



# Advancements in phonetics in the 21st century: Infant speech development <sup>☆</sup>

Elizabeth K. Johnson <sup>a,b,\*</sup>, Katherine S. White <sup>c</sup>

<sup>a</sup> Department of Psychology, University of Toronto, Toronto, Ontario, Canada

<sup>b</sup> Department of Psychology, University of Toronto Mississauga, Mississauga, Ontario, Canada

<sup>c</sup> Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada



## ARTICLE INFO

### Article history:

Received 15 May 2024

Received in revised form 3 May 2025

Accepted 28 May 2025

Available online 14 June 2025

### Keywords:

Infant

Speech perception

Language acquisition

Child development

## ABSTRACT

Infant speech perception emerged as a field late in the 20th century. Early work focused on defining the initial state, and documenting the timecourse of changes in speech perception over the first year of life. At the turn of the century, attention shifted from studying *when* children became attuned to their native language, to asking *how* children achieved this transformation. Statistical learning became the dominant mechanism to explain language development. But, as researchers pushed the bounds of statistical learning, different questions took center stage: given the complexity of spoken language, how do infants determine which regularities to track? And are the patterns infants track influenced by their unique language learning environment? Inspired by these questions, researchers have shifted to studying acquisition across more diverse contexts, and to using dense corpora and big data approaches to examine how individual differences in children's input relate to speech perception in the lab. In this paper, we first review this progression, summarizing how the field has arrived at the current state of the art. We then argue that the time is ripe for the development of new theoretical approaches, and sketch out the loose contours of SLED, a new 21st-century proposal that emphasizes the role of sociophonetic variation and the richness of the speech signal in early development. With advanced tools in hand and data from a wide variety of learning contexts increasingly available, we are excited to see how the field will evolve over the next 25 years.

© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The primary goal of this review is to summarize advances that have been made in the field of infant speech perception during the last 25 years. A secondary – but nonetheless important – goal of this review is to underscore how the rapid recent advances in the field have created the need for new theoretical approaches to understanding how infants acquire spoken language. To illustrate one way of meeting this challenge, we sketch out a new 21st century inspired proposal we have coined SLED. We not only lay out the three primary assumptions of SLED, but also discuss how its predictions differ (or not) from existing models. We then close by discussing how classic and contemporary methodologies can be used in combination to address these and other related predictions. But

before we turn to our two primary goals, we first consider the origins of infant speech research to underscore just how impressive 21st century advances have been.

### 1.1. Before the turn of the century

#### 1.1.1. The birth of a field in the late 20th century

The workings of the infant mind have been shrouded in mystery throughout much of history. With no tools available to reliably assess infants' experience of the world, questions concerning the origins of human cognition were more philosophical than scientific in nature. This was particularly true for questions regarding infants' language capabilities. In the 1900s, diary studies meticulously tracked the words and sounds produced by infants, and researchers theorized about the origins of human language (Leopold, 1949). But diary studies are limited in scope. They are also subject to observer bias and filtered through the adult perceptual system. With no way to ask infants to report on their experience of the world before they began speaking, many assumed the pre-verbal infant mind was quite limited, and that infants could not make sense

<sup>☆</sup> This article is part of a special issue entitled: 'Advancements of Phonetics' published in Journal of Phonetics.

\* Corresponding author at: Department of Psychology, University of Toronto Mississauga, Mississauga, Ontario, Canada.

E-mail address: [elizabeth.johnson@utoronto.ca](mailto:elizabeth.johnson@utoronto.ca) (E.K. Johnson).

of speech until they had mastered some of the motor skills needed to produce words.

But in the latter half of the 20th century, methodological innovations set the stage for change. Animal behavior researchers had developed a looking paradigm to ask non-human primates how they perceived their visual world. Robert Fantz recognized that the same paradigm could be used to tap into the workings of the infant mind. Fantz adapted it for use in human infants (Fantz, 1958) and added a habituation phase, so that infants' ability to tell apart two stimuli could be tested even if they did not have an inherent preference between them (Fantz, 1964). Soon thereafter, this infant habituation paradigm was further adapted for use with auditory stimuli (Clifton & Meyers, 1969). This methodological innovation paved the way for a study that changed how we think about the development of speech and the acquisition of human language.

Speech researchers immediately realized the potential infant habituation paradigms held for addressing longheld mysteries regarding the initial state for infants' perception of linguistic units. The first question they asked was whether infants – like adults – showed categorical perception for speech sounds (Eimas et al., 1971). Young (English-learning) infants were habituated to an English 'ba' syllable and then tested on whether they noticed a shift to 'pa'. The study demonstrated that infants who could not yet babble nonetheless perceived bilabial stop consonants much like adults do (an abrupt shift from 'ba' to 'pa' at a VOT of approximately 30 ms). That is, infants demonstrated something akin to adult-like categorical perception of speech sounds. Numerous studies followed (primarily with English-learning infants), showing that infants carve up many segmental contrasts in the same sort of non-continuous fashion as adults (Eimas, 1974; Eimas, 1975; Eimas & Miller, 1980a, 1980b). Could categorical perception be the key to explaining why humans – and no other animal – could learn spoken language? Could it be evidence that some aspects of language were indeed innately specified (Chomsky, 1965)? And was it possible that motor experience (a key part of adult speech perception theories at the time) might not be necessary for the perception of speech after all? These findings shook the foundation of the 20th century conceptualization of how speech and language development work. We had finally opened the door onto the infant's inner experience, and discovered that it was much richer than we could have imagined. The field of infant speech perception was born.

### 1.1.2. Setting the stage for the 21st century

As time passed, researchers' initial hopes that categorical perception would be the key to unlocking the mysteries of human language development were dashed. Although a growing body of evidence demonstrated that infants carved up speech continua in a fashion that resembled adult speech perception, this did not seem to explain the essence of what made human infants such good language learners. First, it wasn't even clear that what infants were doing was categorical perception. Limitations in infant testing methodologies – which relied on discrimination, but did not get at labeling (or categorizing) – meant that we could not obtain evidence for true categorical perception (McMurray et al., 2000). Second, follow-up work with non-human animals suggested that humans weren't

the only ones who perceived stop contrasts in a non-continuous manner. If a chinchilla or quail performed similarly to a human infant in the perception of human speech contrasts (Kuhl & Miller, 1975), then this deeply undermined the argument that speech was 'special' (i.e., an innately wired perceptual or motor module in the human brain to support the acquisition of spoken language). As a result, the field shifted from viewing infants' performance on categorical perception tasks as evidence for a domain- and species-specific skill to viewing it as a by-product of domain- and species-general auditory perception.

Next, researchers turned their focus from questions of innate endowment to something that was equally remarkable – how quickly infants became attuned to the native language. Researchers established that by 8–10 months of age, infants lost sensitivity to certain speech contrasts that were not used in their native language (e.g., Trehub, 1976; Werker & Tees, 1984). Attunement to language-specific vowel contrasts appeared even earlier (Kuhl et al., 1992; Polka & Werker, 1994). At the same time, work with adults showed that relearning non-native speech contrasts was very difficult, if not impossible (Strange & Dittmann, 1984). These findings merged quite nicely with those in the broader literature suggesting, on the basis of aphasia recovery and cases of late first and second language acquisition, that there was a biologically imposed critical period for the acquisition of language (Newport, 1990). A picture began to emerge where language input led to a pruning of infants' initial universal set of auditory sensitivities and attention to just those speech differences that were contrastive in the native language – potentially due to a loss of cortical plasticity.

These early studies were exciting demonstrations of when infants began tuning into the sound contrasts in their native language, but they left open many questions. First, they were based on a relatively small set of segmental contrasts, with primarily English-learning infants. Other languages carved up the phonetic space differently. For example, whereas English voicing categories are distinguished by short- vs. long-lag VOT, in languages like Spanish and French, they are instead distinguished by negative vs. short-lag VOT, and in Thai, the VOT dimension is carved up into three categories rather than two (Lisker & Abramson, 1964). In Korean, in contrast, there is a three-way distinction among voiceless consonants based on aspiration and tenseness (Cho et al., 2002). Languages also differ widely in the number and type of segments they contain, as well as in the acoustic salience of their contrasts. Would the developmental patterns observed – a loss of sensitivity to non-native distinctions and maintenance of native ones – be found for all types of speech contrasts? Would infants learning languages other than English, or learning more than one language, show different developmental trajectories? And what were the learning mechanisms driving these changes? We also knew very little about infants' perception of speech units above the segment or syllable level, short of some work on newborn infants' preference for familiar versus unfamiliar voices and languages (DeCasper & Fifer, 1980; Mehler et al., 1988). This lack of data on infants' perception of speech units above the segment or syllable was frustrating, especially given developments in our understanding of how adult speech processing worked.

Luckily, as the 20th century drew to a close, additional methodologies were introduced (e.g., the Headturn Preference Procedure; Kemler Nelson et al., 1995 and the Intermodal Preferential Looking Procedure; Golinkoff et al., 1987) that enabled researchers to ask what information infants extracted from continuous speech and – by doing frame-by-frame hand coding of children’s eye movements – to see how their interpretation of spoken words unfolded in real time (Swingley et al., 1998). These developments – coupled with an expanding number of infant speech labs around the world – led to a burst of discoveries as the 20th century drew to a close.

### 1.2. At the turn of the century

With more suitable tools in hand, researchers moved beyond looking at infants’ perception of isolated syllables, and turned to the question of when (and how) infants were able to find words in continuous speech. The results were surprising. Even before they began speaking, infants could not only segment word-sized units from speech, but they did so using language-specific cues to word boundaries (Jusczyk & Aslin, 1995). For example, in English, most content words carry word-initial stress. Remarkably, like English-speaking adults, English-learning 7.5-month-olds appeared to use this information to locate likely word boundaries in speech, segmenting trochaic (but not iambic) words from fluent speech (Jusczyk et al., 1999b). By 9–10 months, infants also segmented words using other language-specific information, like phonotactic structure and the position of allophonic variants (Jusczyk et al., 1999a). The discovery that infants could extract reasonably well-specified word forms from fluent speech led to a tsunami of new questions. Most crucially, the more we learned about how sensitive infants were to the language-specific structure of their native language, the more we were plagued by a fundamental chicken and egg question: how do infants learn how words sound in the first place before knowing any words? Suggestions that infants learned these cues by exposure to isolated words seemed unsatisfactory, since we had growing evidence that most speech directed to infants consists of multi-word utterances (Brent & Siskind, 2001; van de Weijer, 1998).

#### 1.2.1. Statistical solutions to the word segmentation problem

In 1996, a paper was published whose impact on the field of infant speech perception rivaled that of the 1971 Eimas et al. paper on infant categorical perception (Saffran et al., 1996). This study used an artificial language learning paradigm inspired by psycholinguistic studies with adults (Morgan & Newport, 1981) to explore whether infants could use syllable distribution patterns to segment an initial cohort of words from speech. Infants were exposed to a continuous 2-minute stream of speech containing flat prosody and no pauses or other cues to word boundaries besides the likelihood of one syllable following another. The researchers hypothesized that infants would perceive syllable pairs with high transitional probabilities (TPs) between them as words, and those with low TPs as belonging to different words. The artificial language contained four tri-syllabic words that repeated 45 times each, with no word ever repeating in immediate succession and no word containing the same syllables

as any other word. Therefore, the within-word syllable TPs were 1.0 and the between-word syllable TPs were 0.33. After just two minutes of exposure, infants listened longer to trisyllabic strings of syllables that spanned a word boundary than to strings of syllables that formed a word, showing that they had succeeded in segmenting words, despite the absence of any cues to word boundaries other than the syllable TPs. The study had a huge impact on the field. If infants were pulling words out of the fluent speech around them using this general strategy, they could then use these words to extract language-specific patterns (such as the presence of stressed syllables word initially) useful for segmentation. The study also arrived on the scene just as connectionist modeling of cognitive abilities was taking off, leading to exciting interchanges between the infant speech world and various other fields within the cognitive sciences.

### 1.3. The first quarter of the 21st century

The realization that infants as young as 8 months of age could track transitional probabilities between syllables shaped the trajectory of the next two decades of research. Researchers began to explore other problems that might similarly be ‘solved’ by this kind of domain-general learning ability (including the acquisition of word meanings (Smith & Yu, 2008) and grammatical structures (Gómez & Gerken, 1999)). At the level of speech acquisition, researchers wondered whether statistical learning could explain how infants attune to speech contrasts in the native language. By this point, it had become clear that perceptual tuning could take multiple forms, including a loss of sensitivity (narrowing), a heightening of sensitivity to particular contrasts, or even a shifting of category boundaries (Narayan, 2019; Polka et al., 2001). Could statistical learning explain these changes?

#### 1.3.1. New approaches to phonological attunement

Before statistical learning entered the scene, explaining how infants attuned to the segmental contrasts of their native language seemed like an intractable problem. This is because initial attempts to explain infants’ tuning to the speech categories of their native language relied on the notion of minimal pairs (MacKain & Stern, 1985). In much the same way that linguists determine the phoneme inventory of a language, the idea was that learning minimal pairs might clue children in to the relevance of particular speech contrasts. For example, the acquisition of word pairs like ‘pig’/‘big’, ‘pear’/‘bear’, etc. would highlight for the child the importance of the ‘p’/‘b’ distinction, leading to the encoding of those sounds in distinct categories. At the same time, the realization that various tokens of ‘big’ all had the same meaning would lead to within-category differences being ignored.

But this hypothesis did not fit with the documented timing of changes in the perception of native and non-native speech contrasts, which showed that infants’ perception was already being shaped by the native language within the first year of life. Even though the meanings of some words were available earlier than previously imagined – by the age of 6 months (Dale & Fenson, 1996; Tincoff & Jusczyk, 1999) – it seemed unlikely that infants knew enough minimal pairs for word learning to underlie the changes in perception.

Infants' surprising ability to track syllable TPs opened up new possibilities. Perhaps a different type of statistical tracking could be used to learn which sounds signaled meaningful contrasts in the native language. Earlier work in phonetics had documented that a language's voicing categories were mirrored in frequency distributions of tokens along a VOT continuum, providing an interesting possible statistic for infants to track (Lisker & Abramson, 1964). In a landmark study, infants familiarized with a unimodal frequency distribution of tokens along a VOT continuum were less sensitive to the difference between tokens on either side of the midpoint than infants familiarized with a bimodal frequency distribution of the same tokens (Maye et al., 2002). This was followed by demonstrations that exposure to a VOT distribution at one place of articulation influenced infants' perception of contrasts at other places of articulation (Maye et al., 2008) and that this type of learning was possible not just for consonant distinctions, but vowel categories as well (Wanrooij et al., 2014). And finally, after being ignored for the better part of the field's first quarter century (Singh et al., 2022), infants' acquisition of tonal categories also began to receive some attention. Although the picture seemed more complicated than it was with segmental information, the process of attunement to tone categories also appeared to begin within the first year of life (Mattock & Burnham, 2006; Mattock et al., 2008; Yeung et al., 2013; but see Kalashnikova et al., 2024) and there was some evidence that sensitivity to distributional information could underlie the changes (Liu & Kager, 2017).

