Published in Speak Out! – the IATEFL Pronunciation Special Interest Group Journal
Messum, P.R. (2016) How we learn to pronounce the sounds of
speech: (1) Infants. *Speak Out!* 55, 6-31

How we learn to pronounce the sounds of speech: (1) Infants

Piers Messum

Last year, my colleague Ian Howard and I published a paper in the Journal of Phonetics (Messum & Howard 2015) that discussed the mechanism by which young children learn to pronounce the speech sounds of their mother tongue (L1)[1]. The longstanding assumption has been that they do this by some form of imitation. We argued that on current evidence it is more likely that they do this through a mirroring process; with their caregivers as the 'mirror' in which infants and young children discover the linguistic significance of their vocal actions.

This matters for the learning of second language (L2) pronunciation because many of our teaching practices are implicitly based on the idea that learning to produce sounds by listening first and then trying to copy what we have heard is 'natural' (or even that it is the only possible way for the production of new sounds to be learnt). If it is not natural, then we might want to reconsider our use of 'listen first' approaches for teaching speech sounds. These approaches are not notably successful and there is at least one well-developed and proven alternative.

This article summarises the 2015 paper, concentrating on the parts of it that will be of most interest to *Speak Out!* readers. The paper was written for a special issue of the journal which was examining how speech is represented in the brain, hence the paper's title: *Creating the cognitive form of phonological units: the speech sound correspondence problem in infancy could be solved by mirrored vocal interactions in infancy rather than by the imitation of speech sounds.* In a second article, Roslyn Young and I will examine the nature of L2 speech sound learning and the different approaches taken to teaching sounds.

How might young children learn speech sounds?

Speech is something we perceive and something we do, so it is a perceptuo-motor phenomenon. But in its underlying, neural representation, it has seemed to scientists that speech must be more fundamentally one than the other: either a set of sounds

[1] I use the following abbreviations in this article: L1 – first language, L2 – second language, VMS – vocal motor scheme, SBE – similarity based equivalence, ME – mirrored equivalence, AS – awareness of sensation, MP – meaningful perception.

(produced by subordinate vocal gestures) or a set of gestures that are mostly invisible but whose results we hear[2]. One reason why they have felt that this "either/or" choice is forced upon us is an assumption about how children solve the 'speech sound correspondence problem' when learning to pronounce their first language.

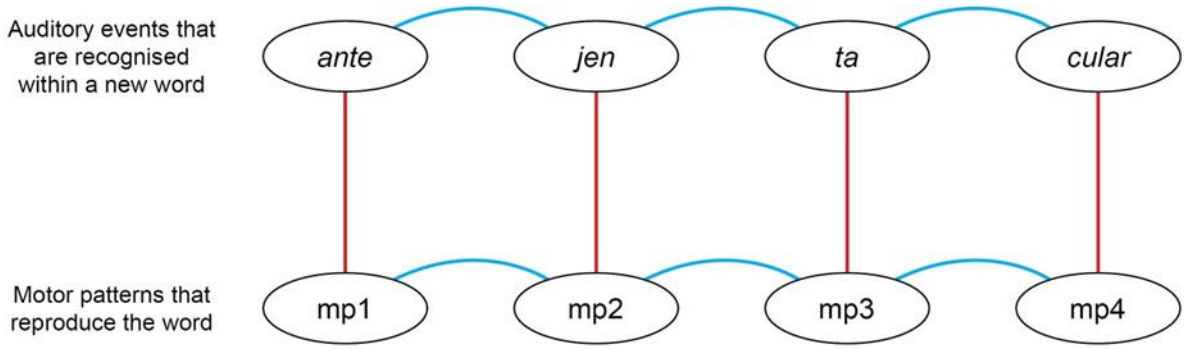

Figure 1. The mature skill of learning the pronunciation of a new word requires learning the identity and ordering of the speech sounds heard; but prior to this, it requires creating the 'vertical' links between speech sounds heard and the motor patterns ('mp1' etc) that can be used to reproduce them.
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. The design of Fig. 1 and the terminology are adapted from Heyes (2001).

The problem exists for both L1 and L2 learners. To understand the issue, we need to distinguish the activities of (i) learning how to pronounce particular words, from (ii) learning how to pronounce speech sounds. The first of these activities is the mature skill of learning the pronunciation of a new L1 (or L2) word: the speaker parses the word he[3] 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

[2] It might also be that we perceive the gestures directly. This is an important and plausible theory of speech perception (Fowler 2003), although it is not widely known about. In this article, I will only take the more mainstream auditory theory of speech perception into consideration.
[3] To avoid continual use of "he or she", "his or her", etc, I use pronouns which describe interactions between a female caregiver and a male child, or a female teacher and male students.

can be greater than phoneme-sized; it might be a syllable, or even a couple of syllables, formed of one or more phonemes.)

We should be aware that the word 'imitation' covers a variety of phenomena. In this case, the process involved is 'serial imitation': the copying of a series of events by producing an equivalent (loosely, 'the same') series of events.

To be able to learn the pronunciation of a new word this way[4], the speaker must first have learnt how to produce speech sounds that will be taken to be equivalent to the ones he hears: process (ii) above. A 'correspondence' must be set up between dissimilar activities: hearing sounds and performing vocal actions.

In learning L1 pronunciation, it has generally been assumed that the child solves the correspondence problem for speech sounds by imitation. That is, the child tries to match what he has heard, and uses his own judgement of similarity to compare what he hears and what he produces. This judgement informs and improves his subsequent production in a 'matching-to-target' process.

