In: **Acorns Workshop 11th September, Interspeech, Brighton, UK, 2009**

Building machines that learn to speak the way infants do

Ian S. Howard

Computational and Biological Learning Laboratory, Department of Engineering,
University of Cambridge, Cambridge CB2 1PZ, UK

ish22@cam.ac.uk

Abstract

Children learn to speak during their first years of life and achieve a level of performance that current technology cannot match. The physical structure of the vocal apparatus aids the discovery of speech sounds and the natural interactions that occur between an infant and his caregiver play a pivotal role in his learning to understand and produce words. Here the basic issues of embodiment and interaction, as well as reward, are introduced. The results from experiments using a computational model are then discussed. This model learns to pronounce through interactions with its human caregiver who is able to teach it the names of objects. Finally the issues relating to building learning machines that acquire speech the way infants do, are examined.

Index Terms: infant, embodiment, interaction, learning cortex, plasticity

1. Introduction

1.1. Motivation

The ability to speak forms an important part of most people’s lives but the learning of speech is often taken for granted, as something that ‘naturally’ occurs. It is only when, as adults or children at school, people need to learn another language or when engineers attempt to build machines to do the same thing, that the issues and difficulties involved become apparent.

I believe is it valuable to study how infants learn to speak for several reasons. Firstly, it is an interesting scientific endeavor in its own right. Secondly, a greater understanding of how infants learn to speak should assist language teachers and speech & language therapists. Thirdly, there is increasing awareness that current approaches to speech technology are reaching their limits, suggesting that there needs to be a reassessment of approaches to building speech recognizers and synthesizers [1]. Currently no machine matches human performance in terms of naturalness of speech output and robustness of recognition performance in noisy environments.

1.2. Current technology

At present, some of the mechanisms and natural interactions that I believe play a vital role in the development of infant speech are not used in the field of speech technology. Building speech recognizers typically involves training statistical models on large speech corpus databases. These are unnatural in that much of the data has previously been labeled by hand. In comparison, human infants are presented with significant quantities of unlabelled data. Infants also actively explore their environment, which involves bidirectional communications (such as asking questions and getting evaluations) from their caregivers. The current state-of-the art speech synthesis systems use the concatenation of labeled real speech sounds from large databases, which are played out in the appropriate order. Although the performances of synthesis systems are closer to human performance than recognition systems, they still have their limitations. In particular the prosody and overall naturalness still leave room for improvement.

If researchers can discover the basic computational principles that lead infants to learning to speak, it may be possible to build machines that can then learn to speak in a similar fashion. In real infants the ability to speak lags behind their ability to understand speech. That is, their speech perception ability leads production. Although speech perception plays an important role in the development of speech production, here the discussion concentrates on production after the initial (10-50 word) stage. However perception is also discussed where appropriate.

2. Infant speech acquisition

Language is normally acquired from a young age and with the child brought up in a coherent linguistic environment (i.e. the parents and peer group all speaking just one or two languages). Given these conditions, there will be many vocal interactions between an infant and his caregivers which help him to develop the ability to speak. It is worth noting that without these interactions, speech may not develop normally. Indeed, the amount of talking that occurs between a young child and its caregivers has been shown to predict intellectual accomplishment [2]. Guidance as to how to build machines that learn to speak can be gleaned from observing these interactions.

During speech acquisition, infants progress through several stages: phonation, primitive articulation, expansion, and the canonical and integrative stages [3]. Observation of these stages also gives guidance on the underling mechanisms involved. In short, an infant appears to explore his vocal apparatus and is rewarded for finding interesting or “good” actions firstly by their sensory consequences and later by encouragement from his caregiver. After basic actions have been learned, these can be used as building blocks for more complex utterances.

Because the infant must be able to distinguish responses from his caregiver, he requires a sophisticated speech perception capability. It is well known that the ability of an infant to distinguish speech utterances leads his ability to produce utterances. To ensure that a model of speech perception can in principle achieve this, it may be appropriate to model at least the initial part of the auditory system to match known psychophysical observations, e.g. using auditory filter banks.

In the first few years of life the words that infants produce differ quite considerably from the adult form, so that sometimes only their mother can understand the meaning of their utterances. This supports the idea that infants generate vocal actions to communicate and do not try to exactly copy adult form. Provided the utterances infants produce are rewarded, they will be reinforced and associated with the effect they have on the caregiver.

