Journal of Phonetics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation

Piers Messum [a,*], Ian S. Howard [b]

[a] Pronunciation Science Ltd, 112 Warner Road, London SE5 9HQ, UK
[b] Centre for Robotics and Neural Systems, School of Computing, Electronics and Mathematics, Plymouth University, Plymouth PL4 8AA, UK

ARTICLE INFO

Article history:
Received 20 December 2014
Received in revised form
28 June 2015
Accepted 27 August 2015

Keywords:
Phonological units
Underlying representation of speech
Speech acquisition
Correspondence problem
Development of pronunciation
Imitation
Mirroring

ABSTRACT

Theories about the cognitive nature of phonological units have been constrained by the assumption that young children solve the correspondence problem for speech sounds by imitation, whether by an auditory- or gesture-based matching to target process. Imitation on the part of the child implies that he makes a comparison within one of these domains, which is presumed to be the modality of the underlying representation of speech sounds. However, there is no evidence that the correspondence problem is solved in this way. Instead we argue that the child can solve it through the mirroring behaviour of his caregivers within imitative interactions and that this mechanism is more consistent with the developmental data. The underlying representation formed by mirroring is intrinsically perceptuo-motor. It is created by the association of a vocal action performed by the child and the reformulation of this into an L1 speech token that he hears in return. Our account of how production and perception develop incorporating this mechanism explains some longstanding problems in speech and reconciles data from psychology and neuroscience.

1. Introduction

This special issue of the Journal of Phonetics focuses on how phonology – the science of speech sound systems – can contribute to the longstanding debate about the nature of speech units in the human brain. Phonology, though, is the outcome of practical processes: a child learning to comprehend and pronounce a particular language. Basic assumptions about how this takes place set the terms of the debate. In particular, in the learning of pronunciation it has always been assumed that the child solves the correspondence problem for speech sounds by imitation. That is, the child uses his[1] own judgement of similarity between what he recovers from the speech input (an acoustic pattern in some theories or a pattern of gestures in others) and what he produces in return to match this. This judgement informs and improves his subsequent production in a 'matching to target' process. A judgement of similarity is only possible between images that are comparable, so the child is assumed to be operating either with auditory or motor primitives, and the favoured one of these is then considered to be the form for the underlying representation of speech in the human brain.

In our work, we have been investigating how the pronunciation of L1[2] is learned. For a number of reasons we argue that it is unlikely that children solve the correspondence problem by acoustic imitation of caregiver speech. Instead we suggest it is plausible that they find a solution in the dynamics of caregiver–infant interaction. Here, imitation takes place, but it is usually the caregiver imitating the child, rather than vice versa. In the gestation period of speech, the form of the imitation is rarely simple mimicry; instead a caregiver reformulates her child's output into L1, giving him evidence of the correspondences between what he does and what she considers its linguistic significance to be.

* Corresponding author. Tel: +44 20 7274 6306.
E-mail addresses: p.messum@pronsci.com (P. Messum), ian.howard@plymouth.ac.uk (I.S. Howard).

[1] We avoid continual use of "he or she", "his or her", etc., by using pronouns which describe interactions between a female caregiver and a male child.

[2] In this article we use the following abbreviations: L1 for 'first language', ME for'mirrored equivalence', SBE for 'similarity based equivalence', AS for 'awareness of sensation', MP for 'meaningful perception', IM for 'inverse model', PM for 'perceptuo-motor' and VMS for 'vocal motor scheme'.

http://dx.doi.org/10.1016/j.wocn.2015.08.005

Like Moulin-Frier, Diard, Schwartz, and Bessière (2015, this issue), we have used a computational model to investigate some aspects of this matter. Our computational agent, Elija (Howard & Messum, 2007, 2011, 2014), models speech acquisition by children. To examine the development of pronunciation, we aimed to endow him with capacities for production, perception and cognitive activity that are no greater than those of a human infant. With these, he was able to develop a repertoire of potential speech sounds, to interact with human subjects taking the role of his caregiver, to learn speech sound correspondences from these interactions, and finally to use these correspondences to learn the pronunciation of simple words in the language of each caregiver.

In this article, we describe the theoretical background to Elija, starting with the issue of how pronunciation is learnt. We discuss previous proposals for how children might solve the correspondence problem, which are principally acoustic matching theories. We describe these as examples of a 'Similarity Based Equivalence' (SBE) mechanism. We then explain how well-attested mirroring behaviour seen on the part of caregivers supports an alternative proposal which we describe as a 'Mirrored Equivalence' (ME) mechanism. We argue that at critical stages of child speech development ME provides a better explanation of observed phenomena that any SBE account. ME would generate an intrinsically perceptuo-motor cognitive form for phonological units and would support a new account of how speech production and perception develop in a child, which we describe.

2. Learning to pronounce

2.1. How the pronunciation of words is learned

To explain our account of the development of pronunciation in a child, we need to distinguish the activities of learning how to pronounce particular words from learning how to pronounce speech sounds. The first of these, the mature skill of learning the pronunciation of a new L1 word, is readily accessible to introspection and is uncontroversial: the speaker parses the word he has just heard into a string of speech sounds and says, in his own voice and in the same order, a string of speech sounds that he knows his listeners will take to be equivalent to the ones he heard (see Fig. 1). A speech sound in this context is a syllable, or even a couple of syllables, formed of one or more phonemes (Guenther, Ghosh, & Tourville, 2006, p. 283) which occurs commonly enough to form part of Levelt's 'mental syllabary' (Levelt, Roelofs, & Meyer, 1999, p. 5). In order to achieve this ability to imitate, the speaker must first learn how to produce speech sounds that will be taken to be equivalent to the ones he hears. To characterise these two distinct learning activities, we use the terminology 'learning to pronounce a word' and 'learning to pronounce' respectively.

It takes several years for a child's pronunciation to approach an adult level of competence (Dodd et al., 2003). Thus learning to pronounce to the point of mastery must represent a significant practical challenge for a child, even though it has been taken to be conceptually straightforward. The general assumption has been that learning to pronounce is a self-supervised process of auditory 'matching to target': having identified a speech sound, the child tries to copy what he hears and then judges for himself the similarity of his output to the target. He uses this to improve his subsequent attempts.

Within this account, 'perception' has been taken as a precursor to the development of production. Therefore it has been assumed that a single phonological lexicon begins to develop as the child's mind grapples with the perceptual data presented to it, and that this lexicon goes on to inform production. As a consequence, scholarly interest has focussed on the question of how phonology is acquired during perception, rather than on the apparently secondary process of how pronunciation is actively learned. However, as we show below, if speech sounds are not learnt by imitation then it is possible for an output phonological lexicon both to develop and to be structured independently from a perceptual lexicon. Thus the learning of the pronunciation of speech sounds is an issue to be addressed in understanding the genesis of phonology.

2.2. The correspondence problem for speech sounds

For a mature speaker to learn to pronounce a word by parsing its component speech sounds and reproducing them using his own voice, he first needs to discover how to produce speech sounds that will be taken to be equivalent to the ones he hears. This requires

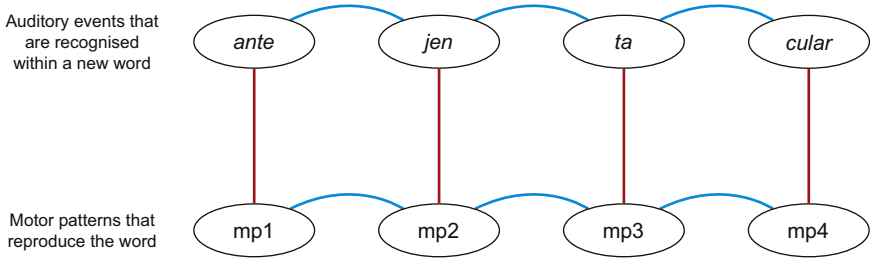

Fig. 1. The mature skill of learning the pronunciation of a new word requires (1) learning the identity and ordering of the speech sounds heard, and (2) prior to this, learning the 'vertical' links between speech sounds heard and the motor patterns that can be used to reproduce them. *source*: When a word is heard for the first time, the speaker parses it into speech sound elements. For example, he may decompose 'antejentacular' into 'ante – jen – ta – cular'. He can reproduce these four auditory events using four motor patterns, each of whose output he knows will be taken by his listeners to be equivalent to what he has heard. Thus he learns to pronounce the word by serial imitation. However, he must have previously learnt the 'vertical' links between speech sounds he hears and their corresponding motor patterns. The correspondence problem for speech sounds is the question of how he achieves this: either using some form of imitation or by some other mechanism. Fig. 1 and terminology adapted from Heyes (2001).

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

a solution to a 'correspondence problem' (Nehaniv & Dautenhahn, 2002) between sounds he hears and the vocal actions he performs. A solution to any correspondence problem is straightforward if the signals that a learner receives and produces are perceptually 'transparent' to him. He can then compare them for himself and solve the problem through a matching to target process. However, if the signals are 'opaque' this is not possible.

Perceptual opacity is the degree to which the sensory experience of observing another individual performing an action overlaps with the sensory experience of observing oneself performing that action (Heyes & Ray, 2000). For example, whistling is essentially perceptually transparent, since hearing oneself whistle and hearing another individual whistling are very similar experiences. Facial gestures, on the other hand, are perceptually opaque since one cannot see one's own face directly. For some actions, signals will be somewhere in between, such as movements involving one's hands or legs. Kicking a football is more opaque than making a ring with one's thumb and forefinger because of the different viewing angles involved in the two tasks.

Imitative accounts of how a child solves the correspondence problem for speech sounds assume that the signals are transparent, and most describe a mechanism of self-supervised auditory matching to target (see Fig. 2). There have been different proposals as to how the child uses target sounds in the linguistic environment, and these proposals are not mutually exclusive. Sounds may be copied when needed (Fry, 1968, p. 18), stored as sound images that are later used to guide production (Kuhl, 2000, p. 11854), or might be discovered in the linguistic environment after the infant has been primed to find them through listening to his own production (Sweeney, 1973, p. 491), in a similar fashion to how Vihman's articulatory filter selects for words to be attempted (Vihman, 1993). Importantly, in all cases the child himself is required to make a judgement of similarity to determine equivalence between what he hears and what he produces. For this reason, we call this class of proposals auditory 'Similarity Based Equivalence' (SBE) accounts.

Other solutions to the correspondence problem also posit the child making a judgment of similarity, but propose that he recovers the vocal gestures of his caregivers from the speech signal to compare to his own (e.g. Goldstein & Fowler, 2003). There are also the accounts described as 'sensory-motor' by Skipper, van Wassenhove, Nusbaum, and Small (2007), Schwartz, Basirat, Ménard, and Sato (2012) and Moulin-Frier et al. (2015), where both auditory and motor representations develop, but where the auditory representation is taken as primary.