There is, though, another way for young children to solve the correspondence problem; they may find a solution within the dynamics of caregiver-infant interaction. Here, vocal imitation is plentiful, but observational studies show that it is predominantly the caregiver who imitates the child, rather than vice versa. Furthermore, the form of the imitation is rarely simple mimicry. Instead, a caregiver reformulates her child's output into L1, interpreting his production as if it was being said by an L1 speaker and saying back to him her interpretation of what he said in L1. This gives him evidence of the correspondence between what he does and what she considers its linguistic significance to be.

In these interactions, the child does not know that he is making L1 sounds. He is playing with his vocal apparatus, discovering what it can do and what he can reliably produce with it. In our experiments with 'caregivers' who spoke English, German and French, described below, there were some occasions when a particular noise produced by our infant was reformulated in strikingly different ways depending on the L1 of the caregiver.

[4] Note that it is not known when the skill of reproducing the pronunciation of a new word by speech sound parsing, as being described, develops. Many scholars suggest that the pronunciation of the very earliest words is learnt differently, by holistic mimicry of the whole word shape. As development progresses, there would be a movement away from this mechanism and towards the use of speech sound parsing and serial reproduction skills. So when a very young child pronounces a word or phrase precociously well, this may be a holistic sequence learnt by mimicry, which will be updated and reproduced by serial imitation of its speech sound elements in due course.

Imitation and similarity based equivalence (SBE)

Let us examine these two mechanisms for solving the correspondence problem in more detail.

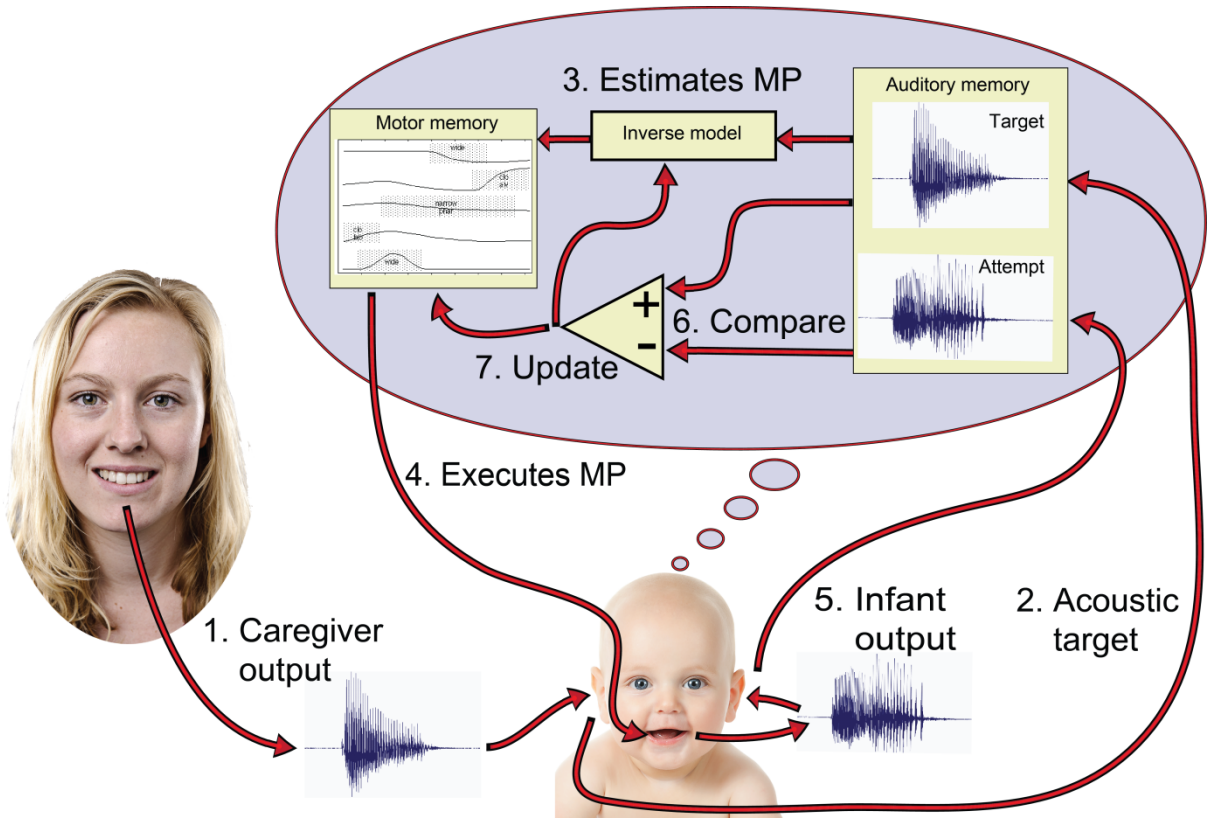

Fig. 2. How a child would solve the correspondence problem for a speech sound in a simplistic auditory Similarity Based Equivalence (SBE) account. (Abbreviation: MP – 'motor pattern'.)
(1) An L1 speech sound is produced by a caregiver (perhaps within a word). (2) The child takes this as a target. (3) Using skills developed prior to and during babbling (an inverse model), (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.

An imitative account must assume that the child can hear both himself and adults correctly. He then makes use of target sounds in the linguistic environment to guide his production, but how he might do this is still unclear: a child may try to match speech sounds when needed, he may store them as sound images that he later

uses to guide production, or he may use them in other ways. (He may also do a combination of these things.) But 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, Messum & Howard called this class of proposals 'Similarity Based Equivalence' (SBE) accounts. Figure 2 illustrates the mechanism.

It should be said that even if the child can hear himself and others correctly, there are difficulties with SBE accounts, including the so-called '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 ever be acoustically identical to that of an adult. There are various proposals for how an infant could judge similarity in these circumstances but the issue is problematic and unresolved.

