MODELLING CAREGIVER TUTORED DEVELOPMENT OF PRONUNCIATION IN A YOUNG CHILD

Ian S. Howard[1] *& Piers Messum*[2]

[1]*Computational & Biological Learning Laboratory, University of Cambridge, UK.*
[2]*Centre for Human Communication, University College London, UK.*
ish22@cam.ac.uk

Abstract: Imitation is almost always assumed to be the mechanism by which infants learn to pronounce speech sounds, which are the elements from which words are made up. Specifically, it is believed that auditory matching enables a child to reproduce speech sounds by copying those that he hears. For several reasons, we believe that this is not the way that this systemic aspect of pronunciation is acquired. We test an alternative account involving a non-imitative mechanism using Elija, a computational model of an infant. Elija started by learning to babble in an unsupervised fashion. Three separate experiments were then run with Elija using one native speaker of English, French and German to play the role of the caregiver. Each caregiver interacted with a different instance of Elija in his or her native language. Using the tutored interactions from each caregiver, which involved their reformulations of his putative speech sounds, Elija learned (1) the importance of his productions, and (2) the correspondence between his and adult speech tokens, thereby developing an ability to imitate a series of such tokens, that is, a word. Finally, using his newly acquired ability to parse input speech sounds in terms of the equivalents to his own tokens, each caregiver taught Elija to say some simple words by serial imitation. We present results from these experiments and discuss the implications of this work.

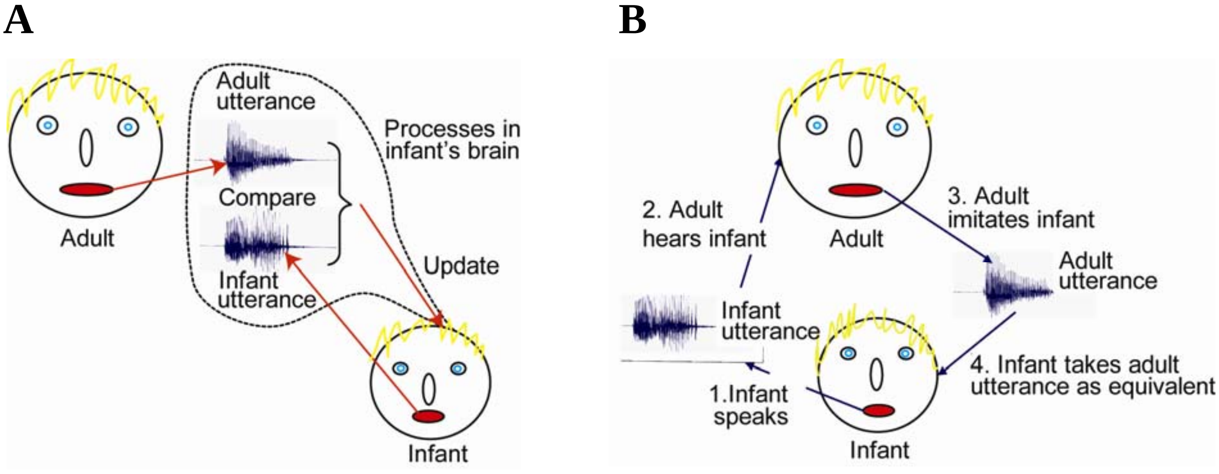

Figure 1 – A Do infants learn to pronounce sounds by imitation using acoustic matching? **B** An alternative hypothesis: Infants learn to pronounce using reinforcement of their productions from caregiver reformulations. These reformulations are also used by the infant to associate his motor actions to adult speech output.

