INFANT VOCAL LEARNING AND SPEECH PRODUCTION

Anne S. Warlaumont

During the first year of life, human infants undergo an extraordinary process of vocal learning, unmatched by other primates. This lays a key foundation for meaningful speech production. The first sections of this chapter describe major milestones and other features of the development of prelinguistic and early speech sounds, including the acquisition of new sound types and of conversational turn-taking skills. The chapter then discusses what we know about the roles of exploratory play, social input, and neural systems in human vocal learning. A section on computational modeling reviews theoretical work that informs our understanding of how these mechanisms interact. Effects of sociocultural and clinical differences on infant vocal development are then discussed. The final section of the chapter discusses policy perspectives on research and interventions in this domain.

Prelinguistic Vocalization Types

At birth, most infant vocalizations are limited to cries, vegetative sounds (such as burps and sucking sounds, produced as byproducts of other processes), and short, quiet sounds where there is vibration of the vocal folds but the upper vocal tract (throat, tongue, mouth, and nasal cavity) is in a neutral position (Oller, 2000). These short, quiet sounds are considered to be the earliest precursors to speech and are considered a type of “protophone”. Protophones can be defined as sounds that are clearly communicative or playful (in contrast with vegetative sounds, which if they serve communicative functions do so only incidentally) and yet do not have a set communicative function (in contrast with cries and laughs, which communicate similar things across all cultures) (Oller et al., 2013).

Within the next two to three months of life, infants begin to produce a much wider variety of vocalization types. These vary more substantially in duration, amplitude, pitch, and vocal quality, to include growls, squeals, yells, and whispers. Infants also begin to posture the tongue and lips. This enables them to produce fully resonant vowels of different types. It also enables the production of primitive consonant-like elements, formed either by a brief pause in phonation (vibration of the vocal folds within the larynx) or by movement of the tongue or lips that closes off the vocal tract (Oller, 1986, 2000; Buder et al., 2013). Interestingly, raspberries, where the lips are set into vibration against each other or against the teeth, are also very common in infancy but are extremely uncommon in adult language. Clicks are also produced at this stage. Not all of these new vocalization types emerge at the same time. Some can take months longer than others to appear, and there appear to be considerable individual differences in which of these early protophones are produced most frequently at any particular age (Stark, 1980).

By about seven months of age, infants begin to consistently (if not very frequently at first) produce canonical syllables. A canonical syllable is a vocalization in which there is at least one full vowel sound following or preceding a consonant, where the transition between consonant and vowel is not overly long (sounding slurred). By some definitions the consonant must be a true consonant, made by movement of the tongue or lips, and not only involving a change in vocal fold vibration (Oller, 1986, 2000; Buder et al., 2013). When canonical syllables are produced without having a clear meaning, the term canonical babble applies. Canonical syllables can be produced alone or in sequence, either repeating the same consonant and vowel elements, a.k.a. reduplicated babbling, or varying them, a.k.a. variegated babbling (Smith et al., 1989). The onset of canonical babbling tends to be a salient transition for caregivers (Oller et al., 2001). Over the next year or more of life (through 18 months of age), canonical syllables become more frequent elements in infants' vocal productions. The types of consonants

and vowels included also become more varied (both overall and within utterances) and rising and falling pitch and amplitude changes are combined with the babbling to create a sense of prosodic structure (this type of babbling has sometimes been referred to as “jargon”) (Oller, 2000). Infant vocal productions thus begin to sound more and more like adult speech. It is important to note that as new sound types are added to the infant’s repertoire, previous sound types do not disappear but typically continue to be produced, if at decreasing rates in some cases. Similarly, once infants begin producing meaningful speech (see the next section), nonword babble continues to be produced at high rates, decreasing only gradually (Robb et al., 1994). See Figure 1 for examples of three different infant vocalization types and for an example of infant-directed adult speech.

Early Meaningful Speech

Often before the first birthday, but sometimes many months later, parents report that infants begin to produce their first words (Schneider et al., 2015). From a motor production standpoint, infants at this point typically have the capability to produce canonical syllables incorporating at least a few different consonant and vowel types, which puts them in a position to produce approximations of words, such as “mama” and “baba” (Vihman et al., 1985; Oller, 2000; McCune & Vihman, 2001). When infants begin to produce a sound sequence reliably correlated with the presence or desire for a particular object, event, or other referent, i.e. when there is a “conventionalized sound-meaning correspondence” (Vihman et al., 1985), these can be considered the infant’s first words. Nouns tend to be more prominent in infants’ early productions than in their caregivers’ speech to them in the same contexts, compared to verbs (Tardif et al., 1999). At first, words are typically produced in isolation, and in instances where an infant appears to produce multiple words, e.g., “what’s this?”, the combination acts more like a single lexical unit, with the component words not yet being used independently or being

recombined with other words (Lieven, Pine, & Barnes, 1992). During the second year of life, many children show a quickening rate of expressive vocabulary growth and begin combining multiple lexical items flexibly into primitive sentences (Clark, 2003).

Children's early speech is typically much more intelligible to their primary caregiver(s) than it is to strangers (Baudonck et al., 2009; Weist & Kruppe, 1977). The sounds that are present in canonical babbling tend to be the same sounds that are used to produce meaningful speech (Vihman et al., 1985; Locke, 1989), with sounds often being deleted, substituted, or sometimes even added in comparison to the adult word form (Oller et al., 1976). Even when particular consonants and vowels are present in the infant's repertoire, some sequences of those sounds may be difficult for the infant produce, leading the child to omit or substitute even some sounds that they can create in other contexts. Consonant clusters can be particularly difficult for young children.

Development of Conversational Turn-Taking

Well before word production or even canonical babbling occur, infants begin to exhibit conversational turn-taking skills. From as early as 2 months, infant-adult vocal interactions tend to occur in distinct turns (Gratier et al., 2015). At early ages, caregivers appear to drive more of the turn-taking than infants, but over the course of the first year of life, infants are increasingly able to contribute to minimizing overlap between their own and their conversation partner's vocalizations (Harder et al., 2015). The ability of an infant and an adult to exhibit such coordination in vocalization timing, particularly matching the timing of pauses between the other speaker's vocalization offset and one's own vocalization onset, at four months has been shown to be correlated with later infant attachment security and cognitive skills at twelve months (Jaffe et al., 2001). Interestingly, at least one study has found that infants have increased lags in their vocal responses to caregiver vocalizations at around nine months, when infants are on the cusp

of producing first words; this may be because of the increased cognitive and motor demands incurred by the process of formulating verbal responses (Hilbrink et al., 2015). Kuchirko et al. (2018) have shown increases from 12 to 24 months of age in the likelihood of infant vocal responses to maternal referential language, indicating that development of conversational turn taking continues after the first year of life.

Mechanisms of Vocal Learning and Speech Production Development

Having provided some description of the types of changes we see in vocalization, speech, and language production over the course of the first two years of life, we turn our attention to some of the factors and mechanisms that underlie these changes.

Intrinsically Motivated Play

A major contributor to infant vocal motor learning may be intrinsically-motivated processes that combine random exploration around an existing skill base with a desire to expand that skill base. Infants often produce vocalizations when they are not actively engaged in social interaction with others (Jones & Moss, 1971). Moreover, infants tend to repeat particular sound types in bouts (Gratier & Devouche, 2011; Oller et al., 2013). Anecdotally, this repetition can appear to some observers to suggest goal-directed behavior, although it is unclear whether a goal-oriented process is actually in place or there just exists momentum in the infant vocalization system. The potential advantage of goal-directed exploration for prelinguistic vocal learning has been demonstrated in computational modeling studies (more on this below). Although intrinsically motivated exploration and learning process can in principle take place without any social input, social influences are certainly also involved.

Social Input

The input infants receive from adult caregivers is clearly related to infants' language development. In terms of promoting the vocalization and early speech milestones described

above, two roles that social input seem to play are (1) to provide positive reinforcement (reward) to infants when they produce relatively advanced behaviors and (2) to provide targets that infants may try to imitate.

Adult responses as positive reinforcers. We know that even young infants can detect sequential contingencies between external stimuli and between their own behaviors and the consequences of those behaviors (Tarabulsy et al., 1996). Converging evidence from naturalistic observation of infant-parent interactions and from experimental studies supports the idea that this contingency learning plays a role in infants' prelinguistic vocal learning.

Parent speech to infants differs acoustically, semantically, and syntactically from speech directed to other adults. For example, infant-directed utterances tend to be higher in pitch, have more exaggerated prosodic contours, be shorter, have longer pause durations, include more repetition, and be semantically and syntactically simpler (Fernald et al., 1989; Soderstrom, 2007). Infant-directed speech is also salient and appealing to infants, being preferred both to adult-directed speech and to non-speech stimuli (Fernald, 1985; Fernald & Kuhl, 1987; Vouloumanos & Werker, 2004). Infants may therefore be motivated to produce behavior that increases the quantity of infant-directed speech they hear. Since infant-directed adult vocalizations also tend to follow infant speech-related vocalization productions (Gros-Louis et al., 2006; Warlaumont et al., 2014) adult vocalizations likely serve to reinforce infants' increased production of speech-related sounds.