Also arguing *contra* SBE accounts, it is possible that the child will not hear himself correctly. For example, an infant may not hear his own vowel sounds well because of interference from bone-conducted sound. Or he may normally 'hear' what he intends to produce (in his 'inner voice') rather than hearing his actual output, making a comparison of the relevant signals difficult[5]. Furthermore, when he is listening to speech for comprehension he may find himself in the wrong attentional set for copying speech sound qualities, as described below after the two ways of experiencing an acoustic signal have been discussed.

Mirroring

SBE is not the only way by which an infant might solve the correspondence problem. Mirroring, which is a general mechanism of social learning, is an alternative possible mechanism.

As we all know, the learner of a motor skill can inform himself about his actions by using a physical mirror to observe himself; ballet dancers look at themselves in mirrors to do this. A learner can also attend to a metaphorical mirror in the form of another person who performs the learner's action back to him. A sports coach might do this: "Look, here's what you're doing." 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.

One sometimes sees the term 'mirroring' used as a simple synonym for copying. In the psychological literature, though, it describes the situation when the response to 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. Note, though, that the term 'mirroring' is not ideal because a real mirror reflects an exact

[5] These potential problems all have counterparts in the learning of L2 pronunciation, of course.

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 even by behaviour that is actually different from that of the learner.

Mirrored interaction in early infancy is considered an important part of the development of affect. Pines (1984:32) described the dynamics of this process:

> *"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 (1985:142) described a spectrum of mirroring behaviour, placing imitation/mimicry at one end of it 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. So if the infant is waving his arms in a happy manner, his mother's response might be an appropriate vocal exclamation rather than waving her arms in return. Stern believed that the infant understood his mother's gesture as an affirmation of his internal, affective state. The transformation of one behaviour into another is actually more meaningful than simple imitation because it makes the baby feel that the mother has understood the feeling behind the behaviour (Galligan, 2006).

Mirroring as a mechanism for solving the correspondence problem

Caleb Gattegno (who is known within language teaching for his Silent Way approach, but who is better known in the wider world for his work on the teaching of mathematics and literacy) was the first person I know of to describe a mirroring paradigm for the child's entry into speech (Gattegno, 1973, 1985). His ideas were elaborated by me in my PhD thesis (Messum, 2007), and implemented in a computer model of an infant by Ian Howard and me (e.g. Howard and Messum, 2014). The learning mechanism that our 'infant', Elija, tested is portrayed in Fig. 3, and Elija is described further below. Two other groups of researchers have made similar proposals.

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 show that in most exchanges it is the caregiver that imitates something that the child has produced[6].

[6] To illustrate this, Pawlby (1977:222) quotes one of the mothers in her study reporting on her daughter's vocal development: "Well, it's a very complicated thing this, … she suddenly discovers a sound and it rather fascinates her and I reinforce it; I make the sound as well. And that tends to make

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.

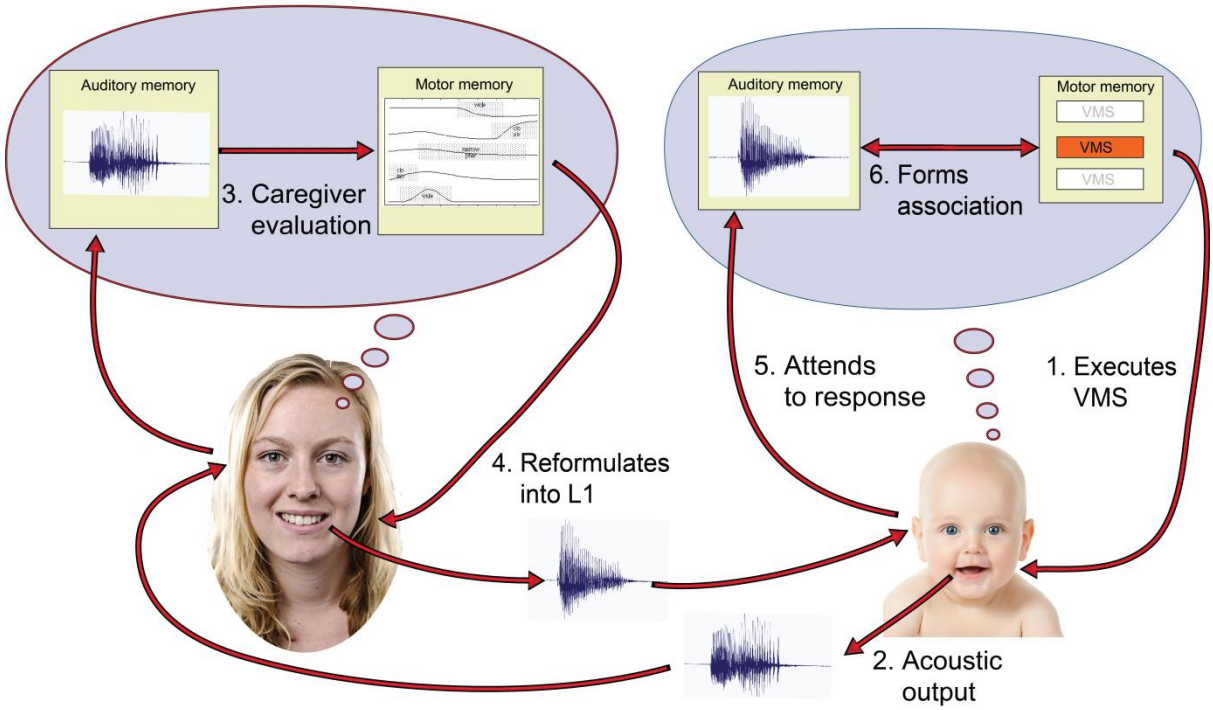

Fig. 3. Solving the correspondence problem for speech sounds using Mirrored Equivalence (ME). (Abbreviation: VMS – 'Vocal Motor Scheme'.)
(1) The child executes a well-practised, sound-making movement, a Vocal Motor Scheme (VMS), (2) which generates acoustic output. (3) The caregiver interprets the output within L1, and (4) reformulates the child's output into an L1 token. One effect of this is that it positively reinforces the child's production. (5) Further, the child understands that this response 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.