1 Introduction

1.1 Classical accounts of learning to pronounce speech sounds

Learning to pronounce speech sounds is conventionally assumed to be an imitative process (see Figure 1A). A child supposedly uses an auditory 'matching to target' mechanism, whereby he compares his output of a given speech sound to what he hears produced by others, or to what he has heard in the past [1, 2]. In this account, he must call upon his own judgment of sound similarity to improve his subsequent performance. This appears to be straightforward, but it relies upon the child perceiving speech sounds from both sources in a manner that is linguistically appropriate. It therefore assumes that to perceive this ‘similarity’, a child has solved the normalisation problem that arises from the differences in child and adult output due to the different sizes of their vocal tracts. In addition, adult and infant word forms can differ considerably [3]. This in itself can make direct comparison difficult. In fact there is no positive evidence to support an imitative account of learning to pronounce. However, many previous computational models of speech development just assume that such an innate imitative mechanism is unproblematic or put to one side the problems posed by normalisation [4-8]. That said, one other research group has recognised the problems posed by the conventional account of direct acoustic imitation and has used mirrored caregiver interactions to train a physical model to learn vowel qualities [9, 10].

1.2 Non-imitative account

As an alternative to learning the pronunciation of speech sounds by imitation, we describe a model that exploits a ubiquitous social interaction between a young child and his caregivers [11]. When an infant starts to make speech-like sounds, his caregiver is invariably willing, from time to time, to vocally ‘imitate’ him, occasionally with responses that are mimicked, but increasingly by reflecting her interpretation of his output within L1 back to him, in well-formed L1 productions. These natural, well-attested interactions, in which a caregiver mirrors back an infant‘s output (Figure 1B), are known as reformulations. Indeed Pawlby showed that in more than 90% of cases of imitative exchanges, it is the mother who imitates the infant [12]. Since a child recognises when he is being imitated, he knows that his caregiver regards the two speech utterances as being equivalent. A mirroring interaction thus solves the correspondence problem [13] without the child needing to perform an auditory match. This gives him the ability to imitate his caregiver. From this point onwards, he can learn words by recognising elementary speech sounds within them and then recalling his motor actions that correspond to these sounds. See [14] for a fuller discussion of this mechanism.

1.3 Problems with the classical account

Among the child speech phenomena which cannot be explained satisfactorily under imitative accounts, there is the well-known “fis”/“fish” [15-17]. Here, a child pronounces “fish” as “fis” and when questioned as to why he did so, insists firstly that he can hear the distinction in the two forms made by the adult and secondly that he did not say the incorrect form himself. Imitative accounts cannot explain this because they assume that the infant learns speech sounds by copying the adult form, in which case it is paradoxical that he can hear a difference in adult speech but not hear it in his own. In our model, on the other hand, production and the parsing of words for speech sound equivalences are initially separate from general speech perception. As such, production forms can differ from those used in general word perception and the “fis”/“fish” phenomenon has an uncomplicated explanation.

Among the problematic adult phenomena, there is the data on speech shadowing, e.g. [18]. Under conventional accounts of speech sound acquisition by acoustic matching, it appears that speech is an exception to the otherwise universal response time differences in simple and choice reaction time tests. In our model, the data are explained more simply, because the production and perception of speech sounds is directly associated rather than going via a common form, whether acoustic or gestural.

1.4 Elija

We model learning to pronounce using Elija, a computational model of infant speech acquisition [11]. Elija has a speech production capability based on a modified Maeda articulatory synthesiser [19] which generates acoustic output using a pair of loudspeakers. Motor actions are modelled akin to the gestural score used in the Task Dynamics model [20]. Elija's hearing is based on an auditory filter bank [21], which receives input from a microphone. In order for Elija to speak, it is necessary for him to "discover" appropriate motor patterns to control his vocal apparatus. Finding a set of appropriate motor patterns corresponding to speech sounds and utterances is the task of the unsupervised and caregiver tutored learning that are described in the next sections. A detailed description of the current implementation of Elija is available in a companion methods paper to the current one [22].