Indeed, experimental work by Nathani & Stark (1996) has found that a few minutes' interaction with a researcher who provides consistent positive vocal responses whenever the infant produces a speech-related (protophone) vocalization leads to the infant producing increased numbers of these vocalizations during an immediately following recording session. Increased frequency of infant vocalization can in turn be expected to increase the opportunities

for adults to provide high quality responses to infant vocal behavior that facilitate infant communication development (Tamis-LeMonda et al., 2001, 2018; Leezenbaum et al., 2013; Warlaumont et al., 2014).

Supporting the idea that contingent adult responses selectively increase infants' rates of protophones as opposed to cries, Warlaumont et al. (2014) studied children ranging in age from 10 to 48 months and found that when speech-related infant vocalizations were followed by adult responses, this was associated with an increased likelihood of the following child vocalization being speech-related as opposed to cry. Moreover, rates of adult responses to the children's vocalizations predicted faster growth, over the three year period, in the increase of speech-related vocalizations relative to cries and other reflexive and vegetative sounds. Looking more closely at different infant protophone types, Gros-Louis and Miller (2018) found that 10 month old infant vocalizations were more likely to be vowel-only (as opposed to consonant-vowel combinations) when the previous infant vowel-only vocalization received an adult response than when it received no response; they found a similar pattern for 12 month old infant consonant-vowel vocalizations, and an opposite tendency for 12 month old vowel vocalizations. Taken together, these studies suggest that when adult responses are contingent on infant vocalizations of a certain type (whether that's a broad category such as speech-related vocalization or a somewhat narrower subcategory of speech-related vocalizations), this can promote subsequent infant vocalizations of that type. On the other hand, the studies also suggest that this may not be true for all vocalization types at all points in development—there must be other factors involved besides just positive reinforcement selectively shaping infant vocalization frequency and type.

Many studies have found a positive relationship between the quantity and diversity of child-directed language a child hears and the child's expressive vocabulary (Hart & Risley,

1995; Ramírez-Esparza et al., 2014; Golinkoff et al., 2015). This positive association may in part reflect the fact that child-directed speech is both contingent on (i.e. responsive to) infant vocalizations and salient to infants, so that quantity of child-directed speech may be a good proxy for quantity of positive reinforcement for infants' productions. However, adult vocalizations themselves also have acoustic and linguistic content, and this content provides a rich source of additional information.

Adult input as targets for imitation. Another way that the content of adult vocalizations may influence infant productions is by creating targets for infant vocal play. As an infant learns that there is some correspondence between the sets of sounds she herself can produce and those that the adults around her produce, this may encourage her to consider any sound types that adults produce as potential additions to her own repertoire. Along these lines, computational modeling work has shown how a single intrinsically-motivated algorithm for choosing acoustic targets can account for an increased interest as an infant gets older in imitating adult vocalizations (Moulin-Frier et al., 2014). This is consistent with work finding that during the first year of life true vocal imitation in response to input in an experimental context is rare (Jones, 2009), but that during the second year children start to actively imitate arbitrary sounds directed to them by adults (Jones, 2007).

Experimental work has shown that learning from the content of adult vocalizations is especially powerful when those adult vocalizations have been produced in response to an infant's own vocalizations. In a study of 9.5-month-olds by Goldstein & Schwade (2008), mothers were told when and how to interact with their infant by a researcher who was observing the interaction from a control room. In one condition the parent was instructed to approach and vocalize to the infant immediately following every speech-related (protophone) vocalization produced by the infant and was also told what kind of sound to make. In this condition, infants'

vocalizations shortly after the controlled interaction period came to take on the broad phonetic properties of the sounds produced by the parent. Moreover, when the parent vocalizations were canonical consonant-vowel syllables, infants tended to produce more canonical syllables; when they were fully resonant vowels without consonant sounds, the infants came to produce more fully resonant vowel sounds. The adaptation did not take place in a yoked-control condition in which the same quantity and type of parental vocalizations occurred but were not timed to immediately follow the infant's vocalizations. Along similar lines, Bloom (1998) found that when adults engaged in verbal but not non-verbal turn-taking with three-month-old infants, subsequent infant vocalizations had a greater tendency to incorporate primitive syllabic elements. (Interestingly, Goldstein et al. (2003) found that even non-verbal contingent responses to 8 month olds' vocalizations led to a subsequent increase in canonical babbling rates.)

Neural Underpinnings

How does an infant's neurophysiology support these exploratory and socially-guided vocal learning processes? One possibility is that speech and pre-speech vocal motor control relies heavily on the recruitment of neural circuitry that previously evolved for the production of reflexive vocal signals such as cries and laughs or other reflexive vocal tract movements, such as those involved in feeding. MacNeilage (1998) has argued that the production of syllabically structured vocalizations relies on existing mechanisms for producing rhythmic feeding movements, particularly the rhythmic jaw movement involved in chewing. Presumably, according to this theory, the fact that the onset of rhythmic babbling does not occur until about 7 months of age would be related to delay in maturation of the chewing jaw movement circuitry, and possibly could be related to the time it takes for children to learn to combine phonation with jaw movement. On the other hand, the timing of syllabic vocalizations, and the increase in rate

of jaw movement with increasing age, indicate that human syllabic vocalization development has more in common with non-human primate lipsmacks than to chewing vocalizations (lipsmacks are a communicative signal produced by alternating mouth closure and opening without concomitant phonation) (Morrill et al., 2012). However, even if lipsmacking and syllabic vocalization are homologous this still leaves many open questions about the neural bases of their development, including whether (gradual or delayed) recruitment of brainstem circuitry for reflexive vocalization or other oral behavior is involved.

A contrasting perspective is that pre-speech vocal motor control is driven primarily by cortical learning that more directly controls vocal motor effectors, in a circuit that bypasses those involved in production of reflexive signals. Key evidence supporting this view comes from anatomical studies comparing the direct pathway tracts between laryngeal regions of primary motor cortex and laryngeal motor neurons of the brainstem between humans and non-human primates (Jürgens, 2002): it appears that in humans there are more robust direct (pyramidal) connections as well as indirect connections from primary motor cortex to circuits in the hindbrain and spinal cord that more immediately control the vocal tract muscles (see Figure 2). In contrast, at least some non-human primates show far less connectivity of this sort and must therefore rely more on the midbrain structures, in particular a region called the periaqueductal gray, that control involuntary vocalizations in both human and non-human primates. Having robust direct motor cortex connections to the neurons that immediately affect vocal tract muscles suggests a greater role of motor cortex in the generation of vocalizations in humans. This is consistent with studies that have electrically stimulated frontal cortex regions and found that while such stimulation can generate vocalizations, including syllabic vocalizations, in adult humans (Penfield & Welch, 1951), it does not reliably elicit vocalization in squirrel monkeys (Jürgens, 1974). More recently, electrocorticography of the lateral sensorimotor cortex has

revealed that when adult humans produce tongue and lip movements to make specific speech sounds, such as /b/, /d/, or /g/, there is a corresponding rise in activation of the primary motor cortex regions associated with the body part (lips, anterior tongue, posterior tongue) most involved in the speech sound (Bouchard et al., 2013). This indicates that vocal tract activity for the production of speech sounds may be driven rather directly by activation of motor cortex.

These neuroanatomical and neurophysiological findings implicating direct pathways from motor cortex to brainstem fit well with findings that while some non-human primates do exhibit substantial vocal learning (e.g., Russell et al., 2013; Perlman & Clark, 2015; Ghazanfar & Zhang, 2016; Giltekin & Hage, 2018), humans demonstrate vocal learning and voluntary control of vocalization to an extreme extent (this has to be the case for the oral language of modern humans to exist at all and may account for some of the difficulties in teaching non-human primates oral language). The neuroanatomical and neurophysiological findings are also consistent with acoustic analyses of adult human laughter. Real and fake laughs have distinctly different acoustic properties, suggesting separate neural mechanisms (Bryant & Aktipis, 2014). In learning to produce fake laughter, humans may learn to produce movement patterns that resemble real laughter rather than simply activating real laughter circuitry. Finally, recent computational modeling studies (see next section) have demonstrated how cortical learning mechanisms can readily account for some of the changes we see in vocal productions across the first year of life in typically developing infants, including the transition to producing syllabically structured vocalizations (Warlaumont & Finnegan, 2016). It seems likely therefore that there is a route for cortical motor learning that does not rely on additional circuitry for generating patterns of vocal tract muscle activation for mastication or for reflexive vocalization.

Regarding the debate about whether or not adult speech involves the (possibly learned) reuse of existing circuitry for reflexive behavior involving the vocal tract or learning new motor

behaviors essentially “from scratch”, it is worth noting that the two possibilities are not mutually exclusive. It is possible that some combination of the two possibilities takes place during development. It is also possible that the two possible pathways represent redundant pathways to mature communicative vocal signals in humans. It is also worth noting that none of the studies mentioned above involved human infant participants. So far, methodological challenges to recording and eliciting neural activity from human infants have forced us to extrapolate from data with adult human participants, non-human animal subjects, and computational modeling to infer what the results might mean for the neurophysiological underpinnings of human infant vocal learning.