Infants already 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

her want to do it more. And then perhaps on a different occasion when she's forgotten all about it, if I make that sound she will imitate it. But the sound seems to have to come from her in the first place."

their activities to be equivalent, and Messum & Howard proposed that the child does indeed accept her judgement. Within the Elija model, Messum & Howard took 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. (This assumption is discussed below.)

There is an obvious parallel between the affect attunement described at the end of the previous section and the vocal reformulation just described. 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 caregiver's interpretation of the child's output is then reflected back to him and can thereby assist his development. Episodes where infants are believed to learn through affect attunement predate the vocal reformulations through which they might learn about speech sounds, so there is no problem of cognitive capacity with respect to the latter.

Messum & Howard use the term 'Mirrored Equivalence' (ME) to describe this mechanism for solving the correspondence problem for speech sounds, since such mirroring behaviour by his caregiver provides an infant with evidence of equivalence.

Elija – a robot infant

Ian Howard and I tested the ME account with Elija, a computational model of infant speech acquisition. In our most recent experiments, Elija modelled a process that started with him 'babbling'. It ended with separate instances of Elija learning to pronounce simple words in three languages during naturalistic interactions with subjects who played the role of 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 (Howard and Messum, 2014).

Even though Elija was physically presented to the caregivers as no more than a computer monitor, microphone and loudspeaker, they found it natural to reformulate the output of most of his motor patterns into well-formed L1 tokens: the behaviour seen in real caregiver-infant interactions.

Learning action-to-sound correspondences from mirrored interactions

The reports on the development of affect described earlier suggest that infants do learn from mirrored interactions with their caregivers. At a slightly later age, Messum & Howard propose that infants learn action-to-sound correspondences in the same

way. They illustrated this through an example, which also addresses the question of what aspect of his activity the infant associates with the caregiver's reformulations.

Consider 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 and whose 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 between the two events, which the child may have the opportunity to test and strengthen on other occasions. Later, he hears his mother say the word "boot". Recognising "boo" within this, he knows that there is something that he can do that (he thinks) she will take to be equivalent to what he has heard her say; so he claps his hands.

In this particular example, his mother will probably not understand what he has done. She may not connect her "boot" to her previous "boo" and the clapping game. However, if his initial action had been a vocal gesture rather than a clap, and this vocal gesture produced a sound which she heard as /bu:/ within L1 and had reformulated as "boo" during previous imitative games, then she will now hear the same sound from him again. She will think 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 the first situation, the mother and child's perspectives of what occurs during the interaction may differ. The mother might conceive the child's clapping as a sound-making activity with the sound as its focus; particularly if, for some reason, the mother could only hear the result of the clapping and not see the action itself. (This is the situation with a vocal gesture performed during early vocal development, which caregivers will therefore conceive in terms of its acoustic output.)

The child, on the other hand, may not yet have mastered clapping to the point where it is automatized. The action of clapping might still require most of the child's attention, and it is likely that the action would then 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, various authorities have suggested that an infant's 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. In Messum & Howard's ME account, therefore, they posit that the child is primarily conceiving his activity as motor rather than sensory.

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

In the 2015 article, Messum & Howard examined the competing SBE and ME hypotheses, demonstrating that various data sets from neuroscience, psychology, child phonology and adult phonology that are anomalous within the SBE paradigm are explained in a straightforward way within the ME one.

One further argument that they made is more directly relevant to the learning of L2 pronunciation. To make this, they considered one fundamental aspect of auditory perception and then the nature of mimicry, since this is a mechanism for recreating the form of a word that is available to a young child and also, of course, to an older learner.

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 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, 1987:65). Thus Thomas Reid (1785) described the senses as having, "a double province – to make us feel, and to make us perceive," and Humphrey (1992) describes the history of this understanding, particularly with respect to vision.

Öhman (1975) described what he called 'ordinary perception', which recovers meaningful events happening in the world from an acoustic signal, and contrasted this with the sensory consequences of sound within the ear: the effect sound has on the listener's "awareness of the developing state of his listening sense." He illustrated this with reference to his wife, who, as he sits at his desk in another room, he can hear moving around the kitchen, opening and closing the refrigerator door etc.

> *"In the way I am [normally] listening, I listen to these events. I do not listen to the sounds of the events. I could listen to the sounds of the events, however, if I wanted to. I would then listen to them as a sort of concrete music, disregarding their physical meaning. This latter sort of listening (…) consists in an immediate awareness of the developing states of my auditory sense. As such it is a form of perception, viz. perception of the states of my own body."*

For speech perception, this duality has been of limited interest because the listener is generally concerned with the meaningful content of the input. To describe it, though, Pisoni (1973) used the terminology of listening in an 'auditory mode' and a 'phonetic mode', and others have used other terminology. In the absence of any consensus, Messum (2007) decided to use the terms 'meaningful perception' (MP)

for Öhman's 'ordinary perception' and 'awareness of sensation' (AS) for what Öhman called 'concrete music' when we are concerned with sound.