These laboratory demonstrations provided tantalizing evidence that distributional learning might explain how infants attune to the sound structure of their native language. There were also intriguing suggestions that the statistics of infants' real-world input were indeed altering their perception. For example, infants' perception of different non-native contrasts changed in an order that matched the frequency of the corresponding native categories in the input (Anderson et al., 2003). And infants in bilingual contexts where distributions of phonetic categories overlapped across languages showed a somewhat different developmental trajectory than monolingual infants, suggesting that they were affected by the overlapping distributions (Bosch & Sebastian-Galles, 2003). In addition, infants whose parents produced more distinct distributions of speech categories showed better discrimination of sounds from those categories (Liu et al., 2003; Cristia, 2011). These findings bolstered claims that infants were tracking the distributions of sounds in their environments and that this drove the organization of phonetic categories in the first year.

Thus began a more concerted effort to document the distributional properties of different types of sound contrasts and in different languages. The field focused on two important questions. First, were phonetic categories clearly marked by the input distributions? And, second, were these distributions more pronounced in infant-directed speech (IDS) compared to speech between adults? If so, given the apparent universality of IDS (Fernald et al., 1989) and infants' attention to it (Cooper & Aslin, 1990; Fernald, 1985), it appeared that there might be an answer to the mysteries of early phonetic category learning.

Initial reports were promising, suggesting that IDS not only provided, but indeed enhanced, the statistical patterns mark-

ing speech categories – and that this was true for other languages in addition to English. For example, examination of the input from mothers to infant learners of Japanese, English, and Catalan found reliable distributional cues to vowel categories that were differentiated by length and quality (Werker et al., 2007; Pons et al., 2012). Moreover, there were reports that distributional cues to vowels were more pronounced in IDS than ADS across multiple languages (Kuhl et al., 1997), and that distributional cues were perhaps even sufficient for bilinguals to simultaneously learn distinct vowel contrasts in their two languages (Danielson et al., 2014). This work suggested that adults, whether consciously or not, were producing a clearer signal for infants to learn from (see also Fritche et al., 2021).

However, the initial enthusiasm about distributional learning alone as a solution to early phonetic attunement has been tempered in recent years by developments on two fronts. The first came from much needed work, supported by advancements in recording technology, exploring language input across additional languages and cultures. This work has demonstrated that IDS is not universal and therefore cannot be essential to infants' speech or language development (Casillas et al., 2020; Cristia et al., 2019). But, even if not essential, that still left the possibility that the distributional properties of IDS were beneficial for learning. Unfortunately, however, it turned out that, in some cases, IDS categories were actually *less* distinct in their distributional properties than categories produced in ADS (Benders et al., 2019; Martin et al., 2015; Cristia & Seidl, 2014; McClay et al., 2022). For example, in some cases, even though the distance between the centers of the speech categories increased in IDS, so too did the variability, leading to more overlapping categories (McMurray et al., 2013).

Even more problematically, there were increasing reports that, regardless of register, the acoustic cues corresponding to some phonetic categories did not lend themselves well to this type of statistical learning approach. For example, in some languages with vowel categories that were differentiated by length (e.g., Japanese and Dutch), the distribution of vowels was not bimodal (Bion et al., 2013; Swingley, 2019). Perhaps the initial work focusing on the distributional properties of point vowels and voicing categories for stop consonants had set expectations too high. When other types of categories were considered, distributional regularities were more difficult to glean from the input (Jones et al., 2012). At the same time, a meta-analysis (Cristia, 2018) cast doubt on the robustness of infants' ability to learn about speech categories from even clear distributional patterns. This uncertainty about whether distributional learning can support phonetic acquisition is magnified for learners of more than one language (Bosch & Sebastian-Galles, 2003), as well as those exposed to multiple varieties of the same language.

If such distributional patterns are not as informative in infants' natural input (and, if infants' ability to track them is more tenuous) than it first seemed, then, clearly, infants' tracking of the statistics of individual sounds cannot be the only thing driving phonetic learning in infancy. As speech researchers came to this realization, similar discussions about real-world scalability were being had regarding the use of syllable TPs to segment words from speech.

### 1.3.2. A closer look at statistical solutions to the word segmentation problem

Work on the utility of transitional probabilities for word segmentation was initially very encouraging (Swingley, 2005). Across labs working with different language learning populations, infants as young as 5 to 6 months succeeded in using TPs to extract words from fluent speech (Johnson & Tyler, 2010; Thiessen & Erickson, 2013), and infants appeared to perform even better with artificial languages produced in IDS (Thiessen et al., 2005). There were also claims that young infants could learn language-specific stress and phonotactic patterns by tracking TPs between syllables (Sahni et al., 2010; Thiessen & Saffran, 2007). The emerging story was that infants under about 7 months of age were fully reliant on TPs between syllables to segment words from speech, but that, by 8–9 months, infants started relying more on language-specific cues to word boundaries (Johnson & Jusczyk, 2001; Johnson & Seidl, 2009; Marimon et al., 2024; Thiessen & Saffran, 2003). Whether or not listeners continued to use TPs to segment words from speech beyond early infancy was debated, but not key to the crucial claim that tracking these statistics provided infants with their first solution to the word segmentation problem.

Nonetheless, nagging questions remained (Johnson, 2012). How would infants – who have not yet learned the phonology of their native language – efficiently extract and categorize syllables in their input? For example, how would young English-learning infants know that stop consonants with and without pre-voicing belonged to the same syllable? How would they know that vowels with and without glottalization were instances of the same vowel? How would an infant know whether to consider syllables that differed in pitch or duration as the same syllable, before they knew whether they were learning a language that had lexical stress or tone (Singh et al., 2008)? How would they even confidently know where syllable boundaries were, given the ambiguity in syllable boundaries, as well as the massive reduction and even deletion of entire syllables in casually produced natural speech (e.g., ‘tomorrow’ being realized as ‘t’morrow’, or even ‘morrah’; Lahey & Ernestus, 2013; Ernestus & Warner, 2011)? Artificial language studies in which a single token of each syllable was presented seemed to assume that these challenges were trivial. There were also questions about whether all languages would be as amenable to syllable tracking segmentation strategies as English. Agglutinative languages (e.g., Hungarian), tone languages with many monosyllabic words (e.g., Mandarin), and polysynthetic languages (e.g., Mohawk) would all seem to require a different type of strategy. Did this mean that there is no universal language-general strategy that infants employ to first break the speech signal down into linguistically relevant units?

Another question was whether the TP-tracking skills infants showed in the lab would be able to scale up to the complexities of the real world. On the one hand, some studies suggested that infants could use TPs to segment words from carefully constructed natural speech passages (Pelucchi et al., 2009; Jusczyk et al., 1999b). There was also evidence that infants recognized strings of syllables that frequently co-occurred in their real world input (Ngon et al., 2012). But on the other hand, some studies suggested that infants’ TP-tracking abilities were

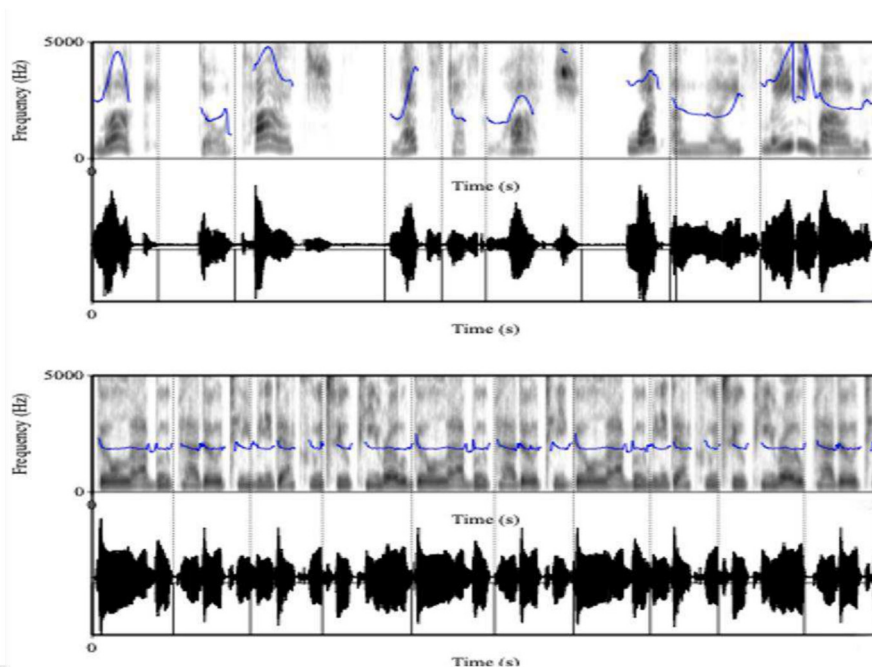
surprisingly fragile. Simply adding the slightest bit of complexity to the artificial languages – like having an artificial language with 2 bisyllabic words and 2 trisyllabic words instead of all trisyllabic words – seemed to throw infants off (Isbilen & Christiansen, 2022; Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012). Other studies suggested that infants were not actually segmenting entire words from the artificial language, but just certain types of syllable pairs (Johnson et al., 2001). And a growing body of work suggested that TPs were very difficult to track without additional convergent speech cues (e.g., Benjamin et al., 2023). With natural speech samples, TP-tracking also seemed constrained by fine-grained prosodic structure; for example, infants did not consider co-occurring syllables spanning a phrase boundary to be part of the same word (Gout et al., 2004; Johnson, 2003; Johnson, 2008). But if TPs alone are not sufficient to solve the word segmentation problem, then how do we explain infants’ early success in segmenting words from fluent speech?

### 1.3.3. The speech signal and segmentation

Traditionally, speech researchers have argued that fluent speech contains no fully reliable cues to word boundaries, and therefore listeners must rely on language-specific cues to locate word boundaries (Cutler, 2012; Cole & Jakimik, 1980). Moreover, with infants receiving mostly multi-word utterances in their input (van de Weijer, 1998; Brent & Siskind, 2001), researchers argued that infants did not hear enough words in isolation to extract language-specific cues to segmentation. Given these claims, it made perfect sense to study infant word segmentation in the lab with streams of continuous speech containing no pausing or prosody. But this is not what real speech is like. It is possible that the simplifications of artificial languages, designed to isolate the word segmentation problem, counterintuitively made it more challenging, by stripping the signal of its natural prosody and variation (see Fig. 1). Could the richness of the natural speech signal mean that infants are not fully dependent on TP-tracking skills to begin pulling words out of speech?

To answer this question, it is useful to reconsider what natural speech really looks like. Even the assumption that the signal contains no pauses does not hold for natural speech. When we speak, we must pause to take a breath, and this intake pause typically coincides with utterance boundaries. Newborns enter the world sensitive to not only utterance boundaries, but also to smaller prosodic units, including intonational (Christophe et al., 2001) and phonological phrase boundaries (Christophe et al., 1994). And although it is true that infants receive predominantly multi-word input, their input consists of relatively short utterances and thus contains far more utterance boundaries than adult speech (e.g., Johnson et al., 2013). IDS itself may also do more than just modulate infants’ attention to the speech signal; it may highlight boundary units in speech (Morgan & Demuth, 1996). Moreover, when we address infants, we likely highlight prosodic boundaries through our visual prosody or gestures even more so than we do when we address adults (Gogate et al., 2000). But how could infants go from sensitivity to major prosodic boundaries to word boundaries?

Critically, phrase and utterance boundaries are coincident with word boundaries. So, one possibility is that prosodically



**Fig. 1.** The top panel shows the waveform and spectrogram for a child-directed utterance 'Look! Don't touch. Let's just watch. What a nice ladybug.' The bottom panel shows the waveform and spectrogram for a string of the artificial language used in Saffran (1996): "golabupadotibidakutupirogolabupadotigolatubidakutupiropadoti." Word boundaries are marked by dotted lines, and the F0 contour is shown in blue. Speech rate and typical utterance length in the top panel approximate that seen in natural child directed speech (~3 words per utterance, at a rate of 4 syllables per second; Johnson, Lahey, Ernestus, & Cutler, 2013). The artificial speech sample in the bottom panel has an infinite utterance length (i.e., no word boundaries) and a speech rate of 3 syllables per second. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

marked utterance and phrase boundaries guide infants' learning about how the edges of linguistic units typically sound in their native language, and constrain the statistical patterns infants track in their input. For example, infants might notice that stressed syllables tend to align with the onset of utterances and phrases, and use this to infer how the edges of linguistically relevant units might be marked in their native language (Johnson et al., 2014). Evidence for the importance of prosodic structure in constraining infants' early speech learning comes from a study showing that six-month-olds do not consider statistically coherent syllables spanning an intonational phrase boundary to be the label for an object, even if those syllables consistently co-occur with the object (Shukla et al., 2011). Additionally, words aligned with utterance edges are more frequent in IDS, due to the shorter utterances. They are also segmented out by infants more readily than utterance-medial words (Johnson et al., 2014) and, in some languages, the types of words that infants tend to learn first are those that tend to be aligned with utterance boundaries (Johnson et al., 2014). Adult speakers have even been found to violate their native language word ordering (e.g., in Turkish) by placing target words in utterance-final position when speaking to infants (Aslin et al., 1996). As infants learn more about the language-specific patterns marking the edges of units aligned with salient prosodic boundaries (i.e., where there are universal cues to prosodic boundaries, such as pauses or large pitch resets), they could become increasingly sensitive to the fine-grained acoustic-phonetic cues marking phrase-internal word boundaries.

The hypothesis that infants use higher-level prosodic boundaries to learn about word boundaries has been termed

the Edge Hypothesis (Johnson et al., 2014; Seidl & Johnson, 2006). Attention to such prosodic boundaries, in combination with some basic assumptions about the possible format of words in human language – for example that all words must contain a vowel (Brent & Cartwright, 1996; Johnson et al., 2003) and that no word can have two syllables with primary stress (Gambell & Yang, 2004) – could get infants quite far in identifying a preliminary set of words (that could then be used to extract other language-specific patterns) useful for segmentation.

#### 1.3.4. Learning multiple levels of structure at once?

Implicit in work up to this point was the assumption that infants acquired each level of language structure independent of other levels (e.g., that infants would first learn the segments of their native language, then find words). However, advances in our understanding of when infants began building a lexicon started to suggest that there is no such linear progression in infant speech development. Indeed, evidence suggested that infants develop a proto-lexicon of frequent word-sized units by the age of 11 months (Ngon et al., 2012; Halle & deBoysse-Bardies, 1994) and a rudimentary set of learned word meanings by the second half of the first year (Bergelson & Swingle, 2012; Tincoff & Jusczyk, 2012), at the same time that they are tuning to the phonetic structure of the language. There was also growing evidence that the learning of these two aspects of language (i.e., sounds and words) might be linked – infants who were more successful in native language speech perception tasks and at segmenting words from fluent speech had larger vocabularies later (Cristia et al., 2014; Kooijman et al., 2013; Newman et al., 2006; Singh

et al., 2012; Tsao et al., 2004). Toddlers also showed better word mapping performance for word forms previously extracted from continuous speech than for novel word forms (Graf Estes et al., 2007). The early assumption was that this relation ran in only one direction – that learning about sounds influenced the learning of word forms and word meanings. But what if the relation also extended in the other direction?