Although functional brain imaging has so far been impractical to study infant vocalization production directly, it can be used to study infant auditory perception, and this has yielded some interesting data on the activity of motor regions during speech perception. In particular, Kuhl et al. (2014) found that infant perception of native and non-native speech sounds stimulated not only the auditory cortex but also the motor cortex, and that the relative activity of the motor cortex compared to that of the auditory cortex was higher for native language sounds for 7-month-olds but for non-native sounds for 11–12-month-olds. Of course, there are many additional open questions about the neurophysiological basis of speech production development in infancy beyond those discussed here.

Computational Models

Computational modeling is used to formulate and test theories about how infants learn to produce both more mature-sounding vocalizations and meaningful speech. The earliest computational models in this domain were connectionist models (neurally inspired but not anatomically or neurophysiologically detailed) consisting simply of a set of perceptual nodes and a set of motor nodes with weighted connections between the two (e.g., Yoshikawa et al., 2003;

Westermann & Miranda, 2004; Kröger et al., 2009). The models learned by producing random movements, sometimes referred to as “motor babbling” in the developmental computational modeling literature. The random movements led to a specific pattern of activation of motor nodes and each motor pattern served as input to a vocal tract simulation that would synthesize a sound based on that motor pattern. Acoustic features of that sound known to be important in human speech perception, most notably formant frequencies, were then measured and input to the perceptual neurons in the model. The co-activation of perceptual neurons and motor neurons allows for self-organized learning of the correspondences between motor commands and perceptual consequences. After experiencing many rounds of this process, a model can build up sufficient knowledge about motor-perceptual mappings to be able to reliably produce a target sound, for example in imitation of a sound input by another individual (Heintz et al., 2009).

One challenge for models of this type is that different individuals have differently shaped vocal tracts and therefore differences in the range of acoustic features they can produce. Adult humans naturally account for this, for example when categorizing a particular combination of formant frequencies as an instantiation of a particular vowel type. However, the procedure just described does not lead to a model that performs this normalization across vocal tracts (Heintz et al., 2009). An approach that has proved successful is to assume that adult caregivers frequently imitate infants, and that this imitation can serve as a cue to the correspondences between one speaker’s vocal tract and another’s; selective imitation only of sounds that fit into categories in the adult interaction partner’s language can also be used to decide which vocal categories and perceptual-motor mappings to retain (Miura et al., 2007; Howard & Messum, 2014).

When dealing with the large number of motor degrees of freedom present in a full vocal tract model and the additional complexity introduced when trying to model the production of

dynamically changing vocal tract movements (for example in order to produce the movements of the vocal tract needed to produce consonant sounds), it has proved important to move beyond a completely random motor babbling approach. This is in part because many combinations of vocal tract movements are not very useful for generating speech-like sounds (Warlaumont et al., 2013). It is better if the model, and by extension the human infant, can first discover which kinds of motor activities are worth the most focused exploration at a given point in the learning process. An approach that has demonstrated both the severity of this problem for vocal learning and a possible solution has been an intrinsically motivated goal-babbling approach (Moulin-Frier et al., 2014).

In Moulin-Frier et al.'s model, vocal exploration at any point in time is guided by a desire to achieve a particular acoustic goal. The model identifies and then executes the combination motor actions that it believes are most likely to achieve the acoustic outcome. At first the selection is based on completely random exploration of motor space to get a rough idea of some of the acoustic consequences associated with movements. After trying the action and observing the actual outcome, the model adjusts its stored knowledge of the relationships between motor patterns and acoustic patterns. Thus, as learning progresses the model's knowledge of motor-acoustic mappings becomes more accurate.The model also learns how likely various acoustic goals are to lead to a high rate of learning. At a given point in the learning process, some goals may be too difficult for the agent to achieve and therefore be unlikely to lead to much helpful refinement of the agent's motor-acoustics knowledge; at that same time point some other goals may already be so easy for the agent to achieve that not much learning is likely to result from pursuing them either. For example, in Moulin-Frier's model, early on the model attempts to produce silence, a goal it quickly masters. It then moves on to primarily choosing goals that consist of a single combination of formant frequencies, somewhat akin to

producing a single vowel in isolation. After improving performance on these goals, it has an increased likelihood of choosing goals that have a combination of two specific sounds in a specific sequence. This model also includes two modes, one in which goals are chosen purely through this endogenous intrinsically motivated process and another in which an external input is supplied as a potential target for imitation. As the model advances in its capabilities, it becomes more interested in using imitation as a means for selecting goals and driving learning.

Thus, the intrinsically-motivated goal-oriented learning model captures the transition from unphonated to phonated to complex sounds and from primarily endogenously-driven exploration to imitation-oriented learning. In other words, the model provides an explanation for why infants' spontaneous vocal productions will initially tend to be simpler sounds with vowel sequences, syllabic consonant-vowel transitions, and variegated babbling only becoming more frequent later in development. It is also important to note that although the model demonstrates the power that goal-babbling has for vocal learning, it does not necessarily imply that infant vocalizations are always goal-directed. The model itself undergoes an initial brief phase of completely random (not goal-driven) exploration in order to initialize its motor-acoustic map. It is conceivable that goal-oriented babbling plays a crucial role despite only operating some of the time.

Such an algorithmic approach is helpful for understanding the possible strategies that might characterize infants' active vocal learning. More detailed models would be needed to link such strategies to possible neural implementations. More biologically detailed computational modeling has not yet achieved this, but has demonstrated some possible neural bases for the dynamic generation of muscle activity patterns and for the shaping of spontaneously generated actions through selective reinforcement (either intrinsic or social). Recently, a spiking neural network model has demonstrated how electrical activity in a small population of cortical neurons

can be summed and low-pass filtered to generate fluctuating activity in muscles that control the opening and closing of the vocal tract (Warlaumont & Finnegan, 2016). As a result of random input as well as random interconnections among neurons in the local cortical network, the model generates spontaneous activity. At first, overall muscle activity and fluctuation in activity levels are too low to consistently generate sounds that alternate opening and closing of the vocal tract, so instead of producing syllabically structured vocalizations, the model produces primarily simple vowel sounds. However, with selective reinforcement for the rare production of a consonant-like sound, the model receives surges of dopamine (see Figure 2). These increase the learning rate between neurons in the cortical network, leading to increased likelihood of those patterns of neural activity that generate consonant-vowel sequences. The selective reinforcement could come either from caregivers' positive contingent responses or from an infant's intrinsic excitement about the sound it just produced. The model thus increases its rate of canonical babbling over the course of learning, using a biologically, psychologically, and socially plausible learning mechanism. While the model lacks many of the neural systems that are known to play a role in motor learning in humans (such as basal ganglia and cerebellum) and even lacks most of the degrees of freedom present in an actual vocal tract, it nevertheless represents an important step toward linking infant vocal learning to neural mechanisms and informs current debates about the origins of syllabic sounds in human evolution and development (see above).

All the models just described have focused exclusively on prelinguistic vocal learning, without addressing infant's development of productive vocabulary. To address meaningful speech production requires that models incorporate some representation of the things in the world that first words tend to refer to. A number of computational models (e.g., Li et al., 2007) have taken a more abstract approach to representing the sound production processes in order

to focus on the development of mappings between sounds and world knowledge. They demonstrate how infants can form word-meaning mappings not only for purposes of word recognition but also to generate appropriate speech sound sequences in the presence of particular referents.

Recently, a model by Forestier and Oudeyer (2017) has integrated vocal learning and learning to obtain objects (by reaching with the arm, reaching with a tool, or asking a caregiver) within the same system. The model includes a three degree of freedom arm placed in a two-dimensional simulated environment. Also in the environment are three objects and a caregiver. The agent also has a seven degree of freedom vocal tract that produces as output trajectories in a two-dimensional acoustic space (acoustic dimensions are the first two formant frequencies, i.e. the lowest two resonant frequencies of the vocal tract, which change as vocal tract shape changes). The caregiver can also produce sounds in this vocal space and knows the labels for all three objects. The infant can learn to produce the names of objects, and uttering an object name has the effect of getting the caregiver to place the object within the infant's reach. Likewise, when the infant grasps an object, the caregiver utters the object's name. On some learning trials, the infant's goal is to imitate caregiver vocalizations, on some trials it is to generate a random sound sequence, and on other trials the goals are to move the hand, the tool, or the objects along a randomly set goal trajectory. An interesting result is that after training, the model is more successful at imitating the sounds that correspond to the three objects' words than it is at imitating the sounds of non-words, despite having experienced equal numbers of imitation trials for the two sound categories. The implication is that infant vocal learning ought to be enhanced by having experience in a physical and social environment in which vocalization can be used as a social tool to manipulate the physical environment. This

implies that research on human infant vocal learning should consider not only the social but also the physical (visual and tactile) environments in which infants vocalize.

Sociocultural Perspectives

Since input from caregivers has been shown to affect both infant vocal learning and early speech production, we would expect differences in infants' social environments to be associated with differences in infant vocalizations and in the pace or trajectory of infant vocal learning. The following sections discuss three dimensions of difference in infants social environments, socioeconomic, linguistic, and cultural, and what is known about how each affects infant vocal development.