Even with a perceptually transparent signal, there are potential difficulties with SBE accounts. One such difficulty is the normalisation problem. This is the result of the mismatch between the sizes of the vocal tract and its articulators in a young child and in an adult, making it impossible for a child's output to acoustically match that of an adult. Various solutions have been proposed as to how an infant could judge similarity in these circumstances (Johnson, 2005; Kuhl, 1987, 1991) but the issue has not been resolved. Note that within those SBE accounts which describe the child matching vocal tract gestures rather than sounds, the problem would not arise since the size of the speaker's vocal apparatus is not considered to affect the linguistic nature of such gestures.

It is possible that the acoustic signal is not transparent to a young learner, or may even be completely opaque (Messum, 2007). For example, an infant may not hear himself adequately, particularly for vowel sounds, because of interference from bone-conducted sound (Porschmann, 2000). Or he may normally 'hear' what he intends to produce (the output of the forward model that, at an older age, generates inner speech when we read) rather than hearing his actual output, making a comparison of the relevant signals difficult. Furthermore, when he is listening to speech for comprehension he may find himself in the wrong attentional set for copying speech sound qualities (as discussed in Section 6.2).

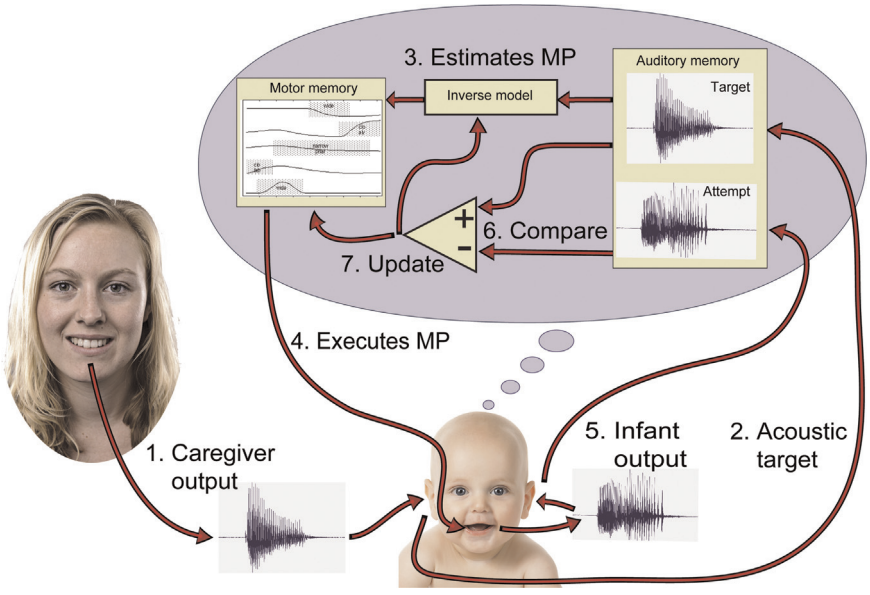

Fig. 2. How a child would solve the correspondence problem for a speech sound in a simplistic auditory Similarity Based Equivalence (SBE) account. (*Abbreviations*: MP – motor pattern, IM – inverse model.) (1) An L1 speech sound is produced by a caregiver (perhaps within a word). (2) The child takes this as a target. (3) Using his previously learned inverse model, developed prior to and during babbling, (4) the child executes a motor pattern to produce a sound to match the target. (5) He listens to his own output and (6) compares his output to the stored target. (7) Depending on the nature of the error signal he updates the motor pattern, the inverse model or both. The comparison mechanism uses his judgement of similarity between the caregiver's output and his own. The steps are repeated until the infant is satisfied with the match. The process is one of auditory matching to target.

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

2.3. Mirroring

SBE is not the only way by which an infant might solve the correspondence problem. For opaque signals an innate, supramodal representation might mediate between perception and action, as postulated in Meltzoff and Moore's Acquired Intermodal Mapping hypothesis (Meltzoff & Moore, 1997), developed in part to explain their evidence of tongue protrusion by newborns in response to similar adult behaviour (Meltzoff & Moore, 1977). This was discussed in relation to speech by Fowler (2004). Alternatively, general mechanisms of social learning may be sufficient for opaque signals, as proposed by Heyes in her Associative Sequence Learning (ASL) paradigm (e.g. Heyes & Ray, 2000, p. 224). In this account, a learner can inform himself about his production by using a physical mirror to observe himself. Alternatively he can attend to a metaphorical mirror in the form of another person who performs his action back to him. In both cases, the 'mirror' informs the learner of what he has just done by 'reflecting' it back. This metaphor can extend to the mirroring of internal states as well as surface behaviour.

The discovery of mirror neurons (Rizzolatti, Fadiga, Gallese & Fogassi, 1996) has led to substantial interest in this issue, as mirror neurons might instantiate the solutions to correspondence problems. See Cook, Bird, Catmur, Press, and Heyes (2014) for a review, and Hickok (2014) for arguments in favour of mirror neurons being the product of associative learning.

Perhaps as a result of the widespread discussion of mirror neurons, the term 'mirroring' is now sometimes used as a simple synonym for copying. Here we restrict its use to when copying of a learner by a social partner provides information for the learner about himself, i.e. when the social partner is acting as a metaphorical mirror for the learner. This is how the term has been used in the psychological literature for many years. Note, though, that the term 'mirroring' is not ideal because a real mirror reflects an exact copy of the object in front of it and does so instantaneously, whereas useful information may be conveyed to a learner from a social partner by selective reflection or by behaviour that is actually different from that of the learner (Pines, 1985; Stern, 1985, p. 144).

Messum (2007) provides a review of mirrored interaction in early infancy, where in particular it is considered in relation to the development of affect. Pines (1984, p. 32) described the dynamics of this process (see also Rochat, 2001, p. 201):

> "It is mother who selects only certain patterns of activity to respond to in her child, thus presenting him with an image of himself through her mirroring behaviour … The child can begin to learn who he is through attending to his mother's response to those aspects of his behaviour which make sense to her. Mother inserts meaning and intentionality into her baby's behaviour and so in this way he begins to recognize himself."

Stern (1985, p. 142) placed imitation/mimicry at one end of a spectrum of mirroring behaviour and so-called 'affect attunement' at the other. In the latter, a caregiver reflects back to the child her understanding of his internal state rather than his overt behaviour. Stern (1985, p. 140) originally suggested that caregivers begin affect attunement behaviours when the infant is around 9 months of age. However Jonsson et al. (2001) found that episodes of affect attunement were seen at just 2 months, were already more common than those of imitation (i.e. mimicry) by 6 months, and increased further in relative importance from then to 12 months, at which point their study finished.

2.4. The role of mirroring in solving the correspondence problem for speech sounds

The educationalist Caleb Gattegno described a mirroring paradigm for the child's entry into speech (Gattegno, 1973, 1985). This was elaborated by Messum (2007), and implemented in the Elija model (Howard & Messum, 2007, 2011, 2014) as portrayed in Fig. 3. Independently, the Asada group arrived at the same potential solution to the correspondence problem (Yoshikawa, Koga, Asada, & Hosoda, 2003) and they have developed this further (e.g. Miura, Yoshikawa, & Asada, 2012). Lacerda (2003, p. 51) also described the mechanism.

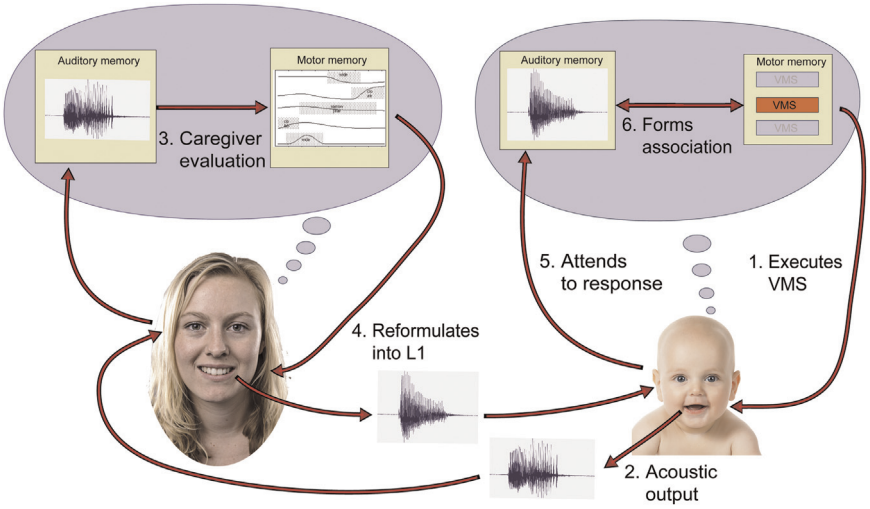

Fig. 3. Solving the correspondence problem for speech sounds using Mirrored Equivalence (ME). (1) The child executes a vocal motor scheme (VMS), (2) which generates acoustic output. (3) The caregiver interprets the output within L1, and (4) reformulates the child's output into an L1 token. (5) The child attends to this response, which reinforces his production, and understands that it is being produced within the context of an imitative interaction. (6) He concludes that his caregiver regards his vocal action and her output as equivalent and associates the two. Mirroring by the caregiver thus informs the child of the linguistic value of his VMS.

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

The basis for these mirroring accounts is the imitative vocal games that caregivers play with infants before their word production starts and for some time afterwards. Studies of these interactions with infants aged from 2 to 21 months (Kokkinaki & Kugiumutzakis, 2000; Otomo, 2001; Papoušek & Papoušek, 1989; Pawlby, 1977b; Veneziano, 1988) show that in most exchanges it is the caregiver that imitates the child. She increasingly imitates not the surface form of his utterances but her interpretation of the utterances within her L1 sound system. That is, she reformulates what he produces into well-formed L1 speech sounds that she considers to be 'similar' to what she heard.

The wealth of imitative interactions from early infancy onwards (Meltzoff & Williamson, 2010) suggests that infants understand the nature of reciprocal imitation games: that B's response to A is something B considers to be equivalent to A's activity. Therefore the caregiver's participation in this interaction is evidence for the child that she considers their activities to be equivalent, and the child then relies upon her judgement. We note that even taking a more cautious view of an infant's cognitive development, such equivalence can arise from simple association between events that are contiguous and contingent (Cook et al., 2014, p. 181). Within the Elija model, we have taken the equivalence to be conceived by the child as between his vocal gestures (rather than his vocal output) and what he hears from the caregiver in return.