How exactly might a burgeoning lexicon be used in the service of phonetic learning? One possibility is that the presence of statistically overlapping phonetic categories in distinct word units could facilitate infants' separation of those categories (Feldman et al., 2013a; Swingley, 2019; Swingley & Alarcon, 2018; Thiessen, 2007). For example, infants might learn to distinguish the highly overlapping [I] and [E] vowels in English by hearing the former in a frequent word like 'milk' and the latter in a frequent word like 'bed'. This is more plausible than the minimal pair account of early phonetic development, given the composition of the early vocabulary and the timing of phonetic tuning. Moreover, there is evidence that the characteristics of natural IDS could potentially support such a strategy (despite its increased variability compared to speech generated by parents in lab settings; Swingley & Alarcon, 2018; Hitczenko & Feldman, 2022). In particular, a classifier trained to identify word and phonetic units simultaneously outperformed a classifier trained to identify only phonetic units (see also Martin et al., 2013).

But this word-level information is only useful to the extent that infants are able to access it and track information at the phonetic and lexical levels simultaneously. There is evidence that they are able to do so – in one study, infants familiarized to a unimodal distribution of a synthesized vowel contrast in two distinct lexical contexts were better able to discriminate the contrast than infants familiarized to the same sounds in overlapping lexical contexts (Feldman et al., 2013b). In order for this to be a viable strategy, of course, its utility will have to be demonstrated across a broader range of phonetic units, including tones, which have not yet been incorporated into such discussions.

An additional possibility is that utterance edges, in addition to being informative about word boundaries, could play a part in infants' learning about sounds. Imagine that an infant is particularly attentive to utterance boundaries. They have a stuffed otter toy that their caregiver often labels. The word 'otter' often corresponds with the physical presence of the otter toy, and the caregiver often produces co-speech gestures with the otter and places the label in utterance-initial or -final position when talking about it (as is common in child-directed speech) (Jesse & Johnson, 2016). The co-occurrence of the otter in visual space with the often edge-aligned label 'otter' in the child's input, reinforced by the caregiver's exaggerated gestures, prosody, and tactile cues (Brand et al., 2002; Ko et al., 2023; Koterba & Iverson, 2009), helps the child pull out the label 'otter'. As the word 'otter' is heard more and more, the child's ability to recognize the word will grow stronger. The same thing happens with other words in their input. At this point, they have a set of words that sound quite suspiciously like another set of words in medial sentence position in their input. Given the similarity between the pairs of word forms across positions, perhaps infants can group them together (much like they have been shown to do with complementary

distributions at the segmental level; White et al., 2008; White & Sundara, 2014). Now, by comparing the variation in the realization of the segments in the utterance-aligned and utterance-medial words, the child will begin to learn about allowable variation in the realization of segments across word forms. They could realize, for example, that words like 'otter' have a glottalized vowel onset when in utterance-initial position but not in medial position. They may also realize that the medial consonant is sometimes released and sometimes flapped. Armed with this generalization, they can then recognize other words in their input, because they will begin to hypothesize that neither glottalization nor the flap/release contrast in stops is contrastive. In line with this proposal, artificial language learning studies have suggested that listeners could use utterance boundaries to begin to learn about native-language phonotactics, by simply noting which segments occur at utterance onsets and offsets (Sohail & Johnson, 2016). Presumably, this type of analysis would work at lower levels of the prosodic hierarchy as well (and not just utterance boundaries). This type of multi-level, multimodal approach to learning would be consistent with a number of other findings in the literature showing that infants' learning of phonetic categories is boosted by redundant information in the input – whether it is visual articulatory information from the speaker (Teinonen et al., 2008) or the speaker's attentional and gestural cues to referents co-occurring with the speech (Kuhl et al., 2003; see also Beech & Swingley, 2024).

### 1.3.5. Coping with variability

From its earliest days, the lack of invariance in the signal has been **the** central issue in the field of adult speech perception. Variability has been seen as a *problem* to overcome – how, despite the differences across contexts and talkers, do we arrive at a stable percept of language-relevant categories? It did not seem that there were invariant acoustic cues that were present in all instances of a particular phone. Direct perception of gestures or engaging in motor simulation could explain how we perceived similarity despite differing acoustic patterns, but not the fact that variability had effects on adults' processing. For example, adults are slower and less accurate at remembering previously presented words when they are heard in a new voice than when they are spoken in the same voice (Bradlow et al., 1999). Similarly, adult listeners show a reduction in accuracy for identifying vowels in mixed vs. single voice contexts (Nearey, 1989). These findings were interpreted as consistent with a costly process of normalization in which such variability was discarded before listeners accessed the abstract 'linguistic' representations.

Were infants able to perceive through the inherent variability of speech in the same way as adults? Initial findings suggested that the answer was yes. For example, both behavioral and neural studies demonstrated that infants recognized the equivalence of syllables that differed in pitch or speech rate (Dehaene-Lambertz & Baillet, 1998; Eimas & Miller, 1980a; Kuhl, 1983). But later work clouded this picture. When familiarized with new words in the lab, young infants sometimes failed to recognize these words when they were spoken in a new voice or pitch – especially when the familiarization words were relatively similar acoustically (Houston & Jusczyk, 2000; Singh et al., 2008). In contrast, when the initial exposure involved

more variable instances, infants were much better at recognizing words through acoustic changes (van Heugten & Johnson, 2012; Singh, 2008). These findings led to suggestions that infants might not yet know which dimensions of variability were important to include in word representations – and that, even in a non-tonal language like English, they entertain the possibility early on that pitch changes alone could distinguish words (Singh et al., 2008). Therefore, when acoustic features like pitch co-occurred with the word form, infants might assume that these features were also lexically relevant. In contrast, when infants encountered a word produced more variably, they would be able to isolate the critical phonetic dimensions (see also Rost & McMurray, 2009). Findings that a broader range of experienced exemplars led to more robust processing were seen as consistent with exemplar approaches to early word representation – and bolstered claims that similar mechanisms (rather than abstract linguistic representations) could explain word recognition in adults. Importantly, these findings also suggested that the nature of infants' real-world exposure to acoustic and talker variation could have significant consequences for their ability to listen through 'irrelevant' information and recognize words spoken by new individuals.

How much talker variability was there in infants' natural environments? For decades, almost all investigations into the input to young language learners proceeded as if a single (female adult) individual provided all of that input – with almost all characterizations of children's early language environments exploring input from mothers alone. Recent work using cross-cultural and big data approaches has begun to examine this implicit assumption, however, documenting much more diversity in infants' language environments than previously acknowledged. For example, in an analysis of five word types in a restricted North American sample of children, children heard a range of 1–13 talkers, with multiple talkers per word (Bulgarelli et al., 2021), and in some societies, children provide more input than adults (Casillas et al., 2020; Loukatou et al., 2022). Even in more commonly studied Western cultures, child speech is a more significant source of input than has previously been acknowledged (Soderstrom et al., 2018), which is noteworthy given that children's productions are far more variable than adult productions (McLeod & Crowe, 2018; Munson, 2004; Romeo, Hazan, & Pettinato, 2013). There is little work so far linking real-world variability in talkers to infants' learning of speech sounds (Bergmann & Cristia, 2018), although the amount of acoustic variability in a learner's environment appears to be related to the age at which words are produced (Bulgarelli & Bergelson, 2024).

Moreover, knowing the number of talkers still does not tell you about the full range of variability in a learner's input. Even adult talkers do not all differ from one another in the same ways. Some have the same language background, differing from one another in characteristics like pitch or other aspects of voice quality, and some aspects of the realization of speech sounds (one monolingual speaker of North American English might produce voiceless sounds with longer VOTs than another monolingual speaker of North American English, even if both fall within the typical range for the language). In other cases, however, the talkers in a learner's environment have different linguistic systems (due to different accents or developmental stages) and sometimes these systems differ in their

phonetic inventories or realizations of the same phonetic categories. For example, an English-learning child listening to a speaker of French-accented English might hear voiceless sounds in English that fall within the voiced range for monolingual English speakers, or a child speaker producing 'r' in place of 'w'. A Korean-learning child may hear different cues to voicing contrasts when listening to their grandmother than when listening to their mother (because of ongoing sound change; Jang et al., 2024).

### 1.3.6. Variation as information

Recognition that even monolingual learners are exposed to multiple language varieties has inspired the next sea change in the field. Although researchers have acknowledged the complexities of phonetic acquisition for infants with bilingual or multilingual input for the past few decades, it is only recently that multi-dialectal input has started to receive similar attention.<sup>1</sup> In seeking to understand how children cope with multiple varieties in their input, there has been a realization that, far from posing a problem to overcome, some variation might be a tool that links children's language knowledge to their burgeoning knowledge of the social world around them. Investigating how learners treat different types of variation could also provide important insights into the nature of their early representations and the learning mechanisms that support them.

If we accept that monolingual children possess a strong bias to assume a one–one mapping between referents and labels (Markman, 1990), exposure to a new variety that introduces phonetic differences *should* be catastrophic for the early word recognition system. But studies suggest otherwise – over the 2nd year of life, children become increasingly more adept at processing words in unfamiliar accents, particularly when they have some exposure to the accent or other supporting context (Johnson et al., 2022) and when they have larger vocabularies (Mulak et al., 2013). In fact, by 24 months of age, eye-tracking studies suggest that toddlers recognize familiar words in an unfamiliar accent as well as they do in a familiar accent (van Heugten et al., 2015).

How are young learners recognizing words despite this type of variation, and what does it tell us about the nature of their early phonological representations? As has been pointed out in the adult literature, there are different strategies that listeners could use to recognize words in different accents. One, consistent with exemplar accounts to representation, is to simply relax category boundaries (long-term) or criteria for recognition in the moment (Schmale et al., 2015). This could allow listeners to recognize varieties that differ in multiple ways, without having to track specific differences between them. For example, one study demonstrated that English-learning toddlers exposed to Spanish-accented speech recognized just trained words in that accent (Schmale et al., 2012). On an exemplar account, one might expect this recognition benefit to generalize to other (untrained) varieties, such as Korean-accented English. To our knowledge, this has not been

<sup>1</sup> Although we focus our discussion on the situation of multi-accent input in monolinguals (given its neglect in the field to date), many of the points we make are intended to apply to all child language learners, regardless of how many languages they are learning. Indeed, since bilinguals are likely to have more exposure to accent variation than monolinguals, ignoring accent variation may have a particularly strong impact on our understanding of bilingual acquisition.

rigorously tested. Similarly, in most work that has presented adult listeners with novel accent varieties, processing improves with exposure, but it is not clear whether listeners have learned about the accent's specific properties or have simply expanded the range of exemplars they accept.

Importantly, however, other work suggests that toddlers *can* learn something specific about the difference between varieties and, importantly, that this occurs at an abstract, sub-lexical level that allows for generalization of a specific accent's properties across the lexicon. This interpretation is supported by studies in which toddlers are exposed to a new accent in the lab in a context that allows them to detect the specific differences between that variety and their own (van Heugten & Johnson, 2014; White & Aslin, 2011). For example, when introduced to an accent in which words in their own dialect produced with /a/ are heard instead with /ae/ (e.g., "sock" as "sack" – in a context depicting an image of a sock) toddlers later recognize other words undergoing that same /a/–/ae/ shift, even if they did not hear those words in that shift previously. But, critically, they do not recognize words in a new shift (for example, /a/ words pronounced with /E/, e.g., "sock" as "seck"). This type of learning, in which toddlers remap specific segments and generalize this across the lexicon, is remarkably similar to lexically driven learning demonstrated in adults (McQueen et al., 2006) and, as has been argued for adults, appears to require representations that contain abstract, sub-lexical information. Although larger vocabularies may boost performance because of better knowledge of these abstract units (e.g., Edwards et al., 2004, Mulak et al., 2013) even 11-month-olds are able to generalize a newly learned vowel shift across lexical items (Weatherhead & White, 2016).

But what happens when learners are exposed in the real world to multiple varieties of their native language(s)? If a child exposed to British, Jamaican, and Canadian English were to compute statistics on the segmental level alone to determine the vowel inventory of their language, it is hard to imagine the vowel system that would result. Does processing in such learners suggest that they have broader exemplar representations than learners exposed to a single variety (Kartushina et al., 2021)? Or does it instead suggest that they have tracked the specific properties of the varieties in their input? If so, this would suggest that learners have the tools to separate these varieties and track their properties individually. It is only now, in the field's 5th decade, that infant speech researchers have begun to take these questions seriously. There is still little work in this area, and what little there is remains somewhat inconsistent, likely reflecting the diversity of learning environments faced by children exposed to multiple accents. For example, some work suggests that bidialectal toddlers, exposed to one British English variant from their parents and a different one in the community, can recognize only the community variant (Flocchia et al., 2012). However, they also appear to accept mispronunciations that are not present in either dialect (Durrant et al., 2014), suggesting overly broad representations that could impact word learning (consistent with findings of a delay in word recognition in van Heugten & Johnson, 2017). But even 5 month olds are sensitive to different accents (Nazzi et al., 2000) and to potential context cues that they could use to separate them (such as talker race, Quinn et al., 2019). Indeed, there are also suggestions that toddlers can

track specific accent differences and recognize words spoken by new talkers in those accents only when they are consistent with the properties of the accent (van der Feest & Johnson, 2016). This pattern is inconsistent with a general broadening of representations or recognition criteria. Instead, it suggests that infants retain specific detail about the variation in their environment that can be used to interpret future productions (as do adults; Kleinschmidt & Jaeger, 2015).

This variation, in turn, might do more than just tell learners about the language system – they might also leverage this information to learn about and act on their social world. In other words, just as learning about sounds and words can bootstrap one another, so, too, might children's language and social knowledge. From infancy and into childhood, listeners use the way people talk to make inferences and predictions about them, making judgments about who they want to befriend or interact with (Kinzler et al., 2007), where someone is from (Weatherhead et al., 2016; Weatherhead et al., 2018; Weatherhead et al., 2019), and what kinds of preferences an individual has (Lieberman et al., 2016), based on their speech variety. Intriguingly, recent work suggests that this relation between speech and social processing runs in the other direction, too – infants and toddlers use social information to interpret the speech they hear. For example, much like adults whose identification of speech sounds differs depending on the social information provided (Johnson et al., 1999), toddlers treat the same acoustic–phonetic information differently depending on information about the speaker, such as their race (Weatherhead & White, 2018) and age (Bernier & White, 2019). Toddlers also expect people of the same race or people who interact with one another to pronounce words in the same way (Weatherhead & White, 2021). These recent findings, based on learners in more complex environments and experimental situations that incorporate information beyond the speech signal itself, have upended our understanding of early language development. Infants are not learning about their social and linguistic worlds separately – rather, from the early stages, learners' representations of speech (like adults') encode dual, intertwined, functions (Kleinschmidt et al., 2018; Sumner et al., 2013), and extracting linguistic and social information is interdependent and mutually informative. What was seen previously as just noise is information, from the identity of the person talking, to other circumstances in the environment – such as a lollipop in the mouth (White, & Daub, 2021) – that can all be used in the service of inferring the underlying structure. The realization that variation is not just noise – and that learners might be linking linguistic and social information far earlier than thought – necessitates new ways of thinking about infant speech perception going forward.