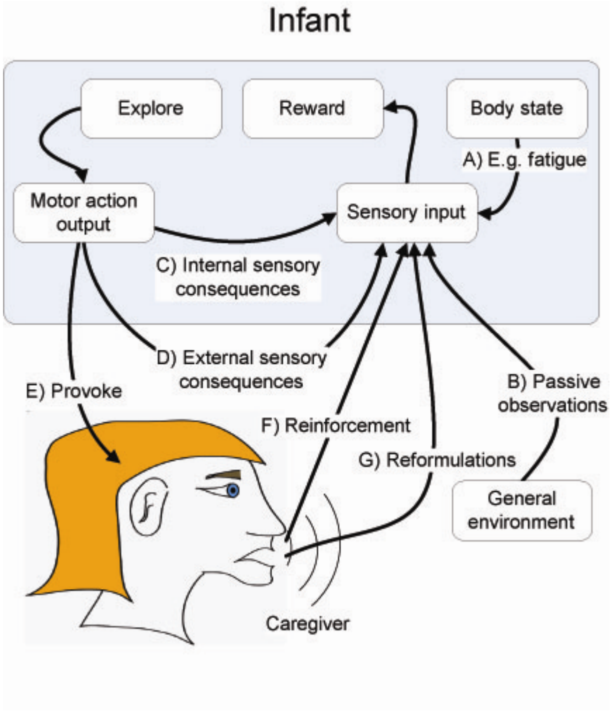

Figure 1. Infant signal flow pathways. ***A*** The state of the body is a sensory consequence of long tem action and could affect long term reward. For example this could signify hunger or fatigue. ***B*** Using passive observation of the environment an infant can self-organize his sensory systems. When he generates an action, there may be sensory consequences. ***C*** There is a proprioceptive signal flow path within the body from motor output to sensory input. ***D*** There is also a path via the external environment representing the infant hearing his own voice. ***E*** There can also be an external path that includes a "caregiver". Because the caregiver has well developed phonological perception and production she can evaluate his utterances in a linguistically appropriate fashion. ***F*** Her response can thus reward certain sounds and ***G*** her reformulation can link his production with hers.

3. Learning to speak

3.1. Current assumptions

It is generally assumed that infants learn to pronounce by imitation [4, 5] That is, they hear adult speech utterances and copy them with their own vocal productions. This may well be true for first words produced by a form of 'whole-word' mimicry, but there are a number of problems with this account if it is extended to the period when a child has started to conceive words as having internal structure. There is no evidence that speech sounds are learnt by imitation, and the developmental data does not support this viewpoint. For a discussion of these issues see [6].

3.2. Initial development

As an alternative to this, the author takes the (widely shared) view that the infant starts by discovering various potentially useful articulations. This illustrates the importance of embodiment, since the physical structure of the vocal apparatus affects what can be discovered and what speech sounds can be generated. In our previous work the author and colleagues have modeled this account of the initial development of infant speech production using an agent. Figure 1 shows the main interaction paths between an infant and his caregiver, taken from [7]. The model ran as a stand alone system and interacted naturally with an external human caregiver. This work showed that actively rewarded exploration of the vocal tract leads to the discovery of potentially useful speech sounds [8]. The computational model also extended the vocal tract model to include the effects of breathing [9]

3.3. Reformulations

In more recent work, the infant agent makes use of caregiver reformulations firstly to reinforce good utterances and also to associate his motor actions that result in vocal gestures to their corresponding caregiver reformulations [7].

Imitative exchanges between the infant and caregiver involving such reformulations of the infant's output lead to the association of the infant's motor actions to the adult judgment of their linguistic value expressed as a vocal form. These natural interactions between the infant and its caregiver give rise to a dataset rich in correspondences between infant motor patterns, infant speech and caregiver speech. This solves the correspondence problem for the sub-word units that are used in learning the pronunciation of words. That is, the infant's vocal action events become associated with the caregiver's corresponding perceptual events. Again, embodiment plays an important role here, to provoke a natural response from the subject/caregiver. If the agent didn't vocally resemble an infant, the caregiver probably wouldn't treat it like one.

3.4. Learning to imitate

As an infant is exposed to the caregiver's speech and develops the ability to generate speech utterances himself, he can link the two together. That is, his vocal actions (which give rise to his speech sounds) become associated with their corresponding caregiver reformulations. This enables the infant agent to learn words by imitation: he can be taught to reproduce words spoken by the caregiver, but using his own vocal forms. During imitation, the infant must generate the sequences of vocal actions that result in the most appropriate vocal output. The caregiver helps him to achieve this goal. Learning to imitate an utterance is generally an iterative process that will involve several exchanges between the caregiver and infant.

3.5. Learning object labels

When the caregiver speaks the names of objects in the context of an object, the infant can learn the names of objects by association. In the last stages of our work, communication was carried out in the context of objects presented visually. In this way, the infant was able to learn the names of common objects by imitating the caregiver.

Note that the field of semi-supervised learning [10] is relevant here, since the infant must initially self categorize caregiver reformulations, and his own actions, on the basis of their sensory consequences. Partially labeled data only becomes available when the caregiver provides evidence for it.

