Motor Control, 2011, 15, 85-117

Modeling the Development of Pronunciation in Infant Speech Acquisition

Ian S. Howard and Piers Messum

Pronunciation is an important part of speech acquisition, but little attention has been given to the mechanism or mechanisms by which it develops. Speech sound qualities, for example, have just been assumed to develop by simple imitation. In most accounts this is then assumed to be by acoustic matching, with the infant comparing his output to that of his caregiver. There are theoretical and empirical problems with both of these assumptions, and we present a computational model—Elija—that does not learn to pronounce speech sounds this way. Elija starts by exploring the sound making capabilities of his vocal apparatus. Then he uses the natural responses he gets from a caregiver to learn equivalence relations between his vocal actions and his caregiver's speech. We show that Elija progresses from a babbling stage to learning the names of objects. This demonstrates the viability of a non-imitative mechanism in learning to pronounce.

Keywords: infant speech development, pronunciation, reformulations, reinforcement, interaction, correspondence problem

Speech Communication

Linguistic communication is considered to be one of the foremost human accomplishments. Speech is the acoustic expression of language, and the most common form in which it is realized. To learn to speak, an infant must master complex movements of his respiratory, laryngeal and articulatory apparatus to produce an acoustic output. From a motor control perspective, the infant learns which activations of the muscles of his vocal tract and breathing apparatus result in somatosensory and auditory sensory consequences. He does this, however, without initially knowing that such activity will have linguistic value (Locke 1996). (On the other hand, the communicative value of some forms of vocal output, e.g., in the form of crying, is discovered early on.)

Howard is with the Computational and Biological Learning Laboratory, Department of Engineering, University of Cambridge, Cambridge, UK. Messum is with the Centre for Human Communication, University College London, London, UK.

The pronunciation of the first L1 words that an infant adopts (i.e., words he hears spoken by others), may be a holistic recreation of their sound images (Studdert-Kennedy 2002). However in due course the child will come to construct words using a repertoire of actions which produce the distinct speech sounds of the ambient language. While traditionally these and many other aspects of infant learning have been seen as projects undertaken by the infant largely on his own, there is now increasing recognition of the potential and actual importance of caregiver interventions and interaction e.g., in general infant development (Zukow-Goldring and Arbib 2007) and in speech development (Goldstein and Schwade 2008; Messum 2007; Yoshikawa et al. 2003).

For convenience in the use of pronouns in this paper, we assume a male infant and a female caregiver in our discussions of caregiver-infant interactions (although a male caregiver was used to run the experiments described in the Experiments Section).

Learning to Pronounce: Cognitive Models

How is the mature skill of word pronunciation developed? As just described, the first words a child adopts from the ambient language may be recreated from adult input by a form of "whole-word mimicry". However, it is uncontroversial that a particulate principle for the phonology of word structure soon emerges. A child then starts to conceive words as being made up of speech sounds (subword units of production forming part of a mental syllabary (Levelt and Wheeldon 1994)). At this point, it is important to draw a distinction between two activities that are required for word adoption: "learning to pronounce" and "learning to pronounce words" (Messum 2008a).

"Learning to pronounce" is the systemic activity of learning to produce speech sounds that will be taken by listeners to be equivalent to the speech sounds that the listeners themselves produce. After the initial stage of "whole word mimicry", "learning to pronounce words" applies this expertise in the adoption of the word forms produced by others: the speech sounds that form a word are identified and reproduced using their equivalents from the child's repertoire. The latter activity is a form of imitation, since the sequence of the speech sounds is reproduced, but it uses elements (the speech sounds that have been learnt to be equivalent) which may or may not have themselves been learnt by imitation.

This distinction was described more generally by Parton (1976) as that between "learning to imitate" and "learning by imitation". Using his terminology, we can say that learning to pronounce is the acquisition of the perceptuo-motor isomorphisms linking the speech sounds that the child hears to the molecular motor behaviors underlying his production of what his listeners will take to be equivalent sounds.

Using a more contemporary formulation of this issue, learning to pronounce can be seen as the child's solution of the "correspondence problem" between speech sounds he hears and speech sounds he makes. Those sounds he makes must be taken by listeners to be equivalent to their own, but for this to happen they need not be identical or even acoustically similar (although their functional equivalence may lead to a "learnt" or "theory-based" (Mompean-Gonzalez 2004) judgment of their similarity).

The general assumption about the mechanism for "learning to pronounce" is that sound qualities are copied, solving the correspondence problem through acoustic matching:

> "Infants learn to produce sounds by imitating those produced by another and imitation depends upon the ability to equate the sounds produced by others with ones infants themselves produce." (Kuhl 1987).

Other cognitive models propose instead that gestures rather than sounds are imitated, e.g., (Goldstein et al. 2003), and further variants on these two possibilities exist; see Messum (2007) for a review. However, models which depend upon imitation are problematic for some theoretical reasons and because observation of infant speech development and adult performance have identified several phenomena that cannot be explained by any imitative account (Messum 2007) [1].

As an alternative to the infant matching his speech sounds to those of his caregivers acoustically, an infant can solve the correspondence problem via the information made available to him by a caregiver when she takes the role of a vocal "mirror" for his output. A caregiver takes this mirroring role whenever she reflects her child's output back to him, either by mimicking it or by reformulating it. There are many such episodes of vocal imitation in mother-child interaction. Pawlby (1977) reported that over 90% of "imitative" exchanges between caregivers and infants between 17 and 43 weeks of age were actually of this type, where a mother "imitates" her child rather than vice versa.

Within this framework of interaction, reformulation rather than mimicry becomes the mother's preferred response and reformulation of a child's vocal output by his mother continues until at least age 4 (Chouinard and Clark 2003). Reformulation transforms the child's output into his mother's interpretation of what he has uttered within the phonology of L1 (the mother's first language). Sound reformulation is therefore analogous to so-called "affect attunement" (Stern 1985) on the part of the mother in more general child development, rather than to simple mimicry. As with reformulations, affect attunement also replaces mimicry of affect in mother-infant interactions (Jonsson et al. 2001). As the mother's response comes within the context of an imitative exchange which the child will recognize as such (Meltzoff 1999), it provides the child with the evidence for him to deduce a correspondence between his output and the speech sound equivalent within L1 that she produces. He understands that his mother regards the two as equivalent, and he relies on her judgment in this matter. In this way, the infant can deduce the linguistic value of what he performs.

Once the correspondence problem is solved, learning to pronounce a word requires recognition and correct sequencing of the speech sound elements that make it up. Heyes (2001) provides a general graphic device for one class of imitation that illustrates this two part process of learning to pronounce and learning to pronounce words. We reproduce this as Figure 1. Here the sequencing problem is represented by horizontal associations, and the correspondence problem is represented by the vertical associations between sensory input and motor output. Thus to learn the pronunciation of a word like "gruffalo", a speaker may parse and then reproduce the word shape as three speech sounds: perhaps corresponding to "gru", "ffa" and "lo".

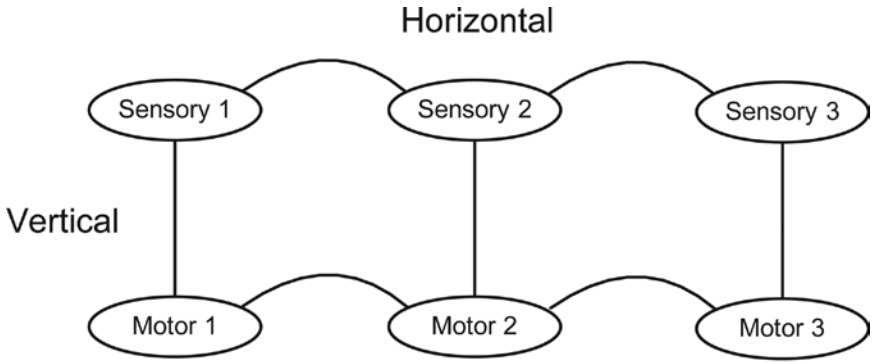

Figure 1 — Mapping between sensory and motor levels of representation. Parsing the input creates a sequencing (horizontal) specification, but the motor equivalents to the sensory elements identified (the vertical specification) must have been established previously. Only then is the mature mechanism of word reproduction possible.

Learning to Pronounce: Previous Computational Models

As well as cognitive models, there are also a number of computational models of how speech production develops, noted below. The main difference between Elija, our model, and these is that Elija interacts with a caregiver. In particular, he makes use of the well-attested caregiver reformulations of child output that are provoked by a child's vocal activity. Such interactions are not used as sources of information in the other models discussed below. In addition, we model development from babbling to the learning of words, with a focus on the learning of pronunciation.

Laboissière created one of the earliest computational models of articulatory skill acquisition, with a connectionist model that learnt to produce vowels (Laboissière 1992). Guenther's DIVA model (Guenther 1994; 1995; Guenther et al. 2006) uses a neural network to investigate the acquisition of speaking skills by infants. DIVA addresses a range of phenomena observed in speech production, such as motor equivalence, contextual variability, coarticulation and speaking rate effects. In HABLAR (Markey 1994; Markey 1993; Menn et al. 1993), Markey modeled the articulatory foundations of phonology with a sensorimotor simulation consisting of an auditory system, an articulatory system and a hierarchical cognitive architecture that bridged the two. Reinforcement learning was employed to train the motor system. Kröger's model (Kröger et al. 2009a; Kröger et al. 2009b) is similar to DIVA but focuses on the neurocomputational issues in speech production. Bailly's model is able to generate speech utterances by learning articulatory to audio-visual mappings (Bailly 1997). Finally, Westermann and Miranda's neural network model concentrates on learning couplings between motor representations and their sensory consequences (Westermann and Miranda 2004).

The Task Dynamic model of speech production (Nam et al. 2004; Saltzman and Munhall 1989) draws on the ideas of articulatory phonology (Browman and Goldstein 1986; Browman and Goldstein 1992; Goldstein et al. 2006) and coordinative

structures (Saltzman and Byrd 2000; Saltzman and Kelso 1987). It does not include perception and is not a model of speech acquisition, but it has been influential in this field. It attempts to explain the continuous movement of the speech articulators in terms of abstract, discrete gestural units. Gestures are activated according to a gestural score, in a relationship that is similar to that between the notes played on a musical instrument and a musical score. The movements of the articulators are modeled using a dynamical system, employing critically damped oscillators that behave as point attractors. The input to control production is specified by orthographic transcriptions of speech.

In the rest of this paper we start by discussing the ways in which sensory information is received and responded to by an infant. Next we describe Elija, a computational model of an infant. We then present the stages by which Elija learns first to pronounce and then to pronounce words, and relate the results to Oller's stages of infant vocal development (Oller 2000). Finally we discuss the implications of our model, its relations with previous models, and proposals for future work.

Signal Flows and Interactions with the Environment

Agent Pathways

An infant interacts with his environment via various signal flow paths operating in parallel, some of which are shown in Figure 2 (see Menn et al. (1993) for a fuller analysis). He receives somatosensory feedback from movement of his articulators, from any contacts that they make, from the vibration created by turbulent airflow and from laryngeal vibrations. He is able to hear sounds. He has basic desires and motives that he tries to meet, represented here in terms of "reward." He can also explore, recognize, remember and associate his sensory inputs and motor outputs.

Passive Observation

It is known that sensory systems can be modeled using self-organization from passive observation of the environment (Figure 2B). For example using the statistics of natural inputs it is possible to develop efficient coding strategies that can explain the structures of sensory processing (Barlow 1961; Olshausen and Field 1996). If the infant's auditory input includes ambient speech sounds, this will help develop speech perception (Saffran et al. 1996). However such passive observation alone will not assist development of the motor system, since it does not require its use.