## 2. Developing new frameworks

Today, advancements in technology and the establishment of more infant research labs around the world have enabled exciting new lines of research that were well out of reach even 20 years ago. But despite the ample data generated in recent years, and methodological innovations, we are still far from an agreement on how infants transition from universal listeners to members of a particular language community, who parse the speech signal with language-specific heuristics. With the solid

foundation of the last 50 years, and the explosion of new testing methodologies and data, where do we go from here? (Fig. 2).

We would argue that a key to continued progress as we move forward into the second quarter of the 21st century is to create new theoretical frameworks better-suited to account for the ever-increasingly complex patterns we are observing (in part due to advances in analytical approaches and technology that allow us to identify and visualize these patterns). New frameworks must consider the diversity of learners' environments, not just from the perspective of the number and type (s) of language(s) being acquired (a point of growing recognition), but also in terms of the complex, dynamic sociolinguistic patterns these language varieties are embedded in. How can we explain infants' success in handling linguistically-diverse input? What strategies might infants use to find the right patterns in their input, rather than becoming sidelined by misleading regularities? How do children learn to speak in the socially dominant accent in their community even when the majority of their input comes from L2 speakers? We believe strongly that the predictive power of these new frameworks will be maximized if we consider speech development broadly, with perspectives from related fields, such as phonetics, sociolinguistics, child development, and the cognitive sciences.

### 2.1. Introducing SLED

What might a 21st century model of infant speech perception, designed from the bottom up with linguistic diversity in mind, look like? In this section, we address this question by sketching out the beginnings of a new framework we will call SLED (Spoken Language Enculturation and Development). By introducing this fledgling framework, we demonstrate one way 21st century advancements in the field can lay the

foundation for models that generate novel and testable predictions.

SLED conceptualizes the goal of speech perception as working out an interlocutor's intended communication, using all available information in the most efficient manner humanly possible. In the SLED framework, the words 'spoken language' are used to admit that the framework is built with the goal of explaining spoken language acquisition, although we would hope that most of the principles we outline would also be relevant to the signed modality. The term 'enculturation' is used to emphasize the important role we think sociolinguistic factors play in speech development. SLED sees diverse language input as the norm, not an exception, and proposes that infants rely on speech variation and contextual factors to make sense of otherwise ambiguous communications, facilitating information uptake and processing. The term 'development' is used to acknowledge the goal of SLED to explain changes as infants transition to children and eventually adults. In SLED, development is seen as a specialization of speech processing abilities, taking infants from a more veridical experience of the world to one that highlights precisely that information that is most suited to children's particular communicative environment.

SLED makes three key assumptions about how infants make sense of their speech input and acquire language. Some of these assumptions reinforce those made by other existing models, whereas others incorporate new perspectives that have been understudied up to now. But all three assumptions feel necessary to us to explain how infants can learn language so quickly and efficiently despite such different input conditions. After laying out these three assumptions, in the following sections, we discuss how SLED differs (or not) from existing models in its approach to fundamental questions in the field and in its predictions – predictions that are now possible for researchers to address, thanks to recent technical develop-

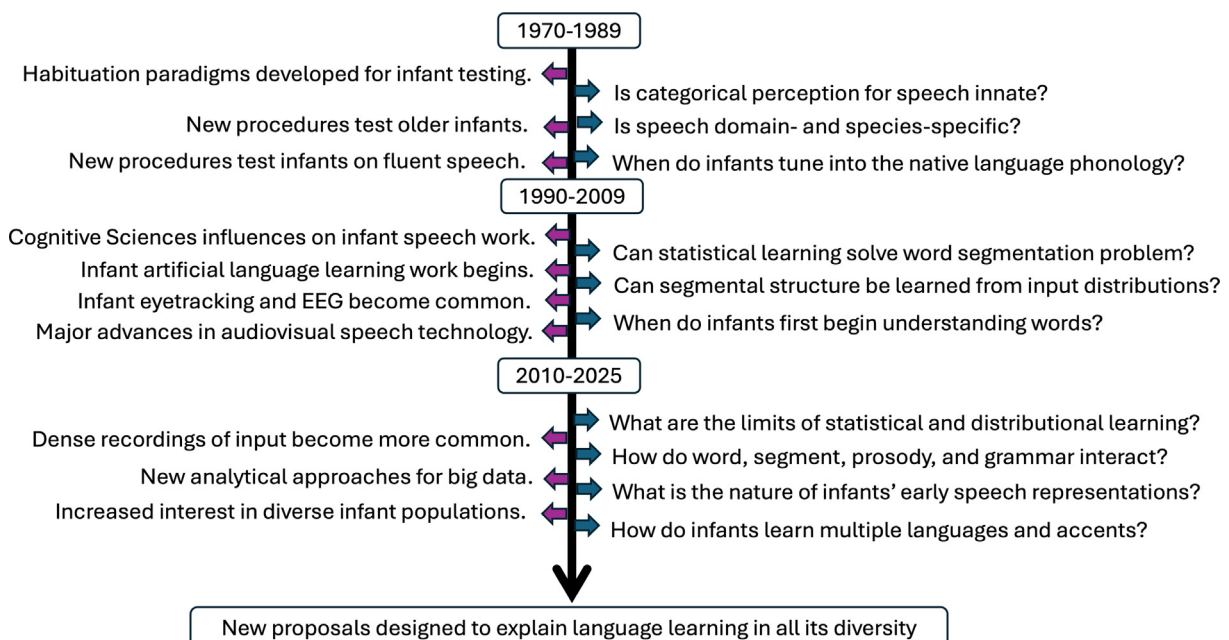


Fig. 2. Important methodological developments (left side) and questions (right side) in the field over the past 50 years.

ments and the expanded scope of the populations being studied.

- (1) **The first assumption of SLED is that linguistic diversity is the norm, rather than a challenge that only a subset of learners need to overcome.** This may at first seem like an uncontroversial statement given recent increases in work on bilingualism, but in fact it is a shift from existing models that either do not directly address the situation of diverse input (e.g., WRAPSA – Jusczyk, 1997; NLM-e – Kuhl et al., 2008; PAM – Best, 1994; DRIBBLER – Morgan, ms.) or that discuss bilingualism as a distinct type of language acquisition (PRIMIR—Curtin et al., 2011). SLED – by default – assumes the learner will probably hear multiple languages or accents in their input, and that a lack of linguistic diversity in a child’s input is the exception rather than the norm. As a corollary of assuming that diverse input is the norm, SLED assumes that linguistically diverse environments will lead to different acquisition trajectories – but importantly, that such diversity should not lead to a significant delay in the discovery of linguistic structure, because infants are equipped to take advantage of the information-rich speech signal.
- (2) **The second assumption of SLED is that the same basic architecture underlies spoken language processing by both infants and adults.** SLED aligns with other proposals that the speech representations infants possess early in infancy are not adult-like (McMurray et al., 2018). Indeed, they will necessarily change, given differences in learners’ perceptual, cognitive, and social abilities, as well as their prior experience. Importantly, however, SLED views many of the differences between young learners and mature language users as a matter of degree, and not kind (White, 2017). Thus, every aspect of SLED is heavily informed by what we know about adult speech processing, and we expect that our understanding of adult speech processing should likewise be informed by our emerging understanding of infant speech processing.
- (3) **The final assumption of SLED is that human speech processing (and learning) cannot be understood without reference to the wide variety of social-cultural contexts within which it is produced.** This assumption is based on insights from sociolinguistics, where language must always be considered within a social or cultural context. SLED assumes that even the development of seemingly low-level perceptual abilities, like distinguishing segmental contrasts, is deeply connected to sociolinguistic awareness. And this is not just because social cues are seen as needed to motivate children to attend to language, as is suggested by many recent proposals in the field of speech development (Nencheva & Lew-Williams, 2022). Nor is it just because social information reduces the noise that variation introduces, or provides more token variability to strengthen exemplar representations. Rather, social information and, contextual information more broadly (White & Daub, 2021), is seen as also providing categorical structure that can constrain computations at various levels of linguistic structure, and help the listener (including infants) make sense out of otherwise ambiguous communications.

## 2.2. Logical consequences of SLED’s assumptions

We now turn to discussing how these assumptions shape the way SLED explains early speech acquisition, focusing on the following questions: Which level of language structure do infants learn first? How do infants extract meaningful patterns from speech? And what types of representations support this process? We see these questions as critical for understanding

how infants learn from a multi-level and highly variable speech signal.

### 2.2.1. What level (or levels) of language do infants track first?

How do infants initially break into the linguistic system(s) of their native language(s)? That is, what level or linguistic unit do infants track first, and in what order do they learn about other levels or units? Do infants proceed from big to small units, or from small to big?

Early approaches tended to assume (either implicitly or explicitly) a linear progression, either from acoustic (or articulatory) features to phonemes to syllables and words, or the reverse, with infants initially attuning to the overall rhythmic and intonation structure of their language and then working their way down to smaller segmental units. However, as discussed in our review above, neither of these progressions – from big to small or small to big – is consistent with infants’ early sensitivity to both segmental and prosodic structure. The state of the art in infant speech work suggests that infants are learning about both the micro (i.e., segmental) and macro (i.e., rhythm and intonation) elements of speech structure from birth (Moon et al., 2013; Mampe et al., 2009; Nazzi & Ramus, 2003).

Therefore, in line with several other contemporary proposals for infant speech processing (Feldman et al., 2013a; Swingley, 2009; Werker & Curtin, 2005), SLED specifies that infants do not proceed sequentially in their acquisition of language, waiting to master one level of linguistic structure before moving on to the next. Rather, because no one level is capable of being interpreted without reference to other levels, infants must track patterns on multiple levels at the same time. In other words, the interpretation of segmental, tonal, word, and prosodic information is inter-dependent and mutually informative. And what is powerful about this type of approach is that learning at one level can constrain learning at other levels (such as the fact that transitional probabilities are not computed across intonational phrase boundaries; Shukla et al., 2011), in an iterative process that will eventually lead to the development of a lexicon and language-specific processing mechanisms. For example, to learn about segmental and tonal contrasts, infants must understand how prosodic structure impacts the realization of these linguistic elements, because many segmental contrasts are difficult to interpret without reference to their placement within a broader prosodic structure (e.g., glottal stops in Maltese; Mitterer et al., 2019). And, since the specific manifestation of prosodic structure (and its impact on the realization of words and segments) varies across languages, attuning to language-specific aspects of the prosodic hierarchy is important to working out all other levels of language structure.

### 2.2.2. What kinds of patterns do infants track, and how do they do it?

In addition to learning multiple levels of language structure at once, SLED also presumes that infants discern meaningful patterns in what other models might consider noise. For example, an exemplar-based model might argue that, because the acoustic–phonetic realization of fricatives is highly variable, infants’ task is to construct a broad representation, so they can be recognized despite these varying realizations. SLED, in contrast, presumes that infants’ goal is to work out what fac-

tors condition variation in the realization of fricatives (e.g., placement in prosodic hierarchy, contextual influences from surrounding segments, accent of the speaker, etc.), and use this information to extract as much information as possible in as efficient a manner as possible from the signal.

Of course, tracking variability as well as the possible sources of this variability increases the complexity of the learning challenge faced by the infant. To account for learners' ability to extract complex patterns from multi-level speech variation, SLED assumes that they are aided by both powerful learning mechanisms and built-in perceptual biases, such as those posited by other 21st century models of infant speech perception (e.g., attention to speech; Lavechin et al., 2024; Shultz & Vouloumanos, 2010). But even with these perceptual constraints/biases, it is hard to explain how infants track the patterns they need to track, especially when acquiring language in the type of linguistically diverse environment that SLED assumes is the norm. Thus, to contend with these challenges, SLED further presumes that there is early emerging or inborn knowledge of the general sound structure of the world's languages, as well as built-in biases that constrain computations. Some evidence in support of this view includes reports that infants seem to expect every word to contain a vowel or sonorant segment (Johnson et al., 2003), and that people who speak alike are in-group members (Lieberman et al., 2017).

In short, SLED does not see initial computations taking place over an undefined acoustic plane, as some 21st century infant speech models suggest (e.g., WRAPSA, PRIMIR, DRIBBLER). Rather, SLED presumes that there are more as of yet to be discovered built-in constraints on how infants package speech input, and one of the most fascinating questions in studying infant speech development is understanding what is wired in, and what is abstracted from the input (and how). This view that infants are powerful learners, but that learning is tightly constrained by in-born biases and constraints, helps explain continuities we see across different language populations.

### 2.2.3. When do abstract representations emerge in development?

As mentioned above, many 21st century models of infant speech perception assume that infants start off with exemplar representations that are built up from an analysis of the acoustic signal (PRIMIR, WRAPSA, DRIBBLER, NLM-e). These models do not view early speech processing as necessarily linguistic. Rather, in many of these models, the development of abstract phonological representations only emerges once children have built up a substantial vocabulary (i.e., presumably around 18 months, when children hit the word spurt). This assumption is also seen in some computational models, where infants do not acquire phonological structure during the first year of life (Schatz et al., 2021). Other models do not posit the development of abstract representations at all (DRIBBLER).

SLED takes a different view on the initial state, presuming that the patterns that infants must track in their input are so complex that they could not be mastered without the use of abstract representations that are formed in the earliest months of life. What evidence exists to support the notion that abstract representations are present far earlier than 18 months of age? Although most evidence for children's abstract representations

emerges well after the first birthday (in alignment with the timing proposed by other contemporary models), there is some evidence compatible with earlier abstraction abilities (e.g., Altuntas et al., 2025; Choi et al., 2017). For example, 6-month-old infants treat CV syllables starting with a particular consonant as the same, even when they differ in the vowel (and therefore, their formant transitions; Hochmann & Papeo, 2014). They also readily recognize words only ever heard in a female voice when presented in a male voice (Johnson et al., 2014). And 8 month olds generalize phonological patterns learned in one place of articulation to other places of articulation (Maye et al., 2008).

This view of early emerging abstract representations is consistent with SLED's assumption of continuity in the basic mental architecture that supports speech processing across development. Many would agree that abstract representations are required to process speech in an adult-like manner. Without abstract representations, it would be difficult to explain how adults handle talker and accent variation (Cutler, 2008; Cutler et al., 2010). Thus, abstract representations are necessary in infants for the same reason they are necessary in adults – it is hard to see how infants could otherwise acquire language so efficiently in diverse settings where multiple talkers and accents are the norm.

### 2.2.4. Are speech representations defined by articulatory or acoustic information?

In the 20th century, many adult models of speech perception saw the units of speech perception as based on motor or gesture representations (Lieberman et al., 1967), because the acoustic speech signal was deemed to be too ambiguous and variable to be used in speech perception. But the emergence of infant speech perception as a field helped bolster the rise of more acoustic-based models of speech perception, since infants' speech perception abilities were considerably more advanced than their production abilities. Current models differ in the roles they propose for articulatory and acoustic information in speech development. PAM defines articulatory gestures as the basic units of speech perception. NLM-e, in contrast, assumes acoustic input is key for speech development, with early-emerging links between perception and production supporting the acquisition of the native language sound structure. Other developmental models (e.g., WRAPSA, DRIBBLER, PRIMIR) assume a receptive lexicon is built largely (if not entirely) from acoustic input, and do not discuss the relationship between production and perception much.