Use of the two terms can be extended to experiences we have in other sensory domains. The duality is most apparent with touch, where it is relatively easy to switch one's perspective from MP to AS. For example, if you hold a pen behind your back you can conceive of the situation in that way, or you can move your presence to the points of light pressure on your fingers caused by holding the pen.

In general, we are concerned with meaningful objects in the world. Living our daily life, MP is the important flow of information: we constantly need and seek information about the changing state of the environment to understand our situation and plan what to do next.

In fact, the mode of perception that does not deliver anything meaningful – AS – is an unusual state to find ourselves in. We may only be regularly attending in this mode in situations like eating (relishing the taste of a dish, perhaps), or listening to music (although even here MP will often create images and structure).

While listening to speech, we are normally in our MP mode and it can be difficult to switch to AS. Bruner et al. (1956:50) described the observation we will all have noted in ourselves as teachers of pronunciation, 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 strongly drawn to do so.

We need labels for the two different ways that listeners relate to acoustic events. In the rest of this article, therefore, we will use the words 'sound' and 'noise' in a technical, defined way. A 'sound' will be something we hear that we recognise as the result of a meaningful event: the sound of a door closing or a speech sound. A 'noise' will be something we hear that we do not recognise to be the result of a meaningful event: 'white noise', Öhman's 'concrete music' or the parts of a string of a spoken language that we are not familiar with.

The MP/AS 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 heard as noises before they have the ability to reproduce words by the mature mechanism of recognising and concatenating speech sounds.

The distinction is also necessary for thinking about older learners of pronunciation: (1) the categorical perception developed in order to understand L1 is an example of MP, which interferes with hearing L2 veridically, and (2) mimicry is a mechanism that

is also available to older learners. To understand it better, it is helpful to have the MP/AS distinction clear.

Mimicry

There are a number of quite distinct behaviours that fall within the ambit of 'imitation'. They are sometimes distinguished by examining what out of three characteristics of the action is being copied: its goals, its form and/or its results (the effects on the environment). Within this framework, mimicry can be considered as B copying the form of A's actions without B adopting A's goal or intending to achieve the results A obtains. This distinguishes mimicry from all the more usual forms of purposive copying in which we want our actions to achieve something in themselves. The defining feature of mimicry is 'copying the form of an action'[7].

This is satisfactory for actions we can see, but must be made more precise to encompass the mimicry of noises. The MP/AS distinction described above allows us to do this. We can now define mimicry as the creation of a signal which perturbs an observer's sensory apparatus, as experienced in the AS mode of perception, 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.

The ability to mimic a noise starts being developed when a baby first makes a noise for himself and listens to it; this process of linking the actions and results of noise making can continue for at least the rest of childhood. To then make use of what he has learnt for mimicry, the child must have an auditory target in mind to copy, either just heard and held in short term memory, or evoked from longer term memory. Deploying a vocal gesture, he can make his best attempt at producing 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.)

[7] There are many fine distinctions that can be drawn when one examines imitation, but to illustrate just some of these, imagine that A is skipping with a rope with a goal to improve her fitness. If B wishes to get fit he might copy A by starting skipping himself. Or, with the same goal but not wishing to purchase a rope he might partly copy her by just jumping up and down rhythmically. Or he might emulate her by getting himself fit but in another way (copying her goal but not the means). But if he has no desire to get fit, he can mimic her by jumping and twisting his arms in a way that looks as if he is skipping with a rope. He will have some meta-purpose for this, perhaps to entertain onlookers.

Although the analogy between speech and writing is imperfect in some important respects, it is helpful to recognise that in both mediums there are two ways that word forms can be adopted.

A graphical word form can be either, (1) recreated by drawing the word seen as an image or, (2) reproduced by writing it as a string of letters, i.e., 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. The second way is much more efficient than the first, which is abandoned when the second becomes available. 'Script-drawing' (Adi Japha and Freeman, 2002) enables the child to enter the written medium, but 'script-writing' is the only long-term approach to communication on paper that is viable.

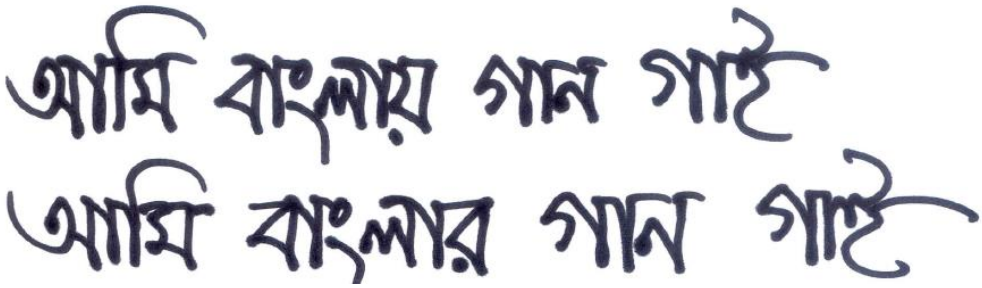

Fig 4. An example of Bangla script. In principle, a non-writer of Bangla could re-create it by drawing the image, but it would be a laborious task. (Note that he could recreate the image equally well whatever the orientation of the original and his copy of it).
A writer of Bangla could reproduce the text in a fraction of the time and effort it would take to draw.

A spoken word form can be either, (1) recreated holistically[8] or, (2) reproduced as a chain of speech sounds, i.e. a child can both *mimic* whole-word shapes and *say* words.

Infants can and do recreate auditory images by mimicry, and this includes the images of whole-word shapes. But the need to evoke and match a target for mimicry means that this takes attentional resource and is thus an inefficient way of saying words. It may enable infants to enter the world of speech through the recreation of early, simple word forms, but it is not a viable way for speech to develop.