Sensory Consequences of Action

As the infant experiments with his vocal apparatus, he receives internal somatosensory feedback arising from proprioception and, if contact occurs, from touch (Figure 2C). This is informative regarding the kinematic and dynamical properties of the vocal apparatus. In particular, tactile feedback reveals vocal tract configurations which may later become the basis for consonants. Activity of the vocal apparatus can also generate acoustic consequences that pass via the external environment (Figure 2D). The infant can evaluate these actions on the basis of the salience of their sensory consequences, leading to the discovery of potential speech sounds, a process we have previously modeled (Howard and Messum 2007).

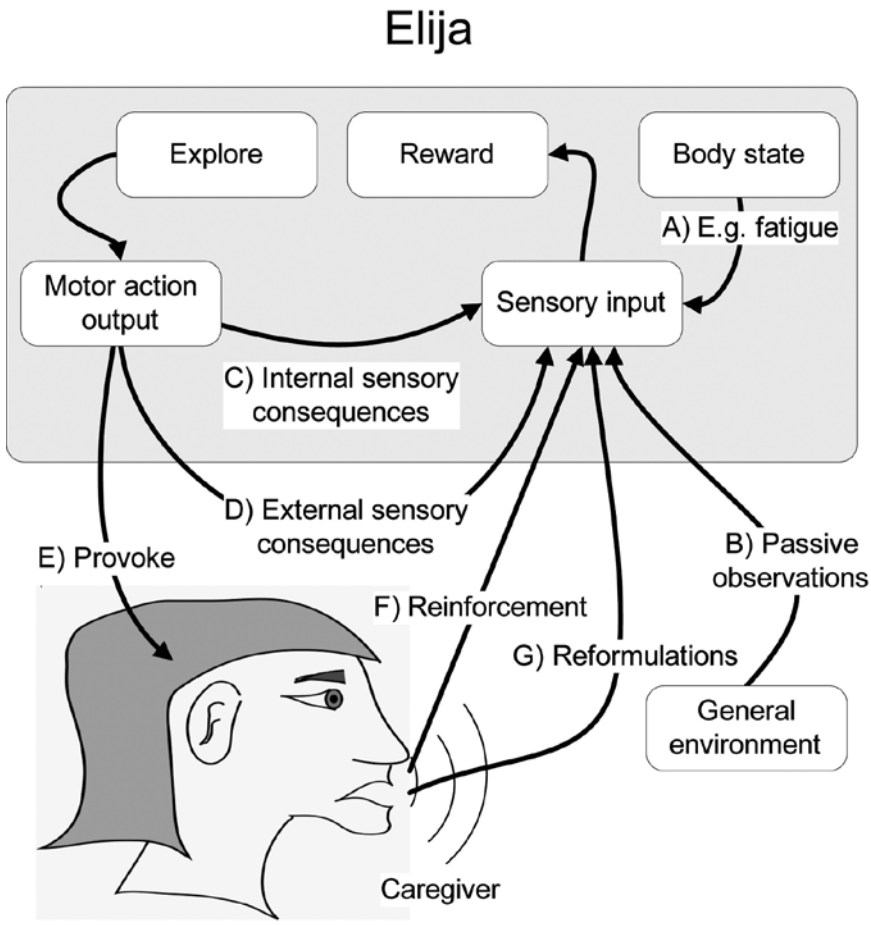

Figure 2 — Infant signal flow pathways. ***A*** The state of the body. ***B*** Using passive observation of the environment an infant can self-organize his sensory systems. ***C*** There is a somatosensory signal flow path within the body arising from motor output fed back to the sensory input. ***D*** There is also a path via the external environment, e.g., the infant hearing his own voice. ***E*** There can also be an external path that includes a caregiver. Because she has well developed phonological perception and production, she can evaluate his utterances in a linguistically appropriate fashion. ***F*** Her response can thus reward certain sounds and ***G*** her reformulations can be associated with his productions.

Response from a Learned Caregiver: Reinforcement and Reformulations

Another signal flow path arises from interaction with a learned caregiver, which is usually the infant's mother (Figure 2 E, F and G). She can evaluate the infant's speech production in terms of the ambient language using her well-developed criteria for speech perception.

During babbling and other vocal play, the infant will produce some sounds that his mother can take to be attempts at linguistic communication. It is normal for caregivers to respond to these vocally or with other forms of encouragement (Newson 1979). This can have several effects. At a simple level it reinforces the infant's original production, encouraging the development of speech sounds (Figure 2 F). Conversely, the absence of a response can be taken as a sign of discouragement.

Among the caregiver's possible vocal responses, we are principally concerned with mimicry and reformulation (Figure 2 G). In mimicry, she produces an acoustically similar utterance. In reformulation, she interprets the infant's utterance within her linguistic system and responds with her equivalent canonical utterance, on the basis of what she has inferred the infant to have said (Otomo 2001).

Both of these responses provide reinforcement, but reformulation enables the infant to connect his vocal action to an acoustic form produced by his mother that need not be acoustically similar. Infants know when they are being imitated (Meltzoff 1999), so he knows that his mother believes her response is equivalent to what he did to provoke it. He can therefore rely upon her judgment to make a strong association between the two events.

Parsing the Input for Reproduction

In the imitation literature, parsing, or "string parsing", is the identification of the sequence of molecular events making up a performance (Byrne 2003). By parsing the caregiver's speech in terms of the speech sounds he has previously learnt, the infant can deduce the actions he must make to reproduce part or all of it using his own vocal actions. This provides a method for more efficient word reproduction than whole word mimicry. That is, recognizing a sequence as being made up from a limited set of speech sounds and then replicating this is more efficient than learning the sound shape of every word in the lexicon discretely. In the same way, it is more efficient to reproduce a written word using a small set of letters than to recreate the whole word shape through drawing.

Using Object Context

Using the speech sound reproduction abilities acquired during the reformulation phase, the infant can now learn the names of objects spoken by the caregiver. For an object within their shared attention, the infant will be able to associate the object, the caregiver's utterance and the sequence of vocal actions he has deduced will correspond to this (Figure 3).

This procedure is likely to involve multiple exchanges, by which the caregiver refines the pronunciation of the infant's labels. If the caregiver likes the infant's production she can signal her approval by congratulation or by simply acting on the meaning she has understood. If she is unhappy with his attempt she can engage him in an iterative loop, in which she repeats the name (possibly with emphasis on a particular element of the pronunciation), inviting him to modify his response. This can continue until she either accepts an attempt or decides that he is unlikely to be able to pronounce the word, and moves on. This procedure further develops the correspondences between caregiver and infant speech sounds, with the shared context providing strong evidence of equivalence to the infant.

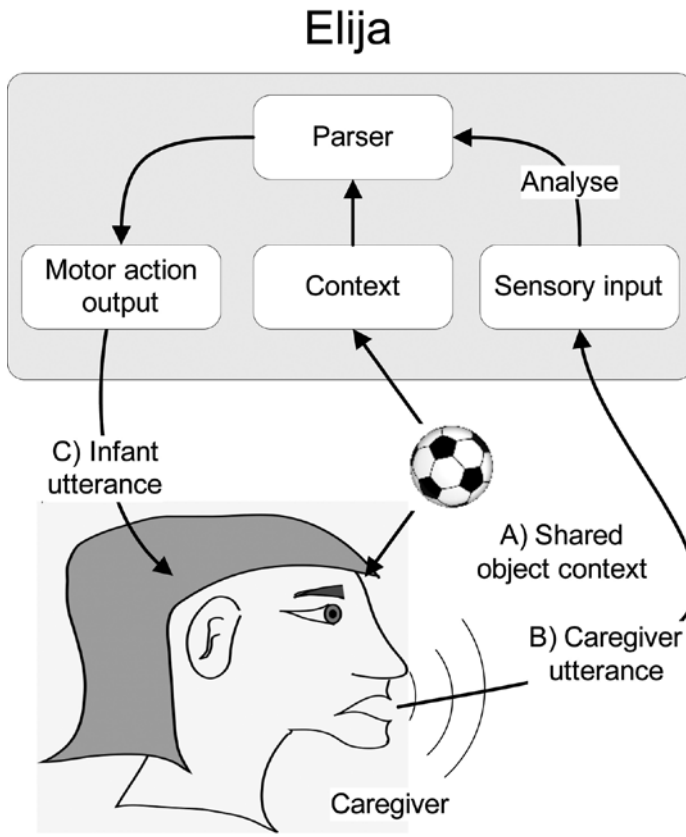

Figure 3 — Learning to pronounce the name of an object. In the presence of an object ***A***, the caregiver pronounces its name. ***B***. Elija analyses the speech signal and parses it on the basis of previous learning to identify a sequence of speech sounds. These have direct associations with vocal actions, and the corresponding sequence of these is generated, resulting in Elija's imitated response. ***C***. The object's context is also associated with the speech sound/vocal action sequence, which can later trigger recall.

Methods

We now describe the design philosophy behind Elija. Then we describe his vocal apparatus, motor system, reward, and memory modules. These are depicted in Figure 4.

Non-imitative Mechanism for Learning Simple Sound Pronunciation

Here we model the development of pronunciation through an agent, Elija, who learns by running experiments on an environment that includes a linguistically

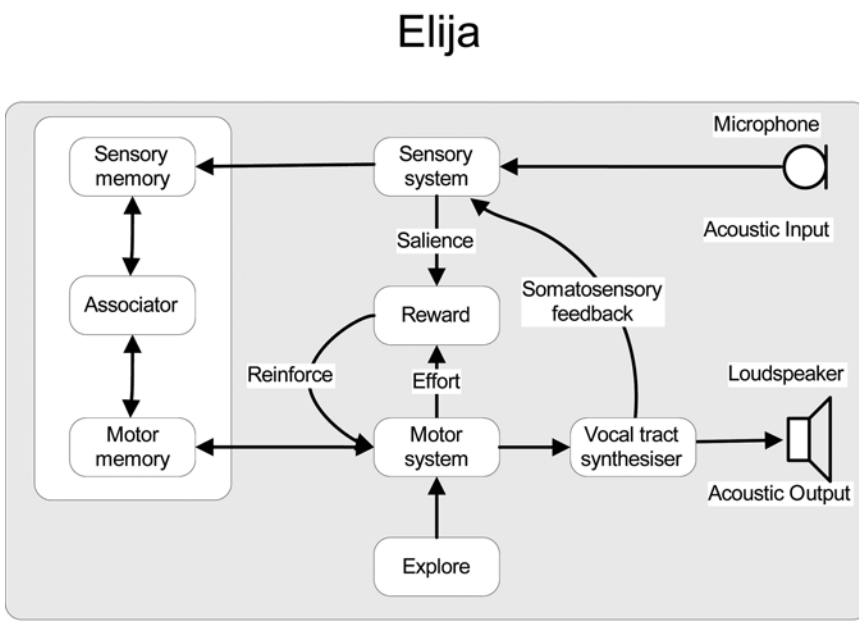

Figure 4 — Inside Elija. Elija listens to his environment and affects it using speech output. The vocal tract is driven by a motor control module which also computes the effort involved in generating a vocal action. The vocal tract generates internal somatosensory feedback from touch. The motor action may arise from motor exploration and can be stored, and later recalled, from motor memory. When an action leads to sensory consequences (e.g., auditory salience or somatosensory feedback) these are evaluated by the reward module. The reward can be used to improve the action using gradient accent, or to reinforce it. Similarly, sensory input can be analyzed in terms of salience and is also recorded in sensory memory. Associations can form between sensory and motor memories, linking action with their sensory consequences.

expert caregiver. Her natural inclinations lead her to respond in ways that assist his development. We model Elija's speech production as initially developing using rewarded exploration of his vocal tract. His own evaluation of the sensory consequences of his actions leads to the discovery of some vocal actions whose acoustic output then attracts the attention of his caregiver. As in real life, her response to sounds similar to those in the ambient language will often be an "imitation"; either a mimicked or reformulated version of his output. This reinforces Elija's actions and thereby biases his production toward the sounds of the ambient language. Importantly, these and all other linguistic judgments of Elija's speech production are made by the caregiver, not by Elija himself.