At this point, there is ample support for both acoustic and articulatory representations being used in infants' speech processing, but neither approach on its own seems to be able to satisfactorily account for the available data. For example, how does an articulatory approach explain that infants' perceptual attunement precedes their production abilities? And how does an acoustic approach explain the fact that motor disruptions (either long-term, as a result of cleft palates, or temporary, as a result of a teether) alter speech perception (Bruderer et al., 2015; Southby et al., 2021)? Thus, SLED takes a different view on the debate over the role of articulatory and acoustic information in the development of speech perception abilities. In line with its assumptions about the continuity of the architecture for speech processing over development,

SLED views speech representations not as acoustic or articulatory, but as amodal and built up from acoustic and articulatory information as well as visual and tactile information. That is, while acoustic and articulatory information are both used to build up abstract speech representations (with acoustic information often providing the most informative channel), the representations themselves are not tied to any particular modality. This design feature helps explain how infants – whether they be blind or deaf or motorically challenged, or simply in a visually challenging or noisy environment – make optimal and flexible use of all available information in communicative settings. Although others have proposed that multisensory information plays a role in speech processing, to our knowledge, SLED is unique in its clear stance that speech representations are amodal.

### 3. Gauging the usefulness of a framework

A good framework does more than summarize the status quo in a field. A good framework outlines a theoretically coherent set of assumptions and makes new concrete testable predictions. In many cases, a strong framework is so well specified that it (or at least sub-components of it) can be realized as a computational model. We have put our toe in the water and outlined some ideas for a fledgling framework we have termed SLED. How does it do along these criteria? Is it specified well enough to be realized (at least partially) in a computational model? Does it generate strong testable hypotheses? And importantly, is it falsifiable? (Table 1).

Obviously SLED, being in its infancy, is not formalized in enough detail to be modeled computationally, but it does make several claims that lend themselves to testing. Below we list some of the major predictions that SLED makes. The predictions we discuss below all follow naturally from the assumptions and issues we outlined above. These predictions include the necessity of abstract representations to explain infant speech capabilities, the amodal nature of speech representations, the general continuity in the handling of speech variation across the lifespan, the multitude of factors affecting attunement to the native language sound structure, the rich and essential scaffolding that sociolinguistic information provides for language acquisition, and the robustness of the system to handle linguistically diverse language environments. Some of these predictions are contrary to the predictions made by one or more of the influential 21st models of infant speech perception; others are simply new or understudied predictions, or alternative ways of explaining well-established findings.

#### 3.1. Abstract speech representations must be present early in life

A key testable prediction that SLED makes is that young infants will show evidence of abstract representations early in development. This contrasts with the more prevalent contemporary claim that children do not develop abstract speech representations until 18 months of age. Support for abstraction earlier than 18 months exists, though the evidence is relatively scant and controversial at this point. One line of evidence supporting this prediction is work showing that at least under some conditions, 15-month-olds can detect mispronunciations in an unfamiliar accent after brief exposure to a talker speaking that

accent in the lab (van Heugten & Johnson, 2014). Another line of evidence comes from speech pattern learning in infants under six months (Altuntas et al., 2025). Additional evidence for early abstract representations could be provided by showing that even younger infants can adapt to accent and talker variation, and by providing more evidence that this adaptation can be targeted rather than just a general expansion of what is considered an acceptable pronunciation of a word. SLED's prediction that early representations are abstract could also be supported by more evidence of young infants' transfer of patterns learned with one set of speech segments to other untrained segments that are related in their phonological status (as in Maye et al., 2008).

#### 3.2. Speech representations draw on multi-sensory input, but are amodal

SLED's assumption of amodal representations is consistent with the data reviewed above showing that input from different modalities influences speech perception. Importantly, however, SLED's assumption of amodal representations leads to three additional closely related predictions.

First, because SLED assumes abstract amodal representations for speech, information provided in one modality (e.g., the visual or tactile domain) should impact perception or detection of patterns in the auditory modality (and vice versa). For example, we would expect that visual articulatory information that influences the interpretation of simultaneously presented auditory information (e.g., Weatherhead & White, 2017) would lead to a subsequent change in the processing of that auditory information when presented alone. Moreover, we would expect that the greater the ambiguity in the auditory channel about the intended message, the greater the influence of other information sources on the resulting representation.

Second, SLED's reliance on amodal representations predicts that speech perception and production are not directly coupled. That is, while allowing for evidence that both articulatory and acoustic information affect speech processing, SLED predicts that there should also be observable disconnects between production and perception. In short, SLED views the observed connections between perception and production as not critical to the success of speech perception. For example, although neural evidence of motor activation during speech processing is intriguing (Kuhl et al., 2014), it does not show that such activation is necessary. And although some have argued that children with cleft palates struggle with aspects of speech perception, these children do appear to develop relatively normal speech processing abilities (Abu-Zhaya, et al., 2023; Southby et al., 2021). Further, studies demonstrate that language users do not find their own utterances easier to understand than others' utterances. Rather, adults find more "canonical" voices easier to understand than their own (Schuerman et al., 2015), and toddlers find adults easier to understand than themselves (Cooper et al., 2018). Finally, contrary to the prediction one would make if our articulatory and perceptual representations were one and the same, 2.5-year-olds do not adjust their productions when the formant structure of their utterances is shifted (MacDonald et al., 2012). Taken together, these findings provide support for SLED's proposal that underlying speech representations are amodal, and

Table 1

How SLED compares to existing frameworks for developmental speech perception. (See above-mentioned references for further information.)

	Abstract phonological representations in first year of life?	What is the primary modality of speech representations?	Focus on sociolinguistic competency as key?
WRAPSA (1997)	no	auditory	no
PRIMIR (2005)	no	auditory	no
NLM-e (2008)	no	auditory	no
PAM (2016)	Not clearly specified	articulatory	no
SLED (2025)	yes	amodal	yes

that, although acoustic and articulatory information can both be used by the listener to decode a speaker's intended communication, there is no reason to presume that successful spoken language comprehension necessitates an essential coupling between perception and production.

And finally, a third prediction relates to flexibility in how speech information is communicated by the talker and received by the listener. Past studies have suggested that listeners shift their reliance on audio and visual cues to speech segments, depending on how reliable those cues are in different contexts (e.g., [Bejjanki et al., 2011](#)). Likewise, talkers rely more on visually salient gestures to successfully communicate in noisy settings (e.g., [Trujillo et al., 2021](#)), or settings where the listener is inexperienced ([Gogate et al., 2000](#)). These findings are in line with SLED's assertion that speech representations are amodal, and information in one sensory channel can compensate for lack of reliable information in another sensory channel. But SLED goes one step further, predicting that a listener's relative reliance on different sources of information in the speech signal will not only rely on environmental information, but will also depend on a speaker or listener's linguistic background. Some evidence already exists to support this prediction, including the fact that bilingual infants might use audio-visual speech information differently than monolingual infants ([Birulés et al., 2024](#); [Weikum et al., 2007](#); but see [Morin-Lessard et al., 2019](#)).

### 3.3. There is general continuity in the handling of speech variation across the lifespan

SLED's prediction of abstract representations early in life is in line with its assumption that the same cognitive architecture and learning mechanisms underlie infant and adult speech perception. A corollary of this assumption of continuity is that infants' and adults' processing of spoken language should be much more similar than contemporary research suggests. This includes their ability to understand unfamiliar accents, something that has been argued to be qualitatively different across development.

Although the fact that infants struggle to recognize familiar words spoken in unfamiliar accents prior to 19 months of age has been provided as evidence that infants under 19 months do not possess phonological constancy ([Mulak et al., 2013](#)), adults also find it more challenging to understand words spo-

ken in unfamiliar accents ([Munro, 1998](#)). Thus, instead of interpreting infants' inability to recognize words in unfamiliar accents as evidence that they process speech in a dramatically different way than adults, one could instead argue that, for both adults and infants, it is more challenging to recognize words in unfamiliar accents (see [Bent, 2014](#), for evidence that adults and pre-school-aged children show the same decrement for unfamiliar vs. familiar-accent word recognition). Evidence that even 15-month-olds adapt to unfamiliar accents when given more context, as well as evidence that infants exposed to multiple accents in their day to day input are not delayed in their vocabulary development ([Fung, 2025](#); [van Heugten & Johnson, 2014](#)), supports this argument.

SLED's assumption of continuity can also be extended to address debates about whether infants recognize words in unfamiliar accents through a specific mapping process (i.e., identifying rules that explain the specific phonological mapping between pronunciations in different accents) or a general expansion process (i.e., relaxing the boundaries for what is considered an acceptable pronunciation of a word). The fact that infants have been shown to engage in general expansion in recognition of other-accented words ([Schmale et al., 2015](#)) could be interpreted as evidence that they do not possess the type of abstract representations that are necessary to engage in more targeted accent adaptation. But adults have also been shown to engage in general expansion under some circumstances ([Babel et al., 2021](#); [Melguy & Johnson, 2022](#)). SLED argues that infants – like adults – will flexibly adapt their speech processing strategies, exhibiting either general expansion or accent-specific targeted re-mapping, depending on the context. But importantly, the fact that infants *do* engage in specific re-mapping in some contexts ([White & Aslin, 2011](#)) supports the notion that infants have abstract representations. More generally, SLED argues that apparent differences between infants and adults (as in the representation and processing of acoustic-phonetic detail; [Stager & Werker, 1997](#); [Walley, 1993](#)) often disappear when the task is better equated ([White & Morgan, 2008](#); [White et al., 2013](#)). Moving forward, SLED recommends greater interaction between infant and adult speech researchers, and predicts that many supposed qualitative differences in how infants and adults process spoken language will upon closer examination be explained by differences in experience, task demands, and/or ecological validity of the test scenario.

### 3.4. Computational complexity impacts attunement to the native language sound structure

Classic data suggest that the infant is a universal listener, who then loses the ability to perceive speech contrasts that do not exist in their native language by 10 to 12 months of age (Werker & Tees, 1984). This loss in sensitivity to non-native contrasts is thought to set the stage for lexical growth and more efficient speech processing (Kuhl et al., 2005). But more recent work has shown that there are some caveats to this general pattern of development. What factors explain the variability we see in the acquisition of different types of contrasts by different types of learners (e.g., Narayan, 2019; Polka et al., 2001; Polka & Bohn, 2011; Werker & Hensch, 2015)?

In alignment with previous claims, SLED predicts that more frequent or more salient contrasts will be learned first. In addition, SLED predicts that computational complexity will also affect ease of acquisition. Regularities requiring reference to fine-grained patterns at multiple levels of language may be harder for infants to master than simpler, more encapsulated patterns. Thus, SLED predicts that infants will show predictable errors over the course of development, reflecting the best solution they can propose based on their current knowledge set. Over the course of development, these errors will disappear as infants refine their understanding of the native language structure. For example, as noted earlier, in Maltese, the phonemic status of glottal stops is dependent on the prosodic context within which they occur (Mitterer et al., 2019). Maltese-learning infants may take time to learn the phonemic status of glottal stops because to do so they must not only track the pattern of glottal stops in their input, but also track the distribution of glottals with respect to prosodic structure. And taking this logic one step further, SLED presumes that the same contrasts in any given language might be more or less difficult for a child to learn depending on a wide array of additional factors. For example, if the child is English-Maltese bilingual, they need to work out that these dependencies are critical in only one of their two languages. And if some caregivers in their environment speak Maltese with an English accent (or vice versa), they need to extract even more complex conditional dependencies across the input they receive.

### 3.5. Sociolinguistic factors play a critical role in language acquisition

SLED assumes that infants' access to the rich multi-layered structure of speech includes social context from a very early age. But SLED views sociolinguistic variation as an information tier that scaffolds (rather than complicates) speech and language development. This leads to a number of predictions. First, we should see that learners' interpretation of speech is influenced by social information. Some evidence for this prediction exists, at least for toddlers. For example, the race of a speaker impacts toddlers' readiness to adapt to unexpected pronunciations (Weatherhead & White, 2018), and toddlers treat mispronunciations typically produced by children differently depending on whether they are produced by a child or an adult (Bernier & White, 2019). There is even some evidence that social affiliation changes expectations for how interlocutors will speak (Weatherhead & White, 2021). Extensions of

this fledgling body of work could look at younger infants, and other social domains. For example, would patterns of speech typically associated with males in a given socio-cultural context be surprising (and thus processed with less fluency) if they were produced by a female? Do children in multi-dialectal situations anticipate, and have an easier time processing, lexical variants that are consistent with a talker's language background (e.g., sweater vs. jumper for North American and British English)? More generally, children's (and even infants') predictions and interpretations of new speakers should be influenced by their burgeoning links between language and social categories (which, in turn, will be influenced by the structure of variation in their own environments).

Second, SLED anticipates that learning trajectories will look quite distinct in different children because acquisition is tightly linked to the specific socio-phonetic environment they find themselves in. That is, beyond differences in the specific phonetic data infants receive, SLED also predicts that we should see that infants' learning is influenced by the *source* of this data – importantly, this should be true not just for bilinguals, but for all language learners. For example, in tracking patterns of stop voicing, infants should be able to sort their input based on whether the variation is attributed to talker-specific idiosyncrasies (some speakers might produce more pre-voicing than others) or group-level variation (some L2 speakers might replace all voiceless dental fricatives with voiceless stops). SLED would predict that it is not the statistical patterns alone that matter (as suggested by some exemplar models), but the source of these patterns – because learners are designed to seek out regularity and link speech variation to social variation, these patterns should be learned more quickly when they can be attributed to group-level factors. For example, returning to the situation of our child learning both English and Maltese, they will have a much easier time learning the conditioning of glottal stops in Maltese if they can separate the input from Maltese sources who speak with a native accent vs. those who speak with an English accent. And a Korean child encountering the ongoing shift in the realization of voicing contrasts in the language (Jang et al., 2024), will learn voicing contrasts much more efficiently by realizing that older and younger generations tend to produce the contrast differently. By understanding the range of inputs humans are designed to learn language in, and the range of solutions humans can tap into, we will better understand the essence of human language itself.

### 3.6. The learning system is robust enough to handle complex multi-variety learning environments

Most 21st century models of infant speech learning (implicitly) viewed monolinguals as the standard learner, with children learning multiple languages seen as a special case. This viewpoint can lead to the prediction that the acquisition of more than one language is a challenge to overcome. Recent developments in the field are driving a strong movement away from the notion that learning more than one language is a burden on the child, but the question of how exposure to more than one accent might affect acquisition has not received as much attention. It is easy to understand how accent variation could also be seen as a challenge for children to overcome. For example,

the standard exemplar model would predict that an infant exposed to only one variety of the native language would process that one familiar variety far more efficiently than any other variety of the native language; in contrast, an infant exposed to multiple varieties of the native language would be less efficient in their handling of all varieties of the native language, relative to the infant who is exposed to only one variety.

Evidence to support this prediction is mixed. On the one hand, there is some evidence that children exposed to more than one accent recognize words more slowly (Buckler et al., 2017), and have less well-specified lexical representations than children who have less accent variation in their input (Durrant et al., 2014). But on the other hand, there is evidence that multi-dialect children can rapidly adapt to different accents in their input (van der Feest & Johnson, 2016). Moreover, if multi-accent exposure really led to slow overall processing of speech, one would surely predict cascading effects on other aspects of language development. For example, one would predict that slower speech processing should lead to slower vocabulary growth. But the little work that has been done on this topic does not support this prediction – a recent study examining development in over 3000 infants and toddlers reported that multi-accent exposure appears to have no effect on children's early vocabulary growth (Fung, 2025). Clearly, understanding how multi-accent exposure impacts both monolingual and multilingual children's development is an open question, and is key to evaluating current frameworks for early infant speech development.