Socioeconomic Status

A number of studies have demonstrated that, in North America and Europe, higher socioeconomic status (SES), usually measured through parental, especially maternal, educational attainment and sometimes including income, is associated with faster language acquisition by infants (Golinkoff et al., 2015). There is more evidence that this matters for early word production than for prelinguistic vocal development. One study that tested for a relationship between SES and infant protophone vocalizations during the first year of life found that higher SES was associated with increased infant volubility (i.e., infants vocalized more often) but did not detect any relationship between SES and the age of onset of canonical babbling (Eilers et al., 1993). This is perhaps surprising, considering that that higher SES is associated with higher rates of sensitive caregiver responding to infant vocal behavior and sensitive responding promotes vocal learning. It is possible that more sensitive measures or larger sample sizes would reveal an association between SES and prelinguistic vocalization milestones. Melvin et al. (2017) found that phonetic perceptual tuning (measured by an infant's insensitivity to phonetic contrasts that don't exist in their native language) at 9 months was

related to features of the home environment that promote language and literacy but was not related to SES. Canonical babbling age of onset might similarly turn out to also be associated fairly strongly with specific features of the home environment but only weakly or not at all with SES more generally. Another possibility is that because intrinsically motivated play drives much of prelinguistic vocal development during the first year, like other motor milestones, achievement of prelinguistic vocal milestones is relatively robust to differences in the social environment that correlate with SES (Eilers et al., 1993; Oller, 2000).

Early productive vocabulary development, on the other hand, is clearly positively correlated with SES. A seminal study by Hart & Risley (1995) found that children of higher SES typically heard a greater number and variety of words, and also received more expansive responses to their own productions. This was reflected in how the children's linguistic productions evolved over the first few years of life. The number of different words the child spoke, both overall and relative to the total number of word tokens spoken, was greater for children of professors than from children receiving welfare. Differences were apparent even at the earliest stages of word production, before infants were 18 months old, and the gap widened as children grew older. More recent research has replicated this SES effect on vocabulary development, has highlighted the importance of interactive child-directed speech and lexical diversity, and has also indicated that additional factors, such as exposure to rich gestural input, may be involved (Hoff, 2003; Huttenlocher et al., 2010; Rowe et al., 2012; Golinkoff et al., 2018).

Linguistic Differences

The question of at what age infants' vocal productions reflect the phonology of the infant's home language(s) has received a fair amount of attention. On the one hand, it has been reported that newborn infant cry acoustics differ for children exposed to French versus German,

two languages with distinctly different stress patterns (Mampe et al., 2009). Follow up studies comparing Swedish to German (Prochnow et al., 2017), and Mandarin to German (Wermke et al., 2017) have found similar results. There is some question though whether these results should be trusted since statistical analyses were at the cry utterance level rather than at the child level, which could mean that the results are primarily driven by individual differences in children's cry acoustics rather than being driven by language-related differences (Gustafson et al., 2017).

If there are indeed language-related differences in newborn cry, this could be due to infants' subconsciously processing the acoustic patterns to which they have been exposed and modifying their reflexive vocalization acoustics to match. An alternative explanation could be that differences in caregiving practices across cultures lead to different intensities of cry and subsequently different average cry acoustics. One reason to doubt the first explanation is that detectable differences in protophone vocalizations have not been reported until 10 months of age, and even at that point, the reported differences are controversial.

The evidence for differences in babbling at 10 months comes from phonetic transcription of infant vocalizations, showing differing distributions of phones in the babble of infants exposed to four different languages (de Boysson Bardies & Vihman, 1991). The controversy stems from the fact that the transcribers were knowledgeable about the infant's home language, had access to the ambient language the infants heard during the naturalistic recordings, and spoke the same language as the home language of the infant whose babble they transcribed, making it possible that the results could be due primarily to transcriber bias. Subsequent studies with tighter controls for transcriber bias have not identified differences in the distribution of sound types across babble from infants exposed primarily to English versus Spanish (Thevenin et al., 1985). On the other hand, there are good reasons to believe that infants at 10 months should be

capable of incorporating the sounds produced by conversation partners into their own productions (Goldstein & Schwade, 2008), so the null results in the latter study could very well be due to small sample size, insufficient phonological differences across the two languages, or a focus on the wrong phonological measures.

As first words begin to be produced we can expect to see differences between infants learning different languages since word recognition by caregivers is by definition language-specific. That said, in early word production, there do appear to be some consistencies that may be driven in part by language-universal constraints on motor control and cognition. For example, stop consonants (such as /b/, /p/, /d/, /t/, /g/), which are common in both infant babble and in early word production are also common to all adult languages, whereas liquids (such as /l/ and /r/) and consonant clusters (such as /st/ and /sp/) are less common in babble, early words, and in adult speech across languages (Vihman et al., 1986).

Cross-Cultural Perspectives

In addition to SES and linguistic differences in infants' social environments, cultural differences appear to affect early language production and may also have effects on prespeech vocalization. Numerous differences have been found across cultures in the degree to which parents engage in protoconversations and conversations with their infants, the contexts and routines in which these conversations take place, the linguistic and semantic features of adults' speech to children, and the diversity and ages of people a child frequently interacts with (Tamis-LeMonda & Song, 2012). These differences are associated with differences in children's verbal productions and various other aspects of their behavior. For example, children whose parents frequently engage them in book reading activities tend to have larger vocabularies (Tamis-LeMonda & Song, 2012). However, it has not yet been determined whether there are

substantial cultural effects on infants' prelinguistic vocalization frequency, types, and developmental timelines.

Of particular relevance to the infant vocal learning milestones and mechanisms discussed above, there are some cultures in which there is an expectation that adults talk frequently even to preverbal infants and other cultures in which adults talk much less to their infants (Richman et al., 1992; Tamis LeMonda & Song, 2012; Cristia et al., 2017). Infant-directed talk is sometimes even being actively avoided due to beliefs about negative effects it could have on the infant and about the infant's mental abilities (Weber et al., 2017). On the one hand, we know that contingent adult responses shape infant vocalization rates and maturity, at least at short timescales. Based on this, we would expect infants to have lower vocalization rates and somewhat later achievement of pre-speech vocalization milestones in cultures in which contingent adult responses to infant protophone vocalizations are less highly valued, and therefore presumably both less frequent and less dependent on infant vocalization type. On the other hand, the fact that differences in canonical babbling age of acquisition have not been found for children from lower SES households suggests that the major prelinguistic infant vocalization type milestones may be relatively unaffected by cultural differences and may be more biologically-driven or at least endogenously-driven (Oller, 2000). And even when adult vocalizations to infants are less frequent, if they are equally contingent on infant vocal maturity (i.e. more likely following more advanced infant sound types than others), the effects on infant vocal development may be minor. It might also be expected that when infants are exposed to less adult speech that has been acoustically modified to accommodate infant perceptual preferences (Tamis-LeMonda & Song, 2012), infants' vocalizations might tend to have proportionally more adult-like as opposed to exaggerated vocalization acoustics. On the other hand, endogenous exploration may play a greater role in prelinguistic vocal development for

such infants and this might drive those infants toward less adult-speech-like and more idiosyncratic vocal productions.

These questions can only be answered by actually studying the frequencies with which different vocalization types are produced by infants at a range of ages across cultures with different behaviors and beliefs concerning infants' prelinguistic vocalizations. And the concern that absence of documented differences in vocalization milestones with SES may be due to insufficiently sensitive methodology will likely also be relevant for studies of cultural differences in prelinguistic vocal milestones.

Clinical Perspectives

We now turn to the question of how clinical differences among infants affect early vocal development. The focus is on the two disorders for which there is the most evidence of effects on prelinguistic vocal development, congenital hearing loss and autism spectrum disorder.

Congenital Hearing Loss

In cases of severe or profound hearing loss without amplification, infants show both delays and differences in their prelinguistic vocal development. In particular, while they do eventually transition from non-canonical to canonical babbling, they typically do so at 10 months or older, which is a considerable delay compared to typically developing infants (Oller & Eilers, 1988). They also may show different distributions of various consonant types within their babble, with more glottal stops (where there is alternation between vibration and no vibration at the larynx but no closure due to tongue, jaw, or lip movement) and glides (such as /j/, the consonant in "yeah", and /w/) (Stoel-Gammon & Otomo, 1986). Less severe hearing loss is associated with lower rates of canonical babbling and with delays in consonant inventory growth (Ertmer & Nathani Iyer, 2010). It is possible that these differences emerge from the fact that much early vocal learning is driven by an intrinsic interest in learning about the sensory consequences of

motor actions. For infants with severe or profound hearing loss, the sensory consequences of vocalization would primarily take the form of tactile and proprioceptive sensations in the neck and head, and the act of phonating (creating sound through vibration of the vocal folds) might generate particularly salient stimuli for these children (Ertmer & Nathani Iyer, 2010). The effects on sound of creating contact between the tongue and the roof of the mouth while phonating might be less salient for these children than for hearing children. This could account for both the delay in canonical babbling onset and for the differences in consonant type frequencies after the onset of canonical babbling. Another possible factor could be that children with hearing loss have different social environments, characterized by fewer adult responses to children's vocalizations (Nittrouer, 2009). Of course, when sign language is used by a fluent caregiver to interact with an infant, this provides an alternate path for language production development that is unimpaired by the hearing loss. Interestingly, infants who receive sign language input exhibit a manual babbling behavior analogous to vocal babbling (Petitto & Marentette, 1991).