Even when the caregiver's response is a reformulation rather than a mimicked version of his output, Elija associates his productions with her adult form, giving him vocal effectivities which generate a set of two-way speech sound correspondences. He will later be able to use these to parse adult speech and generate output that is equivalent to it. This is how he will learn to pronounce words: by firstly

identifying the sequence of speech sounds they contain and then reproducing this sequence with his corresponding motor patterns.

We note that although the caregiver imitates Elija during this process, Elija himself does not imitate the caregiver, in contrast to the assumption made in conventional accounts. For this reason we describe our approach as *"non-imitative"*. Our work was inspired by the observations of child speech development made by Gattegno within his descriptive framework of human learning (Gattegno 1973; 1985; 1987).

Modeling the Vocal Apparatus with an Articulatory Synthesizer

To generate acoustic output, Elija uses an articulatory speech synthesizer. A good model of an infant vocal tract is important to effectively model speech development for several reasons. Firstly, phonology can then develop directly from the basic biomechanical and aerodynamic properties of the vocal apparatus (Lindblom 1999). Secondly, proprioception and touch sensation provide information about distinctive articulator configurations, e.g., touching the tip of the tongue on the back of the teeth or closing the lips, aiding the discovery of those configurations that will be used in the generation of speech sounds. Thirdly, the work of Saltzman and his colleagues (Saltzman and Kelso 1987; Saltzman and Munhall 1992) points to the importance played by the dynamics of the vocal apparatus. Fourthly, a synthesizer that sounds like a real infant will help to provoke natural responses from caregivers.

Elija's vocal tract is based on an implementation of the Maeda articulatory synthesizer (Maeda 1990) and a voice source based on the LF model (Fant et al. 1985)[2]. In all there are 7 articulatory parameters used to specify vocal tract profile: jaw position, tongue dorsum position, tongue dorsum shape, tongue apex position, lip height (aperture), lip protrusion, and larynx height. Our implementation of the LF voice source makes use of two parameters: glottal area and fundamental frequency. The VTCALCS implementation of the Maeda synthesizer (see Acknowledgments) also includes a velopharyngeal port to control nasality. These control parameters are shown on the example trajectories in Figure 5B.

The Maeda vocal tract profile determines an equivalent digital filter which is applied to the excitation from the voice and noise sources, thus leading to an appropriately filtered acoustic output signal. Fricatives are simulated in the model by injecting noise at locations in the vocal tract where turbulent airflow is predicted. In our implementation, the synthesizer operated at an output sampling rate of 24 kHz. To approximate the vocal tract of an infant, the physical dimensions of the original Maeda vocal tract were scaled down by a factor of 0.8 from the default values used for an adult female and the midrange of the fundamental frequency was shifted from 210 Hz to 400 Hz. There are other differences between adult and infant vocal tracts. For example, this scaling does not reflect the real differences in the size of the pharynx (Boë et al. 2007). However, for our study an exact representation of an infant vocal tract was not necessary because Elija does not attempt auditory matching between his infant speech and that of the caregiver. He only matches the caregiver's current speech utterances to her past speech utterances.

The Maeda synthesizer was enhanced to generate contact information, which represents touch feedback arising from the speech production apparatus. The Maeda model operates by first computing the cross sectional area of the vocal tract, which

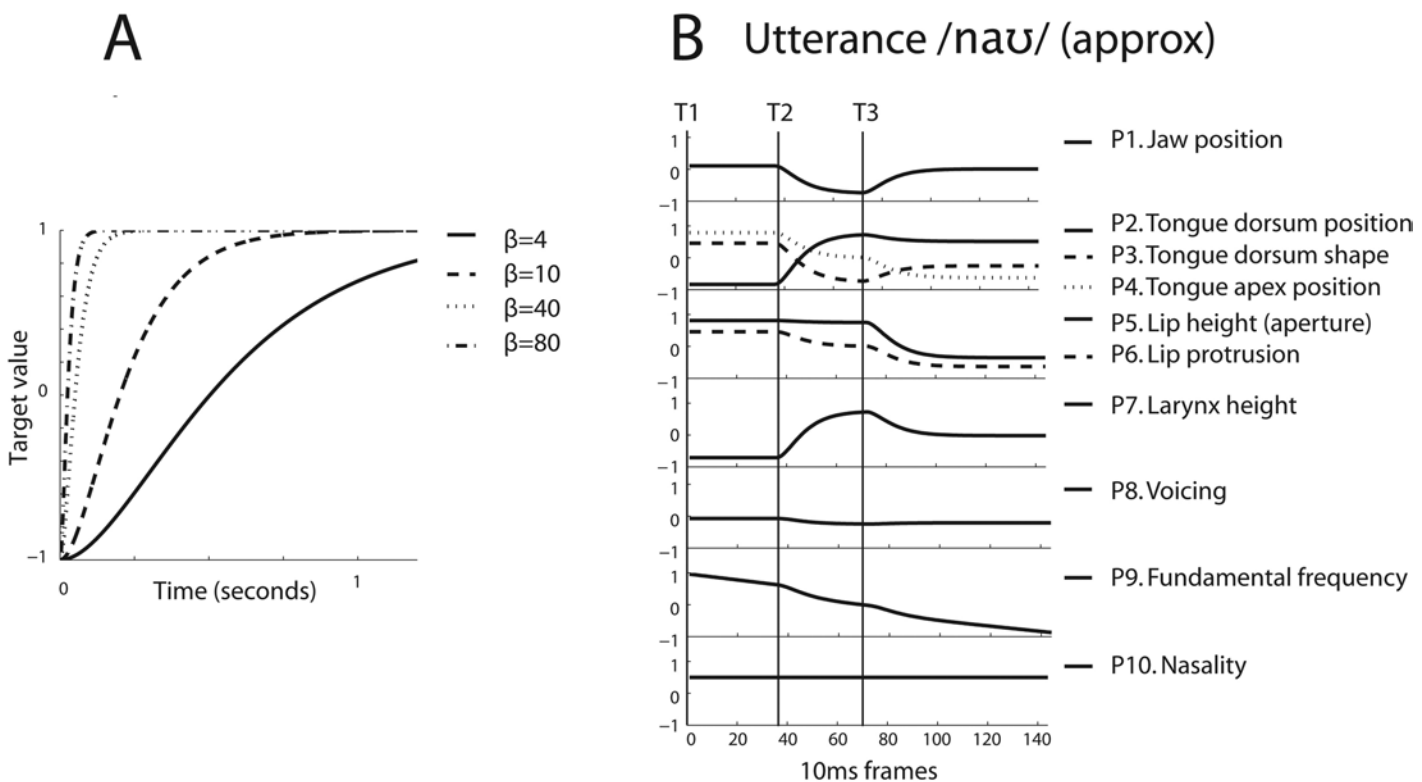

Figure 5 — Examples of vocal action trajectories. A vocal action is defined by starting and ending articulator target locations and the time between them. ***A*** shows an example trajectory generated by a single target value moving from -1 to 1 at time t = 0. The path is defined by a critically damped 2nd order system and is parameterized by the value of β. In the plot the effect of changing the β value between 4 and 80 is shown, which is to increase the speed of the transitions. ***B*** shows the vocal tract control parameter trajectories for a speech utterance generated by Elija, involving three targets. The initial target at T1 results in constant trajectories until the next target occurs at time T2, at which point the trajectories move toward this second target. At time T3 a third target is introduced and the trajectories then move toward this final target.

depends on the values of the control parameters. At points where the cross-sectional area reaches zero, contact has occurred.

In its current form, the articulatory synthesizer generates unnatural acoustic artifacts when the velopharyngeal port is open. To circumvent this deficiency, the sound discovery stages of the experiments were carried out with nasality deactivated (i.e., with the velopharyngeal port closed). To include nasals in the final reformulation repertoire, nasality was only included during the recombination stage for a set of CVs (see the Methods section of the Integrative Stage Experiment).

The Maeda synthesizer was implemented in C++ and all other analyses were written in Matlab.

Modeling a Vocal Motor Action

We use the term motor pattern for the abstract representation of a movement of the vocal tract, for which we use the term vocal action. We model a basic motor pattern as a sequence of up to three vocal tract target positions. Thus motor patterns are defined in terms of articulator position vectors, which specify the 10 vocal tract control parameters. In addition, the time for which a target is maintained is specified. The simplest motor pattern, to produce a vowel V, consists of only a single target vector with 11 elements. More complex motor patterns, such as those producing a CV, VC, or CVV, require two or three target vectors respectively, and contain 22 or 33 elements in total.

A motor pattern generates a vocal action in which the trajectories between targets are determined by articulator dynamics modeled by means of gestural controllers. Here we adopt the approach of Markey by assuming 2nd order dynamics that are critically damped, leading to movements toward targets without overshoot (Markey 1994). The corresponding equation for the trajectories of an articulator is given by:

$$x(t) = x_{endpoint} + \left((x_{startpoint} - x_{endpoint}) + (x_{startpoint} - x_{endpoint})t\beta + v_0 t\right)e^{-\beta t}$$

where

x(t) is the articulator position at time t.

$x_{startpoint}$ is the starting articulator position

$x_{endpoint}$ is the ending articulator target position

v_0 is the initial velocity

the constant β is given by

$$\beta^2 = k / m$$

where k is the spring constant and m the associated mass of the dynamical system. Here we assume $v_0 = 0$ and the constant β is set to 40 to match the range of speeds of human articulators. The effect of β is to change the speed at which the articulators move toward their target positions. Large values of β lead to a rapid movement toward the target, and Figure 5A shows the effect β has on the transition from a target value of -1 to a target value of 1 for a single articulator. An example of the trajectories resulting from a three target CVV motor pattern for all 10 articulator control parameters is shown in Figure 5B.

Because Elija does not learn the articulatory control to move between targets, we use the term "vocal action" to describe his vocal tract movements, rather than using the term "vocal motor scheme" (VMS) (McCune and Vihman 1987). The concepts are similar, but we need to distinguish the two, since low level motor learning is clearly an important part of VMS development in real infants.

We use a simple model of declination to modulate the fundamental frequency, reducing its control parameter by 0.75 each second. The inclusion of this frequency modulation makes the generated utterances sound more natural.

Optimization: the Objective Function

Elija uses rewarded exploration of the vocal tract parameters to find motor patterns that generate vocal actions. This discovery process is formulated as an optimization problem. Optimization is a computational technique that can find the set of parameters of a function that specify its maximum (or minimum) value. Simple gradient ascent (hill climbing) is an iterative process, in which steps are taken in the direction of the gradient. In Newton's method (also known as the Newton–Raphson method), the estimation of the steps needed makes use of the curvature of the objective function. This involves computing its second derivate, or Hessian. For computational reasons, quasi-Newton optimization algorithms are often used in practice, which avoids directly computing such second derivates. In our experiments the parameters to be optimized are those which define the motor patterns, and we use quasi-Newton gradient ascent to find values which maximize their associated objective function or "reward", as described below.