SLED takes a strong stance when predicting how multi-accent exposure will impact speech and language development. While SLED is in alignment with exemplar models in assuming a tight coupling between the input a child receives and the development of speech processing abilities, its assumption that linguistically diverse input is the norm leads to the prediction that exposure to multiple accents should *not* make language acquisition particularly challenging. To be clear, this means there often will be detectable differences in the way children exposed to one versus many varieties of the native language process specific aspects of speech, but overall, these processing differences will not have major consequences for the acquisition of language. In other words, SLED predicts that there is no one-size-fits-all approach and that different environments will lead to different optimal paths. Understanding the wide range of normal paths typically-developing infants take to mastering their native language system(s) is key not only to working out a general theoretical explanation of speech and language development, but also to improving clinicians' ability to detect developmentally atypical speech and language trajectories.

#### 4. Summing up, and moving forward

As is the case in many scientific fields, major advances in our understanding of infant speech perception have been linked to technical innovations. The scientific study of infant speech perception was born when researchers devised behavioral test methods that allowed them to tap into the infant mind, enabling researchers to turn what had previously been philosophical questions into scientifically testable hypotheses. As the 21st century neared, advances in audiovisual technology enabled researchers to ask new questions about infants' input

and their processing of fluent connected speech, leading to transformative insights about the remarkable sensitivities and learning abilities present from the earliest months of life. Researchers were able to ask questions about not just what infants knew about language, but the timecourse of their processing as well. These advances helped spur the development of many exciting new models of infant speech development that still shape the way we think about infant speech development today, such as PRIMIR and WRAPSA.

But despite the availability of a wider array of infant testing paradigms, one possible concern is that much of what we know about early speech perception has been gleaned from studies using highly controlled stimuli, stripped of many types of variability infants contend with in the real world. While presenting infants with simplified speech stimuli in the lab has allowed us to address questions that seem impossible to address otherwise, it comes with the risk of distorting our understanding of how children learn about the sound structure of their native language, because researchers do not necessarily know which aspects of the speech signal contribute to infants' language learning success. For example, although most studies of infants' voicing perception rely on manipulations of VOT, there are other cues to voicing present in natural speech as well (Kim et al., 2018). And, it can be difficult to predict how decisions we make when constructing an artificial language might affect infants' performance. For example, it makes sense to assume that presenting words of all one length in artificial languages simplifies the word segmentation task (Johnson & Tyler, 2010), but more recent studies have suggested that more variable word lengths can be useful when the frequency distribution is Zipfian, as in natural language (Lavi-Rotbain & Aron, 2022). Similarly, adding talker variation to an artificial language may enhance learning in some but not all situations (e.g., Bulgarelli & Weiss, 2021; Seidl et al., 2014). Therefore, when experiments simplify real world learning problems by stripping away information like the natural richness of the speech signal, the properties of natural language structure, and even social context, this means that they define the learning space in a way that may not reflect how infants extract information in the real world.

To summarize, it is hard to predict what dimensions are critical to infants' learning. Moreover, this may change as a function of a child's language environment (e.g., a child who is routinely exposed to multiple languages and/or more accent variation may track patterns differently than a child learning just one language variety). Going forward, it is critical that we not only continue to add back these features of natural speech and language variability that were missing in the field's early approaches, but also consider how added variability is used in real-world contexts as well as in the laboratory context. Fortunately, this is increasingly feasible due to methodological innovations that allow researchers to couple more controlled approaches to studying speech and language development with others that more faithfully represent the richness and diversity of infant learners' environments.

Beyond presenting stimuli that more faithfully track the variation in the real world, our technological advances and the proliferation of infant labs and team science now enable us to look at the rich and diverse social circumstances in which infants acquire language, and to investigate their knowledge *embedded* within these contexts. As we have made clear in our pre-

sensation of SLED, we see consideration of the sociolinguistic context in which infants are learning as key to understanding speech acquisition. We argue that there are many paths to language acquisition, and that humans are designed to find whichever path is most appropriate for them to work out the structure of their native language(s) and dialect(s) (Foushee & Srinivasan, 2024). To document these diverse paths, one of the most exciting developments is the possibility of precisely linking variation along dimensions of the input to individual trajectories of perceptual development, something that has been out of reach until recently (Bergelson et al., 2019; Bergelson et al., 2024; Cristia, 2011; Johnson et al., 2013; Lavechin et al., 2022). The rise of dense corpora, team science, and sophisticated analytical approaches – combined with our more traditional, carefully crafted experimental techniques – will revolutionize the way we understand development and will pave the way for discovery in the years to come.

## 5. In summary

Progress in understanding how infants perceive the speech signal, and attune to their native language phonology, has come a long way in the first quarter of the 21st century. But as we enter the second quarter of the 21st century, thanks to technological developments and the expansion of infant labs around the world (coupled with the rise of big data approaches to science, e.g., Frank et al., 2017), data on infant speech perception is ever increasing. Looking forward, it seems clear to us that our frameworks for understanding infant speech development need to account for the fact that infants are not designed to learn a single system (or even two), but rather to handle multiple language varieties from the beginning (where varieties do not just constitute different languages, but the many variations that exist within a single language). Moreover, infants do not learn the speech system of their language(s) in a vacuum, but embedded within a complex social world. Paradoxically, this added complexity of situating learning within a socio-cultural context may simplify the learning problem, by providing infants with a rich structure that focuses their attention and learning on just those statistics that enable them to learn their speech system in the most efficient way possible. With the growing availability of new and/or more powerful methodologies for studying the complex relationship between specific types of language learning environments and individual differences in speech and language learning trajectories, coupled with an increase in data from around the world on children growing up in linguistically diverse settings, we are excited to see the progress the field will make in the next quarter of a century.

## CRedit authorship contribution statement

**Elizabeth K. Johnson:** Writing – review & editing, Writing – original draft, Conceptualization. **Katherine S. White:** Writing – review & editing, Writing – original draft.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- Abu-Zhaya, R., Goffman, L., Brosseau-Lapr e, F., Roepke, E., & Seidl, A. (2023). The effect of somatosensory input on word recognition in typical children and those with speech sound disorder. *Journal of Speech, Language, and Hearing Research*, 66(1), 84–97. [https://doi.org/10.1044/2022\\_JSLHR-22-00226](https://doi.org/10.1044/2022_JSLHR-22-00226).
- Altuntas, E., Best, C. T., Kalashnikova, M., G tzt, A., & Burnham, D. (2025). Phonological feature abstraction before 6 months: amodal recognition of place of articulation across multiple consonants. *Developmental Science*, 28(2). <https://doi.org/10.1111/desc.13605>.
- 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/10.1177/002383090304600206>.
- Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Lawrence Erlbaum Associates Inc.
- Babel, M., Johnson, K. A., & Sen, C. (2021). Asymmetries in perceptual adjustments to non-canonical pronunciations. *Laboratory Phonology*, 12(1). <https://doi.org/10.16995/labphon.6442>.
- Beech, C., & Swingle, D. (2024). Relating referential clarity and phonetic clarity in infant-directed speech. *Developmental Science*, 27(2). <https://doi.org/10.1111/desc.13442>.
- Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: insights from audio-visual speech perception. *PLoS One*, 6(5). <https://doi.org/10.1371/journal.pone.0019812> e19812.
- Benders, T., Pokharel, S., & Demuth, K. (2019). Hypo-articulation of the four-way voicing contrast in Nepali infant-directed speech. *Language Learning and Development*, 15(3), 232–254. <https://doi.org/10.1080/15475441.2019.1577139>.
- Benjamin, L., Flo, A., Palu, M., Naik, S., Melloni, L., & Dehaene-Lambertz, G. (2023). Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates. *Developmental Science*, 26(2). <https://doi.org/10.1111/desc.13300> e13300.
- Bent, T. (2014). Children's perception of foreign-accented words. *Journal of Child Language*, 41(6), 1334–1355. <https://doi.org/10.1017/S0305000913000457>.
- Bergelson, E., & Swingle, 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>.
- Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, 22(1). <https://doi.org/10.1111/desc.12724> e12724.
- Bergelson, E., Soderstrom, et al. (2024). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, 120(52). <https://doi.org/10.1073/pnas.2300671120> e2300671120.
- Bergmann, C., & Cristia, A. (2018). Environmental influences on infants' native vowel discrimination: the case of talker number in daily life. *Infancy*, 23(4), 484–501. <https://doi.org/10.1111/infa.12232>.
- Bernier, D. E., & White, K. S. (2019). Toddlers process common and infrequent childhood mispronunciations differently for child and adult speakers. *Journal of Speech, Language, and Hearing Research*, 62(11), 4137–4149. [https://doi.org/10.1044/2019\\_JSLHR-H-18-0465](https://doi.org/10.1044/2019_JSLHR-H-18-0465).
- Best, C. T. (1994). The emergence of native-language phonological influences in infants: a perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: the transition from speech sounds to spoken words* (pp. 167–224). The MIT Press.
- Best, C. T., Goldstein, L. M., Nam, H., & Tyler, M. D. (2016). Articulating what infants attune to in native speech. *Ecological Psychology*, 28(4), 216–261. <https://doi.org/10.1080/10407413.2016.1230372>.
- 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). <https://doi.org/10.1371/journal.pone.0051594> e51594.
- Birul s, J., Bosch, L., Lewkowicz, D. J., & Pons, F. (2024). Time course of attention to a talker's mouth in monolingual and close-language bilingual children. *Developmental Psychology*, 60(1), 135–143. <https://doi.org/10.1037/dev0001659>.
- Bosch, L., & Sebastian-Galles, 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>.
- Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, 61(2), 206–219. <https://doi.org/10.3758/BF03206883>.
- Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for 'motionese': modifications in mothers' infant-directed action. *Developmental Science*, 5(1), 72–83. <https://doi.org/10.1111/1467-7687.00211>.
- Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93–125. [https://doi.org/10.1016/S0010-0277\(96\)00719-6](https://doi.org/10.1016/S0010-0277(96)00719-6).
- Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2), B33–B44. [https://doi.org/10.1016/S0010-0277\(01\)00122-6](https://doi.org/10.1016/S0010-0277(01)00122-6).
- Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences*, 112(44), 13531–13536. <https://doi.org/10.1073/pnas.1508631112>.
- Buckler, H., Oczak-Arsic, S., Siddiqui, N., & Johnson, E. K. (2017). Input matters: Speed of word recognition in 2-year-olds exposed to multiple accents. *Journal of*

