 3800–3812. https://doi.org/10.1121/1.412395, PubMed: 7790658

Pajak, B., Bicknell, K., & Levy, R. (2013). A model of generalization in distributional learning of phonetic categories. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (pp. 11–20). Association for Computational Linguistics.

Park, A. S., & Glass, J. R. (2008). Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech and Language Processing*, *16*(1), 186–197. https://doi.org/10.1109/TASL.2007.909282

Pegg, J. E., & Werker, J. F. (1997). Adult and infant perception of two English phones. *Journal of the Acoustical Society of America*, *102*(6), 3742–3753. https://doi.org/10.1121/1.420137, PubMed: 9407666

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

Polka, L., & Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *Journal of the Acoustical Society of America*, *100*(1), 577–592. https://doi.org/10.1121/1.415884, PubMed: 8675849

Polka, L., Colantonie, 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), 2190–2201. https://doi.org/10.1121/1.1362689, PubMed: 11386570

Pollack, I., & Pickett, J. M. (1963). The intelligibility of excerpts from conversation. *Language and Speech*, *6*(3), 165–171. https://doi.org/10.1177/002383096300600305

Ranzato, M., Poultney, C., Chopra, S., & Cun, Y. (2007). Efficient learning of sparse representations with an energy-based model. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 1137–1144). MIT Press.

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, *120*(2), 149–176. https://doi.org/10.1016/j.cognition.2011.04.001, PubMed: 21524739

Räsänen, O., & Blandon, M. C. (2020). Unsupervised discovery of recurring speech patterns using probabilistic adaptive metrics. In *Proceedings of Interspeech* (pp. 4871–4875). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2020-1738

Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, *122*(4), 792–829. https://doi.org/10.1037/a0039702, PubMed: 26437151

Renshaw, D., Kamper, H., Jansen, A., & Goldwater, S. (2015). A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. In *Proceedings of Interspeech*. International Speech Communication Association. https://doi.org/10.21437/Interspeech.2015-644

Riad, R., Dancette, C., Karadayi, J., Zeghidour, N., Schatz, T., & Dupoux, E. (2018). *Sampling strategies in Siamese Networks for unsupervised speech representation learning*. ArXiv. https://arxiv.org/abs/1804.11297

Roark, C. L., & Holt, L. L. (2019). Perceptual dimensions influence auditory category learning. *Attention, Perception, and Psychophysics*, *81*(4), 912–926. https://doi.org/10.3758/s13414-019-01688-6, PubMed: 30761504

Roark, C. L., Plaut, D. C., & Holt, L. L. (2020). A neural network model of the effect of prior experience with regularities on subsequent category learning. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1817–1823). Cognitive Science Society.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x, PubMed: 19143806

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926, PubMed: 8943209

Scharinger, M., Henry, M. J., & Obleser, J. (2013). Prior experience with negative spectral correlations promotes information integration during auditory category learning. *Memory and Cognition*, *41*(5), 752–768. https://doi.org/10.3758/s13421-013-0294-9, PubMed: 23354998

Schatz, T. (2016). *ABX-discriminability measures and applications* (Unpublished doctoral dissertation). Université Paris 6.

Schatz, T., Bach, F., & Dupoux, E. (2018). Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception. *Journal of the Acoustical Society of America*, *143*(5), EL372–EL378. https://doi.org/10.1121/1.5037615, PubMed: 29857692

Schatz, T., & Feldman, N. H. (2018). Neural network vs. HMM speech recognition systems as models of human cross-linguistic phonetic perception. In *Proceedings of the Conference on Cognitive Computational Neuroscience*. Cognitive Science Society. https://doi.org/10.32470/CCN.2018.1240-0

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (2021). Early phonetic learning without phonetic categories. *Proceedings of the National Academy of Sciences*, *118*(7), Article e2001844118. https://doi.org/10.1073/pnas.2001844118, PubMed: 33510040