Computing Reward

In our model, the objective function, or reward R, is defined in terms of the weighted sum of several components, illustrated in Figure 6. Typical signals involved in reward generation during the production of a simple speech utterance are shown in Figure 7. The objective function is defined as sensory salience of the current motor pattern, plus its motor diversity, minus the effort involved in its generation. That is

$$R = \sum (Salience + Diversity - Effort)$$

Three sensory consequences of a vocal action—the acoustic power, the acoustic spectral balance and the sensory feedback from touch—make positive contributions. Specifically, we compute a weighted sum of speech power, ratio of low to high frequency power (above and below 6 kHz), ratio of high to low frequency power (above and below 6 kHz) and high pass filtered touch contact (frequency cut-off = 1Hz). Second order Butterworth filters were used to implement all the low and high pass filters. We compute salience as:

$$Salience = W_{pa}.Power_{acoustic} + W_t.Touch + W_{pHFLF}.Power_{HF/LF} + W_{pLFHF}.Power_{LF/HF}$$

W_{pa} represents the weighting term for acoustic power

W_t represents the weighting term for touch

W_{pHFLF} represents the weighting term for the ratio of high frequency power to low frequency power,

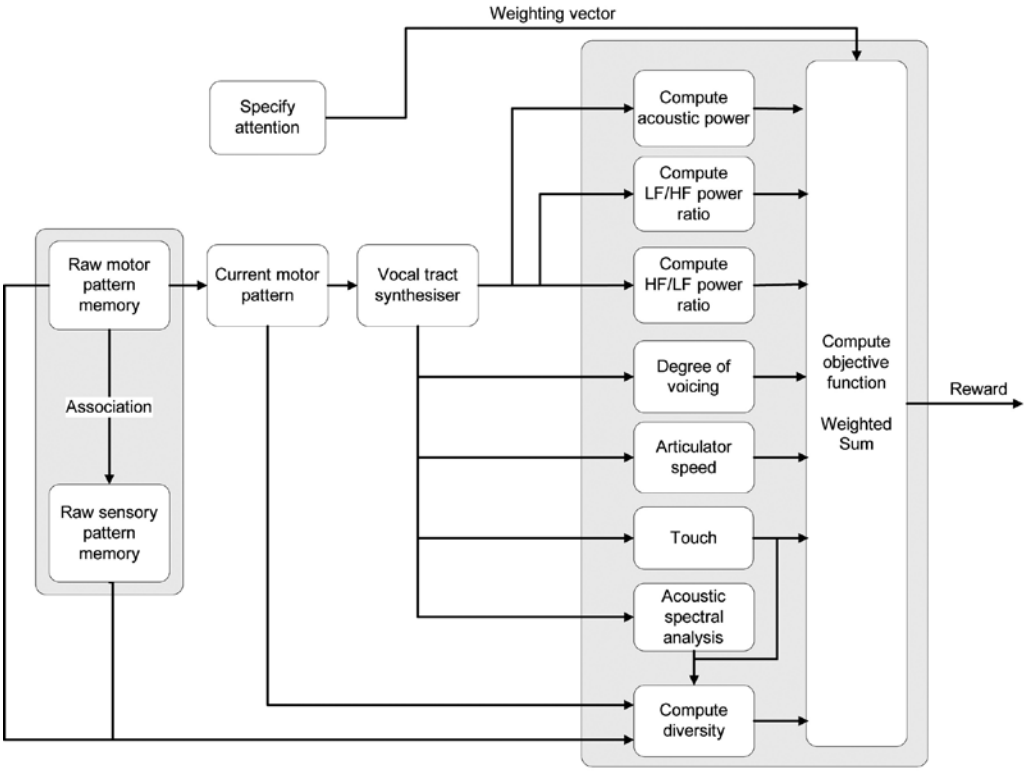

Figure 6 — The computation of reward. The current motor pattern determines the vocal tract configuration and thus affects acoustic and somatosensory output. The auditory consequences from the vocal tract synthesizer are evaluated in terms of acoustic power and spectral balance (the ratio of LF to HF power and the ratio of HF to LF power). Touch arising from vocal tract closure is also calculated. The degrees of voicing and articulator movement are used to estimate effort. Vocal tract closure is used to estimate salience from touch. A diversity measure is computed to estimate how different the current motor pattern and its acoustic and tactile sensory consequences are from the corresponding values for all previous discovered patterns. A weighted sum of these quantities is used to compute overall reward for the current motor pattern (which corresponds to the objective function used in the optimization procedure).

W_{pLFHF} represents the weighting term for the ratio of low frequency power to high frequency power.

The individual terms for acoustic power, touch and spectral balance are computed by averaging the time waveforms for these quantities over the length of each vocal action.

A term is introduced into the reward function using a diversity mechanism which rewards the current motor pattern on the basis of its distance in motor and sensory spaces from the nearest previously discovered motor patterns. This encourages Elija to explore previously unexplored parts of motor pattern space, implementing a simple form of active learning (Mackay 1992). We compute diversity as:

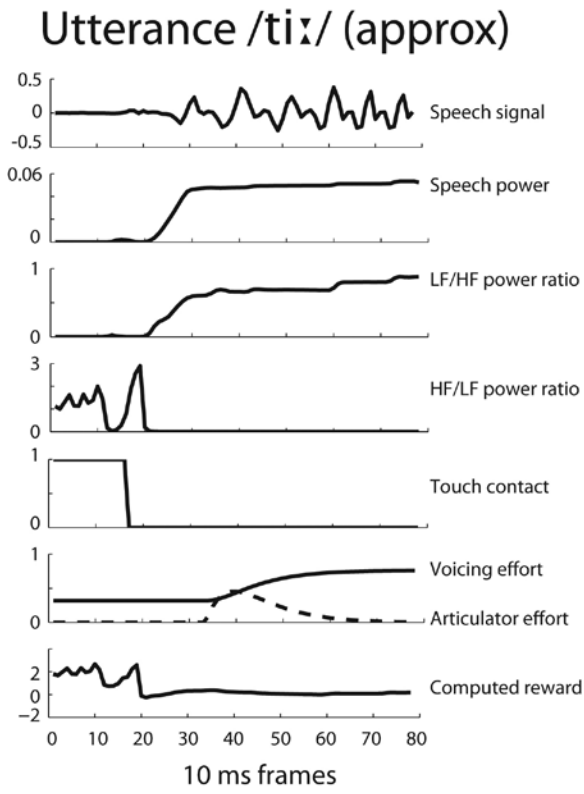

Figure 7 — Reward signals for a speech utterance generated by Elija. The plot shows time traces for the speech signal, its acoustic power, HF/LF power ratio, LF/HF power ratio, touch contact, voicing and articulator effort and the corresponding computed overall reward.

$$Diversity = W_{mpd}.motorpatternDiversity + W_{ppd}.tactileDiversity + W_{sd}.sensoryDiversity$$

where

W_{mpd} represents the weighting term for motor diversity

W_{td} represents the weighting term for tactile diversity

W_{sd} represents the weighting term for sensory diversity

and

$$motorpatternDiversity = \min_{N} \left| currentMotorPattern - existingMotorPattern_{N} \right|$$

$$tactileDiversity = \min_{N} \left| currentTactileConsequences - existingTactileConsequences_{N} \right|$$

$$sensoryDiversity = \min_{N} \left| currentSensoryConsequences - existingSensoryConsequences_{N} \right|$$

where the difference from the current motor pattern and its tactile and acoustic sensory consequences are computed for each of the N motor pattern and sensory consequences that have already been discovered.

The effort required to make a vocal action makes a negative contribution to reward, determined by a combination of loudness and the cost of movement. The latter was calculated as a weighted sum of articulator speed, where jaw movement was made more expensive than other movements. Thus effort is given by:

$$Effort = W_{ae}.ArticulatorEffort + W_{ve}.VoicingEffort$$

where

W_{ae} represents the weighting term for articulator effort and

W_{ve} represents the weighting term for voicing effort

Elija can selectively focus attention on different aspects of sensory feedback by changing the relative contribution to the individual terms in reward using the weighting vector W. Clearly a zero valued element would result in the corresponding quantity being excluded from the optimization procedure. The weights were all set to the range of 0–10. Using different weightings leads to the discovery of different types of speech sound. For example, attending to touch favors configurations where the lips are closed or the tongue touches the roof of the mouth. This attentional set is useful for discovering plosives. Attending to steady state acoustic output with power at lower frequencies favors configurations that lead to vocalic sound production. Attending to acoustic output with a dominant high frequency component favors the discovery of fricatives. This mechanism corresponds to Oller's concept of signal analysis (Oller 2000), in which an infant attends to different aspects of the sensory consequences of his actions.

Initial Discovery of Sounds

To discover the motor patterns that generate sounds that an infant would find of interest as modeled by our reward function, a quasi-Newton optimization algorithm was used, as implemented by the Matlab function ***fmincon***. This function attempts to find a minimum (or a maximum if the sign of the reward term is flipped) of a scalar function of several variables. The optimization was constrained to find control parameters within their valid range. Figure 8 illustrates the optimization process within Elija. Optimization was begun from a random starting point and was run for 3 iterations. Further iterations did not improve the quality of the discovered sounds. The optimization for the motor patterns was used to discover Vs and CVs and ran on a PC for about 50 hr in total.

In the first experiment, motor patterns were discovered in the absence of caregiver interaction. Although it would have been possible for Elija to generate an acoustic output which he then analyzed by listening to himself using a microphone (like a real infant listening to his own babble), we used a direct analysis on the output of the articulatory synthesizer waveform. This enabled the simulation to run several times faster than real time.

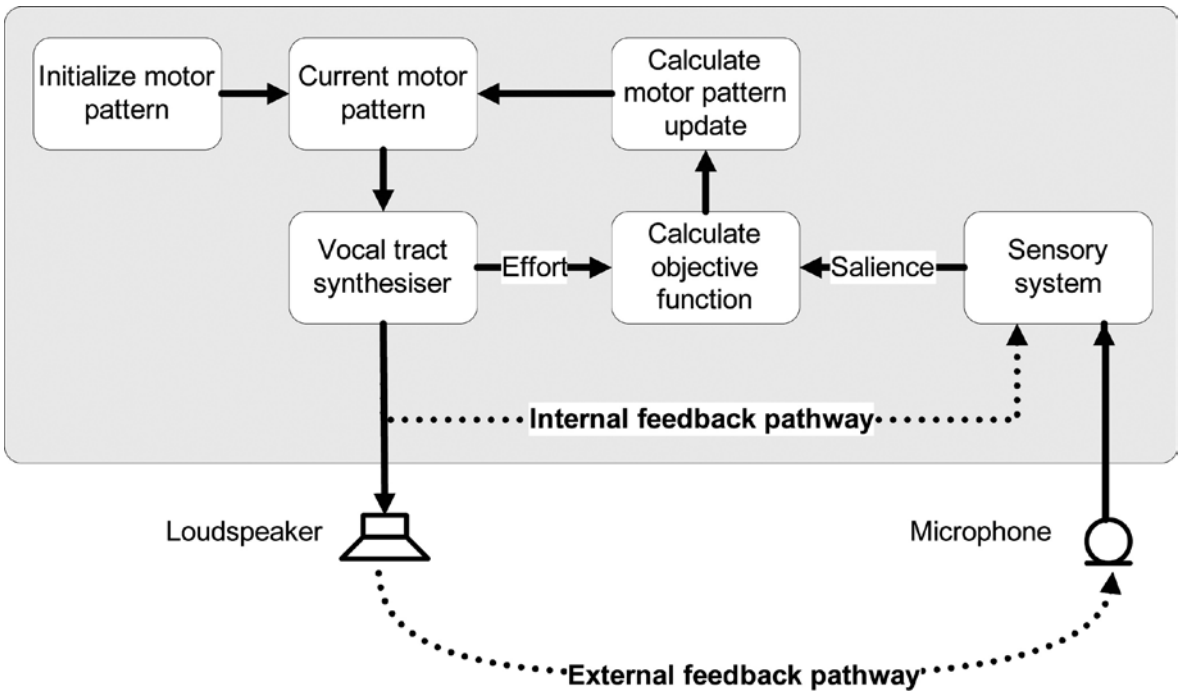

Figure 8 — Optimization procedure used to discover motor patterns. A motor pattern is first randomly initialized and used to drive the synthesizer. The acoustic and somatosensory output is then analyzed by the sensory system to compute salience. This is used, together with effort, to compute the objective function (reward) for the motor pattern, as shown in Figure 6. The optimization procedure computes the changes in the motor pattern needed to improve it using gradient ascent. These changes are then applied to the motor pattern. The process is repeated until the maximum number of iterations have been reached.