[8] 'Holistic' is explained by Studdert-Kennedy (2002:213): "Early words are said to be [produced] holistic[ally] because, although they are formed by combining gestures, gestures have not yet been differentiated as context-free, commutable units that can be independently combined to produce new words."

Steps in the development of production and perception

Whole-word mimicry is a mechanism by which young children can recreate words and even phrases. These renditions can sound precociously accurate.

However, learning the pronunciation of new words only becomes efficient when the child makes use of the associations between the vocal gestures he makes and the L1 speech sounds he hears in return when his caregivers imitate him. Words can then be parsed into units that are efficiently stored, retrieved and run off: a process of motor sequence learning at which a child is already expert.

The learning of pronunciation and other speech skills starts with babbling and ends with the child able to learn the pronunciation of new words efficiently, as shown schematically in Fig. 5.

1 An infant's experimentation (during the periods up to, including and after babbling) gives him increasingly sophisticated skills of mimicry, enabling the recreation of noises heard in the environment.

2 The infant also develops skills in speech comprehension, with the most recent research suggesting that he may start to recognise words and their meanings from as early as 6 months of age.

3 As described earlier, 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 vocal motor schemes and the L1 speech sounds he has heard made in response to them.

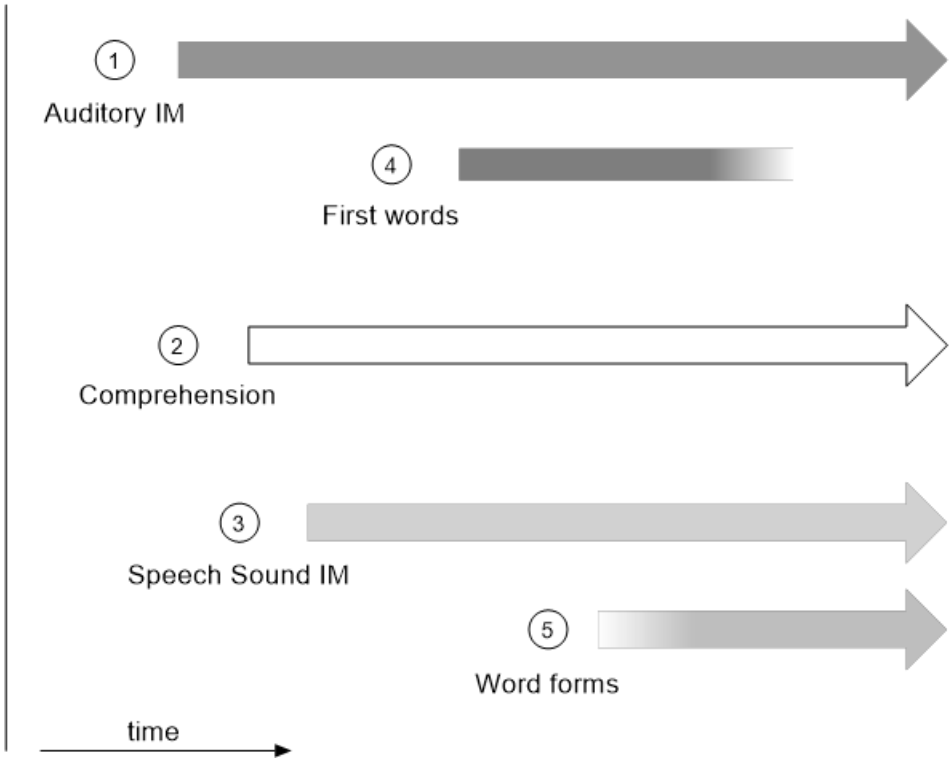

Fig. 5. Initial independence of three separate speech development processes.
Numbers in circles refer to the paragraph numbers here and in the main text:
(1) Development of mimicry skills, supporting (4) first adopted L1 word forms recreated through mimicry;
(2) Speech comprehension;
(3) Development of vocal gesture to speech sound associations, supporting (5) adopted L1 word forms reproduced through serial imitation of speech sounds.
Time shown from birth to around 2 years of age. (Abbreviation: IM – 'Inverse Model', the specification of vocal gestures whose result is equivalent to a given noise.)

4 Most infants start to 'adopt' adult-modelled words from L1 at between 10 and 13 months. These first words are recreated by the child to the best of his ability through mimicry, as whole word forms. (Prior to this, some infants at around 10 to 12 months create protolanguage 'words' – made-up forms – that are not based on L1 models.)

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 (1981:179) 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 the words. As Menn (1983:39) pointed out, "Language is usually used, not contemplated; children expect to listen for meaning, not for sound."

This attentional set allows him to recognize 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. It has been noted in the literature that infants' pronunciation sometimes appears to regress before recovering.

The cognitive nature of phonological units

An ME account of the development of pronunciation describes the direct association of a child's vocal motor scheme with a caregiver speech sound heard in response, implying an intrinsically perceptuo-motor unit as the underlying representation for speech sounds. This perceptuo-motor unit reflects the nature of speech: motoric in production and auditory in perception.

In a related sphere, the philosopher of biology, Ruth Millikan, has called mental representations which both (1) direct an action in the world and (2) describe the world, as 'pushmi-pullyu representations'. (For example, a shopping list can both tell us what to buy, and tell us what we have bought.) The pushmi-pullyu (an imaginary animal created by Hugh Loftus in his Dr Dolittle books for children) gives us a vivid image for the two-headed nature of perceptuo-motor phonological units in the brain.