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proceedings of Interspeech* (pp. 1781–1785). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2013-441

Schertz, J., Carbonell, K., & Lotto, A. J. (2020). Language specificity in phonetic cue weighting: Monolingual and bilingual perception of the stop voicing contrast in English and Spanish. *Phonetica*, *77*(3), 186–208. https://doi.org/10.1159/000497278, PubMed: 31018217

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). *wav2vec: Unsupervised pre-training for speech recognition*. ArXiv. https://arxiv.org/abs/1904.05862v4

Segal, O., Hejli-Assi, S., & Kishon-Rabin, L. (2016). The effect of listening experience on the discrimination of /ba/ and /pa/ in Hebrew-learning and Arabic-learning infants. *Infant Behavior and Development*, *42*, 86–99. https://doi.org/10.1016/j.infbeh.2015.10.002, PubMed: 26708235

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243, PubMed: 3629243

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, *17*(4), 443–464. https://doi.org/10.3758/PBR.17.4.443, PubMed: 20702863

Simon, C., & Fourcin, A. J. (1978). Cross-language study of speech-pattern learning. *Journal of the Acoustical Society of America*, *63*(3), 925–935. https://doi.org/10.1121/1.381772

Slawinski, E. B., & Fitzgerald, L. K. (1998). Perceptual development of the categorization of the /r-w/ contrast in normal children. *Journal of Phonetics*, *26*(1), 27–43. https://doi.org/10.1006/jpho.1997.0057

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382. https://doi.org/10.1038/41102, PubMed: 9237755

Stilp, C. E., & Kluender, K. R. (2012). Efficient coding and statistically optimal weighting of covariance among acoustic attributes in novel sounds. *PLoS ONE*, *7*(1), Article e30845. https://doi.org/10.1371/journal.pone.0030845, PubMed: 22292057

Stilp, C. E., Rogers, T. T., & Kluender, K. R. (2010). Rapid efficient coding of correlated complex acoustic properties. *Proceedings of the National Academy of Sciences*, *107*, 21914–21919. https://doi.org/10.1073/pnas.1009020107, PubMed: 21098293

Streeter, L. A. (1976). Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature*, *259*(5538), 39–41. https://doi.org/10.1038/259039a0, PubMed: 1256541

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of /d-ð/ in monolingual and bilingual acquisition of English. *Cognition*, *100*(2), 369–388. https://doi.org/10.1016/j.cognition.2005.04.007, PubMed: 16115614

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, *364*(1536), 3617–3632. https://doi.org/10.1098/rstb.2009.0107, PubMed: 19933136

Swingley, D. (2019). Learning phonology from surface distributions, considering Dutch and English vowel duration. *Language Learning and Development*, *15*(3), 199–216. https://doi.org/10.1080/15475441.2018.1562927, PubMed: 31607832

Taniguchi, T., Nagasaka, S., & Nakashima, R. (2016). Nonparametric Bayesian double articulation analyzer for direct language acquisition from continuous speech signals. *IEEE Transactions on Cognitive and Developmental Systems*, *8*(3), 171–185. https://doi.org/10.1109/TCDS.2016.2550591

Thiollière, R., Dunbar, E., Synnaeve, G., Versteegh, M., & Dupoux, E. (2015). A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In *Proceedings of Interspeech* (pp. 3169–3173). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2015-640

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x, PubMed: 21339861

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience*, *18*(6), 1029–1043. https://doi.org/10.1162/jocn.2006.18.6.1029, PubMed: 16839308

Tripp, A., Feldman, N. H., & Idsardi, W. J. (2021). Social inference may guide early lexical learning. *Frontiers in Psychology*, *12*, Article 645247. https://doi.org/10.3389/fpsyg.2021.645247, PubMed: 34093326

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Vandenhoeck und Ruprecht.