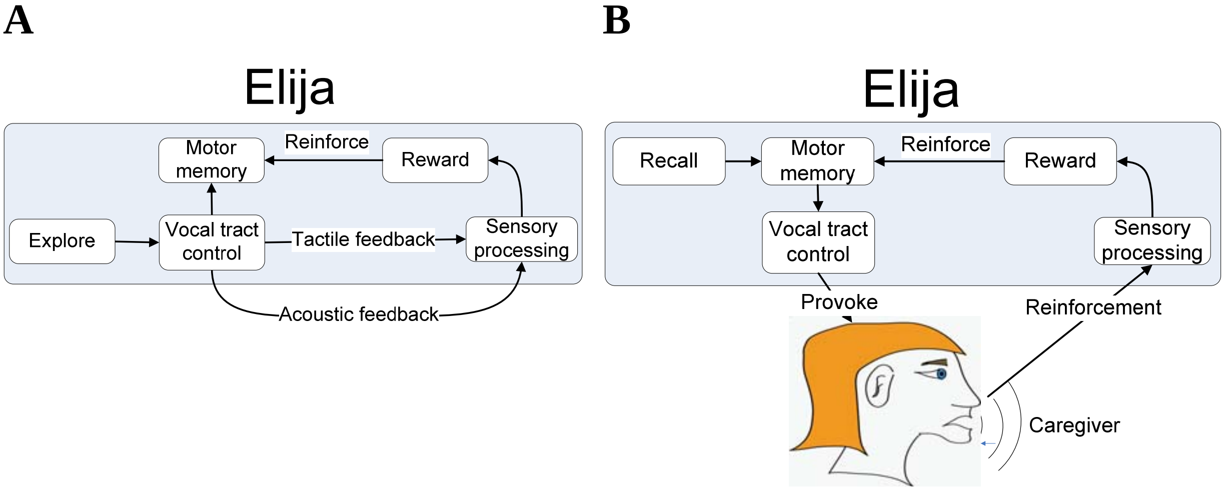

Figure 2 – A Active learning of salient configurations. **B** Reinforcement from caregiver responses.

1.5 Unsupervised sound discovery

During their acquisition of speech production, infants progress through several identifiable developmental stages [23]. Within a few months of birth, infants gurgle, sigh, chuckle and coo. Between the ages of three and six months they start to make syllabic sounds and then to babble. Much of this initial development appears to occur in an unsupervised fashion, with the infant apparently experimenting with his speech apparatus. To model this process, Elija first discovers sounds in an unsupervised manner (Figure 2A) by finding motor patterns that are solutions to an optimization problem [24]. The objective function of a motor pattern is defined as a sum of its sensory salience, diversity and stability, less the energetic effort involved in action generation.

1.6 Caregiver reformulations

During later stages of pronunciation development, interaction with learned speakers of the ambient language (L1) becomes important. As infant sound production progresses, it starts to attract his caregiver's attention. When it provokes reactions, these can constitute evaluations of the infant's productions and this reinforces some of the sounds he can make. In addition, through reformulation of the child output into well-formed L1 speech sounds, his caregivers provide the child with an L1 interpretation of his production. We simulate this second phase of learning by using the sounds Elija discovers to potentially provoke a caregiver to respond. The responses are typically reformulations and are used (1) to reinforce speech sounds that will often be appropriate for L1 (Figure 2B), and (2) to learn equivalence relations between Elija's vocal actions and his caregiver's responses (Figure 3A).

1.7 Learning words by imitation

Since Elija has now learned some caregiver reformulations that correspond to his productions, he can attempt to analyse incoming speech into these component speech sounds and perform serial imitation of those elements he recognises. That is, he can first parse input in terms of the speech sounds he has heard previously and then select and generate their corresponding motor actions, thereby imitating the caregiver's speech (Figure 3B). In this way, the caregiver teaches Elija words by speaking to him and Elija then repeats the words using serial imitation.