Autism Spectrum Disorder (ASD)

Lately there has been strong interest in identifying earlier risk factors for ASD, so the prelinguistic babble of infants with ASD or at high familial risk for ASD has received attention. Patten et al. (2014) retrospectively analyzed home videos of infants later diagnosed with ASD and found that most of the infants produced less canonical babbling than typically developing infants. Interestingly, two of the infants with ASD produced more canonical babble than typically developing infants; the researchers suspected that for these children, canonical babbling was a type of motor stereotypy. Similarly, Swanson et al. (2018) found increased vocalization rates among a subgroup of infants who had older siblings with ASD, with that subgroup not showing the high rates of conversational turns with adults that would be expected for typically developing infants with high vocalization rates. Paul et al. (2011) also found differences in the vocalizations

of high risk infants, namely reduced rates of speech-like vocalizations, reduced inventories of consonant types produced, and reduced numbers of different syllable shapes, compared to low risk infants. Paul et al. also found that differences in early productions were associated with differences in ASD symptoms during the second year of life for the high risk infants. It thus seems that differences in preverbal infants' speech-related vocalizations may be potential early indicators of autism risk. Moreover, given the relationship between prelinguistic non-cry vocalizations and later meaningful language production, for infants identified as high risk, early intervention around vocal communication may be justified.

Policy Perspectives

The final section of this chapter discusses some of the currently recommended approaches to encouraging timely speech production development in infants, particularly those with clinical disorders or from socioeconomically disadvantaged communities. The section then discusses how issues of cultural sensitivity affect both researchers and interventionists.

Interventions to Promote Infant Vocal Learning and Speech Production

Beginning at birth, and even for premature infants (Caskey et al., 2011), clinicians and scientists encourage caregivers to provide lots of verbal input to infants and to be attentive and responsive to their infants' vocalizations, facial expressions, and gestures. In most of the clinical cases mentioned above the primary treatments are behavioral interventions that include educating the infant's primary caregivers on methods for creating a physical and social environment that promotes language development. Even in cases of hearing loss, while amplification and/or cochlear implants are also provided, behavioral interventions to create environments especially rich in input and sensitive responding are a key component of treatment (Moeller et al., 2013).

In cases where infants are typically developing but are members of (usually lower SES) communities where children's language development is slower compared to other groups, behavioral interventions are often provided. The goal is to create an environment that provides ample high-quality child-directed language input, not too much background noise, and frequent sensitive, positive responses to infants' communication acts (Hirsh-Pasek et al., 2015; Yazejian et al., 2017). A randomized controlled study with low-SES American participants found that such interventions can, at least in the short-term, be effective in increasing child vocalization rates, adult speech heard by children, and conversational turns between children and adults (Suskind et al., 2016). A similar approach has been shown effective in increasing maternal speech to infants and infant volubility and expressive language skills in a culture that traditionally discourages infant-directed speech and gaze (Weber et al., 2017). The larger purpose is to even the playing field for infants, regardless of background, to have the skills that help prepare them for formal schooling to promote cognitive development and economic success.

Relatedly, there has been a push for all families, regardless of clinical or socioeconomic status, to limit infants' screen time, such as television viewing and smartphone or tablet usage, experienced by infants. The American Academy of Pediatrics recommends no screen time at all before the age of two years and only a very limited (1 hr or less) daily upper limit of screen time after that age (Council on Communications And Media, 2013). There are other reasons for such recommendations, such as promoting physical activity, but a large part of the motivation comes from evidence that increased screen time is associated with both reduced adult-infant conversations and slower rates of language development (Christakis et al., 2009; Zimmerman et al., 2009). Experimental work supports this recommendation, finding that 9-month-old infants cannot learn phonetic properties of a non-native language, at least not as well, from television or pre-recorded audio-only input, while they can learn from an equivalent duration of live exposure

to lessons by a speaker of that language (Kuhl et al., 2003). There every reason to expect that this effect would also hold for infant vocal production. We know that contingent responses affect infant vocal productions and recorded input and even more interactive games and toys (at least at this point in history) are not contingent on infant vocalization and infant vocalization type. Moreover, even the presence of interactive toys (those that make an “electronic sound or automatic movement in response to manipulation”) as opposed to more traditional toys (such as nesting cups and balls) has been shown to be associated with reduced infant vocalizations and parent responses to infant vocalizations (Miller et al., 2017). Relatedly, caregivers’ own use of electronic devices can also be cause for concern when it distracts adults from attending to their infants. For older children, maternal use of mobile devices is associated with decreased mother-child interactions (Radesky et al., 2015) and interruptions to word learning (Reed et al., 2017).

Culturally Sensitive Policies

Despite the ubiquity of interventions designed to encourage certain types of adult interactions with infants, such as frequent, positive responses to infants’ speech-related vocalizations that reinforce the infants’ emerging communication skills by incorporating imitation, expansion, object labeling, and so on, these recommendations are not without controversy (Weber et al., 2017). Often the controversies center around the perspective that members of groups considered “at-risk” should not be pressured to conform to a pattern of development that is typical with reference to some particular cultural standpoint. For example, the neurodiversity movement has argued that being on the autism spectrum should not be considered a disorder but rather a difference, and that individuals on the spectrum should be valued for the diversity they bring to society, culture, and the workplace (Kapp et al., 2013). Perspectives and policies that imply or seem to imply that parents of children with or at risk for autism are not interacting

optimally with their children from some mainstream cultural perspective might harm families. The may interfere with parents' more natural, intuitive styles of interaction, which may already be essentially optimized from the perspective of that child's and family's happiness as well as from the perspective of a society that values diversity (Akhtar et al., 2016). This being acknowledged, some families may appreciate having information about what helps promote oral language development in infancy. Parents may be eager to modify their caregiving practices and/or their infant's physical and social environment accordingly (Warlaumont et al., 2016), and such interventions may prove important for maximizing individuals' and communities' economic success (Weber et al., 2017). The challenges are to conduct research that is culturally sensitive and appreciates the multidimensionality of infant development and the tradeoffs of risks and benefits of various perspectives and practices, and to present research findings in a way that is descriptive and informative but not prescriptive or judgmental. Likewise, interventions that target the natural caregiving practices of parents, whether they are members of subgroups of a larger society, such as is often the case for families of lower SES in the US for example, or whether members of a non-industrial society, can and should be questioned as to whether they are truly in the best interests of an infant and her family and community or whether they are instead impositions of cultural values of a more dominant culture or subculture.

A methodological challenge to achieving this goal has been the difficulty in obtaining data from diverse populations, so that our understanding of typical development can be more comprehensive and less biased (Henrich et al., 2010). Fortunately, data sharing initiatives, such as the well-established CHILDES (CHIld Language Data Exchange System) (MacWhinney, 2000) and newer initiatives such as the Databrary system for sharing video data of research studies with children (Gilmore & Adolph, 2017) and the HomeBank resource for sharing long-form audio recordings from child-worn recorders (VanDam et al., 2016) provide a range of

resources for researchers to contribute to and access repositories of raw data that are larger and more diverse.

Conclusion

Infant vocal behavior changes dramatically over the course of the first year of life, and this is considered by many to be one of the defining features of our species. At birth, human infants produce only cries, vegetative sounds, and short, quiet initial precursors to speech. By the end of the first year most children can produce long sequences that include a variety of speech sounds, can coordinate their vocalizations in turn-taking patterns with other humans, and in many cases are beginning to use their vocal sound-making apparatus to produce meaningful, recognizable (if not completely correctly pronounced) words. It appears that this process of vocal learning is the product of endogenous exploration influenced by social input from responsive caregivers. We are beginning to gain an understanding, aided by computational modeling studies and by neuroscientific research on humans and other animals, of some of the neurophysiological mechanisms underlying infant vocal learning. In particular, there is some evidence that spontaneous activity and reinforcement-driven learning in cortical regions play a major role. It also appears that while there are some sociocultural and clinical differences in vocal learning and early speech production (particularly evident for individuals with severe congenital hearing loss and increasingly documented for individuals at risk of later autism diagnosis), many aspects of the prelinguistic vocal learning process are fairly robust. Interventions to increase infant vocalization rates and language learning more generally tend to focus on increasing adult caregivers' verbal input, particularly in the form of sensitive responding to infant vocalization and other behaviors.

There are a number of areas that future research should especially target. Intrinsic motivation appears to play a large role in infant vocal development, yet has not received as

much attention as social input has. This may be due in part to methodological challenges. For example, as a researcher, it is easier to track moments when a positive social response is received than it is to track moments when a child is intrinsically rewarded for producing a particular sound. Of course, social input and intrinsic motivations likely mutually influence each other, and these mutual influences are an interesting and important topic for future research. Neurophysiological bases of human infant vocal learning are also difficult to study given the methodological limitations to studying neural activity in awake, healthy infants. Improvements in imaging technology and in computational models may help overcome some of these challenges in the future. Studies of sociocultural differences may also benefit from future advances in methods for characterizing and classifying infant vocalizations of different types, and from increases in sample size and diversity thanks to data sharing. Data sharing also has the advantages of promoting replicable science, reducing research costs, and facilitating interdisciplinary collaboration for example with audio processing and speech recognition experts who can advance methods to study infant vocal communication development. Finally, besides informing our understanding of how typically developing infants acquire vocal communication skills, future research on these and other topics covered in this chapter can be expected to facilitate the development of culturally-appropriate interventions to reduce gaps in school-readiness and better enable children with communication disorders.