3.6. Training perception

Note that reformulation and subsequently learning the names of objects also provide a means to train word perception. Reformulations are clearly just as informative to perception as well they are to production. By interacting with the caregiver, who reacts appropriately, the infant will be able to better identify and define the categorical boundaries for speech sounds. Thus, the infant can learn perceptual contrasts using the same mechanisms that are also used to learn actions, to reinforce speech production and associate infant speech with the adult equivalent form.

4. Discussion

4.1. Summary

This paper described some of the issues involved in an infant learning to speak. Then it described some experiments demonstrating a non-imitative account of learning to pronounce. The author does not dispute that imitation plays a role in word learning, but suggests that initial speech sounds are discovered and then associated with the caregiver’s reformulations. After these associations have developed, they can be used to identify such sounds and thus to imitate sequences of them.

The work involves the use of an articulatory synthesizer, auditory and tactile sensory systems. The agent was able to learn to pronounce simple utterances following interaction with an external human caregiver. A real infant would also have access to multi-modal input, including vision. This would provide additional visual cueing from the caregiver and also objects of interest in the environment.

In our account, the use of a good model of the vocal apparatus plays a key role. Its kinematics, structure and simulated dynamics constrain the effect that control signals can have. Thus using a good model of a vocal apparatus, it is straightforward to produce speech-like sounds. Of course, a real physical model would have the potential to perform even better and give an even more convincing demonstration of learning to speak, and some researchers are already making progress in this direction [11, 12].

Concatenative speech synthesis shares some similarities to our account of human speech production. In both cases, a series of segments must be identified and concatenated. The main different is that in the former, interpolation is performed in the acoustic domain. In the latter case it is performed on the articulator domain, taking advantage of the natural interpolation between target positions that result from the dynamic and kinematic properties of the vocal apparatus. This should be “optimal” for speech production, since this is the mechanism by which natural speech is generated.

4.2. Inspirations from biology

The way in which engineers typically design and build machines is quite different from how biological systems develop. Machines are often built completed, with a structure that is hardwired from the onset. In contrast, biological systems, such as a human infant, start off as single cells which divide and differentiate as they grow and form a physical body. Even at birth, this developmental process still continues, although by then basic sensory and motor systems are in place. Cellular interactions play a fundamental role in this development and much self-organization appears to be involved. At some stage, nerve cells develop that can record and recall past sensory experience, as well as generate and record motor actions. In addition to the cellular developmental process, it is known that the development of the sensory system is also affected by sensory inputs from the environment [13] and recent work also suggests that motor system development may be similarly affected by the tasks it is required to perform [14].

The human nervous system is thus highly plastic and is continually affected by its inputs and outputs, as it learns to interpret and control its environment. From this perspective, learning to explore and control the vocal apparatus is a natural extension of the developmental processes. It may therefore be beneficial to design machines using similar principles.

Clearly, in learning vocal motor actions that lead to the generation of utterances, the ability to recognize and distinguish acoustic input and the ability to form associations between the two are only some of the abilities needed to learn to understand and produce speech. There is also the need to develop more complex relationships between input and output and also to relate them to internal rewards and desires. It has long been suggested that the human brain implements a model of its world [15]. Such an internal model appears useful for many reasons. It can use the model as a prior belief on observations, so that it is possible to interpret sensory input even when little sensory evidence is present. For example, people can walk down a familiar staircase in almost complete darkness because they have an expectation of where the stairs are. Such a model is valuable in control where time delays are involved. If this includes a model of a caregiver, this understanding makes it possible for the infant to say the right thing to get the caregiver to do what is required by the infant.

The human nervous system has several identifiable regions ranging from the neo-cortex to the brainstem and spinal cord. The neo-cortex is considered to be the seat of high level information processing and can learn complex relationships and models of the environment. A deeper understanding of the cortex in particular may well shed light on how to build more sophisticated computer models of speech perception and production and research is being carried our in these areas [16]. Generative models are currently being developed along these lines as a means to model the deep structure in sensory data [17]. However how goal directed performance evaluation, such as reward, can be incorporated into such models is not yet clear.

In the future it should be possible to formulate the infant's learning and response generation in a Bayesian framework which can take the evidence (including reliability) of each data source into account. This could include self monitoring, which could then be used to stabilize production in the event of changes in the vocal apparatus due to muscle fatigue, or perturbations.

4.3. Conclusions

Infants learn complex articulator movements when they learn to speak. There are still fundamental and unresolved issues as to how motor patterns are represented within the human brain. Similarly, although people are able to learn complex skills using their articulators as well as their limbs and extensions of them (e.g. tools), work still needs to be done to investigate the underlying mechanisms and strategies employed by the motor system which enables it to learn such arbitrary tasks. Designing machines that acquire the ability to speak by infant-like learning is likely to have wider applications than just speech. Leaning to control the arms and hand, as well as locomotion may well also benefit from a similar approach.