There is an obvious parallel between affect attunement and vocal reformulation. In the former, infant behaviour is interpreted by his caregiver to be the expression of a particular inner state, his affective disposition. In the latter, the infant's vocal output is interpreted by his caregiver as if it was expressed from within the L1 sound system. In both cases the interpretation of the child's output is reflected back to him and can thereby assist his development.

Affect attunement behaviours precede the reformulation of sounds during and after babbling. Therefore when a child encounters reformulations in imitative exchanges he already has experience of picking up information about both his behaviour and his inner state from the mirroring behaviour of others. We use the term 'Mirrored Equivalence' (ME) to describe the mechanism being proposed by ourselves and the Asada group for solving the correspondence problem for speech sounds, since such mirroring behaviour by his caregiver provides an infant with evidence of equivalence.

2.5. Learning the pronunciation of first words

The principal issue we have addressed with Elija is how an infant solves the correspondence problem for speech sounds. However, it is not known when the mature skill of reproducing the pronunciation of a new word by speech sound parsing develops. Although this mechanism might be used by a child from the start, many scholars suggest that the pronunciation of early words is learnt differently, by holistic mimicry of the whole word shape. The term 'holistic' is appropriate because although the words are formed by combining gestures, "the gestures have not yet been differentiated as context-free, commutable units that can be independently combined to produce new words" (Studdert-Kennedy, 2002, p. 213). As development progresses, there would then be a movement away from this mechanism and towards the use of speech sound parsing and reproduction skills (Ferguson & Farwell, 1975, p. 422; Locke & Kutz, 1975, p. 185; Nazzi & Bertoncini, 2003; Snow, 1988, p. 348).

We note that progressive and regressive phonological idioms (Moskowitz, 1980) may be holistic sequences learnt by mimicry, co-existing with the child's developing 'particulate' lexicon until their content, too, becomes updated and reproduced by serial imitation of their speech sound elements.

3. Overview of Elija

We have tested the ME account with Elija, a computational model of infant speech acquisition (Howard & Messum, 2007, 2011, 2014). In our recent experiments, Elija models a process that starts with him 'babbling'. It ends with separate instances of Elija learning to pronounce simple words in three languages during naturalistic interactions with caregivers. His pronunciation of typical first words in English, French and German reached a level of competence that is comparable to that of a young child of around two years of age (Howard & Messum, 2014).

Elija's development begins by him discovering how to produce potential speech sounds by means of unsupervised active learning. In this 'babbling' phase, he develops a repertoire of motor patterns analogous to an infant's vocal motor schemes (VMSs). McCune and Vihman (2001, p. 152) describe a VMS as, "a generalized action plan that generates consistent phonetic forms." It is usually used to refer to consonantal forms, but for this article we use the term to cover forms that are analogous to all speech sounds.

During Elija's first interactions with a caregiver, ME provides the mechanism by which motor actions can be rewarded, and motor and auditory representations can become associated. Caregivers find it natural to reformulate the output of some of Elija's motor patterns into well-formed L1 tokens, as seen in real caregiver–infant interactions. This mirroring activity has two results. It selectively reinforces his range of potential speech sounds, with Elija retaining motor patterns to which a caregiver responded, and it enables him to associate his vocal actions with the speech sounds he hears in response to them. Importantly, the equivalence between them is based on a judgment of sound similarity made by the caregiver rather than by Elija. The association operates bi-directionally and solves the correspondence problem for speech sounds. The process is illustrated in Fig. 3.

The correspondences between motor patterns and caregiver reformulations are used by Elija during a second, word imitation phase of interaction. The caregiver says a word, and Elija parses it in terms of those sounds he has heard before from the same caregiver. He responds using a sequence of the motor actions he has associated to her speech sounds. In this way a caregiver can teach Elija to say some simple words through Elija's serial imitation of the words' component speech sounds.

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

We aimed to endow Elija with no more capacity than a human infant, although the practical demands of running experiments with human subjects demanded some changes from complete naturalism. We credited him with various powers of the mind (Gattegno, 1986; Young & Messum, 2011) including curiosity, selective attention, and the ability to infer equivalence during imitative games with a caregiver. When Elija interacted with human caregivers, they were allowed to ignore him or respond to him as they felt appropriate, but overall they did in fact do what caregivers are observed to do with real children. In previous articles (Howard & Messum, 2011, 2014) we have described the technical workings of Elija and do not repeat them here. We have also compared Elija to other computational models of speech development, including those of the Asada group (e.g. Yoshikawa et al., 2003; Miura et al., 2012).

4. Support for Mirrored Equivalence (ME)

Currently no studies have been reported that conclusively determine how children solve the correspondence problem for speech sounds. However, here we present some reasons why an ME mechanism represents a plausible hypothesis (only briefly summarising the reasons we have described before).

4.1. Previous observations made regarding ME

In previous work (Howard & Messum, 2007, 2011, 2014; Messum, 2007) we have argued in favour of an ME account in which the judgment of equivalence is made by the adult, rather than by the child. This makes speech sound development more straightforward for the child than in SBE accounts: he has less cognitive work to perform, his perceptual system need not be as developed, and he does not need to solve the normalisation problem (whether via innate brain mechanisms or otherwise). We also discussed and referenced studies which provide the following data that is either needed in support of the ME account or is consistent with it (Howard & Messum, 2014; Messum, 2007; Messum & Howard, 2012).

- The reformulation exchanges demanded by an ME account have been documented in studies of natural interaction between caregivers and children from 2 to 21 months.
- Even with Elija, a non-human model, we found that subjects readily responded to his potential speech sounds and that when they did so, this was almost always with a reformulation of his output into tokens of L1.
- There is a parallel to be found in nature, where acoustic matching is not always the mechanism by which complex vocal behaviour is learnt. In some birds, song development is not imitative, for example in the American cowbird (which is a brood parasite like a cuckoo). Instead, non-singing female cowbirds tutor young males by means of beak gapes and wing strokes. This reinforcement feedback improves the young males' performance (Smith, King, & West, 2000; West & King 1988).
- Experiments with human infants show that caregiver behaviour is perceived and acted upon by young learners to generate more advanced forms of vocalisations. Contingent social feedback creates changes in babbling, increasing its vocal quality in ways that reflect the specific maternal behavior (Goldstein & Schwade, 2010; Locke, 1993, p. 163).

Relevant to the first point are some statistics from Pawlby (1977a) that we have not previously discussed. Pawlby recorded 8 dyads for 10 min per week for 26 weeks starting with children at 4 months old, coding for all types of imitation. She found that: (a) 48% of the imitations made by the mothers were of speech-like sounds. (b) Vocal imitation by the mother was much more common than imitation by the infant. Mothers imitated an infant speech-like sound 625 times (on average once every 3.3 min) and infants imitated a mother's speech sound 60 times (once every 34.7 min). (c) Within the 625 episodes when a mother imitated an infant's speech-like sound, the infant maintained the interchange with a second vocalisation 199 times. The interchange contained 3 or more infant vocalisations 81 times. These figures demonstrate the high frequency of infant–caregiver interactions that can potentially solve the correspondence problem. Masur and Olson (2008) present complementary data about sustained exchanges for the period from 10 to 21 months.

Messum and Howard (2012) discuss data presented by MacDonald, Johnson, Forsythe, Plante, and Munhall (2012) that bears on the different predictions about the self-regulation of sound production that auditory SBE and ME accounts would make. To perform auditory matching of speech sounds, young children would have to monitor their acoustic output. In a speech feedback alteration study, MacDonald et al. found that toddlers (mean age 2 years 6 months) did not self-regulate, whereas young children (mean age 4 years 3 months) did, but not as effectively as adults. The absence of self-regulation in toddlers seems inconsistent with the SBE mechanism for speech sound learning. However, within an ME account, an infant is not required to use his own acoustic output and therefore need not monitor it.

An ME mechanism is consistent with the results of natural and experimental studies of the "fis/fish" and similar phenomena, in which a child perceives a word correctly but pronounces it incorrectly without any realisation that he is doing so (Locke, 1979; Priestly, 1980). This is only puzzling if it is assumed that a child must monitor his own acoustic output to improve production. If, on the other hand, the production develops using ME, then the phenomenon is an example of a persistent use of an erroneous speech sound correspondence that has been accepted by those around the child in the past. The persistence may be surprising in the face of the explicit contrary evidence being provided, but acting on this would involve changing an existing association, something that does eventually happen since children do act to rectify their pronunciation errors of this type. Menn, Markey, Mozer, and Lewis (1993, p.

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

430) point out that erroneous pronunciations should result in error correction in [self-] supervised learning accounts, but that error correction in practice only occurs in response to differential reinforcement, as expected in the ME account.

The children referred to above, who are apparently not monitoring the mispronunciation of sounds in words, are older than the infants on which the ME account focusses. However pronunciation develops throughout early childhood, so if auditory matching is the mechanism used for solving the correspondence problem these children should also be self-monitoring. On the other hand, with respect to the requirements of an ME mechanism over the entire period during which pronunciation is learnt, Chouinard and Clark (2003) demonstrate that reformulations of all types, including phonological reformulations, are plentiful until at least the age of four.

4.2. Learning action–sound correspondences from mirrored interactions

The reports on the development of affect described in Section 2.3 suggest that infants learn from mirrored interactions with their caregivers. At a later age we propose that infants learn action to sound correspondences in the same way. To illustrate this and to address the question of what aspect of his activity the infant associates with the caregiver's reformulations, consider a thought experiment involving clapping, an activity that is analogous to speech in that it involves a motor activity that produces sound.

Imagine an infant who claps his hands. His mother responds by saying "boo". The child performs the action again, and the mother responds the same way. It is clear that an initial association may quickly be built, which the child may have the opportunity to test and strengthen on other occasions. Later, he hears his mother say the word 'boot', in a context of shared attention. Recognising 'boo' within this, he knows that there is something that he can do that she will take to be equivalent to (part of) what he has heard her say; so he claps his hands.

In this particular example, his mother will probably not understand what he is doing. She may not connect her "boot" to her previous "boo" and the clapping game. However, if his initial action had been a vocal gesture, a VMS rather than a clap, which produced a sound which she interpreted as /buː/ within L1 and reformulated as "boo", then she will now hear the same sound again and might well understand that her child is referring to the object within their field of shared attention. She may signal her approval, and thus reinforce the vocal action he performed, which produces what he will come to understand to be a word.

In this example the mother and child's perspectives of what occurs during the interaction may differ. It is probable that the mother would conceive the child's clapping as a sound-making activity; particularly if, for some reason, the mother could only hear the result and not see the action of clapping itself. This is likely to be the situation with a VMS performed during vocal development, which caregivers will conceive in terms of its output. The child, on the other hand, may not yet have mastered clapping to the point where it is automatised. Clapping still requires some of the child's attention, and it is likely that the action would be a more vivid aspect of the experience to him than the sound it produced. So the fact that his mother responded vocally, by saying "boo" to his clap, does not mean that the child's association would be between the noise originally made by his clapping and this sound. It seems likely, instead, that he would associate his action, the clapping movement, with the speech token she has produced.