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. *Journal of the Acoustical Society of America*, *120*(4), 2285–2294. https://doi.org/10.1121/1.2338290, PubMed: 17069324

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191. https://doi.org/10.1002/dev.21179, PubMed: 24273029

Underbakke, M., Polka, L., Gottfried, T. L., & Strange, W. (1988). Trading relations in the perception of /r/-/l/ by Japanese learners of English. *Journal of the Acoustical Society of America*, *84*(1), 90–100. https://doi.org/10.1121/1.396878, PubMed: 3411058

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*(33), 13273–13278. https://doi.org/10.1073/pnas.0705369104, PubMed: 17664424

van den Oord, A., Li, Y., & Vinyals, O. (2018). *Representation learning with contrastive predictive coding*. ArXiv. https://arxiv.org/abs/1807.03748v1

van Niekerk, B., Nortje, L., & Kamper, H. (2020). Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. In *Proceedings of Interspeech* (pp. 4836–4840). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2020-1693

Versteegh, M., Thiollière, R., Schatz, T., Cao, X.-N., Anguera, X., Jansen, A., & Dupoux, E. (2015). The zero resource speech challenge 2015. In *Proceedings of Interspeech* (pp. 3169–3173). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2015-638

Wanrooij, K., Boersma, P., & van Zuijen, T. L. (2014). Fast phonetic learning occurs already in 2-to-3-month old infants: An ERP study. *Frontiers in Psychology*, *5*(77), 1–12. https://doi.org/10.3389/fpsyg.2014.00077, PubMed: 24701203

Werker, J. F., Byers-Heinlein, K., & Fennell, C. T. (2009). Bilingual beginnings to learning words. *Philosophical Transactions of the Royal Society B*, *364*(1536), 3649–3663. https://doi.org/10.1098/rstb.2009.0105, PubMed: 19933138

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

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

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. https://doi.org/10.1016/j.cognition.2006.03.006, PubMed: 16707119

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*(1), 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3

Westermann, G., & Reck Miranda, E. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, *89*(2), 393–400. https://doi.org/10.1016/S0093-934X(03)00345-6, PubMed: 15068923

Yang, M., & Sundara, M. (2019). Cue-shifting between acoustic cues: Evidence for directional asymmetry. *Journal of Phonetics*, *75*, 27–42. https://doi.org/10.1016/j.wocn.2019.04.002

Yeung, H. H., Chen, K. H., & Werker, J. F. (2013). When does native language input affect phonetic perception? The precocious case of lexical tone. *Journal of Memory and Language*, *68*(2), 123–139. https://doi.org/10.1016/j.jml.2012.09.004

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*(2), 234–243. https://doi.org/10.1016/j.cognition.2009.08.010, PubMed: 19765698

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Naätänen, R. (2009). Training the brain to weight speech cues differently: A study of Finnish second-language users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332. https://doi.org/10.1162/jocn.2009.21272, PubMed: 19445609

Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy*, *15*(4), 420–433. https://doi.org/10.1111/j.1532-7078.2009.00024.x, PubMed: 32693519

Yu, D., Deng, L., & Dahl, G. (2010, December 10). *Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition* [Paper presentation]. Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning, Whistler, BC.

Zevin, J. D. (2012). A sensitive period for shibboleths: The long tail and changing goals of speech perception over the course of development. *Developmental Psychobiology*, *54*(6), 632–642. https://doi.org/10.1002/dev.20611, PubMed: 22714710

Zhao, J., Al-Aidroos, N., & Turk-Browne, N. B. (2013). Attention is spontaneously biased toward regularities. *Psychological Science*, *24*(5), 667–677. https://doi.org/10.1177/0956797612460407, PubMed: 23558552

Zlatin, M. A., & Koenigsknecht, R. A. (1975). Development of the voicing contrast: Perception of stop consonants. *Journal of Speech and Hearing Research*, *18*(3), 541–553. https://doi.org/10.1044/jshr.1803.541, PubMed: 1186163