- Experimental Child Psychology*, 164, 87–100. <https://doi.org/10.1016/j.jecp.2017.06.017>.
- Bulgarelli, F., & Bergelson, E. (2024). Linking acoustic variability in the infants' input to their early word production. *Developmental Science*, 27(6). <https://doi.org/10.1111/desc.13545> e13545.
- Bulgarelli, F., & Weiss, D. J. (2021). Desirable difficulties in language learning? How talker variability impacts artificial grammar learning. *Language Learning*, 71(4), 1085–1121. <https://doi.org/10.1111/lang.12464>.
- Bulgarelli, F., Mielke, J., & Bergelson, E. (2021). Quantifying talker variability in North-American infants' daily input. *Cognitive Science*, 46(1). <https://doi.org/10.1111/cogs.13075> e13075.
- Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tselal Mayan village. *Child Development*, 91(5), 1819–1835. <https://doi.org/10.1111/cdev.13349>.
- Cho, T., Jun, S.-A., & Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, 30, 193–228. <https://doi.org/10.1006/jpho.2001.0153>.
- Choi, J., Cutler, A., & Broersma, M. (2017). Early development of abstract language knowledge: evidence from perception-production transfer of birth-language memory. *Royal Society Open Science*, 4(1). <https://doi.org/10.1098/rsos.160660> 160660.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Christophe, A., Mehler, J., & Sebastian-Galles, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy*, 2(3), 385–394. [https://doi.org/10.1207/S15327078IN0203\\_6](https://doi.org/10.1207/S15327078IN0203_6).
- Christophe, A., Dupoux, E., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *The Journal of the Acoustical Society of America*, 95(3), 1570–1580. <https://doi.org/10.1121/1.408544>.
- Clifton, R. K., & Meyers, W. J. (1969). The heart-rate response of four-month-old infants to auditory stimuli. *Journal of Experimental Child Psychology*, 7(1), 122–135. [https://doi.org/10.1016/0022-0965\(69\)90091-5](https://doi.org/10.1016/0022-0965(69)90091-5).
- Cole, R. A., & Jakimik, J. (1980). A model of speech perception. *Perception and production of fluent speech*, 133(64), 133–142.
- Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, 61(5), 1584–1595. <https://doi.org/10.1111/j.1467-8624.1990.tb02885.x>.
- Cooper, A., Fecher, N., & Johnson, E. K. (2018). Toddlers' comprehension of adult and child talkers: adult targets versus vocal tract similarity. *Cognition*, 173, 16–20. <https://doi.org/10.1016/j.cognition.2017.12.013>.
- Cristia, A. (2011). Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category. *The Journal of the Acoustical Society of America*, 129(5), 3271–3280. <https://doi.org/10.1121/1.3562562>.
- Cristia, A. (2018). Can infant learn phonology in the lab? a meta-analytic answer. *Cognition*, 170, 312–327. <https://doi.org/10.1016/j.cognition.2017.09.016>.
- 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>.
- Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: a time allocation study. *Child Development*, 90(3), 759–773. <https://doi.org/10.1111/cdev.12974>.
- Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoot, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, 85(4), 1330–1345. <https://doi.org/10.1111/cdev.12193>.
- Curtin, S., Byers-Heinlein, K., & Werker, J. F. (2011). Bilingual beginnings as a lens for theory development: PRIMIR in focus. *Journal of Phonetics*, 39(4), 492–504. <https://doi.org/10.1016/j.wocn.2010.12.002>.
- Cutler, A. (2008). The abstract representations in speech processing. *The Quarterly Journal of Experimental Psychology*, 61(11), 1601–1619. <https://doi.org/10.1080/13803390802218542>.
- Cutler, A. (2012). *Native listening: language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.
- Cutler, A., Eisner, F., McQueen, J., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougerson, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory Phonology* (Vol. 10, pp. 91–112). Berlin, New York: De Gruyter Mouton. <https://doi.org/10.1515/9783110224917.1.91>.
- Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, 28, 125–127. <https://doi.org/10.3758/BF03203646>.
- Danielson, D. K., Seidl, A., Onishi, K. H., Alamian, G., & Cristia, A. (2014). The acoustic properties of bilingual infant-directed speech. *The Journal of the Acoustical Society of America*, 135(2), EL95–EL101. <https://doi.org/10.1121/1.4862881>.
- DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: newborns prefer their mothers' voices. *Science*, 208, 1174–1176. <https://doi.org/10.1126/science.737592>.
- Dehaene-Lambertz, G., & Baillet, S. (1998). A phonological representation in the infant brain. *Neuroreport*, 9(8), 1885–1888.
- Durrant, S., Delle Luche, C., Cattani, A., & Floccia, C. (2014). Monodialectal and multidialectal infants' representation of familiar words. *Journal of Child Language*, 42(2), 447–465. <https://doi.org/10.1017/S0305000914000063>.
- Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, 47(2), 421–436. [https://doi.org/10.1044/1092-4388\(2004\)034](https://doi.org/10.1044/1092-4388(2004)034).
- Eimas, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception & Psychophysics*, 16, 513–521. <https://doi.org/10.3758/BF03198580>.
- Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-] distinction by young infants. *Perception & Psychophysics*, 18, 341–347. <https://doi.org/10.3758/BF03211210>.
- Eimas, P. D., & Miller, J. L. (1980a). Contextual effects in infant speech perception. *Science*, 209(4461), 1140–1141. <https://doi.org/10.1126/science.7403875>.
- Eimas, P. D., & Miller, J. L. (1980b). Discrimination of information for manner of articulation. *Infant Behavior & Development*, 3, 367–375. [https://doi.org/10.1016/S0163-6383\(80\)80044-0](https://doi.org/10.1016/S0163-6383(80)80044-0).
- 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.171.3968.30>.
- Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39, 253–260. [https://doi.org/10.1016/S0095-4470\(11\)00055-6](https://doi.org/10.1016/S0095-4470(11)00055-6).
- Fantz, R. L. (1958). Pattern vision in young infants. *Psychological Record*, 8, 43–47.
- Fantz, R. L. (1964). Visual experience in infants: decreased attention to familiar patterns relative to novel ones. *Science*, 146(3664), 668–670. <https://doi.org/10.1126/science.146.3664.668>.
- Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013a). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4), 751–778. <https://doi.org/10.1037/a0034245>.
- Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013b). 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>.
- Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior & Development*, 8(2), 181–195. [https://doi.org/10.1016/S0163-6383\(85\)80005-9](https://doi.org/10.1016/S0163-6383(85)80005-9).
- Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16(3), 477–501. <https://doi.org/10.1017/S0305000900010679>.
- Floccia, C., Delle Luche, C., Durrant, S., Butler, J., & Goslin, J. (2012). Parent or community: where do 20-month-olds exposed to two accents acquire their representations of words? *Cognition*, 124, 95–100. <https://doi.org/10.1016/j.cognition.2012.03.011>.
- Foushee, R., & Srinivasan, M. (2024). Infants who are rarely spoken to nevertheless understand many words. *Proceedings of the National Academy of Sciences*, 121(23). <https://doi.org/10.1073/pnas.2311425121> e2311425121.
- Frank, M. C. et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, 22(4), 421–435. <https://doi.org/10.1111/inf.12182>.
- Fritche, R., Shattuck-Hufnagel, S., & Song, J. Y. (2021). Do adults produce phonetic variants of /l/ less often in speech to children? *Journal of Phonetics*, 87. <https://doi.org/10.1016/j.wocn.2021.101056> 101056.
- Fung, P. (2025). *Navigating variability: language development in linguistically diverse environments*. University of Toronto. PhD Thesis.
- Gambell, T., & Yang, C. (2004). Statistics learning and universal grammar: modeling word segmentation. In *Proceedings of the workshop on psycho-computational models of human language acquisition* (pp. 51–54). Geneva, Switzerland: COLING.
- Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: the role of temporal synchrony between verbal labels and gestures. *Child Development*, 71(4), 878–894. <https://doi.org/10.1111/1467-8624.00197>.
- Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, 14, 23–45. <https://doi.org/10.1017/S030500090001271X>.
- Gómez, R. L., & Gerken, L. A. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, 70, 109–135. [https://doi.org/10.1016/S0010-0277\(99\)00003-7](https://doi.org/10.1016/S0010-0277(99)00003-7).
- Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, 51(4), 548–567. <https://doi.org/10.1016/j.jml.2004.07.002>.
- Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, 18(3), 254–260. <https://doi.org/10.1111/j.1467-9280.2007.01885.x>.
- Halle, P., & deBoysson-Bardies, B. (1994). Emergence of an early receptive lexicon: infants' recognition of words. *Infant Behavior & Development*, 17(2), 119–129. [https://doi.org/10.1016/0163-6383\(94\)90047-7](https://doi.org/10.1016/0163-6383(94)90047-7).
- Hitzenko, K., & Feldman, N. H. (2022). Naturalistic speech supports distributional learning across contexts. *Proceedings of the National Academy of Sciences*, 119(38). <https://doi.org/10.1073/pnas.2123230119> e2123230119.
- 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>.
- Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 1570–1582. <https://doi.org/10.1037/0096-1523.26.5.1570>.
- Isbilen, E. S., & Christiansen, M. H. (2022). Statistical learning of language: a meta-analysis into 25 Years of Research. *Cognitive Science*, 46(9). <https://doi.org/10.1111/cogs.13198> e13198.
- Jang, J., Kim, S., & Cho, T. (2024). Voice quality distinctions of the three-way stop contrast under prosodic strengthening in Korean. *Phonetics and Speech Sciences*, 16(1), 17–24. <https://doi.org/10.13064/KSSS.2024.16.1.017>.
- Jesse, A., & Johnson, E. K. (2016). Audiovisual alignment of co-speech gestures to speech supports word learning in two-year-olds. *Journal of Experimental Child Psychology*, 145, 1–10.
- Johnson, E. K. (2012). Bootstrapping language: are infant statisticians up to the job? In P. Rebuschat & J. Williams (Eds.), *Statistical learning and language acquisition* (pp. 55–90). Mouton de Gruyter.

- Johnson, E. K. (2003). *Word segmentation during infancy: the role of subphonemic cues to word boundaries*. Baltimore: Johns Hopkins Univ. PhD thesis, Dep. Psychol. Brain Sci..
- Johnson, E. K. (2008). Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech. *The Journal of the Acoustical Society of America*, 123, EL144–EL148. <https://doi.org/10.1121/1.2908407>.
- Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*, 44(4), 548–567. <https://doi.org/10.1006/jmla.2000.2755>.
- Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, 13(2), 339–345. <https://doi.org/10.1111/j.1467-7687.2009.00886.x>.
- Johnson, E. K., Jusczyk, P. W., Cutler, A., & Norris, D. (2003). Lexical viability constraints on speech segmentation by infants. *Cognitive Psychology*, 46(1), 65–97. [https://doi.org/10.1016/S0010-0285\(02\)00507-8](https://doi.org/10.1016/S0010-0285(02)00507-8).
- Johnson, E. K., Lahey, M., Ernestus, M., & Cutler, A. (2013). A multimodal corpus of speech to infant and adult listeners. *The Journal of the Acoustical Society of America*, 134(6), EL534. <https://doi.org/10.1121/1.4828977>.
- Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Developmental Science*, 12(1), 131–141. <https://doi.org/10.1111/j.1467-7687.2008.00740.x>.
- Johnson, E. K., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PLoS One*, 9(1). <https://doi.org/10.1371/journal.pone.0083546> e83546.
- Johnson, E. K., van de Weijer, J., & Jusczyk, P. W. (2001). Word segmentation by 7.5-month-olds: Three words do not equal one. *BUCLD 25: proceedings of the 25th annual Boston University conference on language development* (Vol. 2, pp. 389–400). Somerville, MA: Cascadilla Press.
- Johnson, E. K., van Heugten, M., & Buckler, H. (2022). Navigating accent variation: a developmental perspective. *Annual Review of Linguistics*, 8, 365–387. <https://doi.org/10.1146/annurev-linguistics-032521-053717>.
- Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, 27(4), 359–384. <https://doi.org/10.1006/jpho.1999.0100>.
- Jones, C., Meakins, F., & Muwiyath, 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. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29(10), 1–23. <https://doi.org/10.1006/cogp.1995.1010>.
- Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999a). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, 61, 1465–1476. <https://doi.org/10.3758/BF03213111>.
- Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999b). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39(3–4), 159–207. <https://doi.org/10.1006/cogp.1999.0716>.
- Kalashnikova, M. et al. (2024). The development of tone discrimination in infancy: evidence from a cross-linguistic, multi-lab report. *Developmental Science*, 27(3). <https://doi.org/10.1111/desc.13459> e13459.
- Kartushina, N., Rosslund, A., & Mayor, J. (2021). Toddlers raised in multi-dialectal families learn words better in accented speech than those raised in monodialectal families. *Journal of Child Language*, 13, 1–26. <https://doi.org/10.1017/S0305000921000520>.
- Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. A. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior & Development*, 18(1), 111–116. [https://doi.org/10.1016/0163-6383\(95\)90012-8](https://doi.org/10.1016/0163-6383(95)90012-8).
- Kim, S., Kim, J., & Cho, T. (2018). Prosodic-structural modulation of stop voicing contrast along the VOT continuum in trochaic and iambic words in American English. *Journal of Phonetics*, 71, 65–80. <https://doi.org/10.1016/j.wocn.2018.07.004>.
- Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 12577–12580. <https://doi.org/10.1073/pnas.0705345104>.
- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203. <https://doi.org/10.1037/a0038695>.
- Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Cognitive Science*, 10(4), 818–834. <https://doi.org/10.1111/tops.12331>.
- Ko, E.-S., Abu-Zhaya, R., Kim, E.-S., Kim, T., On, K.-W., Kim, H., Zhang, B.-T., & Seidl, A. (2023). Mothers' use of touch as infants' development and its implications for word learning: Evidence from Korean dyadic interactions. *Infancy*, 28(3), 597–618. <https://doi.org/10.1111/inf.12532>.
- Kooijman, V., Junge, C., Johnson, E. K., Hagoort, P., & Cutler, A. (2013). Predictive brain signals of linguistic development. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00025>.
- Koterba, E. A., & Iverson, J. M. (2009). Investigating motionese: the effect of infant-directed action on infants' attention and object exploration. *Infant Behavior & Development*, 32(4), 437–444. <https://doi.org/10.1016/j.infbeh.2009.07.003>.
- 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.116630>.
- Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior & Development*, 6, 263–285. [https://doi.org/10.1016/S0163-6383\(83\)80036-8](https://doi.org/10.1016/S0163-6383(83)80036-8).
- 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>.
- Kuhl, P. K., Andruski, J. E., Chistovich, I. 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>.
- 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.173636>.
- Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 363, 979–1000. <https://doi.org/10.1098/rstb.2007.2154>.
- Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: implications for the "critical period." *Language Learning and Development*, 1(3–4), 237–264. <https://doi.org/10.1080/15475441.2005.9671948>.
- Kuhl, P. K., Ramirez, R. R., Bosseler, A., Lotus Lin, J.-F., & Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proceedings of the National Academy of Sciences*, 111(31), 11238–11245. <https://doi.org/10.1073/pnas.1410963111>.
- Lahey, M., & Ernestus, M. (2013). Pronunciation variation in infant-directed speech: phonetic reduction of two highly frequent words. *Language Learning and Development*, 10(4), 308–327. <https://doi.org/10.1080/15475441.2013.860813>.
- Lavechin, M., De Seyssel, M., Gautheron, L., Dupoux, E., & Cristia, A. (2022). Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics*, 8(1), 389–407. <https://doi.org/10.1146/annurev-linguistics-031120-122120>.
- Lavechin, M., de Seyssel, M., Métais, M., Metzke, F., Mohamed, A., Bredin, H., & Cristia, A. (2024). Modeling early phonetic acquisition from child-centered audio data. *Cognition*, 245. <https://doi.org/10.1016/j.cognition.2024.105734> 105734.
- Lavi-Rotbain, O., & Amon, I. (2022). The learnability consequences of Zipfian distributions in language. *Cognition*, 223. <https://doi.org/10.1016/j.cognition.2022.105038> 105038.
- Leopold, W. (1949). *Speech development of a bilingual child*. Evanston, Ill: Northwestern University Press.
- Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: expectations about word length guide infant statistical learning. *Cognition*, 122(2), 241–246. <https://doi.org/10.1016/j.cognition.2011.10.007>.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of speech code. *Psychological Review*, 74, 431–461. <https://doi.org/10.1037/h0020279>.
- Liberman, Z., Woodward, A. L., Sullivan, K. R., & Kinzler, K. D. (2016). Early emerging system for reasoning about the social nature of food. *Proceedings of the National Academy of Sciences*, 113, 9480–9485. <https://doi.org/10.1073/pnas.1605456113>.
- Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). The origins of social categorization. *Trends in Cognitive Sciences*, 21(7), 556–568. <https://doi.org/10.1016/j.tics.2017.04.004>.
- 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. (2017). Statistical learning of speech sounds is most robust during the period of perceptual attunement. *Journal of Experimental Child Psychology*, 164, 192–208. <https://doi.org/10.1016/j.jecp.2017.05.013>.
- Liu, H.-M., Kuhl, P. K., & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, 6(3), F1–F10. <https://doi.org/10.1111/1467-7687.00275>.
- Loukatou, G., Scaff, C., Demuth, K., Cristia, A., & Havron, N. (2022). Child-directed and overheard input from different speakers in two distinct cultures. *Journal of Child Language*, 49(6), 1173–1192. <https://doi.org/10.1017/S0305000921000623>.
- MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P., & Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Current Biology*, 22, 113–117. <https://doi.org/10.1016/j.cub.2011.11.052>.
- MacKain, K., & Stern, D. (1985). *The concept of experience in speech development*. In K. Nelson (Ed.), *Children's language* (Vol. 5). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Mampe, B., Friederici, A. D., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology*, 19(23), 1994–1997. <https://doi.org/10.1016/j.cub.2009.09.064>.
- Marimon, M., Langus, A., & Hohle, B. (2024). Prosody outweighs statistics in 6-month-old German-learning infants' speech segmentation. *Infancy*. <https://doi.org/10.1111/inf.12593>.
- Markson, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, 14, 57–77.
- Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a protollexicon. *Cognitive Science*, 37(1), 103–124. <https://doi.org/10.1111/j.1551-6709.2012.01267.x>.
- Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: a comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, 26(3), 341–347. <https://doi.org/10.1177/09567976145624>.
- 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](https://doi.org/10.1207/s15327078in1003_3).

- Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, 106(3), 1367–1381. <https://doi.org/10.1016/j.cognition.2007.07.002>.
- Maye, J., Werker, J. F., & Gerken, L. A. (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](https://doi.org/10.1016/S0010-0277(01)00157-3).
- 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>.
- McClay, E. K., Cebiolglu, S., Broesch, T., & Yeung, H. H. (2022). Rethinking the phonetics of baby-talk: differences across Canada and Vanuatu in the articulation of mothers' speech to infants. *Developmental Science*, 25(2). <https://doi.org/10.1111/desc.13180>.
- McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: a cross-linguistic review. *American Journal of Speech-Language Pathology*, 27(4), 1546–1571. [https://doi.org/10.1044/2018\\_AJSLP-17-0100](https://doi.org/10.1044/2018_AJSLP-17-0100).
- McMurray, B., Spivey, M., & Aslin, R. (2000). The perception of consonants by adults and infants: categorical or categorized? Preliminary results. In K. M. Crosswhite & J. S. Magnuson (Eds.), *University of Rochester Working Papers in the Language Sciences* (Vol. 1(2), pp. 215–256).
- McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchorn, 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>.
- McMurray, B., Danelz, A., Rigler, H., & Seedorf, M. (2018). Speech categorization develops slowly through adolescence. *Developmental Psychology*, 54(8), 1472–1491. <https://doi.org/10.1037/dev0000542>.
- McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, 30, 1113–1126. [https://doi.org/10.1207/s15516709cog0000\\_79](https://doi.org/10.1207/s15516709cog0000_79).
- Mehler, J., Jusczyk, P. W., Lambert, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29, 143–178. [https://doi.org/10.1016/0010-0277\(88\)90035-2](https://doi.org/10.1016/0010-0277(88)90035-2).
- Melguy, Y. V., & Johnson, K. (2022). Perceptual adaptation to a novel accent: phonetic category expansion or category shift? *The Journal of the Acoustical Society of America*, 152(4), 2090–2104. <https://doi.org/10.1121/1.50014602>.
- Mitterer, H., Kim, S., & Cho, T. (2019). The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language*, 108. <https://doi.org/10.1016/j.jml.2019.104034>.
- Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experienced in utero affects vowel perception after birth: a two-country study. *Acta Paediatrica*, 102(2), 156–160.
- Morgan, J. L. (unpublished manuscript). *DRIBBLER: Dimensionally-Reduced Item Based LExical Recognition*. Providence, USA: Department of Cognitive, Linguistic, and Psychological Sciences, Brown University.
- Morgan, J. L., & Demuth, K. (1996). *Signal to syntax: bootstrapping from speech to grammar in early acquisition*. New York: Lawren Erlbaum Associates Inc.
- Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior*, 20(1), 67–85. [https://doi.org/10.1016/S0022-5371\(81\)90312-1](https://doi.org/10.1016/S0022-5371(81)90312-1).
- Morin-Lessard, E., Poulin-Dubois, D., Segalowitz, N., & Byers-Heinelein, K. (2019). Selective attention to the mouth of talking faces in monolinguals and bilinguals aged 5 months to 5 years. *Developmental Psychology*, 55(8), 1640.
- Mulak, K. E., Best, C. T., Tyler, M. D., Kitamura, C., & Irwin, J. R. (2013). Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words in a non-native regional accent. *Child Development*, 84(6), 2064–2078. <https://doi.org/10.1111/cdev.12087>.
- Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, 20, 139–154.
- Munson, B. (2004). Variability in /s/ production in children and adults: evidence from dynamic measures of spectral mean. *Journal of Speech, Language, and Hearing Research*, 47, 58–69. [https://doi.org/10.1044/1092-4388\(2004\)006](https://doi.org/10.1044/1092-4388(2004)006).
- Narayan, C. R. (2019). An acoustic perspective on 45 years of infants' speech perception, Part 1: consonants. *Language and Linguistics Compass*, 13(10). <https://doi.org/10.1111/lnlc3.12352>.
- Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, 43(1), 1–19. <https://doi.org/10.1006/jmla.2000.2698>.
- Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, 41(1), 233–243. [https://doi.org/10.1016/S0167-6393\(02\)00106-1](https://doi.org/10.1016/S0167-6393(02)00106-1).
- Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, 85, 2088–2113. <https://doi.org/10.1121/1.397861>.
- Nencheva, M. L., & Lew-Williams, C. (2022). Understanding why infant-directed speech supports learning: a dynamic attention perspective. *Developmental Review*, 66. <https://doi.org/10.1016/j.dr.2022.101047>.
- Newman, R., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: a retrospective analysis. *Developmental Psychology*, 42(4), 643–655. <https://doi.org/10.1037/0012-1649.42.4.643>.
- Newport, E. L. (1990). Maturation constraints on language learning. *Cognitive Science*, 14(1), 11–28. [https://doi.org/10.1016/0364-0213\(90\)90024-Q](https://doi.org/10.1016/0364-0213(90)90024-Q).
- Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2012). (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, 16(1), 24–34. <https://doi.org/10.1111/j.1467-7687.2012.01189.x>.
- Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, 80(3), 674–685. <https://doi.org/10.1111/j.1467-8624.2009.01290.x>.
- Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2), 421–435. <https://doi.org/10.1037/0096-1523.20.2.421>.
- Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-linguistic comparison of /d-/r/ perception: evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, 109(5), 2190–2201. <https://doi.org/10.1121/1.1362689>.
- Polka, L., & Bohn, O. S. (2011). Natural Referent Vowel (NRV) framework: an emerging view of early phonetic development. *Journal of Phonetics*, 39(4), 467–478. <https://doi.org/10.1016/j.wocn.2010.08.007>.
- Pons, F., Biesanz, J. C., Kajikawa, S., Fais, L., Narayan, C. R., Amano, S., & Werker, J. F. (2012). Phonetic category cues in adult-directed speech: evidence from three languages with distinct vowel characteristics. *Psicologia: International Journal of Methodology and Experimental Psychology*, 33(2), 175–207.
- Quinn, P. C., Lee, K., & Pascalis, O. (2019). Face processing in infancy and beyond: the case of social categories. *Annual Review of Psychology*, 70(1), 165–189. <https://doi.org/10.1146/annurev-psych-010418-102753>.
- Romeo, R., Hazan, V., & Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *The Journal of the Acoustical Society of America*, 134, 3781–3792. <https://doi.org/10.1121/1.4824160>.
- Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12, 339–349. <https://doi.org/10.1111/j.1467-7687.2008.00786.x>.
- 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>.
- Sahni, S. D., Seidenberg, M. S., & Saffran, J. R. (2010). Connecting cues: overlapping regularities support cue discovery in infancy. *Child Development*, 81(3), 727–736. <https://doi.org/10.1111/j.1467-8624.2010.01430.x>.
- Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (2021). Early phonetic learning without phonetic categories: insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7). <https://doi.org/10.1073/pnas.2001844118>.
- Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science*, 15, 732–738. <https://doi.org/10.1111/j.1467-7687.2012.01175.x>.
- Schmale, R., Seidl, A., & Cristia, A. (2015). Mechanisms underlying accent accommodation in early word learning: evidence for general expansion. *Developmental Science*, 18, 664–670. <https://doi.org/10.1111/desc.12244>.
- Schuerman, W. L., Meyer, A., & McQueen, J. M. (2015). Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker. *PLoS One*, 10(7). <https://doi.org/10.1371/journal.pone.0129731>.
- Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, 9, 566–574. <https://doi.org/10.1111/j.1467-7687.2006.00534.x>.
- Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker variation aids young infants' phonotactic learning. *Language Learning and Development*, 10(4), 297–307. <https://doi.org/10.1080/15475441.2013.858575>.
- Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-month-old infants. *Proceedings of the National Academy of Sciences*, 108(15), 6038–6043. <https://doi.org/10.1073/pnas.1017617108>.
- Shultz, S., & Vouloumanos, A. (2010). Three-month-olds prefer speech to other naturally occurring signals. *Language Learning and Development*, 6(4), 241–257. <https://doi.org/10.1080/15475440903507830>.
- Singh, L., White, K. S., & Morgan, J. L. (2008). Building a phonological lexicon in the face of variable input: influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, 4(2), 157–178. <https://doi.org/10.1080/15475440801922131>.
- Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, 106(2), 833–870. <https://doi.org/10.1016/j.cognition.2007.05.002>.
- Singh, L., Reznick, J. S., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: a longitudinal analysis. *Developmental Science*, 15(4), 482–495. <https://doi.org/10.1111/j.1467-7687.2012.01141.x>.
- Singh, L., Rajendra, S. R., & Mazuka, R. (2022). Diversity and representation in studies of infant perceptual narrowing. *Child Development Perspectives*, 16(4), 191–199. <https://doi.org/10.1111/cdep.12468>.
- Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568. <https://doi.org/10.1016/j.cognition.2007.06.010>.
- Soderstrom, M., Grauer, E., Dufault, B., & McDivitt, K. (2018). Influences of number of adults and adult:child ratios on the quantity of adult language input across childcare settings. *First Language*, 38(6), 563–581. <https://doi.org/10.1177/014272371878501>.
- Sohail, J., & Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, 12(2), 105–115. <https://doi.org/10.1080/15475441.2015.1073153>.
- Southby, L., Harding, S., Phillips, V., Wren, Y., & Joinson, C. (2021). Speech input processing in children born with cleft palate: A systematic literature review with narrative synthesis. *International Journal of Language & Communication Disorders*, 56(4), 668–693. <https://doi.org/10.1111/1460-6984.12633>.
- 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>.

- Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-/l/ by Japanese adults learning English. *Perception & Psychophysics*, 36, 131–145. <https://doi.org/10.3758/BF03202673>.
- Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2013). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.01015>.
- Swingle, D., Pinto, J. P., & Fernald, A. (1998). Assessing the speed and accuracy of word recognition in infants. *Advances in Infancy Research*, 12, 257–277.
- Swingle, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50(1), 86–132. <https://doi.org/10.1016/j.cogpsych.2004.06.001>.
- Swingle, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 364, 3617–3622. <https://doi.org/10.1098/rstb.2009.0107>.
- Swingle, D., & Alarcon, C. (2018). Lexical learning may contribute to phonetic learning in infants: a corpus analysis of maternal Spanish. *Cognitive Science*, 42(5), 1618–1641. <https://doi.org/10.1111/cogs.12620>.
- Swingle, 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>.
- Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, 108(3), 850–855. <https://doi.org/10.1016/j.cognition.2008.05.009>.
- Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, 56(1), 16–34. <https://doi.org/10.1016/j.jml.2006.07.002>.
- Thiessen, E. D., & Erickson, L. C. (2013). Discovering words in fluent speech: the contribution of two kinds of statistical information. *Frontiers in Psychology*, 3. <https://doi.org/10.3389/fpsyg.2012.00590>.
- Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, 3(1), 73–100. [https://doi.org/10.1207/s15473341ld0301\\_3](https://doi.org/10.1207/s15473341ld0301_3).
- Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39(4), 706.
- Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, 7(1), 53–71. [https://doi.org/10.1207/s15327078in0701\\_5](https://doi.org/10.1207/s15327078in0701_5).
- Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10(2), 172–175. <https://doi.org/10.1111/1467-9280.0012>.
- Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, 17, 432–444. <https://doi.org/10.1111/j.1532-7078.2011.00084.x>.
- Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development*, 47(2), 466–472. <https://doi.org/10.2307/1128803>.
- Trujillo, J., Özyürek, A., Holler, J., & Drijvers, L. (2021). Speakers exhibit a multimodal Lombard effect in noise. *Science Reports*, 11(1), 16721. <https://doi.org/10.1038/s41598-021-95791-0>.
- Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Development*, 75(4), 1067–1084. <https://doi.org/10.1111/j.1467-8624.2004.00726.x>.
- van de Weijer, J. (1998). *Language input for word discovery*. MPI Series in Psycholinguistics.
- van der Feest, S. V. H., & Johnson, E. K. (2016). Input driven differences in toddlers' perception of a disappearing phonological contrast. *Language Acquisition*, 23, 89–111. <https://doi.org/10.1080/10489223.2015.1047096>.
- van Heugten, M., & Johnson, E. K. (2012). Infants exposed to fluent natural speech succeed at cross-gender word recognition. *Journal of Speech, Language, and Hearing Research*, 55, 554–560. [https://doi.org/10.1044/1092-4388\(2011\)10-0347](https://doi.org/10.1044/1092-4388(2011)10-0347).
- van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy: benefits of brief speaker exposure. *Journal of Experimental Psychology: General*, 143, 340–350. <https://doi.org/10.1037/a0032192>.
- Van Heugten, M., Krieger, D. R., & Johnson, E. K. (2015). The developmental trajectory of toddlers' comprehension of unfamiliar regional accents. *Language Learning and Development*, 11(1), 41–65. <https://doi.org/10.1080/15475441.2013.879636>.
- van Heugten, M., & Johnson, E. K. (2017). Input matters: Multi-accent language exposure affects word form recognition in infancy. *The Journal of the Acoustical Society of America*, 142(2), EL196–EL200. <https://doi.org/10.1121/1.4997604>.
- Walley, A. C. (1993). The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review*, 13(3), 286–350. <https://doi.org/10.1006/drev.1993.1015>.
- 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. <https://doi.org/10.3389/fpsyg.2014.00077>.
- Weatherhead, D., & White, K. S. (2016). He says potato, she says potahto: young infants track talker-specific accents. *Language Learning and Development*, 12(1), 92–103. <https://doi.org/10.1080/15475441.2015.1024835>.
- Weatherhead, D., & White, K. S. (2017). Read my lips: visual speech influences word processing in infants. *Cognition*, 160, 103–109. <https://doi.org/10.1016/j.cognition.2017.01.002>.
- Weatherhead, D., & White, K. S. (2018). And then I saw her race: race-based expectations affect infants' word processing. *Cognition*, 177, 87–97. <https://doi.org/10.1016/j.cognition.2018.04.004>.
- Weatherhead, D., & White, K. S. (2019). Toddlers link social and speech variation during word learning. *Developmental Psychology*, 57(8), 1195. <https://doi.org/10.1037/dev0001032>.
- Weatherhead, D., White, K. S., & Friedman, O. (2016). Where are you from? Preschoolers infer background from accent. *Journal of Experimental Child Psychology*, 143, 171–178. <https://doi.org/10.1016/j.jecp.2015.10.011>.
- Weatherhead, D., Friedman, O., & White, K. S. (2018). Accent, language, and race: 4-6-year-old children's inferences differ by speaker cue. *Child Development*, 89(5), 1613–1624. <https://doi.org/10.1111/cdev.12797>.
- Weatherhead, D., Friedman, O., & White, K. S. (2019). Preschoolers are sensitive to accent distance. *Journal of Child Language*, 46(6). <https://doi.org/10.1017/S0305000919000369>.
- Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, 316(5828), 1159. <https://doi.org/10.1126/science.1137686>.
- 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>.
- Werker, J. F., & Curtin, S. (2005). PRIMIR: a developmental model of speech processing. *Language Learning and Development*, 1, 197–234. <https://doi.org/10.1080/15475441.2005.9684216>.
- Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: new directions. *Annual Review of Psychology*, 66, 173–196. <https://doi.org/10.1146/annurev-psych-010814-015104>.
- Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, 7(1), 49–63. [https://doi.org/10.1016/S0163-6383\(84\)80022-3](https://doi.org/10.1016/S0163-6383(84)80022-3).
- White, J., & Sundara, M. (2014). Biased generalization of newly learned phonological alternations by 12-month-old infants. *Cognition*, 133(1), 85–90. <https://doi.org/10.1016/j.cognition.2014.05.020>.
- White, K. S. (2017). Listening to (and listening through) variability during word learning. In *Early word learning* (pp. 83–95). Routledge.
- White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, 14(2), 372–384. <https://doi.org/10.1111/j.1467-7687.2010.00986.x>.
- White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, 59, 114–132. <https://doi.org/10.1016/j.jml.2008.03.001>.
- White, K. S., & Daub, O. (2021). When it's not appropriate to adapt: toddlers' learning of novel speech patterns is affected by visual information. *Brain and Language*, 222. <https://doi.org/10.1016/j.bandl.2021.105022>.
- White, K. S., Peperkamp, S., Kirk, C., & Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, 107(1), 238–265. <https://doi.org/10.1016/j.cognition.2007.11.012>.
- White, K. S., Yee, E., Blumstein, S. E., & Morgan, J. L. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, 68(4), 362–378. <https://doi.org/10.1016/j.jml.2013.01.003>.
- 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>.