In vocal development, the possession of a VMS means that an infant can produce a sound reasonably reliably, but it does not mean that this is automatic to the point of requiring none of his attention. Rather, his primary sense of "what he does" is likely to be his vocal gesture, not the sound output that he or an adult hears as a result. Supporting this notion, Locke (1986, p. 245), Kent (1992, p. 84), Davis, MacNeilage, and Matyear (2002, p. 102) and others have argued that the representations or phonetic plans of children for first words are motoric in nature, rather than acoustic. In our ME account, therefore, we posit that the child is primarily conceiving his activity as motor rather than sensory.

4.3. Are ME interactions universal?

Ochs and Schieffelin (1984) reported that adults only interact minimally with their pre-linguistic children in Kaluli and Samoan cultures. This appears to undermine the universality of an ME account; however we make two observations in response.

Firstly, Zukow-Goldring (1996, p. 208, footnote 1) notes that "older siblings take care of younger sisters and brothers during much of the day in agrarian societies … [where] sibling caregivers sensitively adjust what they say and do so infants can understand them." Ochs and Schieffelin (1994, p. 484) state that older siblings help to care for Kaluli 6–12 month-olds, and that adults and older children occasionally repeat vocalisations back to 12–16 month-olds. Older siblings also take care of Samoan infants. See also similar comments to ours made by Messer (1994, p. 229) and Ramsdell, Oller, Buder, Ethington, and Chorna (2012, p. 1636).

Secondly, only a small number of imitative exchanges may be sufficient for an infant to form the correspondences required for the ME mechanism to operate. As Cook et al. (2014, p. 186) point out, "A common misconception about associative learning is that it always occurs slowly. … [S]tudies show that, when the contingency is high, infants can learn action-effect associations in just a few trials."

4.4. SBE vs. ME: determining the mechanism by which the correspondence problem is solved

As we described in Section 2.2 (and illustrated in Fig. 2), almost all theories of how a child solves the correspondence problem for L1 pronunciation rely on an SBE mechanism: they suppose that he determines equivalence by making a judgement of similarity between what he perceives and what he produces or does. Within this class of accounts, the child might be actively working towards matching an auditory target, he might be discovering that his output is auditorily similar to what is already in use by others, or he might be making the judgement of similarity between his perception of vocal gestures rather than between acoustic tokens. The self-supervised nature of the process, however, has not been a contentious point, and we are not aware of it having been defended and

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

justified. On the other hand, we have described potential problems with SBE accounts in Section 2.2, and these are further discussed in Messum (2007).

Since we have been discussing mirroring behaviour on the part of caregivers, it is worth noting that SBE accounts have attributed no role to it. It might, though, function as 'assisted imitation' (Zukow-Goldring & Arbib, 2007), by preparing and encouraging infants to imitate vocal output by modelling this behaviour to them.

Another potential effect of mirroring behaviour is to reinforce the infants' output of tokens that are close enough to L1 to be reformulated by their caregivers. Twenty years ago, Vihman (1996, p. 118) reported that "the role of social context in facilitating advances in vocal production is intriguing but unresolved." However, it is now well established experimentally that the vocal responses of an infant's social partners are used by the infant to shape his output (Goldstein & Schwade, 2010; Hsu & Fogel, 2001, p. 104; Pelaez, Virues-Ortega, & Gewirtz 2011).

We believe that reformulation goes beyond encouraging vocal development and has phonological significance. The child gets more from a vocal reformulation of his output than he would from a non-linguistic reinforcement like a smile of approval. In principle, reformulations of VMSs into well-formed tokens of L1 allow the child to solve the correspondence problem. We gave reasons to believe that this happens in practice earlier in this section. We now suggest how this might be better established in the future.

1. Ideally, the issue would be addressed with direct experimental work undertaken within child phonology. However, a longitudinal design would be needed, perhaps extending over many months. In practice, this would lead to sparse sampling and the risk, therefore, of key interactions being missed, unless massive data collection and automated analysis techniques were employed, as pioneered in the Speechome project (Roy, 2009). Even then, it is not certain that the learning moves being made by the child would be identifiable by observers.
2. The work of the Asada group as well as our own work with Elija demonstrates that cognitive developmental robotics can help test theories of speech acquisition. Asada (2015, p. 251) points out that a constructivist approach using computational modelling enables researchers to, "generate completely new understanding through cycles of hypothesis testing and verification, targeting the issues that are very hard or impossible to solve under existing scientific paradigms."
3. Gattegno arrived at his insights into child speech development by applying his model of learning to the world of infants. This was work done in parallel to his main professional activity as an educationalist, where he developed and applied his model to the teaching of mathematics, literacy and foreign languages. Further progress on how infants learn, including how they co-construct communication with their caregivers (Lock & Zukow-Goldring, 2010), will certainly shed light on their possible paths of phonological development.
4. Finally, we can look at the coherence of any hypothesis by examining it within wider contexts: the rest of child phonology, the final form of L1 phonology, and related fields such as psychology and neuroscience. This will lend support to one hypothesis or another by the principle of 'inference to the best explanation' (Lipton, 2004), or the explanatory power of each. We take this approach in Sections 5–7.

5. A Mirrored Equivalence (ME) account of how production and perception develop

An ME account would not be plausible unless it could form part of a coherent account of the development of speech production and perception in a child. To describe this, we first need to consider one fundamental aspect of auditory perception and to clarify the nature of mimicry, since this is a mechanism for recreating the form of a word that is available to a young child.

5.1. Awareness of sensation (AS) and meaningful perception (MP)

Events in the world that impact our senses create two flows of information. We normally attend to the event itself, the distal cause of the stimulus, but we can also attend directly to the effect that the stimulus is having on us, "the pattern of sensory stimulation" (MacKay, 1987, p. 65). Thomas Reid (1785) described the senses as having, "a double province – to make us feel, and to make us perceive." Humphrey (1992) describes the history of this understanding, particularly with respect to vision.

Öhman (1975) described the sensory consequences of sound as being the effect it has on the listener's "awareness of the developing state of his listening sense", contrasting this with what he called 'ordinary perception' which recovers meaningful events from the signal. For speech perception, this duality has been of limited interest because the listener is generally concerned with meaningful objects in the input. However the issue has been addressed: Durlach and Braida (1969) described the distinction as being between two modes of auditory attention, 'sensory-trace mode' and 'context-coding mode', and Pisoni (1973) used the terminology of 'auditory mode' and 'phonetic mode'. There is no consensus on the terminology and we use the terms 'awareness of sensation' (AS) and 'meaningful perception' (MP) to maintain the link with other sensory modalities.

In normal life, and while listening to speech, we are normally in our MP mode and it can be difficult to switch to AS (Repp, 1984, p. 321). Bruner, Goodnow, and Austin (1956, p. 50) described a common observation, that, "having learned a new language, it is almost impossible to recall the undifferentiated flow of voiced sounds that one heard before one learned to sort the flow into words and phrases." When there is something in the signal that can be recognised, we are drawn to do so. Once in our MP mode, it may not be possible to simultaneously attend to the signal in AS mode (Linell, 1982, p. 67), although we can, of course, recognise multiple events on different planes of understanding within MP (Werker & Curtin, 2005).

The distinction is necessary for thinking about infant speech development because (1) infants do hear words while they are still speech 'noises' to them rather than being strings of speech sounds, and (2) mimicry allows them to recreate words before they have

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

the ability to reproduce words by the mature mechanism of recognising and concatenating speech sounds. In this paper, we keep different processes of imitation clearly distinct by using 'recreate' and 'reproduce' as in the previous sentence. The implications of the AS and MP modes of attention are described in Section 6.1 by reference to Fig. 5.

5.2. Mimicry

Call and Carpenter (2002) describe mimicry as B copying the form of A's actions without B adopting A's goal or intending to achieve the results A obtains. If the defining feature of mimicry is 'copying the form of an action' then this distinguishes it from all the forms of purposive copying that are also called 'imitation'. In normal 'copying' and emulation we want our actions to achieve something in themselves.

While not disagreeing with Call and Carpenter, we think it is more insightful in the present context to describe mimicry as creating a signal which perturbs an observer's sensory apparatus in a way that resembles the perturbation caused by the signal from the target behaviour (hence 'impressionist' as a synonym for 'mimic'). Mimicry is possible because we can attend to a signal in our AS mode of perception not just in our MP mode. We can also recognise the resemblance between present and earlier experiences of this type in the same way as for other experiences, and mimicry is the name given to the deliberate activity that leads to such recognition of resemblance for sensory perturbations.

Mimicry, then, is concerned with recreating surface form/sensory perturbation, and it is performed by driving an Auditory Inverse Model. This inverse model starts being developed when a baby first makes and listens to any noises for himself (including crying), and the model is elaborated for at least the rest of childhood. To drive the inverse model, the user must have an auditory target in mind to copy, either held in short term memory, or evoked from longer term memory. This is run through the inverse model thereby generating a vocal gesture and mimicked output. Note that if we have not committed a sound image to long-term memory, we may find that we can mimic something in the moment but later be unable to repeat this.

Although the analogy between speech and writing is imperfect in some important respects, it is helpful to compare and contrast two forms of spoken word form adoption (recreation by mimicry and reproduction by concatenating sequences of VMS's) with two forms of graphical word adoption (recreation by drawing and reproduction by writing strings of letters). A child may draw his first words (starting, perhaps, with his own name), until he learns how to form letters, at which point he can write words. 'Script-drawing' (Adi-Japha and Freeman, 2001) enables him to get into the written medium, but 'script-writing' is the only long-term approach to writing that is viable.

Infants can and do recreate sound images by mimicry, but the need to evoke and match a target means that this is an attentionally demanding way of saying words. It may enable infants to get into word form recreation and meaningful word form use, but it is not a viable way for speech to develop.

5.3. Steps in the development of production and perception

We now present an integrated account of child production and perception that incorporates the ME mechanism. The adoption of word forms from L1 starts with the child mimicking words and sometimes phrases that he hears. For this, he uses the Auditory Inverse Model he developed through associating his own vocal activity to its acoustic consequences (Howard & Huckvale, 2005a, b).

However, learning the pronunciation of new words only becomes efficient when he makes use of a Speech Sound Inverse Model, which (bi-directionally) maps discrete caregiver speech sounds he has heard previously to the movements he needs to make to produce their equivalents in his own voice. Unlike the Auditory Inverse Model, the development of this Speech Sound Inverse Model depends upon social interaction, namely the mirroring of the child's speech behaviour by his caregivers.