We allow Elija and his caregiver to engage in repetitive loops. A word spoken by the caregiver can be repeated a few times until an appropriate production is performed by Elija, or the caregiver chooses to move on. This mechanism reduces the effect of any previous incorrect associations and, in principle, provides an effective means to train not only production, but also perception. All recognition is performed on templates using a DTW speech recognizer [25, 26] with an auditory filter bank front end [21]. The identity of the words spoken by the caregiver acted as contextual cues to Elija. Such contextual cues reduce uncertainly in the interpretation of what the other said, since it must be related to the context. In a more realistic model of an infant, its visual system could be used to ground the context of the interactions in a more natural way. However this was beyond the scope of the current study.

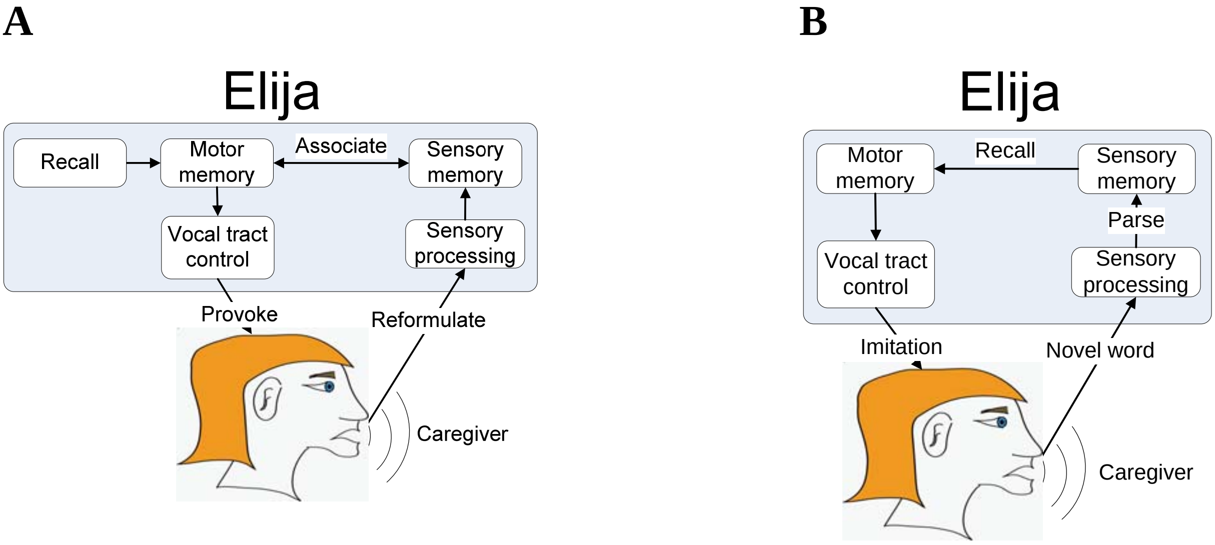

Figure 3 A Motor-acoustic associations arising from caregiver reformulations. **B** Parsing and reproducing caregiver speech using acoustic-motor associations.

2 Results

After providing written informed consent, a male German (G), a female French Canadian (FC) and a female American English (AE) speaker took part in the experiments. All subjects were native speakers of the languages they taught Elija. A local ethics committee at the University of Cambridge approved the experimental protocol. Examples of Elija's acoustic output for the different subjects at various stages of the experiment are available online at: www.ianhoward.de/publications/HowMesESSV2011SupMat.ppt.

2.1 Unsupervised sound discovery experiment

During the initial sound discovery phase, no caregiver involvement was required. Elija was run in several sessions to discover different sounds types. This involved separate runs with an emphasis on low frequency power (for vowels), high frequency power (for fricatives) and touch (for plosives) respectively. For computational reasons, clustering was employed to limit the number of discovered articulations in each type class. This maintained variety but limited redundancy and ensured there was no subsequent combinatorial explosion of VV, CV and VC configurations. The number of vocalic sounds discovered was limited to 15, the number of plosives was limited to 15 and the number of fricatives limited to 10. In order to increase the variety of available sounds, voicing was explicitly enabled or disabled in each plosive and fricative articulation. Similarly, plosive articulations were generated with or without voicing. The voiced plosives were also generated with and without nasality. In this way, Elija was able to identify distinct articulations useful for the generation of a wide range of potential speech sounds.