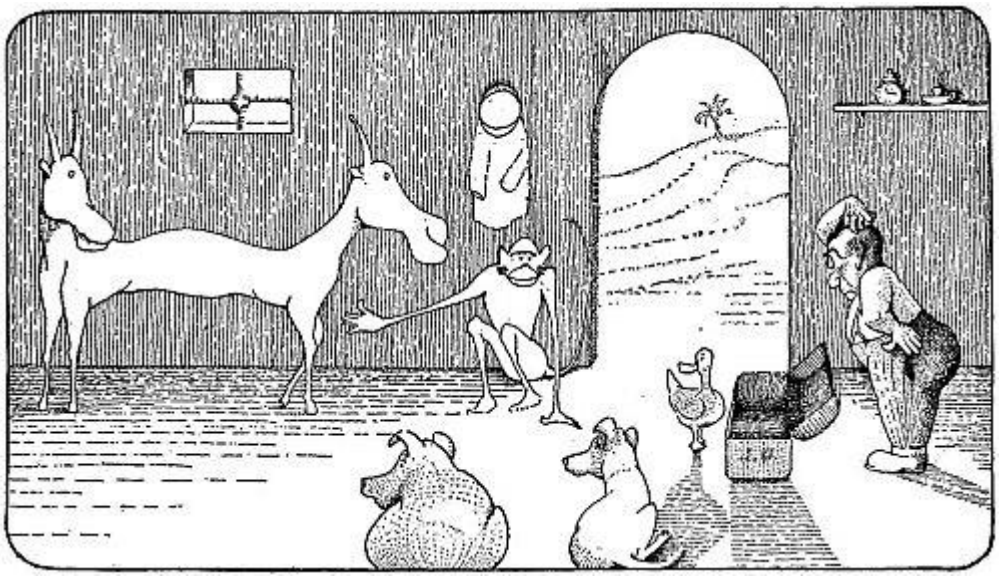

Fig. 5. A pushmi-pullyu (left). As illustrated by Hugh Lofting in The Story of Dr Dolittle (1920)[9]. Original caption: "'Lord save us!' cried the duck. 'How does it make up its mind?'"

The debate about the underlying nature of speech has a long history. For example, Stetson (1951) argued that speech is "gestures made audible." Most phoneticians and speech scientists, though, have followed the idea that the auditory aspect of speech is primary (e.g. Sapir, 1921). One major consequence of speech sounds being learnt by the ME mechanism is that the representation of speech would be both motoric and auditory at the same time, i.e. intrinsically perceptuo-motor. The discovery of mirror-neurons in primates and the understanding that direct neural associations can be formed by contiguous motor and perceptual activity (Heyes, 2013) makes this idea neurologically plausible. In the 2015 article, Messum & Howard also showed how it can explain various longstanding problems in speech research.

A further potential problem with the auditory SBE account

Having drawn the distinction between the AS and MP modes of auditory attention, we can see that there is 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

[9] http://www.gutenberg.ca/ebooks/lofting-story/lofting-story-00-h-dir/lofting-story-00-h.html

recreate the speech sound elements within it as noises, has disappeared (Linell, 1982:67).

Adults demonstrate this when learning L2 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 noises within a word, and then attempt to recreate these noises. Infants do not have the capacity to engineer this kind of presentation.

This new understanding of what might be natural in the learning of pronunciation can inform our pedagogical practice as L2 teachers. It would support existing approaches which emphasise motor system experimentation combined with feedback on performance given by the teacher.

Summary

Messum & Howard (2015) argue against the idea that young children learn to pronounce L1 speech sounds by imitation, i.e. by what we called a Similarity Based Equivalence mechanism. In the work I did for my PhD thesis (Messum, 2007), it was surprising to discover that this idea has only ever been an assumption, without any good evidence to back it up. Interestingly, the idea was only asserted sporadically in the literature (e.g. Fry, 1968; Kuhl, 2000), but it has certainly been a widespread, if not universal, belief among researchers, others who work on child speech and the general public.

It was, perhaps, an easy assumption to make:

- Children must learn the pronunciation of L1 words by imitation (by serial imitation, as described above) since the only source for these words is the children's linguistic environment. It is easy for this evident truth to obscure the fact that learning the pronunciation of the speech sounds that make up words is a quite different task, which must involve a different learning mechanism. (One possible mechanism being 'imitation', of course, but not serial imitation: auditory matching, instead.)
  I am not aware of any speech literature that clearly identifies learning to pronounce speech sounds and learning the pronunciation of a word as two different processes, and it seems that it was easy for researchers to merge these processes into one in their minds.

- It must also have seemed like common sense: "The speech signal is available to a young child; surely he learns to reproduce it by copying what he hears," must have been a thought that went through many minds. This could draw support from the fact that children do mimic noises in the environment and chunks of speech produced by others. So if one has not distinguished mimicry from serial

imitation, this would seem to be evidence that children do learn to pronounce by auditory matching.

On the other hand, there have been the anomalies which are so important to paradigm change (Kuhn, 1962) and which are, perhaps, 'counter-instances' to the SBE mechanism rather than the 'puzzles' which speech science has hoped them to be and treated them as. Among these, discussed in more detail in Messum & Howard (2015) and Messum (2007), are:

- There have been no observations of children actually practising their production of speech sounds on their own. They do this for the production of words, so why not for speech sounds if these, too, are being copied?

- No theoretical significance could be found for what are, surely, the most significant vocal events in the lives of young children after the babbling phase: the imitative vocal games they play with their caregivers, where it was repeatedly demonstrated in observational studies that caregivers imitate their charges far more than the infants imitate their caregivers. A staggering amount of vocal interplay occurs, dwarfing the numbers of instances of imitative games with gestures and objects that also occur and clearly advance child development. Yet this form of vocal interplay was ignored by child phonology, which has historically preferred to view the infant as the isolated consumer of the input[10], even if this view has been increasingly changing.