The sequence of events starts with babbling and ends with the child able to learn the pronunciation of new words efficiently, as shown schematically in Fig. 4.

1 An infant's experimentation (during the periods up to, including and after babbling) gives him an increasingly sophisticated Auditory Inverse Model for vocal production. This enables the mimicry (recreation) of noises heard in the environment.

2 The infant develops increasing skills in processing speech input for comprehension. He becomes familiar with particular word forms and their meanings by acquaintance (arising from repeated exposure to them in context), perhaps from as early as 6 months (Bergelson & Swingley, 2012).

At this stage, word perception need not be fine grained, as it only has to serve recognition. It does not need to retrieve categories of sounds that will later inform word production. Speech perception in children, as in adults, can be 'multi-modal' (Hawkins, 2003) and also make use of distinctions found at any level of granularity. 'Portions' and 'parts' (Kuhl, 1987, pp. 351–355) of the signal are perceived to differ between utterances heard, but the differences are found in the patterning of the acoustic signal rather than in the recognition of linguistic elements within it (Warren, 1999, pp. 169–173; Walley, Metsala, & Garlock, 2003).

3 Caregivers play imitative games with their infant during which they reformulate the output from his motor vocal actions into well-formed tokens of L1. As portrayed in Fig. 3, this mechanism of Mirrored Equivalence solves the correspondence problem for him; he learns bi-directional equivalence pairings between some of his VMSs and L1 speech sounds he has heard made in response to them.

4 Many infants at around 10–12 months (Ferguson, 1978) produce vocables (or protolanguage 'words') that are not based on L1 forms (e.g. Halliday, 1975, p. 9). Separately from this, most infants start to 'adopt' adult-modelled words from L1 between 10 and 13

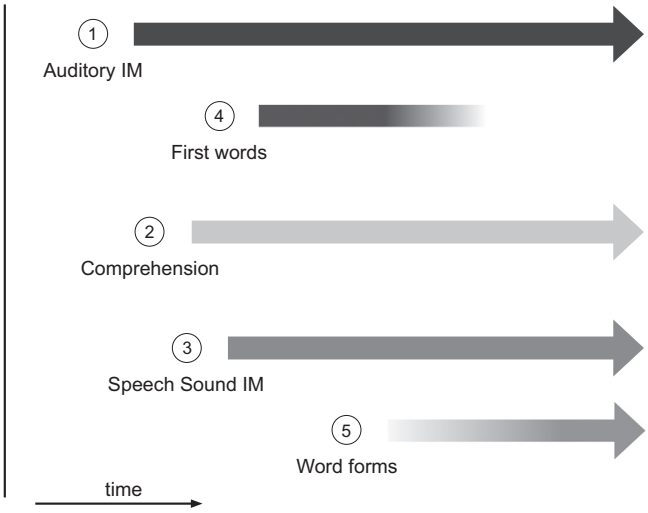

Fig. 4. Initial independence of three separate speech development processes. Numbers in circles refer to the paragraph numbers in the main text: (1) development of Auditory Inverse Model, supporting (4) first adopted L1 word forms recreated through mimicry; (2) speech comprehension; (3) development of Speech Sound Inverse Model, supporting (5) adopted L1 word forms reproduced through serial imitation of speech sounds. Time shown from birth to around 2 years of age.

months. These are recreated by the child to the best of his ability through mimicry, using the Auditory Inverse Model, as whole word forms (Ferguson & Farwell, 1975). Many researchers have commented on how learning first words is difficult and slow (e.g. Snow, 1988, p. 348).

Learning the pronunciation of words this way is not well adapted to wholesale L1 word form adoption because mimicry of a form heard previously requires that the child evoke a holistic model in order to recreate it. (Alternatively, or additionally, it is not suited for wholesale word adoption for the reason given by Kent (1981, p. 179) that “the child is forced to a segmental (phonetic) motor organization through sheer force of economy and manageability.”)

5 In interchanges with his caregivers, the young child’s normal attentional set towards words being said to him has been that of MP for some time; he is trying to retrieve meaning from them. As Menn (1983, p. 39) pointed out, “Language is usually used, not contemplated; children expect to listen for meaning, not for sound."

This attentional set allows him to recognise those elements within the words which have become vocal objects in their own right. So a second route to word production presents itself. He recognises portions of the speech signal that form part of the inventory of equivalence pairs formed by ME. He tries out the corresponding motor vocal gestures, and is successful at approximating the pronunciation of a word or phrase (as demonstrated by Elija). This route to word production is highly efficient. No evocation of sound images is necessary; the child can encode words as sequences of gestures, something at which he is practised and expert.

6 The accuracy of the words reproduced this way will depend upon the quality of the speech sound equivalences previously learnt, which may initially be poor. It is therefore unsurprising that first words learnt by mimicry will often be closer to L1 word forms than the early words learnt by this second route. Thus the later adoption of this route explains the appearance of a “regression” (Ferguson, 1978) in the accuracy of word forms, also discussed in the literature as “U-shaped” development curves (Vihman, 1996).

In the practical business of learning to pronounce words, other processes are also important: word selection (e.g. Vihman, 1993), the adaption of word forms to a child’s current effectivities (e.g. Vihman & Croft, 2007), the perception and production of correct accentual patterning (e.g. Vihman, Nakai, DePaolis, & Hall, 2004), and so on. But we propose that one central process, that of developing the mature skill of learning the pronunciation of words, is as described above.

6. The cognitive nature of phonological units

The various SBE accounts of how children solve the correspondence problem for speech sounds imply different structures for the underlying or neural representation of speech. If a child judges similarity between the acoustic images he hears and produces, this suggests the underlying representation will be acoustic. There are other arguments to believe that this is the case (e.g. Coleman, 1998). If the judgement of similarity is made between gestural images, this suggests the underlying representation will be motoric. Again, there are arguments to support this viewpoint (e.g. Goldstein & Fowler, 2003). However, such SBE accounts require that the underlying representation of speech is either acoustic or motor: that is, there is a single, primary modality within which the judgment of similarity made by the child takes place.

In contrast, an ME account describes the direct association of a child’s vocal motor scheme with a caregiver speech sound heard in response, implying an intrinsically perceptuo-motor (PM) unit as the underlying representation for speech sounds. The motor and auditory components of this PM unit are developed independently. The former by discovery during the babbling phase (and later by similar articulatory exploration), and the latter from the categorisation of the contiguous and contingent acoustic input from other

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

speakers heard during imitative interactions. A direct association between production and perception is made. This PM unit reflects the nature of speech: motoric in production and auditory in perception.

The proposed PM representation could be instantiated within the so-called human 'mirror system' (Hickok, 2014, p. 20), since it would be active both when hearing and producing what are, linguistically, 'the same' speech sounds within a given language. There are two leading accounts for how the mirror system is created, the 'genetic' and 'associative' accounts (Heyes, 2013). We find the evidence for the latter hypothesis more compelling, but it is not necessary for the ME account that the mirror system should arise one way or the other. If the mirror system is innate, then it could presumably be adapted for speech and primed in the ME interactions we have described. If the mirror system develops through experience, due to correlated sensory and motor activity, then the mechanism we have described for speech sounds operates in the way that one leading account describes the general development process (Cook et al., 2014, p. 181). While our theoretical perspective on the infant learner credits him with awareness and the mental capacity to come to judgements about the equivalence of his caregiver's activities and his own during imitative games, the Cook et al. description of learning is less demanding. For them, simple association is all that is necessary for learning when there is correlated excitation of sensory and motor neurons. The result is that motor neurons become mirror neurons (Heyes, 2010).

6.1. *Three routes available to a speaker for saying words*

Within our account, Fig. 5 summarises the three routes available to a speaker for word reproduction (going via the Speech Sound Inverse Model) and recreation (mimicked as a whole-word form via the Auditory Inverse Model).

1 If the speaker can recognise the whole or any parts of the input, his attention will normally be drawn to one or more of these percepts; he will be in the MP mode of perception.
2 However, during early speech acquisition, an infant may not yet have developed either the speech sound equivalences needed or have a sufficient understanding of word structure to parse a word he hears in terms of PM units. The input will instead be experienced as a noise or noises without linguistic structure, with the infant operating in the AS mode of perception. The image recovered can drive the general Auditory Inverse Model (route 3 in Fig. 5, described as "*Mimicry* …"). Generally the first 15–25 words that are adopted from L1 by the infant are recreated this way, but "are dependent on the work done by the adult, are co-constructed and negotiated as communicative acts by adult and child" (Snow, 1988, p. 349). As well as first words, phonological idioms are mimicked in this way. This is also the route that enables some forms of echolalia.
3 Conversely, the first time a new word is encountered at a later stage of development, it will not be recognised (matched with a form in the word form recognition lexicon) but it can be parsed for its component speech sounds, and the form added to the output phonological lexicon structured in PM units (route 1a). This will create its first representation in this lexicon. Not shown are the additional associations between the child's semantic hypotheses about the word and the word's new representations in the input and output lexicons. Subsequent encounters with the word form will be both recognised and parsed, and will improve the stored output representation. In normal speech, the word will be reproduced using the pathway starting with Expression (route 1b).

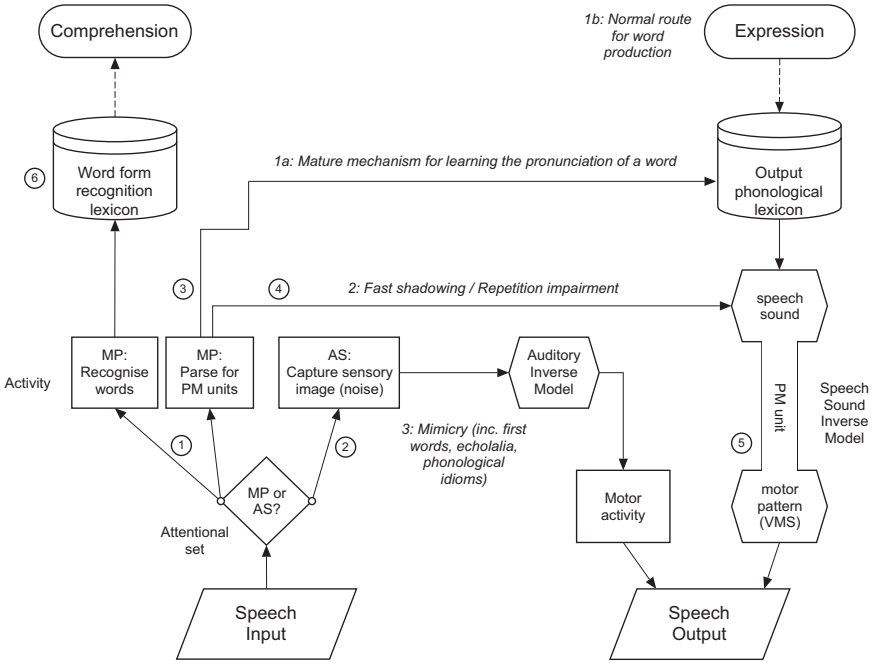

Fig. 5. Three routes to word production and word recreation, labelled in italic text. Numbers in circles refer to the paragraph numbers in the main text. (*Abbreviations*: MP – meaningful perception, AS – awareness of sensation, PM – perceptuo-motor, VMS – vocal motor scheme.)