The basic phonetic elements discovered are listed here in SAMPA phonetic transcriptions [27]. They included a wide range of vowels /i:, 3:, u:, A:/, diphthongs /aI, @U, OI, eI, U@, i@, aU /, plosives /b, p, d, t, g, k /, nasals and liquids /j, y, w, m, n, l, r /, and fricatives /f, h, s, S, z, dZ /. See the online supplementary material for Section 3.1 for examples of these sounds.

2.2 Reformulation experiment

After single articulations had been discovered, they were combined to generate VVs, CVs and VCs. Implausible sounds that included synthesiser artefacts such as clicks were removed by the authors. In total, 927 sounds were generated by Elija. These were then played to each of the 3 subjects, who took the role of Elija's caregiver. Each caregiver ran with a different instance of Elija, so only their interactions would affect Elija's learning. They only reformulated a subset of Elija's productions and this pruned down Elija's repertoire. The subjects (G, CF, AE) reformulated 53%, 81% and 87% of Elija's outputs respectively. They experienced no difficulty in performing this task. They provided good adult forms corresponding to their interpretations of Elija's utterances.

For sounds responded to by all three caregivers, the reformulations could be quite different, as a result of each interpreting Elija's output within their native language This can be heard in the examples of these sounds in the supplementary material Section 3.2.

2.3 Word imitation experiment

The reformulations enabled Elija to recognise some speech sounds embedded in words and then to attempt to pronounce the words himself. Without prompting, each caregiver used motherese to emphasise certain parts of the words they were trying to teach. All three caregivers succeeded in teaching Elija to pronounce around 30 typical first words in their language to a level of performance that is characteristic for young infants aged around 2 years.

Interestingly, the imitation mechanism was able to correct the effect from previous errors that arose in the reformulation stage. That is, sometimes an inappropriate reformulation would be associated with Elija's production. However, during word imitation, the criterion for acceptance of his vocal action was only on the basis of what he said back to the caregiver. By changing how she spoke, the caregiver could sometime provoke a better and more appropriate response by matching against a different and correctly associated reformulation. Similarly, although Elija sometimes made errors in recognition of the sounds in caregiver's speech, these could also be corrected by the caregiver as she could prompt him for another attempt at recognition or even speak again until he generated an appropriate response. Some of the better examples of imitation are given in the supplementary material Section 3.3.

3 Discussion

3.1 Summary

We demonstrated an alternative to a purely imitative account of how infants learn to pronounce words. Our model learned to produce simple utterances using only reinforcement feedback and association and did not employ acoustic imitation to learn to produce speech sounds. Elija was taught to speak words in English, French and German. In these three languages he was able to pronounce some basic words with a level of competence that matches that of a young child. His ability to learn words in these three languages demonstrates that our account of speech acquisition is language independent.

Initially, actions underlying simple sounds were discovered using reward based on the salience of their sensory consequences. Reinforcement provided by an evaluation from a learned caregiver quickly suppressed some sounds, and retained only those that the caregiver approved (generally those present in L1). We note that it was a learned caregiver who judged Elija’s speech sound production. By virtue of being a fluent native speaker, the caregiver is in a much better position to perform this function than Elija would be. The caregiver reformulations then provided the means to associate Elija’s motor actions with the caregiver’s adult L1 interpretations, solving the correspondence problem between simple sub-word acoustic units and his motor productions. This developed an ability to learn words by imitation. Importantly, the imitation stage was able to correct for previous reformulation errors, demonstrating the overall robustness of the mechanism.