Consolidation of Motor Patterns

We deliberately limited the number of motor patterns that were represented in memory by partitioning the data set into a limited number of clusters, and then only retaining the most central exemplar in each. This procedure removes redundancy from the repertoire of vocal actions without sacrificing diversity. An example is given in Figure 9. This is important because a caregiver will interact with Elija to reinforce and reformulate a wide variety of possible speech sounds. If the set of motor patterns was highly redundant this would rapidly lead to a combinatorial explosion of sounds and unnecessarily increase the number of interactions required in our experiments. By removing redundancy, we limited the length of interaction with a caregiver in the experiment to about 8 hr in total.

Categorization of Motor Patterns

Elija could potentially categorize his motor patterns using any of three datasets associated with them. On the basis of:

- direct similarity of the motor patterns in vocal tract control parameter space.
- similarity of the resulting acoustic outputs, computed using the DTW algorithm described below.
- similarity of the caregiver's corresponding acoustic outputs (usually reformulations) again using the DTW algorithm.

This is illustrated in Figure 10. Initially, the first two datasets are the only means by which Elija can cluster his motor patterns. When the acoustic consequences of his vocal actions have been reformulated by a caregiver, this third dataset can also be used.

Categorization based on each dataset will lead to different results. For example, on the basis of similarity in articulator space, a vocal action that generates a fricative and one that generates an approximant may fall into the same category, because only a small change in articulator position differentiates them, whereas an acoustic categorization would be likely to separate them.

The categories that Elija will find will not necessarily reflect the phonological structure of the ambient language. However, such a result is more likely to occur in the third case, i.e., by clustering caregiver reformulations, because the caregiver is linguistically competent. Her productions within a category will therefore be more consistent than his (i.e., they will show low intertoken variability) and her reformulations will more correctly and reliably express the phonological contrasts of the ambient language. Of course, phonological boundaries will only be definitively learned when semantic contrasts give the infant direct evidence of their locations. Currently we do not implement this procedure in our experiments.

Implementation of Pattern Clustering

As described above, after Elija has acquired a set of motor patterns in an experimental run, he uses clustering to consolidate them. Elija can consolidate speech utterances either on the basis of their motor properties or acoustic properties. For the latter, the utterance is analyzed using a 21 channel filterbank described in the section Implementing Utterance Recognition based on the channel vocoder (Gold and Rader 1967) .

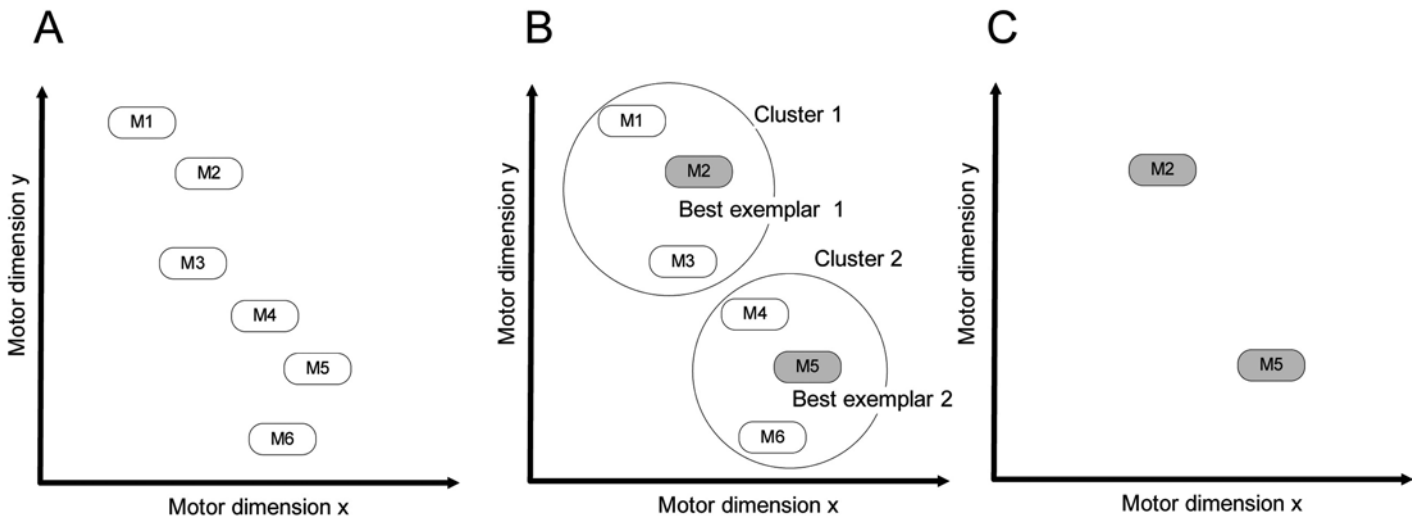

Figure 9 — Motor pattern consolidation. ***A*** Example of six motor patterns, plotted here in only two-dimensions for clarity. ***B*** The K-means algorithm is used to identify clusters of patterns. In this case two clusters are found. This process also identifies the best exemplar in each cluster. ***C*** The best exemplars are retained and all other patterns are discarded, thus reducing the size of the dataset while maintaining variety.

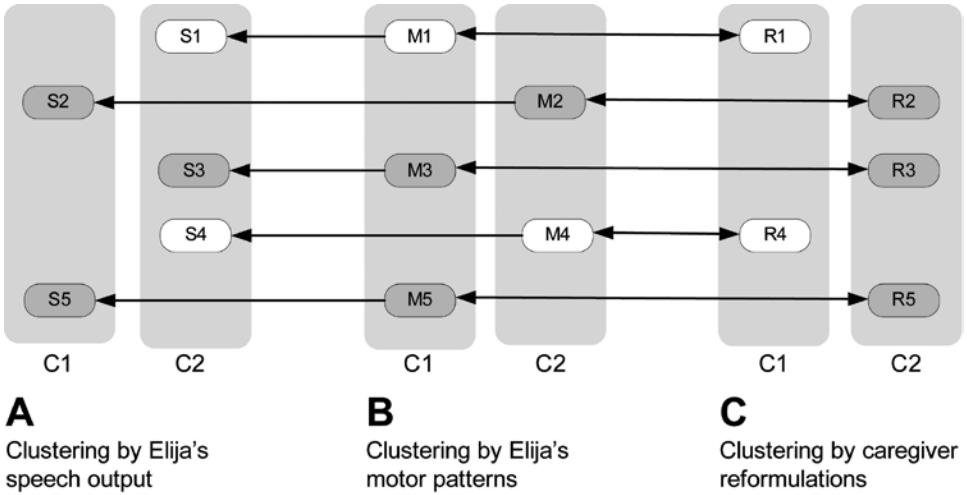

Figure 10 — Three alternative criteria for categorization. Each of Elija's motor patterns B are associated with his speech output A and the corresponding caregiver reformulations C. Tokens in each data set can be categorized by similarity, but the categories obtained from the caregiver reformulations are preferred for two reasons. Firstly, the categories are more distinct and the tokens within them more similar because the caregiver is an expert speaker of L1. Secondly, the infant is aware that the caregiver is setting the rules of the game; her judgments are more consistent and authoritative than his and she will not be influenced by any counter proposals that he makes. In the diagram, tokens are shaded according to the reformulation categories and, as shown, these do not always coincide with Elija's speech output or vocal action categories.

Motor patterns are clustered directly using a standard K-means algorithm, as available in Matlab. For acoustic clustering of utterances, which will vary in length (different utterances from Elija will typically have different time durations, as will the caregiver's utterances), the standard K-means algorithm is not appropriate, since it requires a fixed pattern length (see the K-means implementation in NETLAB for further details (Nabney 2004)). Therefore we perform clustering using a modified version of the standard algorithm, which we call DTW K-means. This is similar to the standard K-means algorithm except that 1) it represents a cluster using the best exemplar rather than its mean and 2) it uses a DTW distance metric. It operates in two steps. Let us assume we have already decided on the number of clusters, K. First the algorithm randomly chooses a best exemplar pattern to define each of the K clusters. It then begins an iterative loop. It processes each utterance in the dataset, assigning them to their nearest cluster exemplar. In standard K-means, a Euclidian distance metric is often used to directly compute distance. However, in the DTW K-means algorithm, dynamic time warping is used to determine the distance between utterances (as described in the section on Implementing Utterance Recognition). After all utterances have been assigned to a cluster, we then use all the utterances within each cluster to recompute the best exemplar, which is defined as the utterance that is on average closest to all other utterances. It is found

simply by adding up the distances to all other utterances for each utterance in turn, and choosing the utterance with the minimum summed distance. Then we once again assign each utterance to the closest exemplar. The assignment/recomputation process is repeated until no further change of assignment occurs.

Motor and Sensory Memory

The organization of Elija's motor and sensory memory is shown in Figure 11.

As motor patterns are discovered, they are recorded in Elija's current motor memory. When Elija uses a vocal action to generate a speech-like sound to which his caregiver responds, her corresponding acoustic response is retained in current sensory memory. In addition, an association is formed between these motor and

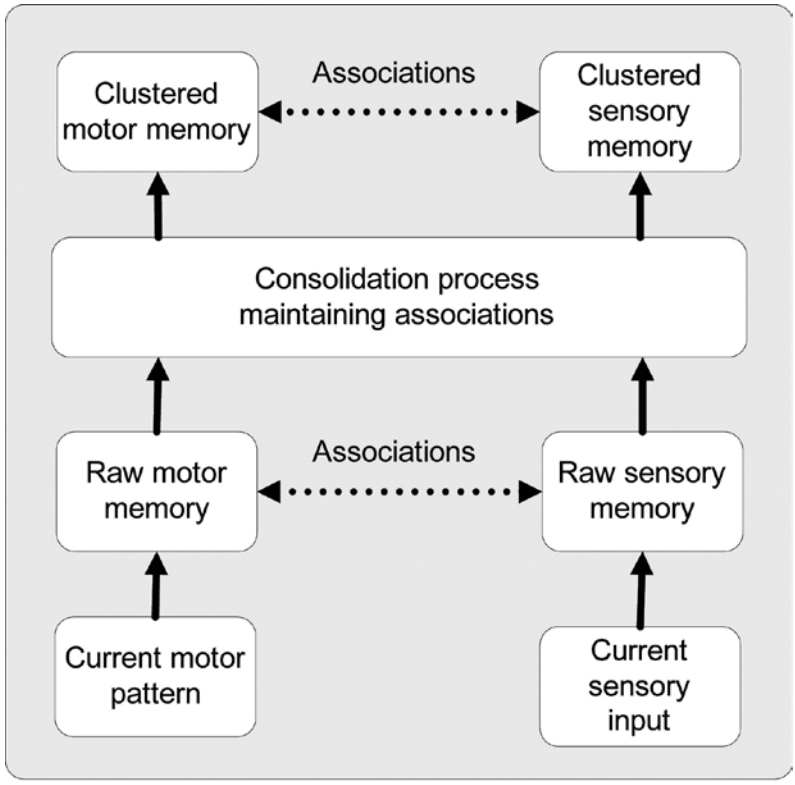

Figure 11 — Organization of Elija's motor and sensory memory. Newly discovered motor patterns and associated responses from the caregiver are shown at the bottom. These are recorded in raw motor and sensory memory. During a consolidation phase, clustering is performed to reduce redundancy in the motor patterns. This consolidation process maintains the associations between the motor and sensory memories. Finally clustered motor and sensory memories are recorded.

sensory patterns, which is also retained during clustering. Motor patterns which generate no response are discarded.