4 As described shortly in Section 7.1, during fast shadowing tasks some subjects can suppress the word recognition process and drive the Speech Sound Inverse Model directly, via route 2 in Fig. 5. This is also the pathway that we propose is used by subjects in Choice Reaction Time experiments with speech material, and by patients who achieve word repetition despite semantic processing impairments. The normal lexical route to word repetition would be via Comprehension and subsequent Expression.
5 The Speech Sound Inverse Model has been created by the ME process. Since vocal gestures (VMSs) were directly associated to sounds heard in response, the perceptual and articulatory information about speech sounds is stored together.
6 Note that the Word Form Recognition Lexicon corresponds to what is often labelled the 'Input Phonological Lexicon' or similar in other accounts. However, in an ME account the representations stored here need not have any internal phonological structure. They will have 'portions' and 'parts' (Kuhl, 1987, pp. 351–355) that distinguish them from each other, but the speaker need not conceive of these as phonetic concepts like phonemes or distinctive features.

6.2. *A further potential problem with the auditory SBE account*

Having now drawn and illustrated the distinction between the AS and MP modes of auditory attention, we can present another reason why auditory matching to target may not be a mechanism with which an infant can solve the correspondence problem. To both understand a word and to identify a speech sound within it, he must attend to the signal as informing him of meaningful events, in MP mode. But as speech is ephemeral, the opportunity to then attend to it as a sensory experience, in order to recreate the speech sound elements within it, disappears (Linell, 1982, p. 67). Adults demonstrate this when learning foreign language pronunciation by asking for problematic words to be said to them again. Knowing what they are about to hear they can set themselves to deliberately listen to the sound of the word, and then attempt to recreate the sounds within it. Clearly infants do not have the capacity to engineer this kind of presentation.

7. Implications of Mirrored Equivalence (ME)

We believe that an ME account can reconcile the evidence that speech is "gestures made audible" (Stetson, 1951) with the evidence that speech is best characterised as an acoustic code (e.g. Coleman, 1998). As a result of being learnt by the ME mechanism, the representation of speech is both motoric and auditory at the same time. With this understanding, we now examine two problematic areas in speech research and show how an ME account explains the data.

7.1. *Repetition impairment and speech shadowing*

Support for the 3-route model of word reproduction and recreation portrayed in Fig. 5 comes from word repetition studies with aphasic patients and from speech shadowing experiments.

The normal processing route for word repetition is verbal comprehension followed by semantic/phonological transcoding (a process which would pass from the Comprehension box to the Expression box in Fig. 5). Researchers including McCarthy and Warrington (1984), Hanley, Dell, Kay, and Baron (2004) and Nozari, Kittredge, Dell, and Schwartz (2010) have reported on aphasic patients who lack this route. They argue that the success of their subjects in repeating speech material despite their impairment demonstrates the existence of an alternative pathway for repetition that McCarthy and Warrington called the 'non-semantic route' and described as performing simple auditory/phonological transcoding. This function would be performed by route 2 in Fig. 5. McCarthy and Warrington (p. 482) commented that the "biological necessity and *modus operandi* of dual processing routes in speech production remains obscure." Our account explains how the non-lexical route develops as a by-product of the ME mechanism.

McLeod and Posner (1984) provided evidence for the same route being used in speech shadowing, a "privileged loop" which allows an articulatory programme to be activated by aural words. Marslen-Wilson (1985) summarised a series of experiments with longer stretches of speech, which identified so-called 'close shadowers', who are able to reduce the delay between hearing and production to 250 ms or less. The difference between close shadowers and 'distant shadowers' (latencies averaging over 500 ms) was that the former were able to speak before they were fully aware of what they were hearing. This both supports the existence of our route 2, and supports the underlying representation of speech being in PM units, which specify the articulation of words directly, helping to make this fast shadowing possible.

The latencies involved in route 2 have been investigated more directly in reaction time experiments. As a general principle, simple reaction time tasks that involve detecting stimulus change are performed faster than those that require making a choice (Luce, 1986, p. 208). This is because the former only involves detection of a stimulus change whereas the latter requires that a choice of response be made depending on the change in stimulus identity. However, data from speech shadowing experiments (e.g. Porter & Lubker, 1980; Porter & Castellanos, 1980) show that choice reaction times with speech material are not significantly longer than simple reaction times.

Fowler, Brown, Sabadini, and Weihing (2003) have extended and confirmed the earlier results on speech shadowing. To explain them, they favoured either a Motor Theory or Direct Realist account of perception. They proposed that a subject perceives the vocal tract gestures of the model and is thus informed of the articulation required in the choice reaction time test as part of his process of perception. This takes place without the need for an additional stage of processing after recognition of an auditory category. However, they acknowledged that an "augmented" acoustic theory might also explain their initial data, if it allowed for articulatory properties as well as acoustic ones to be associated with phonological categories.

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

In other words, information for producing a token of the category would be part of the perceptual category. Discussing this possibility, Shockley, Sabadini, and Fowler (2004, p. 422) said that they were unaware of any theory in which this was actually proposed. However, the ME account proposes exactly this, and would therefore be able to account for data which argues against the underlying representation of speech being auditory.

7.2. The PM unit implies a 2-lexicon model

It is a natural implication from the comparison mechanism in SBE accounts that a child develops a single phonological lexicon (or stream of processing) serving input and output, based on either auditory or motor representations. One implication of the ME account is, instead, that separate input and output lexicons develop.

Within the field of child phonology, both 1- and 2-lexicon models have been developed (Baker, Croot, McLeod, & Paul, 2001; Menn & Matthei, 1992; Menn, Schmidt, & Nicholas, 2013). However the output lexicon in the ME account functions differently to those in previous 2-lexicon proposals, being structured as PM units that are used for both parsing the input and for word reproduction. Support for this aspect of our model comes from two other disciplines.

Firstly, psycholinguists point to the very different nature of perception (recognition of a word-form and recall of a concept) and production (recognition of concepts and recall of word-forms). They argue that this implies two very different forms of phonological knowledge (e.g. Clark & Malt, 1984, p. 200; Huttenlocher, 1974, p. 335; Straight, 1980). Second, there is evidence for dissociable routes of phonological processing from patients with neurological disorders. For example, Jacquemot, Dupoux, and Bachoud-Lévi (2007) report on a patient with conduction aphasia who has no production, perception or real-word repetition deficits but performs badly on pseudoword repetition tasks. They argue that this cannot be accounted for by a model with just a single phonological code serving both perception and production, but requires two separate but connected phonological codes, one in perception and one in production, as argued for by previous neurological research. See Martin, Lesch, and Bartha (1999) for a fuller review of earlier evidence for this.

8. Discussion and conclusion

The cognitive nature of phonological units has been the subject of debate for many years. For example, Sapir (1921, p. 17) argued that the auditory aspect of speech was primary; Stetson (1951) that speech is “gestures made audible.” Other proposals have been made (Nolan, 1990), and the conflicting evidence from phonetics, the other speech sciences and other disciplines has not yet been reconciled.

We approach the problem developmentally. As Tinbergen (1963) argued, an ontogenetic perspective – along with consideration of cause (mechanism), adaptive value and phylogeny – is necessary in order to gain a comprehensive, coherent understanding of any biological behavior. The ontogenetic perspective is equally important for phonology (Ferguson & Farwell, 1975, p. 437; Lindblom, 2000). Our proposal is a departure from previous ones because it questions the assumption that children solve the correspondence problem for speech sounds by imitating their caregivers. Instead, we have argued that the mechanism they use is more likely to be Mirrored Equivalence (ME), based on the reformulation by caregivers of children’s vocal output into L1 tokens during imitative exchanges. From this evidence of equivalence (or from simple association arising from an experience of correlated motor and sensory events) a child would form direct mappings between his vocal motor schemes (VMSs) and the tokens of L1 presented to him, which are a linguistic interpretation of his output. This would create an underlying representation of speech that is intrinsically perceptuo-motor, matching the nature of speech itself: motor in production, auditory in perception. The discovery of mirror neurons in primates and the deepening of our understanding of the human mirror system show that from a neurological perspective there is nothing implausible about such a proposal.

The ME account invokes well-attested interactions on the part of children and caregivers. It involves either a learning task for the child that appears to be well within his competence or, more cautiously, a context that is highly favourable for associative learning. It avoids the problematic aspects of accounts that involve acoustic matching by a child, particularly the normalisation problem. It simplifies the child’s task by giving a significant role to his caregivers, including them making the judgement of similarity between the results of what he does and what they say.

If ME solves the correspondence problem and the underlying representation of speech is perceptuo-motor, then this would be of obvious importance to phonetics, the speech sciences and the teaching of L2 pronunciation. This new understanding would illuminate theoretical issues and inform therapeutic interventions and pedagogical practice: existing approaches which emphasise motor system experimentation combined with feedback on performance given by the clinician or teacher would be given a strong theoretical justification. In fact, a basic change in the paradigm has even wider implications than in these fields, since speech is also studied in psychology, neuroscience and other disciplines.

The integrated account of production and perception development in a child that we have presented in Sections 5 and 6 is not a necessary implication of the ME mechanism, but it is a natural development from it. It must be possible to create such an account from any proposed solution to the correspondence problem, and we have shown how ours is consistent with the developmental data and resolves some longstanding problems in speech development (“fis/fish”, phonological idioms, ‘U’-shaped developmental curves). As described in Section 7, the account also resolves problems with adult experimental data that are hard to explain under auditory SBE accounts, including data from speech shadowing, repetition impairment and choice reaction time experiments. It is consistent

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

with data from psychology and neurology which argue for the existence of input and output phonological lexicons, while SBE accounts imply the development of just a single lexicon.

In conclusion, there is no decisive evidence against the Similarity Based Equivalence (SBE) accounts of how a child solves the correspondence problem for speech sounds and in favour of a Mirrored Equivalence (ME) one, but we have argued that the ME account provides the best fit with the current data. At this point, therefore, we claim that the ME account has what Dewey (1941) called 'warranted assertability' and that it justifies the critical attention of researchers in child phonology and related fields. The major implication of the ME mechanism for Phonetics is that the cognitive form for phonological units would not be either auditory or motor but intrinsically perceptuo-motor. This would allow for straightforward and economical explanations of some phenomena in speech and speech acquisition that have been problematic for many years.