Imitation is frequently cited as the means by which we acquire complex actions [28-30]. Indeed it has been claimed previously that infants have an innate ability to learn complex actions by imitation, such as the ability to stick out the tongue when the action is observed [31]. However this issue, and other assumptions regarding imitation, are discussed at length by Heyes [32, 33], who points out that there is little evidence for an innate ability for imitation and a much evidence for an alternative associative hypothesis. Our account of the acquisition of pronunciation by infants is in line with Heyes’ view. Our results show that communicative interaction with a caregiver can lead to the development of speech using only reinforcement and associative mechanisms. In our account no innate ability to imitate need exist. Rather it develops due to the learned association between production and perception. That is, imitative abilities emerge as the associations between word pronunciation and perception are learned.

As a final point, we remark that a role for mirror neurons is sometimes invoked in explanations of the ability to imitate. They are supposed to provide the link between perception and production that is needed in learning by imitation [34-38]. We believe that the existence of mirror neurons is consistent with our model, but rather than those implicated in speech existing innately, we provide an account of how they could develop during learning to pronounce [14].

Acknowledgments

VTCALCS is a vocal tract model written by Shinji Maeda, modified and distributed by Satrajit Ghosh. Our version has been further extended by Mark Huckvale and ISH. We thank Daniel Wolpert for allowing us to use facilities in the CBL, University of Cambridge.

References

[1] P. K. Kuhl, “A new view of language acquisition,” Proc Natl Acad Sci U S A, vol. 97, no. 22, pp. 11850-7, Oct 24, 2000.

[2] D. B. Fry, “The phonemic system in children's speech,” Br J Disord Commun, vol. 3, no. 1, pp. 13-9, Apr, 1968.

[3] N. Waterson, “Child phonology: a prosodic view,” Journal of Linguistics, vol. (1971) no. 7, pp. 179-211, 1971.

[4] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, “Neural modeling and imaging of the cortical interactions underlying syllable production,” Brain and Language., vol. 96, no. 3, pp. 280-301, Mar, 2006.

[5] G. Bailly, “Learning to speak. Sensori-motor control of speech movements,” Speech Communication, no. 22, pp. 251–267, 1997.

[6] K. L. Markey, “The sensorimotor foundation of phonology; A computational model of early childhood articulatory development,” University of Colorado at Boulder, 1994.

[7] B. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, “Towards a neurocomputational model of speech production and perception,” Speech Communication, vol. 51, no. 9, pp. 793-809, Sep, 2009.

[8] G. Westermann, and E. Miranda, “A new model of sensorimotor coupling in the development of speech,” Brain and Language, 89, 393-400., 2004.

[9] Y. Yoshikawa, M. Asada, K. Hosoda et al., “A constructivist approach to infants' vowel acquisition through mother-infant interaction,” Connection Science 14 (4), 245-258, 2003.

[10] K. Miura, Y. Yoshikawa, and M. Asada, “Unconscious anchoring in maternal imitation that helps finding the correspondence of caregiver's vowel categories,” Advanced Robotics, vol. 21, no. 13, pp. 1583-1600, 2007.

[11] I. S. Howard, and P. Messum, “Modeling the development of pronunciation in infant speech acquisition ” Motor Control, 2011.

[12] S. J. Pawlby, "Imitative interaction," Studies in Mother-Infant Interaction, H. R. Schaffer, ed., pp. 203-223 London: Academic Press, 1977.

[13] C. L. Nehaniv, and K. Dautenhahn, "The correspondence problem," Imitation in Animals and Artifacts, C. L. Nehaniv and K. Dautenhahn, eds., pp. 41–61, Cambridge, MA The MIT Press, 2002.

[14] P. R. Messum, “The Role of Imitation in Learning to Pronounce,” PhD., University of London, 2007.

[15] E. V. Clark, and H. H. Clark, "First sounds in the child's language," Psychology and language: an introduction to psycholinguistics, pp. 375-404, New York: Harcourt Brace Jovanovich, 1977.

[16] J. L. Locke, "The child's processing of phonology," Child Language and Communication: Minnesota Symposium on Child Psychology Volume 12, W. A. Collins, ed., pp. 83-119, Hillsdale, NJ: LEA, 1979.