Reinforcement and Recombination of Motor Patterns

After simple V and CV motor patterns discovered by the optimization procedure are consolidated, they are played to the caregiver and reinforced (retained) if the caregiver responds acoustically. Otherwise they are discarded.

The reinforced motor patterns are used as building blocks for other motor patterns. By decomposing them into C and V targets and then recombining them, the repertoire of plausible CVs was expanded. This is because some Cs only occurred with a limited number of Vs, and vice versa. This procedure corresponds to the activity described by Oller as segmentation, at the end of his Integrative stage (Oller 2000). Similarly, using this procedure, more complex motor patterns were added to Elija's repertoire, such as VC, CVV and VV.

Implementing Utterance Recognition

Elija has no a priori phonetic or phonological knowledge but he must learn to discriminate sounds in his environment.

To recognize speech it is usual to first extract features of the speech signal. Many representations are possible, ranging from spectrograms to Mel-frequency cepstral coefficients (Mermelstein 1976). Here we employ an auditory filterbank front-end based on a 21 channel vocoder, which generates an output frame every 16 ms. Our analysis incorporates elementary amplitude normalization by employing a logarithmic scale to encode intensity, from which the total power is subtracted.

We implemented a recognition capability using a template-based dynamic time warping (DTW) algorithm. This algorithm aligns and locally warps the input speech utterances to account for differences in timing between them. It compares each frame in the input data with the corresponding ones in a set of reference templates that comprise the vocabulary of the recognizer, and returns a metric of similarity for each. By using dynamic programming (DP), this procedure can be computed efficiently. DP has formed the basis for many speech recognition systems (Sakoe and Chiba 1978). The implementation of the DP used in our experiments was due to Ellis (Ellis 2003). Although this algorithm was originally used for music recognition (Turetsky and Ellis 2003), it is equally suitable for speech recognition since the underlying DP algorithm required is the same in both cases.

As mentioned above, the DTW algorithm is also used as the similarity metric in the DTW K-means algorithm.

Recognizing Caregiver Sounds

A two-stage procedure was used to recognize caregiver reformulations. This firstly identifies the category of an input sound produced by the caregiver based on acoustic similarity and then the best matching sound within that category. This procedure required the caregiver reformulations to be partitioned into 100 clusters, a value chosen by experimentation. This was performed using the DTW K-means algorithm described above. The associations with vocal motor patterns were maintained during clustering, so that identification of a reformulation also identified Elija's corresponding motor pattern. Figure 12 illustrates this process.

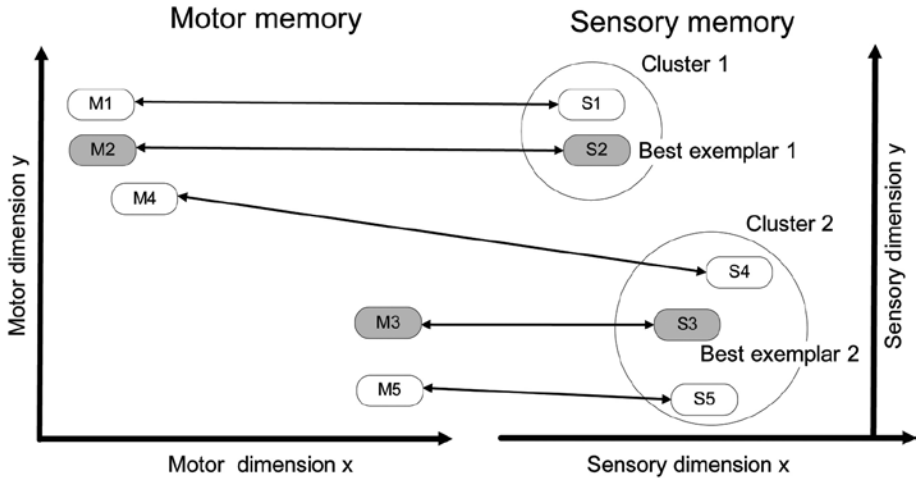

Figure 12 — Clustering process used to implement the two-stage DTW speech recognizer. Examples are limited to two dimensions for clarity. Speech reformulations S1—S5 are clustered into two groups. The best exemplar in each cluster is also identified. Notice that the links to the associated motor patterns are maintained during this process.

During sound recognition, the DTW recognizer first uses the best exemplars in each cluster as the templates to identify the sound category. The recognizer then uses the members of the best category as templates, to identify the best specific matching sound. Figure 13 shows a schematic of this process. This step is also valuable because it identifies Elijah's set of corresponding motor actions, which can then be offered as suggestions during the later labeling phase (see the later Object Labeling Experiment).

Experiments

A single subject (the author ISH) played the role of caregiver. For simplicity, we modeled developmental stages in series, rather than as the parallel and overlapping processes that occur in a real infant.

In all interactions, the caregiver imagined that Elija was a real infant and responded accordingly to his output. This usually meant that the caregiver reformulated any utterance that sounded like a speech sound or word from Southern British English and ignored other utterances. Such reformulations are typical interactions observed between young infants and their caregivers (Pawlby 1977). In the final object labeling experiment, the caregiver spoke the name of an object to Elija, who responded with an attempted imitation. Again, if the caregiver liked Elija's response it could be accepted, or rejected if not.

Elija developed the ability to pronounce and then pronounce words in discrete experiments which correspond to Oller's five stages of protophone development in real infants (Oller 2000; Oller et al. 1999): phonation, primitive articulation, expansion, and the canonical and integrative stages. Because the articulatory synthesizer was unable to reliably generate nasal sounds, these were not initially

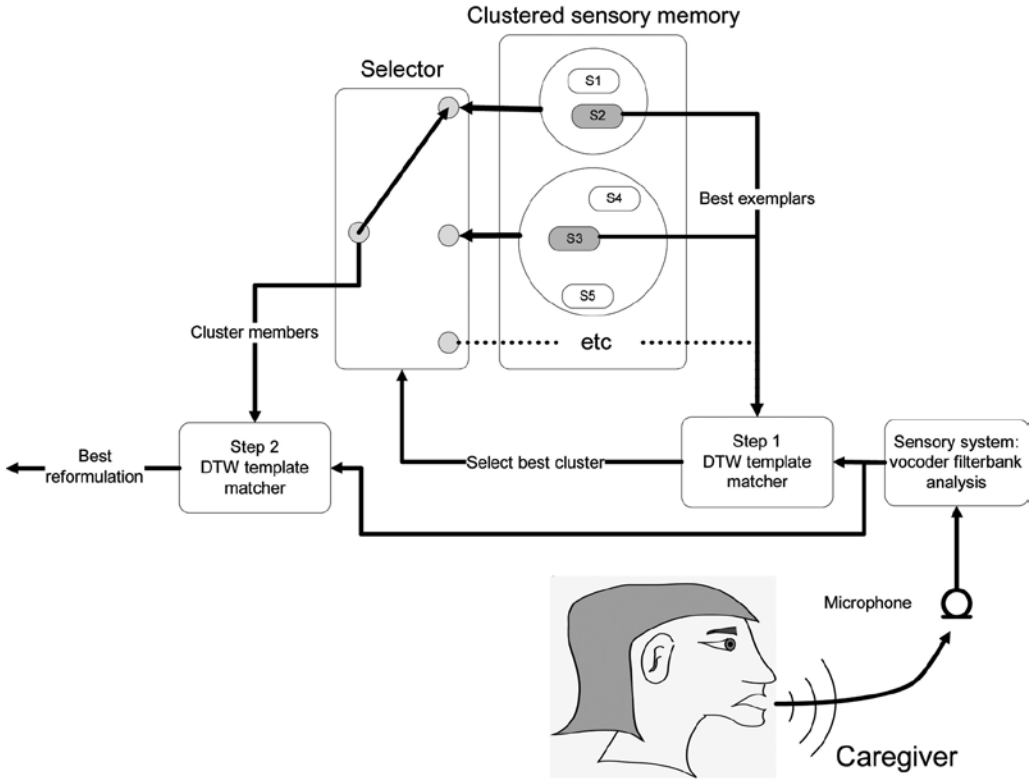

Figure 13 — Two-level DTW utterance recognition. Template based recognition of the input utterance spoken by the caregiver is performed to identify the best matching caregiver reformulations. The reformulations have already been clustered and the best exemplar in the cluster has already been identified (shown in Figure 12). First, the input utterance is matched against all the best cluster exemplars. This identifies the cluster that best matches the input. A second match is then made to the reformulations within that cluster, leading to the identification of the overall best matching reformulation. This process requires less computation than directly matching against the entire set of reformulations.

generated. Nasalization was only introduced to a subset of CV patterns in the final recombination stage.

Examples of the results are available in online supplementary material, which provides WAV files of Elija's and the caregiver's utterances. They are identified by the names of the experimental stages they relate to. The supplementary material is available at:

www.ianhoward.info/MCSupplementaryMaterial.ppt

Phonation Stage Experiment

Method. Oller's phonation stage describes an infant's development of the ability to control his vocal folds and breathing for voicing. This leads to the production of quasi-vowels.

Elija begins learning sounds by discovering vocal actions that generate steady-state vowels. Initial values for 2000 single target motor patterns were created by random sampling from a uniform distribution of permitted values. Each motor pattern was then optimized on the basis of reward, by encouraging overall acoustic power while penalizing high frequency acoustic power and touch. Vocal actions which generated more high frequency than low frequency acoustic output were discarded, leaving 800 potentially useful vowel-like sounds. These motor patterns were clustered on the basis of their acoustic sensory consequences and the best exemplar in each was identified and retained. We found that using 50 clusters maintained a good balance between removing redundancy and repetition in the repertoire while maintaining variety.

Results. Elija found a wide variety of vowel qualities, including some tokens which were unlike any used in English.

Twenty-two of the 50 clustered "vowels" that Elija found could be readily recognized as tense vowels in English: four as /i:/, five as /a:/, seven as /u:/ and six as /ɜ:/. There were no close counterparts to /ɔ:/. 26 tokens either had some similarities to the tense vowels, might be taken as lax vowels if shortened and embedded in a word, might be taken as diphthongs in some contexts, or might not exist in English but might resemble vowels in other languages. Two tokens had qualities that made them unlike natural vowels.

Primitive Articulation, Expansion and Canonical Stages Experiment

Method. During Oller's primitive articulation stage, the first limited articulations are made simultaneously with phonation. During this "gooing", the vocal tract is often closed by the tongue at the rear of the mouth. In his expansion stage, infants articulate from a closed vocal tract to a postured full vowel while producing normal phonation. In this canonical stage, infants produce well-formed syllables. We model all of these stages in one experiment because Elija's optimization discovers sounds sequentially rather than in parallel.

To find CV structures, Elija added a preceding target to the 50 previously discovered vowel motor patterns. These were used to "seed" the V target in 1000 CV patterns. The C targets were randomly initialized. Target durations were randomly set in the range 300–600 ms.