Acknowledgments

Thanks to Kim Oller for many helpful discussions on this topic. The writing of this chapter was facilitated by a James S. McDonnell Foundation Scholar Award in Understanding Human Cognition and by National Science Foundation grants SMA-1539129 and BCS-1529127 funding related work. Any opinions, findings, and conclusions or recommendations expressed in this

material are those of the author and do not necessarily reflect the views of the National Science Foundation or the James S. McDonnell Foundation.

References

Akhtar, N, Jaswal, V. K., Dinishak, J., & Stephan, C. (2016). On social feedback loops and cascading effects in autism: A commentary on Warlaumont, Richards, Gilkerson, and Oller (2014). *Psychological Science*, *27*, 1528–1530.

Baudonck, N. L. H., Buekers, R., Gillebert, S., & Van Lierde, K. M. (2009). Speech intelligibility of Flemish children as judged by their parents. *Folia Phoniatrica et Logopaedica*, *61*, 288–295.

Bloom, K. (1988). Quality of adult vocalizations affects the quality of infant vocalizations. *Journal of Child Language*, *15*, 469-480.

Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 327–332.

Bryant, G. A., & Aktipis, C. A. (2014). The animal nature of spontaneous human laughter. *Evolution and human behavior*, *35*, 327–335.

Buder, E. H., Warlaumont, A. S., & Oller, D. K. (2013). An acoustic phonetic catalog of prespeech infant vocalizations from a developmental perspective. In B. Peter & A. N. MacLeod (Eds.), *Comprehensive perspectives on child speech development and disorders: Pathways from linguistic theory to clinical practice*. Nova Science Publishers.

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2011). Importance of parent talk on the development of preterm infant vocalizations. *Pediatrics*, *128*, 910–916.

Christakis, D. A., Gilkerson, J., Richards, J. A., Garrison, M. M., Xu, D., Gray, S., & Yapanel, U. (2009). Audible television and decreased adult words, infant vocalizations, and

conversational turns: A population-based study. *Archives of Pediatrics & Adolescent Medicine*, *163*, 554–558.

Council on Communications And Media. (2013). Children, adolescents, and the media. *Pediatrics*, *132*, 958–961.

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2017). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. doi: 10.1111/cdev.12974

de Boysson Bardies, B., & Vihman, M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, *67*, 297–319.

Eilers, R. E., Oller, D. K., Levine, S., Basinger, D., Lynch, M. P., & Urbano, R. (1993). The role of prematurity and socioeconomic status in the onset of canonical babbling in infants. *Infant Behavior and Development*, *16*, 297–315.

Ertmer, D. J., & Nathani Iyer, S. (2010). Prelinguistic vocalizations in infants and toddlers with hearing loss: Identifying and stimulating auditory-guided speech development. In M. Marschark & P. E. Spencer (Eds.), *The Oxford handbook of deaf studies, language, and education*. Oxford University Press.

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

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

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*, 477–501.

Forestier, S., & Oudeyer, P.-Y. (2017). A unified model of speech and tool use early development. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.),

Proceedings of the 39th Annual Meeting of the Cognitive Science Society. Cognitive Science Society.

Ghazanfar, A. A., & Zhang, Y. S. (2016). The autonomic nervous system is the engine for vocal development through social feedback. *Current Opinion in Neurobiology*, *40*, 155–160.

Gilmore, R. O., & Adolph, K. E. (2017). Video can make behavioural science more reproducible. *Nature Human Behaviour*, *1*, 0128.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 8030–8035.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*, 515–523.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby) talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science*, *24*, 339–344.

Golinkoff, R. M., Hoff, E., Rowe, M. L., Tamis-LeMonda, C. S., & Hirsh-Pasek, K. (2018). Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child Development*. doi: 10.1111/cdev.13128

Gratier, M., & Devouche, E. (2011). Imitation and repetition of prosodic contour in vocal interaction at 3 months. *Developmental Psychology*, *47*, 67–76.

Gratier, M., Devouche, E., Guellai, B., Infanti, R., Yilmaz, E., & Parlato-Oliveira, E. (2015). Early development of turn-taking in vocal interaction between mothers and infants. *Frontiers in Psychology*, *6*, 1167.

Gros-Louis, J., & Miller, J. L. (2018). From ‘ah’ to ‘bah’: Social feedback loops for speech sounds at key points of developmental transition. *Journal of Child Language*, *45*, 807–825.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants’ prelinguistic sounds. *International Journal of Behavioral Development*, *30*, 509–516.

Gultekn, Y. B., & Hage, S. R. (2018). Limiting parental interaction during vocal development affects acoustic call structure in marmoset monkeys. *Science Advances*, *4*, eaar4012.

Gustafson, G. E., Sanborn, S. M., Lin, H.-C., & Green, J. A. (2017). Newborns’ cries are unique to individuals (but not to language environment). *Infancy*. doi: 10.1111/infa.12192

Harder, S., Lange, T., Foget Hansen, G., Væver, M., & Køppe, S. (2015). A longitudinal study of coordination in mother-infant vocal interaction from age 4 to 10 months. *Developmental Psychology*, *51*, 1778–1790.

Hart, B., & Risley, T. R. (1995). Meaningful differences in the everyday experience of young American children. Paul H. Brookes Publishing Co.

Heintz, I., Beckman, M., Fosler-Lussier, E., & Ménard, L. (2009). Evaluating parameters for mapping adult vowels to imitative babbling. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 688–691.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*, 29.

Hilbrink E. E., Gattis, M., & Levinson, S. C. (2015). Early developmental changes in the timing of turn-taking: A longitudinal study of mother-infant interaction. *Frontiers in Psychology*, *6*, 1492.

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., Yust, P. K. S., & Suma, K. (2015). The contribution of early communication quality to low-income children's language success. *Psychological Science*, *26*, 1071–1083.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*, 1368–1378.

Howard, I. S., & Messum, P. (2014). Learning to pronounce first words in three languages: An investigation of caregiver and infant behavior using a computational model of an infant. *PLoS ONE*, *9*, e110334.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*, 343–365.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., & Jasnow, M. D. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the Society for Research in Child Development*, *66*, 1–32.

Jones, S. J., & Moss, H. A. (1971). Age, state, and maternal behavior associated with infant vocalizations. *Child Development*, *42*, 1039–1051.

Jones, S. S. (2007). Imitation in infancy: The development of mimicry. *Psychological Science*, *18*, 593–599.

Jürgens, U. (1974). On the elicitability of vocalization from the cortical larynx area. *Brain Research*, *81*, 564–566.

Jürgens, U. (2002). Neural pathways underlying vocal control. *Neuroscience and biobehavioral reviews*, *26*, 235–258.

Kapp, S. K., Gillespie-Lynch, K., Sherman, L. E., & Hutman, T. (2013). Deficit, difference, or both? Autism and neurodiversity. *Developmental Psychology*, *49*, 59–71.

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 of the United States of America*, *100*, 9096–9101.

Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, *51*, 793–809.

Koopmans-van Beinum, F. J. & van der Stelt, J. M. (1986). Early stages in the development of speech movements. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech*. Stockton Press.

Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J.-F. L., & Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proceedings of the National Academy of Sciences*, *2014*, *111*, 11238–11245.

Leezenbaum, N. B., Campbell, S. B., Butler, D., & Iverson, J. M. (2013). Maternal verbal responses to communication of infants at low and heightened risk of autism. *Autism*, *18*, 694–703.

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, *31*, 581–612.

Lieven, E. V. M., Pine, J. M., & Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of Child Language*, *19*, 287–310.

Locke, J. L. (1989). Babbling and early speech: Continuity and individual differences. *First Language*, *9*, 191–206.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*, 499–511.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Mampe, B., Friederici, A., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology*, 1994–1997.

McCune, L., & Vihman, M. M. (2001). Early phonetic and lexical development: A productivity approach. *Journal of Speech, Language, and Hearing Research*, *44*, 670–684.

Melvin, S. A., Brito, N. H., Mack, L. J., Engelhardt, L. E., Fifer, W. P., Elliott, A. J., & Noble, K. G. (2017). Home environment, but not socioeconomic status, is linked to differences in early phonetic perception ability. *Infancy*, *22*, 42–55.

Miller, J. L., Lossia, A., Suarez-Rivera, C., & Gros-Louis, J. (2017). Toys that squeak: Toy type impacts quality and quantity of parent-child interactions. *First Language*, *37*, 630–647.

Miura, K., Yoshikawa, Y., & Asada, M. (2007). Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories. *Advanced Robotics*, *21*, 1583–1600.

Moeller, M. P., Carr, G., Seaver, L., Stredler-Brown, A., & Holzinger, D. (2013). Best practices in family-centered early intervention for children who are deaf or hard of hearing: An international consensus statement. *Journal of Deaf Studies and Deaf Education*, *18*, 429–445.