- Experiments suggest that toddlers do not monitor their own speech output, while older children and adults do. Yet toddlers are supposed to be learning sounds by copying them, which would require self-monitoring. Similarly, some slightly older children who mispronounce a sound (saying "fis" for "fish", for example) can hear the mispronunciation and distinction between the sounds in the speech of others, but insist that their own production is correct. If they learn sounds by imitation, how can they not hear the mistake in their own speech?

- And, most embarrassingly, having assumed that speech must be either a motor or an auditory phenomenon at its deepest, underlying level, speech science has had to accept that there is apparently contradictory evidence about which of

[10] I may be stating this too strongly, and the charge may now be firmly part of history rather than the present, but one underlying assumption of child language research in the past is indicated in this quote from Elbers and Wijnen (1992:341):
"… a 'production-based' approach has the important advantage of bringing together language learning and other kinds of learning that occur in childhood. For instance, no one would seriously defend the idea that a child learns how to build with blocks primarily by analysing the block constructions produced by others. Rather, one would assume that the child learns from his or her own constructive operations and their outcome …. Yet theories of language acquisition, of whatever signature, mainly acknowledge the role of input in the learning process, not that of children's constructive production."

these it is. More than one hundred years of debate has failed to decide the matter[11].

All this said, the question of how children learn the pronunciation of L1 speech sounds is not settled. What evidence there is, though, favours the Mirrored Equivalence (ME) mechanism. It provides the best fit with the current data and therefore has what Dewey (1941) called 'warranted assertability'.

As I wrote in my PhD thesis (Messum, 2007), the question of which account to prefer, SBE or ME, at this point in time[12] is not one that pits a challenger against a champion who has earned his position. The incumbent has never demonstrated any claim to the title whatsoever. In this situation, scientists can afford to sit on the fence, but as language teachers, our responsibility is to do the best we can from this Monday morning onwards. If we base our practice of teaching pronunciation to any degree on what is natural in children (as I think we do) then we have to come to a judgement about which account is more likely, given what we presently know.

If that judgement is that ME is the more likely mechanism for speech sound learning in children, then how that might feed into our understanding of L2 learning is the subject of the next article.

References

Adi-Japha, E., & Freeman, N. H. (2001). Development of differentiation between writing and drawing systems. *Developmental Psychology*, *37*(1), 101–114.

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

Fowler, C. A. (2003). Speech production and perception. In *Handbook of Psychology: 4 Experimental Psychology* (pp. 237–266). New Jersey: John Wiley.

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

[11] As we have seen above, the ME mechanism finesses the issue completely. SBE demands that at its deepest level speech be either one or the other; ME leads to it being both motor and perceptual - simultaneously and unproblematically.

[12] Postal (2005): "One needs to ask for particular scientific hypotheses not only, grandly, whether they are true, but in more limited and procedural terms, what the currently available evidence for them is. A hypothesis might ultimately be true without there being enough or even any evidence at a certain point which could be taken to scientifically justify it, a hypothesis might be false without it being possible to show that definitively at a given historical point. … the only viable way to achieve a reasoned judgment about [a scientific hypothesis] must involve a consideration of the degree of scientific evidence now available for (or against) it."

Galligan, R. (2006, April 24). Stern and affect attunement. Retrieved from http://psychopathology2.blogspot.co.uk/2006/04/stern-affect-attunement.html

Gattegno, C. (1973). *The Universe of Babies*. New York: Educational Solutions.

Gattegno, C. (1985). Chapter 13: The Learning and Teaching of Foreign Languages. In *The Science of Education*. New York: Educational Solutions.

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

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

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.

Humphrey, N. (1992). *A History of the Mind: Evolution and the Birth of Consciousness*. New York: Simon & Schuster.

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.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences USA*, *97*(22), 11850–11857.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Linell, P. (1982). The concept of phonological form and the activities of speech production and speech perception. *Journal of Phonetics*, *10*, 37–72.

MacKay, D. G. (1987). Perceptual sequencing and higher level activation. In *The Organization of Perception and Action*. New York: Springer Verlag.

Menn, L. (1983). Development of articulatory, phonetic, and phonological capabilities. In B. Butterworth (Ed.), *Language Production, 2* (pp. 3–50). London: Academic Press.

Messum, P. R. (2007). *The Role of Imitation in Learning to Pronounce* (PhD). University College London.

Messum, P. R., & Howard, I. S. (2015). 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. *Journal of Phonetics*, *53*, 125–140.

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.

Pawlby, S. J. (1977). Imitative interaction. In H. R. Schaffer (Ed.), *Studies in Mother-Infant Interaction* (pp. 203–223). London: Academic Press.

Pines, M. (1984). Reflections on mirroring. *International Review of Psycho-Analysis*, *11*, 27–42.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*(2), 253–260.

Reid, T. (1785). *Essays on the intellectual powers of man*. Dublin: L. White.

Sapir, E. (1921). *Language*. New York: Harcourt, Brace and World.

Stern, D. N. (1985). The sense of a subjective self: Affect attunement. In *The Interpersonal World of the Infant* (pp. 138–145). London: Karnac Books.

Stetson, R. H. (1951). *Motor Phonetics*. Amsterdam: North Holland.

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.

Piers Messum teaches and trains teachers through Pronunciation Science Ltd (www.pronsci.com). His previous articles for *Speak Out!* can be found at their website. He is the Treasurer of PronSIG and has taught in the UK, France and Japan. He has a PhD in Phonetics from University College London.

Email: p.messum@gmail.com