When Elija attended to touch, he rewarded closure of the vocal tract and this mainly lead to the discovery of plosives. Similarly, when he attended to high frequency acoustic power he discovered fricatives. In both cases, only the initial consonantal part of the motor pattern was optimized.

Until now, Elija had performed self-supervised evaluation of his productions. From here onwards, Elija also interacted with a caregiver. He performed each vocal

action and the corresponding acoustic output was played to the caregiver via a loudspeaker. Elija then listened for two seconds for any vocal response the caregiver chose to make. The caregiver found it natural to reformulate certain sounds and ignore others. When Elija detected the presence of acoustic power he recorded the reformulation. Vocal actions that provoked no response were discarded.

Elija's self-supervised search procedure lead to the discovery of 1000 CVs and diphthongs. Reinforcement on the basis of caregiver reformulations cut this down to 890, removing non-English utterances. The CV's were decomposed into their C and V components. The surviving V patterns were acoustically clustered into 15 groups and the surviving C patterns into 40 groups on the basis of motor similarity. The fricative sounds were extracted from the CV's and clustered into 10 groups. Once again, cluster group sizes were selected to maintain variety without redundancy.

Results. Five of the 15 clustered "vowels" that Elija found in the primitive articulation and expansion stage could be readily recognized as tense vowels in English: one as /i:/, one as /a:/, two as /u:/ and one as /ɜ:/. The remainder either had some similarities to the tense vowels, might be taken as lax vowels if shortened and embedded in a word, or might not exist in English but might resemble vowels in other languages.

Fifteen of the 40 syllables produced using clustered consonants with a following vowel were either clearly or recognizably similar to English syllables. Two of the tokens were unlike natural syllables. Most of the remainder could be assimilated to an English syllable by a generous listener. The consonant sounds that appeared were /w j p-b t-d g/.

Six of the clustered fricatives might be interpreted as /f/ when heard contextually. One might be heard as /θ/. Three were too unnatural to be heard as speech sounds.

Four of the clustered diphthongs could be heard as /ɪə/ /aʊ/ /aɪ/ and /əʊ/. The others did not resemble English diphthongs.

Integrative Stage Experiment

Method. By the end of Oller's integrative stage, many infants recombine segment-sized elements of well-formed syllables to form new syllables. By recombining C and V targets a wide variety of new motor patterns were generated. The V targets were recombined to generate VVs (diphthongs). The Cs and Vs were recombined to generate CVs, VCs, and CVVs. Vs were recombined with Fs (fricatives) to generate FVs and VFs. The CV patterns were copied and nasalized to generate NVs. Simple Fs and Vs were also present in the new expanded repertoire. In total 1535 new motor patterns were generated, which were pruned to 915 by caregiver reformulation. The reformulations themselves were used to form the basis of Elija's speech sound recognition, enabling him to now parse the caregiver's speech.

The recombination procedure was fruitful because it meant that a given articulation only had to occur once for it to be reused in all other contexts. This expanded the production repertoire without the need for further exploration.

Results. The large majority of Elija's recombinations can be recognized as similar to the caregiver's tokens and dissimilar to other syllables that Elija produces. Because of the difficulty of transcribing child speech, no objective

measure of this has been attempted but the utterances are available for review in the online Supplementary Material.

Object Labeling Experiment

Method. In the final experiment, Elija was taught the names of 25 objects by the caregiver. The majority of these were chosen from a list of the first words produced by infants (Morrison et al. 1997). A picture of each object was presented to the caregiver, who spoke its name. This "meaning" provided additional evidence to link the name spoken by the caregiver with the attempt generated by Elija. As Elija does not have a visual system, an identifier index was presented to Elija, so he could also form an association between his motor action, the object context and the associated caregiver speech utterance. Elija performed recognition on this input, and responded using a sequence of corresponding vocal actions. If the caregiver accepted this response, the caregiver indicated approval by moving on to the next picture. If not, the procedure was repeated iteratively with Elija searching through all the vocal actions in the category and offering other candidate imitations.

Results. Some single syllable words were reproduced accurately because the syllable existed in Elija's gestural repertoire. Others were approximated with recognizable results. In the case of two-syllable words, Elija's reproductions were recognizable provided at least one syllable of the word could be approximated.

For 14 of the 35 words that Elija attempted to imitate (*Piers*, *Ian*, *car*, *chair*, *door*, *fish*, *flower*, *house*, *shoe*, *spoon*, *table*, *tree*, *ball* and *(ba)nana)*, his production is similar to the typical words produced by early speakers. His pronunciation of these words would be acceptable to sympathetic (e.g., family) listeners. Although these results do not represent the standards of production achieved by an adult, they are typical of a young infant.

Discussion

Summary

Conventional accounts of how infants learn to pronounce posit a purely imitative mechanism. Using Elija, a computational model, we demonstrated that an alternative, in which an infant makes use of natural interactions with a caregiver, could also solve the correspondence problem between infant and adult speech sounds, and thereby enable word learning to develop. Notably, it was always the learned caregiver, not Elija, who judged his vocal performance. Elija followed the developmental stages described by Oller (Oller 2000).

Elija first discovered vocal actions that lead to the production of sounds on his own, in the absence of a caregiver. Initially these were simple, single target configurations corresponding to vowels, which were then followed by simple syllables. His recombination of their constituent Cs and Vs expanded his sound repertoire. Responses from a caregiver suppressed non speech-like sounds and biased production toward speech sounds found in the ambient language. Caregiver reformulations of Elija's output allowed him to develop associations between his vocal actions and adult speech sounds. During a final object labeling task, Elija learnt sequences of vocal actions which reproduced some of the object names spoken by the caregiver.

During this final labeling experiment, the caregiver's objective was for Elija to generate an appropriate pronunciation for the object. To do so, the caregiver encouraged him to find the best sequence of vocal actions that he was capable of producing. In this, it helped when the caregiver became familiar with how Elija recognized and produced words. Then the caregiver could sometimes prompt him with modified speech utterances which evoked responses which were considered acceptable at this stage in his development. For this, the caregiver emphasized or simplified speech output away from the conventional pronunciation of the word in question.

Comparison with Other Computational Speech Acquisition Models

We now compare Elija with some other computational models.

Guenther's DIVA model (Guenther 1994; 1995; Guenther et al. 2006) and that of Kröger (Kröger et al. 2009a; Kröger et al. 2009b) focus on the sensorimotor transformations involved in the control of articulator movements during speech production. They use both feed-forward and feedback control pathways and address the learning of low level motor control. The models by Laboissière (Laboissière 1992) and Bailly (Bailly 1997) also learn an acoustic forward model.

In Elija, we instead follow the approach taken in HABLAR (Markey 1994), which was in turn inspired by Articulatory Phonology and the Task Dynamic model (Saltzman and Munhall 1989). Thus Elija does not learn the low level motor control of his articulators and can immediately repeat the motor pattern for any vocal action he discovers. In the Task Dynamic model, the atomic unit is a "gesture", defined as the goal directed movement of a single articulator resulting in a vocal tract constriction (Goldstein et al. 2006). In Elija, the atomic unit is a simple articulation that is defined in terms of the articulatory synthesizer control parameters. In both models, several atomic units must be organized appropriately to build up speech utterances, e.g., a syllable. Thus Elija's motor patterns are similar to the gestural score used in the Task Dynamic model, with the targets similarly implemented as point attractors and the movement of the articulators affected by attractor dynamics.

Most other models use a babbling phase, based on either a random or exhaustive search of articulatory configurations, to learn a mapping between articulatory and auditory representations. Babbling also plays an important part in the development of Elija's speech. However, we use a more natural approach in which speech sounds are discovered by rewarded exploration. In addition, in contrast to other models, babbling is not used to develop inverse and forward models linking the trajectories of motor actions and their sensory consequences. Rather, babbling allows associations to be formed between the discrete events corresponding to Elija's motor actions and the caregiver's responses.

Most other models learn speech sounds by imitation, without either addressing the normalization problem (the fact that an infant's production is objectively very different from an adult's due to its different size vocal tract) or the fact that in real life infant utterances sometimes bear little resemblance to their linguistically equivalent adult forms. In contrast, although Elija also uses a form of imitation at later stages of development (copying the serial order of speech sounds in a word), he "learns to imitate" speech sounds, rather than this being an innate ability.

Future Work

In our model we chose to concentrate on the natural discovery of what will become speech sounds and on the role played by caregiver interactions. We made no attempt to model motor development of low level articulator control. If we took such development into account, we would expect some simpler vocal actions to emerge before more complex ones. For example, some of the syllables Elija developed would, in real life, develop before others due to the ease of their generation.

The articulatory synthesizer we used had limitations and in particular the quality of nasal sounds generated was poor. Improvements to the articulatory model are therefore needed, since nasals sounds are created early by real infants.

Our memory model uses associations between actions and acoustic utterances. A real infant will develop more complex relationships between input and output, and relate them to internal rewards and desires. It should be possible to formulate Elija's learning and response generation in a Bayesian framework which can take the evidence (and reliability) of each data source into account. In addition, latent variables in such models are able to model hidden structure, such as cooperation between the agent, the caregiver and meaning (Frank et al. 2009). This is important because cooperation of the caregiver must be taken into account by Elija to take advantage of reformulations.

Including the constraints imposed by speech breathing would also be beneficial since they play a role in the development of timing and prosody in pronunciation (Messum 2008b), and we have already made preliminary steps in this direction (Howard and Messum 2008). To improve the experience of interacting with Elija, he would benefit from having an animated face synchronized to his acoustic output. A real infant would also have access to multimodal input, including vision. This would provide additional visual cueing from the caregiver (Huckvale et al. 2009).

In the field of speech technology, it is increasingly recognized that current engineering solutions are reaching limits of performance (Moore 2007). We believe that a deeper understanding of how infants learn to perceive and produce speech, in particular as embodied agents that interact with their caregivers, offers a new way forward through building systems which learn to speak and listen rather than having these capabilities specified by their designers.

Finally, our non-imitative account of learning to pronounce incorporates principles that are likely to apply to the development of a wide range of motor abilities, such as learning the control of skilled hand and arm movements.

Notes

1. For example, among the theoretical problems, Messum distinguishes the ecological situations of child first language learners and older second language learners (Messum 2008a; Messum 2007). The latter are able to engineer situations to improve their pronunciation. In these situations, they are presented either with speech sounds spoken in isolation or with words that they have requested and/or expect to hear. The older learners can then listen to these sounds/words with the attentional set required to hear the acoustic signal veridically, i.e., in Pisoni's "phonetic mode" (Pisoni 1973). This contrasts with a listener's normal attentional set (Pisoni's "auditory mode") which is to listen to recognize words, in which case the veridical signal is not retained.

A child learner must almost always listen with this second attentional set, since his verbal interactions with his caregivers are the result of one or both sides wishing to express or

communicate something. Thus he rarely gets the opportunity to compare his own speech sound production with that of others.

Among the child speech phenomena which cannot be explained satisfactorily under imitative accounts, there is the well-known "fis"/"fish" phenomenon (Clark and Clark 1977; Locke 1979; Priestly 1980). Here, a child pronounces "fish" as "fis" and when questioned as to why he did so, insists firstly that he can hear the distinction in the two forms made by the adult and secondly that he did not say the incorrect form himself. Imitative accounts cannot explain this because they assume that the infant learns speech sounds by copying the adult form, in which case it is paradoxical that he can hear a difference in adult speech but not hear it in his own. In our model, on the other hand, production and the parsing of words for speech sound equivalences are initially separate from general speech perception. As such, production forms can differ from those used in general word perception and the "fis"/"fish" phenomenon has an uncomplicated explanation.