Morrill, R. J., Paukner, A., Ferrari, P., & Ghazanfar, A. A. (2012). Monkey lipsmacking develops like the human speech rhythm. *Developmental Science*, *15*, 557–568.

Moulin-Frier, C., Nguyen, S. M., & Oudeyer, P.-Y. (2014). Self-organization of early vocal development in infants and machines: The role of intrinsic motivation. *Frontiers in Psychology*, *4*, 1006.

Nathani, S., & Stark, R. E. (1996). Can conditioning procedures yield representative infant vocalizations in the laboratory? *First Language*, *16*, 365–387.

Nittrouer, S. (2009). *Early development of children with hearing loss*. Plural publishing.

Oller, D. K. (1986). Metaphonology and infant vocalizations. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech*. Stockton Press.

Oller, D. K. (2000). The emergence of the speech capacity. Lawrence Erlbaum Associates.

Oller, D. K., Buder, E. H., Ramsdell, H. L., Warlaumont, A. S., Chorna, L., and Bakeman, R. (2013). Functional flexibility of infant vocalization and the emergence of language. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 6318–6323.

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development*, *59*, 441–449.

Oller, D. K., Eilers, R., & Basinger, D. (2001). Intuitive identification of infant vocal sounds by parents. *Developmental Science*, *4*, 49–60.

Oller, D. K., Eilers, R. E., Neal, A. R., & Cobo-Lewis, A. B. (1998). Late onset canonical babbling: A possible early marker of abnormal development. *American Journal of Mental Retardation*, *103*, 249–263.

Oller, D. K., Wieman, L. A., Doyle, W. J., & Ross, C. (1976). Infant babbling and speech. *Journal of Child Language*, *3*, 1–11.

Patten, E., Belardi, K., Baranek, G. T., Watson, L. R., Labban, J. D., & Oller, D. K. (2014). Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency. *Journal of Autism and Developmental Disorders*, *44*, 2413–2428.

Paul, R., Fuerst, Y., Ramsay, G., Chawarska, K., & Klin, A. (2011). Out of the mouths of babes: Vocal production in infant siblings of children with ASD. *Journal of Child Psychology and Psychiatry*, *52*, 588–598.

Perlman, M., & Clark, N. (2015). Learned vocal and breathing behavior in an enculturated gorilla. *Animal Cognition*, *2015*, 1–15.

Penfield, W., & Welch, K. (1951). The supplementary motor area of the cerebral cortex: A clinical and experimental study. *A.M.A. Archives of Neurology and Psychiatry*, *66*, 289–317.

Petitto, L. A., & Marentette, P. F. (1991). Babbling in the manual mode: Evidence for the ontogeny of language. *Science*, *251*, 1493–1496.

Prochnow, S., Hesse, V., & Wermke, K. (2017). Does a 'musical' mother tongue influence cry melodies? A comparative study of Swedish and German newborns. *Musicae Scientiae*. doi: 10.1177/1029864917733035

Radesky, J., Miller, A. L., Rosenblum, K. L., Appugliese, D., Kaciroti, N., & Lumeng, J. C. (2015). Maternal mobile device use during a structured parent-child interaction task. *Academic Pediatrics*, *15*, 238–244.

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*, 880–891.

Reed, J., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). Learning on hold: Cell phones sidetrack parent-child interactions. *Developmental Psychology*, *53*, 1428–1436.

Richman, A. L., Miller, P. M., & LeVine, R. A. (1992). Cultural and educational variations in maternal responsiveness. *Developmental Psychology*, *28*, 614-621.

Robb, M. P., Bauer, H. R., & Tyler, A. A. (1994). A quantitative analysis of the single-word stage. *First Language*, *14*, 37–48.

Rowe, M. L., Raudenbush, S. W., & Goldin-Meadow, S. (2012). The pace of vocabulary growth helps predict later vocabulary skill. *Child Development*, *83*, 508–525.

Russell, J. L., McIntyre, J. M., Hopkins, W. D., & Taglialatela, J. P. (2013). Vocal learning of a communicative signal in captive chimpanzees, *Pan troglodytes*. *Brain and Language*, *127*, 520–525.

Schneider, R. M., Yurovsky, D., & Frank, M. C. (2015). Large-scale investigations of variability in children's first words. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Smith, B., Brown-Sweeney, S., & Stoel-Gammon, C. (1989). A quantitative analysis of reduplicated and variegated babbling. *First Language*, *9*, 175–189.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*, 501–532.

Stark, R. E. (1980). Stages of speech development in the first year of life. *Child Phonology, Vol. 1: Production*. Academic Press.

Stoel-Gammon, C. (2001). Down syndrome phonology: Developmental patterns and intervention strategies. *Down Syndrome, Research, and Practice*, *7*, 93–100.

Stoel-Gammon, C., & Otomo, K. (1986). Babbling development of hearing-impaired and normally hearing subjects. *Journal of Speech and Hearing Disorders*, *51*, 33–41.

Suskind, D. L., Leffel, K. R., Graf, E., Hernandez, M. W., Gunderson, E. A., Sapolich, S. G., Suskind, E., Leininger, L., Goldin-Meadow, S., & Levine, S. C. (2016). A parent-directed

language intervention for children of low socioeconomic status: A randomized controlled pilot study. *Journal of Child Language*, *43*, 366–406.

Swanson, M. R., Shen, M. D., Wolff, J. J., Boyd, B., Clements, M., Rehg, J., Elison, J. T., Paterson, S., Parish-Morris, J., Chappell, J. C., Hazlett, H. C., Emerson, R. W., Botteron, K., Pandey, J., Schultz, R. T., Dager, S. R., Zwaigenbaum, L., Estes, A. M., Piven, J., & the IBIS Network. (2018). Naturalistic language recordings reveal "hypervocal" infants at high familial risk for autism. *Child Development*, *89*, e60–e73.

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Development*, *72*, 748–767.

Tamis-LeMonda, C. S., Kuchirko, Y., & Suh, D. D. (2018). Taking center stage: Infants' active role in language learning. In M. Saylor & P. Ganea (Eds.), *Active learning from infancy to childhood*. Springer.

Tamis-LeMonda, C. S., & Song, L. (2012). Parent-infant communicative interactions in cultural context. In R. M. Lerner, E. Easterbrooks, & J. Mistry (Eds.), *Handbook of Psychology (2nd Ed.), Volume 6: Developmental Psychology*.

Tarabulsy, G. M., Tessier, R., & Kappas, A. (1996). Contingency detection and the contingent organization of behavior in interactions: Implications for socioemotional development in infancy. *Psychological Bulletin*, *120*, 25–41.

Tardif, T., Gelman, S., & Xu, F. (1999). Putting the noun bias in context: A comparison of English and Mandarin. *Child Development*, 70, 620–635.

Thevenin, D. M., Eilers, R. E., Oller, D. K., & Lavoie, L. (1985). Where's the drift in babbling drift? A cross-linguistic study. *Applied Psycholinguistics*, *6*, 3–15.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language*, *37*, 128–142.

Vihman, M. M., Ferguson, C. A., & Elbert, M. (1986). Phonological development from babbling to speech: Common tendencies and individual differences. *Applied Psycholinguistics*, *7*, 3–40.

Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A re-assessment of the continuity issue. *Language*, *61*, 397–445.

Vouloumanos, A, & Werker, J. F. (2004). Tuned to the signal: The privileged status of speech for young infants. *Developmental Science*, *7*, 270–276.

Warlaumont, A. S., & Finnegan, M. K. (2016). Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PLOS ONE*, *11*, e0145096.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., Messinger, D. S., & Oller, D. K. (2016). The social feedback hypothesis and communicative development in autism spectrum disorder: A response to Akhtar, Jaswal, Dinishak, and Stephan (2016). *Psychological Science*, *27*, 1531–1533.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological Science*, *25*, 1314–1324.

Warlaumont, A. S., Westermann, G., Buder, E. H., & Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, *38*, 64–75.

Weber, A., Fernald, A., & Diop, Y. (2017). When cultural norms discourage talking to babies: Effectiveness of a parenting program in rural Senegal. *Child Development*, *88*, 1513–1526.

Weist, R. M., & Kruppe, B. (1977). Parent and sibling comprehension of children's speech. *Journal of Psycholinguistic Research*, *6*, 49–58.

Wermke, K., Ruan, Y., Feng, Y., Dobnig, D., Stephan, S., Wermke, P., Ma, L., Chang, H., Liu, Y., Hesse, V., & Shu, H. (2017). Fundamental frequency variation in crying of Mandarin and German neonates. *Journal of Voice*, *31*, 255.e25–255.e30.