Machine implementations of speech recognition and synthesis currently differ from how infants learn to produce and understand speech. Modeling how infants do this may well lead to systems that perform better. Of course, if a machine that learns to speak using exploration and interactions with caregivers as discussed above is built, it is not necessary for each new instance of the machine to start from scratch and learn for itself. They could just be duplicated.

The extent to which everything related to speech production and perception in an infant needs to be exactly modeled to produce useful technology is unclear. On one side, it is known that the approach taken by nature does result in systems (people) that currently work better than machines. However, once researchers understand more of the principle of such systems, there may well be alternative engineering approaches that will also work very well, maybe even better. For example, is it necessary to use an auditory filter bank representation as a basis of speech recognition? It may well be that other representations are equally good or better. Similarly, although using an articulator synthesizer facilitates learning by exploration and leads to natural sounding speech it may be possible to adopt other approaches that still maintain the essential characteristics. For example it seems unlikely that vocal tract modeling must extend to the problems that arise from real muscles, which actually perform quite poorly compared to mechanical actuators in terms of their accuracy and precision.

Of course, if the primary research goal is to explain natural speech data and infant development, with all its idiosyncrasies, it is probably necessary to accurately model many aspects of the human vocal system including the vocal tract, the respiratory system, neural control and speech perception. This is currently where my interest lies.

Acknowledgements

I thank Piers Messum for helpful discussion and comments on the manuscript. I also thank James Ingram for proof-reading the manuscript.

References

1. Moore, R.K., *PRESENCE: A human-inspired architecture for speech-based human machine interaction.* IEEE Trans. Computers, 2007. **56, No.9** (Special Issue on Emergent Systems, Algorithms an Architectures for Speech-Based Human-Machine Interaction).
2. Risley, T. and B. Hart, *Promoting early language development.* In N. F. Watt, C. Ayoub, R. H. Bradley, J. E. Puma & W. A. LeBoeuf (Eds.), The crisis in youth mental health: Critical issues and effective programs, Volume 4, Early intervention programs and policies 2006: p. 83-88.
3. Oller, D., *The emergence of the speech capacity*. 2000: Lawrence Erlbaum Associates, Mahwah, NJ.
4. Kuhl, P.K., P. Salapatek, and L. Cohen, *Perception of speech and sound in early infancy*, in *Handbook of Infant Perception, Vol 2*. 1987, AP: New York. p. 275-382.
5. Kuhl, P.K. and A.N. Meltzoff, *Infant vocalizations in response to speech: Vocal imitation and developmental change.* Journal of the Acoustical Society of America 100 (4), 2425-2438, 1996.
6. Messum, P.R., *The Role of Imitation in Learning to Pronounce*. 2007, University of London: London.
7. Howard, I. and P. Messum, *Modeling the development of pronunciation in infant speech acquisition.* Motor Control In Review.
8. Howard, I. and P. Messum. *A Computational Model of Infant Speech Development*. in *XII International Conference "Speech and Computer" (SPECOM'2007)*. 2007. Moscow, Russia: Moscow State Linguistics University.
9. Howard, I. and P. Messum. *Modeling motor pattern generation in the development of infant speech production*. in *8th International Seminar on Speech Production – (ISSP'08)*. 2008. Strasbourg, France: INRIA
10. Zhu, X. (2005) *Semi-Supervised Learning Literature Survey*. http://pages.cs.wisc.edu/ jerryzhu/research/ssl/semireview.html.,
11. Hofe, R. and R. Moore, *Towards an investigation of speech energetics using 'AnTon': an animatronic model of a human tongue and vocal tract.* Connection Science, 2008. **20**(4): p. 319-336.
12. Yoshikawa, Y., et al., *A constructivist approach to infants' vowel acquisition through mother-infant interaction.* Connection Science 14 (4), 245-258, 2003.
13. Barlow, H.B., *Possible principles underlying the transformation of sensory messages.*, in *Sensory Communication*, W. Rosenblith, Editor. 1961, M.I.T. Press, Cambridge MA. p. 217.
14. Howard, I.S., et al., *The Statistics of Natural Movements are Reflected in Motor Errors.* J Neurophysiol, 2009.
15. Craik, K.J.W., *The Nature of Explanation*. 1943: Cambridge University Press
16. Hinton, G.E., *Learning multiple layers of representation.* Trends Cogn Sci, 2007. **11**(10): p. 428-34.
17. Hinton, G.E., S. Osindero, and Y.W. Teh, *A fast learning algorithm for deep belief nets.* Neural Comput, 2006. **18**(7): p. 1527-54.