Acknowledgements

We are grateful to the two reviewers and to Roslyn Young and Susan Attwood, all of whose perceptive comments and suggestions helped us to improve the manuscript.

References

Adi-Japha, E., & Freeman, N. H. (2001). Development of differentiation between writing and drawing systems. *Developmental Psychology*, *37*(1), 101–114, http://dx.doi.org/10.1037/0012-1649.37.1.101.

Asada, M. (2015). Toward language: vocalization by cognitive developmental robotics. In G. Cheng (Ed.), *Humanoid robotics and neuroscience: Science, engineering and society* (pp. 251–274). Boca Raton, FL: CRC Press Taylor & Francis Group.

Baker, E., Croot, K., McLeod, S., & Paul, R. (2001). Psycholinguistic models of speech development and their application to clinical practice. *Journal of Speech, Language and Hearing Research*, *44*, 685–702, http://dx.doi.org/10.1044/1092-4388(2001/055).

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

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Call, J., & Carpenter, M. (2002). Three sources of information in social learning. In K. Dautenhahn & C. L. Nehaniv (Eds.), Imitation in Animals and Artifacts (pp. 211–228). Cambridge, MA: MIT Press.

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

Clark, H. H., & Malt, B. C. (1984). Psychological constraints on language: A commentary on Bresnan and Kaplan and on Givón. In W. Kintsch, J. R. Miller, & P. G. Polson (Eds.), *Method and tactics in cognitive science* (pp. 191–214). Hillsdale, NJ: LEA.

Coleman, J. (1998). Cognitive reality and the phonological lexicon: A review. *Journal of Neurolinguistics*, *11*(3), 295–320, http://dx.doi.org/10.1016/S0911-6044(97)00014-6.

Cook, R., Bird, G., Catmur, C., Press, C., & Heyes, C. (2014). Mirror neurons: From origin to function. *Behavioral and Brain Sciences*, *37*(02), 177–192, http://dx.doi.org/10.1017/S0140525X13000903.

Davis, B. L., MacNeilage, P. F., & Matyear, C. L. (2002). Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. *Phonetica*, *59*, 75–107, http://dx.doi.org/10.1159/000066065.

Dewey, J. (1941). Warranted assertability, and truth. *The Journal of Philosophy*, *38*(7), 169–186.

Dodd, B., Holm. A., Hua, Z., and Crosbie, C. (2003). Phonological development: a normative study of British English-speaking children. Clinical Linguistics & Phonetics, 17 (8), 617–43.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception: A preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, *46*, 372–383, http://dx.doi.org/10.1121/1.1911699.

Ferguson, C. A. (1978). Learning to pronounce: the earliest stages of phonological development in the child. In F. D. Minifie, & L. L. Lloyd (Eds.), *Communicative and cognitive abilities*. Maryland: University Park Press.

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language*, *51*(2), 419–439, http://dx.doi.org/10.2307/412864.

Fowler, C. A. (2004). Speech as a supramodal or amodal phenomenon. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. MIT Press.

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, http://dx.doi.org/10.1016/S0749-596X(03)00072-X.

Fry, D. B. (1968). The phonemic system in children's speech. *British Journal of Disorders of Communication*, *3*, 13–19, http://dx.doi.org/10.3109/13682826809011436.

Gattegno, C. (1973). *The universe of babies*. New York: Educational Solutions.

Gattegno, C. (1985). *The learning and teaching of foreign languages*. New York: Educational Solutions.

Gattegno, C. (1986). A working model for health. *Educational Solutions Newsletter*, *16*(2).

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. In N. O. 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. (2010). From birds to words: perception of structure in social interactions guides vocal development and language learning. In M. S. Blumberg, J. H. Freeman, & S. R. Robinson (Eds.), *The Oxford handbook of developmental and comparative neuroscience* (pp. 708–729). OUP.

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, http://dx.doi.org/10.1016/j.bandl.2005.06.001.

Halliday, M. A. K. (1975). *Learning how to mean*. London: Edward Arnold.

Hanley, J. R., Dell, G., Kay, J., & Baron, R. (2004). Evidence for the involvement of a nonlexical route in the repetition of familiar words: A comparison of single and dual route models of auditory repetition. *Cognitive Neuropsychology*, *21*(2), 147–158, http://dx.doi.org/10.1080/02643290342000339.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3/4), 373–405, http://dx.doi.org/10.1016/j.wocn.2003.09.006.

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences*, *5*(6), 253–261, http://dx.doi.org/10.1016/S1364-6613(00)01661-2.

Heyes, C. 2010. "Mesmerising Mirror Neurons." Neuroimage 51: 789–91.

Heyes, C. (2013). A new approach to mirror neurons: Developmental history, system-level theory and intervention experiments. *Cortex*, *49*(10), 2946–2948, http://dx.doi.org/10.1016/j.cortex.2013.07.002.

Heyes, C. M., & Ray, E. D. (2000). What is the significance of imitation in animals?. *Advances in the Study of Behavior*, *29*, 215–245, http://dx.doi.org/10.1016/S0065-3454(08)60106-0.

Hickok, G. (2014). *The myth of mirror neurons: The real neuroscience of communication and cognition*. New York: W. W. Norton & Company.

Howard, I. S., & Huckvale, M. (2005a). Learning to control an articulatory synthesizer by imitating real speech. *ZAS Papers in Linguistics*, *40*, 63–78.

Howard, I. S., & Huckvale, M. (2005b). *Training a vocal tract synthesizer to imitate speech using distal supervised learning*. Patras: University of Patras In *10th international conference on speech and computer* (pp. 159–162).

Howard, I. S., & Messum, P. R. (2007). *A computational model of infant speech development*. Moscow: Moscow Linguistics University In *Proceedings of SpeCom XII* (pp. 756–765).

Howard, I. S., & Messum, P. R. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, *15*, 85–117.

Howard, I. S., & Messum, P. R. (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*(10), e110334, http://dx.doi.org/10.1371/journal.pone.0110334.

Hsu, H.-C., & Fogel, A. (2001). Infant vocal development in a dynamic mother–infant communication system. *Infancy*, *2*(1), 87–109, http://dx.doi.org/10.1207/S15327078IN0201_6.

Humphrey, N. (1992). *A history of the mind: Evolution and the birth of consciousness*. New York: Simon & Schuster.

Huttenlocher, J. (1974). The origins of language comprehension. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 331–368). Potomac, MD: LEA.