[17] T. M. S. Priestly, “Homonymy in child phonology.,” Journal of Child Language., vol. 7, pp. 413-427, 1980.

[18] C. A. Fowler, J. M. Brown, L. Sabadini et al., “Rapid access to speech gestures in perception: evidence from choice and simple response time tasks,” Journal of Memory and Language, vol. 49, pp. 396-413, 2003.

[19] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," Speech production and speech modelling, W. J. Hardcastle and A. Marchal, eds., pp. 131-149, Boston: Kluwer Academic Publishers, 1990.

[20] E. Saltzman, and K. Munhall, “A dynamical approach to gestural patterning in speech production,” Ecological Psychology, vol. 1, pp. 333-382, 1989.

[21] M. Slaney, “An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank,” Apple Computer Technical Report #35, 1993.

[22] I. S. Howard, and P. Messum, "The computational architecture of Elija: A model of a young child that learns to pronounce," Studientexte zur Sprachkommunikation; ESSV 2011, Aachen, Germany.

[23] D. Oller, The emergence of the speech capacity, Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[24] I. S. Howard, and P. R. Messum, “A Computational Model of Infant Speech Development,” in XII International Conference "Speech and Computer" (SPECOM'2007). 2007, pp. 756-765

[25] R. Turetsky, and D. Ellis, “Ground-truth transcriptions of real music from force-aligned MIDI syntheses,” 4th International Symposium on Music Information Retrieval ISMIR-03, pp. 135-141, 2003.

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

[27] J. C. Wells, "SAMPA computer readable phonetic alphabet," Handbook of Standards and Resources for Spoken Language Systems, D. In Gibbon, Moore, R. and Winski, R., ed., Berlin and New York: Mouton de Gruyter, 1997.

[28] R. Grush, “The emulation theory of representation: motor control, imagery, and perception,” Behav Brain Sci, vol. 27, no. 3, pp. 377-96; discussion 396-442, Jun, 2004.

[29] E. Kohler, C. Keysers, M. A. Umilta et al., “Hearing sounds, understanding actions: action representation in mirror neurons,” Science, vol. 297, no. 5582, pp. 846-8, Aug 2, 2002.

[30] M. A. Arbib, "The Mirror System, Imitation, and the Evolution of Language," Imitation in Animals and Artifacts, C. Nehaniv and K. Dautenhahn, eds., 2000.

[31] A. N. Meltzoff, and M. K. Moore, “Imitation of Facial and Manual Gestures by Human Neonates,” Science, vol. 198, no. 4312, pp. 75-78, 1977.

[32] C. M. Heyes, and G. Bird, "Mirroring, association and the correspondence problem," Sensorimotor Foundations of Higher Cognition, Attention & Performance, Y. R. M. K. P. Haggard, ed.: Oxford University Press, 2007.

[33] E. Ray, and C. Heyes, “Imitation in infancy: the wealth of the stimulus,” Dev Sci, vol. 14, no. 1, pp. 92-105, Jan, 2011.

[34] F. Vargha-Khadem, D. G. Gadian, A. Copp et al., “FOXP2 and the neuroanatomy of speech and language,” Nat Rev Neurosci, vol. 6, no. 2, pp. 131-8, Feb, 2005.

[35] G. Rizzolatti, and M. A. Arbib, “Language within our grasp,” Trends Neurosci, vol. 21, no. 5, pp. 188-94, May, 1998.

[36] G. Rizzolatti, and L. Craighero, “The mirror-neuron system,” Annu Rev Neurosci, vol. 27, pp. 169-92, 2004.

[37] M. A. Arbib, “From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics,” Behav Brain Sci, vol. 28, no. 2, pp. 105-24; discussion 125-67, Apr, 2005.

[38] J. F. Prather, S. Peters, S. Nowicki et al., “Precise auditory-vocal mirroring in neurons for learned vocal communication,” Nature, vol. 451, no. 7176, pp. 305-10, Jan 17, 2008.