Among the problematic adult phenomena, there is the data on speech shadowing (e.g., Fowler et al. 2003). Under conventional accounts of speech sound acquisition by acoustic matching, it appears that speech is an exception to the otherwise universal response time differences in simple and choice reaction time tests. In our model, the data are explained more simply, because the production and perception of speech sounds is directly associated rather than going via a common form, whether acoustic or gestural.

These and other possible problems with conventional accounts of how pronunciation develops are discussed further in Messum (2007).

2. The Maeda articulatory synthesizer uses a 2-dimensional model to represent the cross-sectional profile of the vocal tract along the midsagittal plane. The parameters in the model were estimated (by Maeda) using factor analysis of a dataset. This consisted of cine-radiographic vocal tract profiles and frontal lip shape recordings of 2 female French speakers producing 10 French sentences. The vocal tract was divided up into 3 sections—lip aperture, principal vocal tract and pharynx. The principal vocal tract was characterized in semi polar coordinates, the lips by an ellipse and the larynx by its height. A jaw model (Lindblom and Sundberg 1971) was used to represent the dataset in terms of the parameters jaw, tongue-body, tongue-tip, lip height, lip width and larynx height. The vocal tract shape is determined by a linear combination of these primitive elements found using a directed factor analysis. In contrast to normal factor analysis, this method allowed Maeda to represent the data in terms of the parameters he had selected.

Acknowledgments

We thank an unknown reviewer for constructive criticisms of the initial version of the manuscript. VTCALCS is a vocal tract model written by Shinji Maeda, modified and distributed by Satrajit Ghosh. We used a version that had been further extended by Mark Huckvale and ISH. ISH & PM have no financial conflicts of interest. We thank Daniel Wolpert for allowing us to use facilities in the Computational and Biological Learning Laboratory, Department of Engineering, University of Cambridge.

References

Bailly, G. (1997). Learning to speak. Sensorimotor control of speech movements. *Speech Communication,* 251–267.

Barlow, H.B. (1961). Possible principles underlying the transformation of sensory messages. In W.M.I.T. Rosenblith (Ed.), *Sensory Communication*. Cambridge, MA: Press.

Boë, L.J., Heim, J.L., Honda, K., Maeda, S., Badin, P., & Abry, C. (2007). The vocal tract of newborn humans and Neanderthals: Acoustic capabilities and consequences for the debate on the origin of language. A reply to Lieberman (2007a). *Journal of Phonetics, 35,* 564–581.

Browman, C., & Goldstein, L. (1986). *Towards an articulatory phonology*. Cambridge University Press.

Browman, C.P., & Goldstein, L. (1992). Articulatory Phonology - an Overview. *Phonetica, 49,* 155–180.

Byrne, R.W. (2003). Imitation as behaviour parsing. *Philosophical Transactions of the Royal Society of London, 358,* 529–536.

Chouinard, M.M., & Clark, E.V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language, 30,* 637–669.

Clark, E.V., & Clark, H.H. (1977). *Psychology and language: an introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.

Ellis, D. Dynamic Time Warp (DTW) in Matlab. http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/. 2003.

Fant, G., Liljencrants, J., & Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR, 4,* 1–13.

Fowler, C.A., Brown, J.M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: evidence from choice and simple response time tasks. *Journal of Memory and Language, 49,* 396–413.

Frank, M.C., Goodman, N.D., & Tenenbaum, J.B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20,* 578–585.

Gattegno, C. (1973). *In the beginning there were no words: The universe of babies*. New York: Educational Solutions.

Gattegno, C. (1985). *The Learning and Teaching of Foreign Languages*. New York: Educational Solutions.

Gattegno, C. (1987). *The Science of Education. Part 1: Theoretical considerations*. New York: Educational Solutions.

Gold, B., & Rader, C.M. (1967). Channel Vocoder. *IEEE Transactions on Audio and Electroacoustics, AU-15,* 148–161.

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural units in understanding the evolution of phonology. In M.A. Arbib (Ed.), *Action to language via the mirror neuron system* (pp. 215–249). Cambridge University Press.

Goldstein, L., & Fowler, C.A. (2003). Articulatory phonology: a phonology for public language use. In N.S. Schiller & A.S. Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and Production* (pp. 159–207). Mouton de Gruyter.

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

Guenther, F.H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics, 72,* 43–53.

Guenther, F.H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review, 102,* 594–621.

Guenther, F.H., Ghosh, S.S., & Tourville, J.A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96,* 280–301.

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences, 5,* 253–261.

Howard, I.S., & Messum, P. R. (2008). Modeling motor pattern generation in the development of infant speech production. In: *8th International Seminar on Speech Production – (ISSP'08)*, edited by Sock R, Fuchs S, and Laprie Y. Strasbourg, France: INRIA, pp. 165-168.

Howard, I.S., & Messum, P.R. (2007). A Computational Model of Infant Speech Development. In: *XII International Conference "Speech and Computer" (SPECOM'2007)* Moscow State Linguistics University, pp. 756-765.

Huckvale, M.A., Howard, I.S., & Fagal, S. (2009). KLAIR: a Virtual Infant for Spoken Language Acquisition Research. In: *Proceedings of InterSpeech*. Brighton, UK.

Jonsson, C.O., Clinton, D.N., Fahrman, M., Mazzaglia, G., Novak, S., & Sörhus, K. (2001). How do mothers signal shared feeling-states to their infants? An investigation of affect attunement and imitation during the first year of life. *Scandinavian Journal of Psychology, 42,* 377–381.

Kröger, B., Kannampuzha, J., Lowit, A., & Neuschaefer-Rube, C. (2009a). Phonetotopy within a neurocomputational model of speech production and speech acquisition. In S. Fuchs, H. Loevenbruck, D. Pape, & P. Perrier (Eds.), *Peter Lang* (pp. 59–90). Berlin.

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

Kuhl, P.K. (1987). Perception of speech and sound in early infancy. In P. Salapatek & L. Cohen (Eds.), *Handbook of Infant Perception* (pp. 275–382). New York: Academic Press.

Laboissière R. (1992). Préliminaires pour une robotique de la communication parlée: inversion et contrôle d'un modèle articulatoire du conduit vocal. l'Institut National Polytechnique de Grenoble.

Levelt, W.J., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50,* 239–269.

Lindblom, B. (1999). Emergent phonology. In: *25th Annual Meeting of the Berkeley Linguistics Society*. U. California, Berkeley.

Lindblom, B.E., & Sundberg, J.E. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America, 50,* 1166–1179.

Locke, J.L. (1979). The child's processing of phonology. In: *Child Language and Communication: Minnesota Symposium on Child Psychology Volume 12*, edited by Collins WA. Hillsdale, NJ: LEA, pp. 83-119.

Locke, J.L. (1996). Why do infants begin to talk? Language as an unintended consequence. *Journal of Child Language, 23,* 251–268.

Mackay, D.J.C. (1992). Information-Based Objective Functions for Active Data Selection. *Neural Computation, 4,* 590–604.

Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 131–149). Boston: Kluwer Academic Publishers.

Markey, K.L. (1994). *The sensorimotor foundation of phonology; A computational model of early childhood articulatory development*. University of Colorado at Boulder.

Markey KL. (1993). A sensorimotor model of early childhood phonological development. University of Colorado at Boulder: Department of Computer Science.

McCune, L., & Vihman, M.M. (1987). Vocal Motor Schemes. *Papers and Reports in Child Language Development*. *Stanford University Department of Linguistics, 26,* 72–79.

Meltzoff, A.N. (1999). Origins of theory of mind, cognition and communication. *Journal of Communication Disorders, 32,* 251–269.

Menn, L., Markey, K.L., Mozer, M., & Lewis, C. (1993). Connectionist modeling and the microstructure of phonological development: a progress report. In B. de Boysson-Bardies (Ed.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 421–433). Dordrecht: Kluwer.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. In C.H. Chen (Ed.), *Pattern Recognition and Artificial Intelligence*. New York: Academic Press.

Messum, P. (2008a). What if children don't learn to pronounce by imitation? *Speak Out, 39,* 16–21.

Messum, P.R. (2008b). Embodiment, not imitation, leads to the replication of timing phenomena. In *Proceedings of Acoustics* (pp. 2405–2410). Paris: SFA/ASA/EAA.

Messum PR. (2007). The Role of Imitation in Learning to Pronounce, PhD Thesis. University of London.

Mompean-Gonzalez, J.A. (2004). Category overlap and neutralization: the importance of speakers' classifications in phonology. *Cognitive Linguistics, 15,* 429–469.

Moore, R.K. (2007). PRESENCE: A human-inspired architecture for speech-based human machine interaction. *IEEE Transactions on Computers, 56*(9).

Morrison C.M., Chappell T.D., and Ellis A.W. (1997). Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables. *The Quarterly Journal of Experimental Psychology* Section A, 50:3, 5: 28-559.

Nabney, I. (2004). *NETLAB: Algorithms for Pattern Recognition*. Springer.
Nam H., Goldstein L., Saltzman E., and Byrd D. (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *Journal of the Acoustical Society of America* 115(5,2): 2430-2430.
Newson, J. (1979). The growth of shared understandings between infant and caregiver. In M. Bullowa (Ed.), *Before speech: the beginning of interpersonal communication* (pp. 207–222). Cambridge University Press.
Oller, D. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates.
Oller, D.K., Eilers, R.E., Neal, A.R., & Schwartz, H.K. (1999). Precursors to speech in infancy: the prediction of speech and language disorders. *Journal of Communication Disorders, 32,* 223–245.
Olshausen, B.A., & Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607–609.
Otomo, K. (2001). Maternal responses to word approximations in Japanese children's transition to language. *Journal of Child Language, 28,* 29–57.
Parton, D.A. (1976). Learning to imitate in infancy. *Child Development, 47,* 14–31.
Pawlby, S.J. (1977). Imitative interaction. In H.R. Schaffer (Ed.), *Studies in Mother-Infant Interaction* (pp. 203–223). London: Academic Press.
Pisoni, D.B. (1973). Auditory and phonetic memory codes in discrimination of consonants and vowels. *Perception & Psychophysics, 13,* 253–260.
Priestly, T.M.S. (1980). Homonymy in child phonology. *Journal of Child Language, 7,* 413–427.
Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.
Sakoe, H., & Chiba, S. (1978). Dynamic-programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26,* 43–49.
Saltzman, E., & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science, 19,* 499–526.
Saltzman, E., & Kelso, J.A. (1987). Skilled actions: a task-dynamic approach. *Psychological Review, 94,* 84–106.
Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1,* 333–382.
Saltzman, E.L., & Munhall, K.G. (1992). Skill acquisition and development: the roles of state-, parameter-, and graph dynamics. *Journal of Motor Behavior, 24,* 49–57.
Stern, D.N. (1985). The sense of a subjective self: Affect attunement. In *The Interpersonal World of the Infant* (pp. 138–145). London: Karnac Books.
Studdert-Kennedy, M. (2002). *Mirror neurons, vocal imitation, and the evolution of particulate speech* (pp. 207–227). Amsterdam: John Benjamins.
Turetsky, R., & Ellis, D. (2003). Ground-truth transcriptions of real music from force-aligned MIDI syntheses. *4th International Symposium on Music Information Retrieval ISMIR-03* 135-141.
Westermann, G., & Miranda, E. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language, 89,* 393–400.
Yoshikawa, Y., Asada, M., Hosoda, K., & Koga, J. (2003). A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science, 14*(4), 245–258.
Zukow-Goldring, P., & Arbib, M.A. (2007). Affordances, effectivities, and assisted imitation: Caregivers and the directing of attention. *Neurocomputing, 70,* 2181–2193.