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

Jacquemot, C., Dupoux, E., & Bachoud-Lévi, A.-C. (2007). Breaking the mirror: Asymmetrical disconnection between the phonological input and output codes. *Cognitive Neuropsychology*, http://dx.doi.org/10.1080/02643290600683342.
Johnson, K. (2005). Speaker normalization in speech perception. In D. Pisoni, & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Oxford: Blackwell.
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, http://dx.doi.org/10.1111/1467-9450.00249.
Kent, R. (1992). The biology of phonological development. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: models, research, implications* (pp. 65–89). Timonium, MA: York Press.
Kent, R. D. (1981). Sensorimotor aspects of speech development. In R. N. Aslin, J. R. Alberts, & M. R. Peterson (Eds.), *Development of perception*, *Volume 1* (pp. 162–185). New York: Academic Press.
Kokkinaki, T., & Kugiumutzakis, G. (2000). Basic aspects of vocal imitation in infant–parent interaction during the first 6 months. *Journal of Reproductive and Infant Psychology*, *18*(3), 173–187, http://dx.doi.org/10.1080/713683042.
Kuhl, P. K. (1987). Perception of speech and sound in early infancy. In P. Salapatek, & L. Cohen (Eds.), *Handbook of infant perception*, *Vol 2* (pp. 275–382). New York: AP.
Kuhl, P. K. (1991). Perception, cognition, and the ontogenetic and phylogenetic emergence of human speech. In S. E. Brauth, W. S. Hall, & R. J. Dooling (Eds.), *Plasticity of development* (p. 79). Cambridge, MA: MIT Press.
Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences USA*, *97*(22), 11850–11857, http://dx.doi.org/10.1073/pnas.97.22.11850.
Lacerda, F. (2003). Phonology: An emergent consequence of memory constraints and sensory input. *Reading and Writing*, *16*, 41–59.
Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *The Behavioral and Brain Sciences*, *22*, 1–75, http://dx.doi.org/10.1017/S0140525X99001776.
Lindblom, B. (2000). Developmental origins of adult phonology: the interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica*, *57*, 297–314, http://dx.doi.org/10.1159/000028482.
Linell, P. (1982). The concept of phonological form and the activities of speech production and speech perception. *Journal of Phonetics*, *10*, 37–72.
Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London, New York: Routledge/Taylor and Francis Group.
Lock, A., & Zukow-Goldring, P. (2010). Preverbal communication. In J. G. Bremner, & T. D. Wachs (Eds.), *The Wiley-Blackwell handbook of infant development* (pp. 394–425). Oxford, UK: Wiley-Blackwell.
Locke, J. L. (1979). The child's processing of phonology. In W. A. Collins (Ed.), *Child language and communication: Minnesota symposium on child psychology Volume 12* (pp. 83–119). Hillsdale, NJ: LEA.
Locke, J. L. (1986). Speech perception and the emergent lexicon: An ethological approach. In P. Fletcher, & M. Garman (Eds.), *Language acquisition* (pp. 240–250). CUP.
Locke, J. L. (1993). *The child's path to spoken language*. Cambridge, MA: Harvard University Press.
Locke, J. L., & Kutz, K. J. (1975). Memory for speech and speech for memory. *Journal of Speech and Hearing Research*, *18*, 176–191, http://dx.doi.org/10.1044/jshr.1801.176.
Luce, R. D. (1986). *Response times*. New York: OUP.
MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P., & Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Current Biology*, *22*, 1–5, http://dx.doi.org/10.1016/j.cub.2011.11.052.
MacKay, D. G. (1987). The Organization of Perception and Action. New York: Springer Verlag.
Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. *Speech Communication*, *4*, 55–73, http://dx.doi.org/10.1016/0167-6393(85)90036-6.
Martin, R. C., Lesch, M. F., & Bartha, M. C. (1999). Independence of input and output phonology in word processing and short-term memory. *Journal of Memory and Language*, *41*, 3–29, http://dx.doi.org/10.1006/jmla.1999.2637.
Masur, E. F., & Olson, J. (2008). Mothers' and infants' responses to their partners' spontaneous action and vocal/verbal imitation. *Infant Behavior and Development*, *31*(4), 704–715, http://dx.doi.org/10.1016/j.infbeh.2008.04.005.
McCarthy, R., & Warrington, E. K. (1984). A two-route model of speech production. *Brain*, *107*, 463–485, http://dx.doi.org/10.1093/brain/107.2.463.
McCune, L., & Vihman, M. M. (2001). Early phonetic and lexical development: a productivity approach. *Journal of Speech, Language and Hearing Research*, *44*, 670–684, http://dx.doi.org/10.1044/1092-4388(2001/054).
McLeod, P., & Posner, M. I. (1984). Privileged loops from percept to act. In H. Bouma, & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 55–66). London: LEA.
Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*, 75–78, http://dx.doi.org/10.1126/science.198.4312.75 (7 October 1977).
Meltzoff, A. N., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting*, *6*, 179–192.
Meltzoff, A. N., & Williamson, R. A. (2010). The importance of imitation for theories of social-cognitive development. In J. G. Bremner, & T. D. Wachs (Eds.), *The Wiley-Blackwell handbook of infant development* (pp. 345–364). Oxford, UK: Wiley-Blackwell.
Menn, L. (1983). Development of articulatory, phonetic, and phonological capabilities. In B. Butterworth (Ed.), *Language production*, *Vol. 2* (pp. 3–50). London: Academic Press.
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.
Menn, L., & Matthei, E. (1992). The "Two-Lexicon" account of child phonology: Looking back, looking ahead. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 211–247). Timonium, MA: York Press.
Menn, L., Schmidt, E., & Nicholas, B. (2013). Challenges to theories, charges to a model: The Linked-Attractor model of phonological development. In M. M. Vihman, & T. Keren-Portnoy (Eds.), *The emergence of phonology: Whole-word approaches and cross-linguistic evidence* (pp. 460–502). CUP.
Messer, D. (1994). *The development of communication*. Chichester: John Wiley.
Messum, P., & Howard, I. S. (2012). Speech development: Toddlers don't mind getting it wrong. *Current Biology*, *22*(5), R160–R161, http://dx.doi.org/10.1016/j.cub.2012.01.032.
Messum, P. R. (2007). *The role of imitation in learning to pronounce* (Ph.D.). University College London.
Miura, K., Yoshikawa, Y., & Asada, M. (2012). Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver. *Advanced Robotics*, *26*(1–2), 23–44, http://dx.doi.org/10.1163/016918611X607347.
Moskowitz, B. A. (1980). Idioms in phonology acquisition and phonological change. *Journal of Phonetics*, *8*, 69–83.
Moulin-Frier, C., Diard, J., Schwartz, J. L., & Bessière, P. (2015). COSMO ("Communicating about Objects using Sensory-Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics*.
Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition?. *Developmental Science*, *6*(2), 136–142, http://dx.doi.org/10.1111/1467-7687.00263.
Nehaniv, C. L., & Dautenhahn, K. (2002). The correspondence problem. In K. Dautenhahn, & C. L. Nehaniv (Eds.), *Imitation in animals and artifacts*. Cambridge, MA: MIT Press.
Nolan, F. (1990). Who do phoneticians represent?. *Journal of Phonetics*, *18*, 453–464.
Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language*, *63*(4), 541–559, http://dx.doi.org/10.1016/j.jml.2010.08.001.
Ochs, E., & Schieffelin, B. B. (1984). Language acquisition and socialization: three developmental stories. In R. Shweder, & R. LeVine (Eds.), *Culture theory: Essays on mind, self and emotion* (pp. 276–320). New York: CUP.
Ochs, E., & Schieffelin, B. B. (1994). Language acquisition and socialization: Three developmental stories and their implications. In B. G. Blount (Ed.), *Language, culture and society* (pp. 470–512). Prospect Heights, Illinois: Waveland Press, Inc.
Öhman, S. E. G. (1975). What is it that we perceive when we perceive speech?. In A. Cohen, & S. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 36–47). Berlin: Springer.
Otomo, K. (2001). Maternal responses to word approximations in Japanese children's transition to language. *Journal of Child Language*, *28*, 29–57, http://dx.doi.org/10.1017/S0305000900004578.
Papoušek, M., & Papoušek, H. (1989). Forms and functions of vocal matching in interactions between mothers and their precanonical infants. *First Language*, *9*, 137–158, http://dx.doi.org/10.1177/014272378900900603.
Pawlby, S. J. (1977a). *A study of imitative interaction between mothers and their infants* (Ph.D.). University of Nottingham.
Pawlby, S. J. (1977b). Imitative interaction. In H. R. Schaffer (Ed.), *Studies in mother–infant interaction* (pp. 203–223). London: Academic Press.
Pelaez, M., Virues-Ortega, J., & Gewirtz, J. L. (2011). Reinforcement of vocalizations through contingent vocal imitation. *Journal of Applied Behavior Analysis*, *44*(1), 33–40, http://dx.doi.org/10.1901/jaba.2011.44-33.
Pines, M. (1984). Reflections on mirroring. *International Review of Psycho-Analysis*, *11*, 27–42.
Pines, M. (1985). Mirroring and child development. *Psychoanalytic Inquiry*, *5*, 211–231, http://dx.doi.org/10.1080/07351698509533585.
Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*(2), 253–260, http://dx.doi.org/10.3758/BF03214136.

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005

Porschmann, C. (2000). Influences of bone conduction and air conduction on the sound of one's own voice. *Acustica*, *86*, 1038–1045.
Porter, R. J., & Castellanos, F. X. (1980). Speech-production measures of speech perception: Rapid shadowing of VCV syllables. *The Journal of the Acoustical Society of America*, *67*(4), 1349, http://dx.doi.org/10.1121/1.384187.
Porter, R. J., & Lubker, J. F. (1980). Rapid reproduction of vowel–vowel sequences: evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research*, *23*(3), 593–602, http://dx.doi.org/10.1044/jshr.2303.593.
Priestly, T. M. S. (1980). Homonymy in child phonology. *Journal of Child Language*, *7*, 413–427, http://dx.doi.org/10.1017/S0305000900002713.
Ramsdell, H. L., Oller, D. K., Buder, E. H., Ethington, C. A., & Chorna, L. (2012). Identification of prelinguistic phonological categories. *Journal of Speech Language and Hearing Research*, *55*(6), 1626, http://dx.doi.org/10.1044/1092-4388(2012/11-0250).
Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice*, *Vol. 10* (p. 244). Academic Press.
Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*(2), 131–141, http://dx.doi.org/10.1016/0926-6410(95)00038-0.
Rochat, P. (2001). Origins of self-concept. In G. Bremner, & A. Fogel (Eds.), *Blackwell handbook of infant development* (pp. 191–212). Oxford.
Roy, D. (2009). New horizons in the study of child language acquisition. In *Proceedings of Interspeech* (Vol. 1, pp. 13–20). Brighton, UK: ISCA.
Sapir, E. (1921). *Language*. New York: Harcourt, Brace and World.
Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336–354, http://dx.doi.org/10.1016/j.jneuroling.2009.12.004.
Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception and Psychophysics*, *66*(3), 422–429, http://dx.doi.org/10.3758/BF03194890.
Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387–2399, http://dx.doi.org/10.1093/cercor/bhl147.
Smith, V. A., King, A. P., & West, M. J. (2000). A role of her own: female cowbirds, *Molothrus ater*, influence the development and outcome of song learning. *Animal Behaviour*, *60*, 599–609, http://dx.doi.org/10.1006/anbe.2000.1531.
Snow, C. E. (1988). The last word: questions about the emerging lexicon. In J. L. Locke, & M. D. Smith (Eds.), *The emergent lexicon: the child's development of a linguistic vocabulary* (pp. 341–353). New York: Academic Press.
Stern, D. N. (1985). The Interpersonal World of the Infant. London: Karnac Books.
Stetson, R. H. (1951). *Motor phonetics*.
Straight, H. S. (1980). Auditory versus articulatory phonological processes and their development in children. In G. Yeni-Komshian, J. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Vol. 1, Production* (pp. 43–71). NY: Academic Press.
Studdert-Kennedy, M. (2002). Mirror neurons, vocal imitation, and the evolution of particulate speech. In M. I. Stamenov, & V. Gallese (Eds.), *Mirror neurons and the evolution of brain and language* (pp. 207–227). Amsterdam: John Benjamins.
Sweeney, S. (1973). The importance of imitation in the early stages of speech acquisition: A case report. *Journal of Speech and Hearing Disorders*, *38*(4), 490–494, http://dx.doi.org/10.1044/jshd.3804.490.
Tinbergen, N. (1963). On aims and methods in ethology. *Zeitschrift Fur Tierpsychologie*, *20*, 410–433, http://dx.doi.org/10.1111/j.1439-0310.1963.tb01161.x.
Veneziano, E. (1988). Vocal–verbal interaction and the construction of early lexical knowledge. In J. L. Locke, & M. D. Smith (Eds.), *The emergent lexicon: the Child's development of a linguistic vocabulary* (pp. 109–147). New York: Academic Press.
Vihman, M., & Croft, W. (2007). Phonological development: Toward a "radical" templatic phonology. *Linguistics*, *45*(4)), http://dx.doi.org/10.1515/LING.2007.021.
Vihman, M. M. (1993). Variable paths to early word production. *Journal of Phonetics*, *21*, 61–82.
Vihman, M. M. (1996). *Phonological development*. Cambridge, MA: Blackwell.
Vihman, M. M., Nakai, S., DePaolis, R. A., & Hall, P. A. (2004). The role of accentual pattern in early lexical representation. *Journal of Memory and Language*, *50*, 336–353, http://dx.doi.org/10.1016/j.jml.2003.11.004.
Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing*, *16*, 5–20.
Warren, R. M. (1999). *Auditory Perception*. CUP.
Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234, http://dx.doi.org/10.1080/15475441.2005.9684216.
West, M. J., & King, A. P. (1988). Female visual displays affect the development of male song in the cowbird. *Nature*, *334*(6179), 244–246, http://dx.doi.org/10.1038/334244a0.
Yoshikawa, Y., Koga, J., Asada, M., & Hosoda, K. (2003). A constructive model of mother–infant interaction towards infant's vowel articulation. In *Proceedings of the 3rd international workshop on epigenetic robotics* (pp. 139–146).
Young, R., & Messum, P. R. (2011). *How we learn and how we should be taught: An introduction to the work of Caleb Gattegno*. London: Duo Flumina.
Zukow-Goldring, P. (1996). Sensitive caregiving fosters the comprehension of speech: When gestures speak louder than words. *Early Development and Parenting*, *5*(4), 195–211.
Zukow-Goldring, P., & Arbib, M. A. (2007). Affordances, effectivities, and assisted imitation: Caregivers and the directing of attention. *Neurocomputing*, *70*(13–15), 2181–2193, http://dx.doi.org/10.1016/j.neucom.2006.02.029.

Please cite this article as: Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics* (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.005