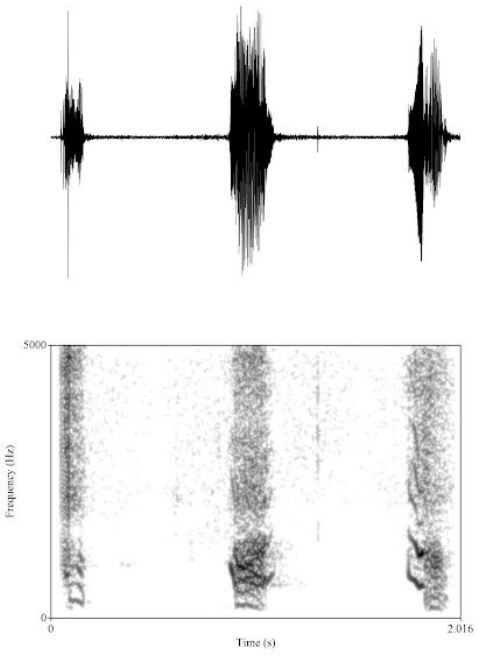
Westermann, G., & Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, *89*, 393–400.

Yazejian, N., Bryant, D. M., Hans, S., Horm, D., St. Clair, L., File, N., & Burchinal, M. (2017). Child and parenting outcomes after 1 year of Educare. *Child Development*, *88*, 1651–1688.

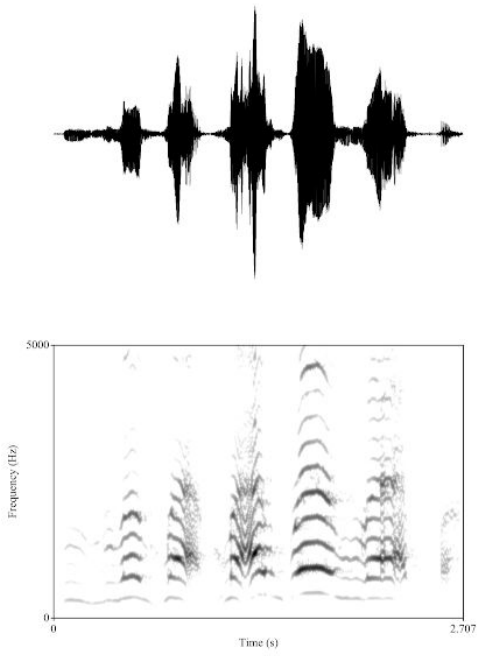
Yoshikawa, Y., Asada, M., Hosoda, K., & Koga, J. (2003). A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science*, *15*, 245–258.

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, *124*, 342–349.

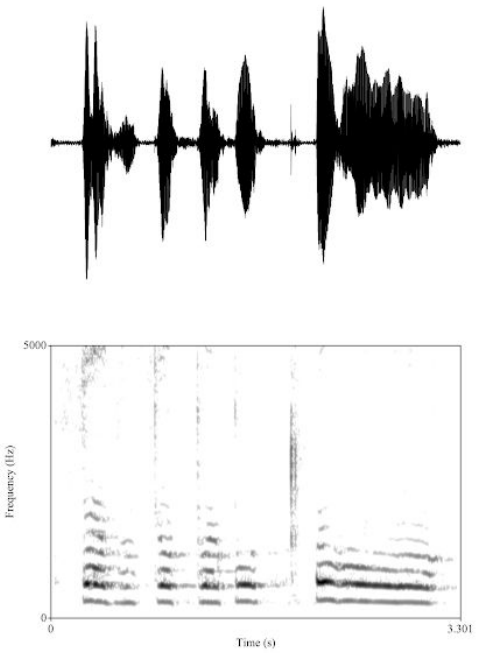
A Newborn protophones



B Early canonical babble with initial “b”



C Early word repetition: “bubble”



D Mother: “are those bubbles?”

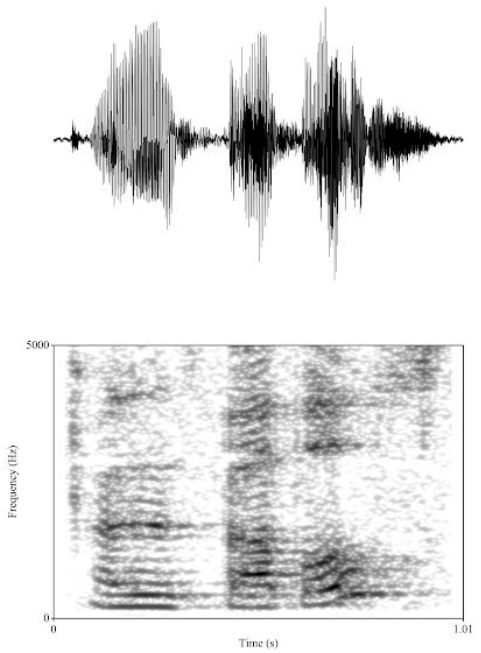


Figure 1. A: Waveform and spectrogram illustrating a protophone sequence produced by a 19 day old infant. The sounds are short, quiet, and do not contain consonant margins. The first sound could be coded as vowel-like and the second and third as growls. B: A sequence of syllable vocalizations produced by the same infant at 4 months 20 days. This is one of the first canonical babble sequences produced by the infant. C: Early word production by the same infant at 1 year 2 months 20 days. The sequence could be transcribed as “bubble buh buh buh bubble”. D: The maternal utterance, “are those bubbles?” that preceded the infant vocalization is shown in panel C. The sound files corresponding to the images are available at https://doi.org/10.6084/m9.figshare.7119578

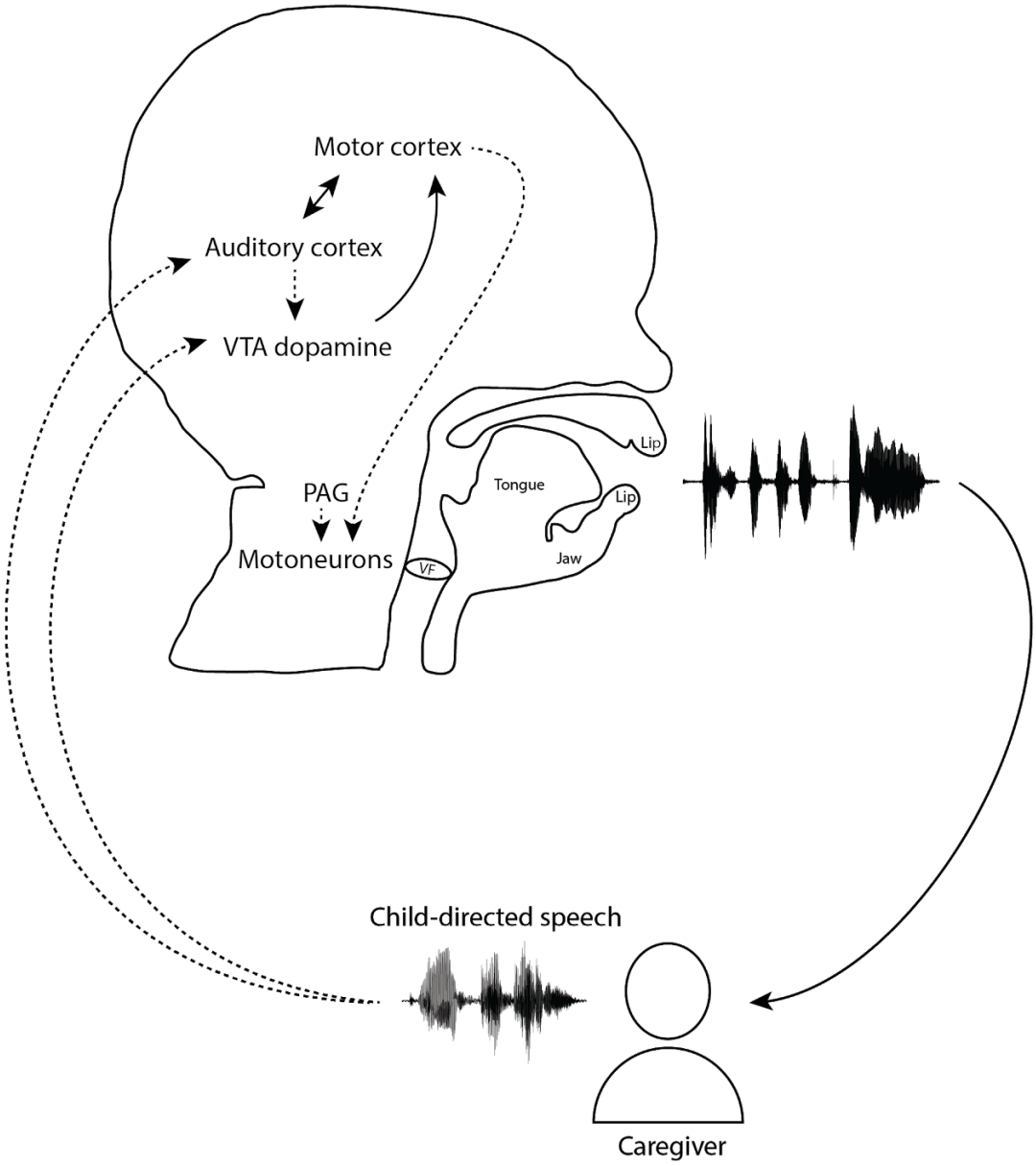


Figure 2. Schematic illustration of some of the major anatomical structures and neural and social pathways involved in infant vocalization and vocal learning. Dashed lines illustrate pathways where various subcortical regions make up part of the pathway but are not shown. Abbreviations: VTA = ventral tegmental area; PAG = periaqueductal gray; VF = vocal folds.