Journal of Phonetics 53 (2015) 61–65

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Modeling the emergence of syllabic structure

Anne S. Warlaumont

Cognitive and Information Sciences, University of California, Merced, 5200 N. Lake Rd., Merced, CA 95343, USA

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form
1 May 2015
Accepted 20 June 2015
Available online 15 July 2015

Keywords:
Syllabicity
Canonical babbling
Learning
Computational modeling
Vocal development
Speech sound emergence

ABSTRACT

Most computational models that have addressed the development of consonant–vowel syllable systems have assumed a preexisting tendency to produce syllabically structured utterances. For instance, the COSMO model assumes a consistent base of jaw movement upon which finer-grained articulatory patterns are learned, motivated by MacNeilage and Davis's frame-then-content theory (FCT). While research operating under this assumption has provided much useful information on infant speech development, it does not address the gradual transition from non-syllabic to syllabic sounds that occurs during the first seven months of human infancy. It is important to consider the role that learning plays in this very significant transition. This paper discusses two computational models that address how infants may learn to produce syllabic utterances. Future work should develop agent-based models of sound production that show how syllabicity itself can emerge in a population of agents.

1. Introduction

In this issue's target paper, Moulin-Frier, Diard, Schwartz, and Bessière (hereafter MDSB) introduce their COSMO model of the "emergence of phonological systems". It is an ambitious model that attempts to provide a computational account for a variety of different features of phonological systems and of phonological processing. One highlight is the elegant, comprehensive framework for formally implementing, from a Bayesian standpoint, both sensory and motor theories of speech perception. For example, using this framework, it was found that motor constraints reduce the variability in the converged upon sound systems. Another nice finding from the work is that the quantity and distribution of noise in formant space has a substantial effect on which specific vowels the population of agents converge on. This provides a nice complement to previous models (e.g. de Boer, 2001; Oudeyer, 2006). There is much that could be discussed and debated. I will focus here on the consonant–vowel syllable portion of the authors' modeling effort.

In the last section of their paper, MDSB present their work using COSMO to model the emergence of consonant–vowel syllable systems. As the authors point out, this is something that has not received enough attention in the modeling literature. One possible reason is that syllabic production is an inherently dynamic process, involving transitions between states of the vocal tract. This may be a big reason that so much speech sound evolution and development modeling, at least that which has included articulatory synthesis, has chosen to focus on vowel systems rather than include consonants (Oudeyer, 2005). To deal with this dynamic nature of syllable production, MDSB's model makes the assumption that all syllables have CV form and that individuals begin with a pre-existing jaw oscillation "frame". The jaw oscillation frame ensures that consonant–vowel (CV) syllables will be produced, allowing the authors to focus on which *particular* CV combinations emerge in their agents' inventories.

2. The jaw oscillation assumption

The motivation for the assumption of a jaw oscillation frame comes from the "frame then content" theory (FCT) of speech acquisition and evolution (Davis & MacNeilage, 1995; Davis, MacNeilage, & Matyear, 2002; MacNeilage, 1998; MacNeilage, Davis, Kinney, & Matyear, 2000). FCT proposes that rhythmic jaw movement forms the initial basis for producing syllabic utterances. For

E-mail address: awarlaumont2@ucmerced.edu
URL: http://www.annewarlaumont.org

0095-4470/$ - see front matter © 2015 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2015.06.004

infants who have not yet learned fine-grained articulatory control, depending on where the tongue happens to be positioned at the start of the rhythmic jaw movement, different vowels and consonants will be produced. Over the course of development, there is refinement of movement by other articulators such as the tongue and lips, leading to volitional control over which specific vowels and consonants are produced and to a greater variety of consonant–vowel combinations being produced.

One of FCT theory's main predictions is that during the canonical babbling period, within infants' syllabic vocalizations, certain vowels and consonants will co-occur with greater likelihood than would be expected by chance (Davis & MacNeilage, 1995; Davis et al., 2002; MacNeilage, 1998; MacNeilage et al., 2000). In particular, it was proposed that front vowels and alveolar consonants would co-occur at high rates, mid vowels and labial consonants would co-occur at high rates, and back vowels and velar consonants would co-occur at high rates; other combinations, for example front vowels and velar consonants would co-occur less frequently. This prediction has been supported by phonetic transcription studies of syllabic infant vocalizations. The CV combinations that were predicted to co-occur at high rates do show higher rates of production than would be expected from the base rates of production of any of the individual consonants and individual vowels (Davis & MacNeilage, 1995; Giulivi et al., 2011). This has been demonstrated across many studies cross-linguistically, and is observed in infant babbling, including babbling of typically developing, hearing impaired, and developmentally delayed infants, in first words, and in adult speech (Davis et al., 2002; MacNeilage et al., 2000).

Despite these strengths, FCT theory leaves some significant features of infant babbling still to be explained. FCT only applies to infants who have already reached the canonical babbling stage, which on average occurs at roughly 7 months of age. Even among infants who have reached this stage, FCT only applies to syllabic babbling, not to the other sounds the infant still produces that are not canonical. While learning to fine-tune the articulator movements required for generating specific, distinct consonant and vowel types is certainly an important aspect of speech development, we need an answer to the question, how do children acquire syllabic vocalizations in the first place? This question will be the focus of the rest of this paper.

Previous authors have proposed that the jaw oscillation frame has evolutionary roots in feeding movements, such as chewing, perhaps with lip smacking as an intermediate step (MacNeilage, 1998). Work by Ghazanfar and colleagues on the dynamics of hyoid, tongue, and lip movements during chewing and lip smacking in macaque monkeys supports the connection between lip smacking and speech (Ghazanfar & Takahashi, 2014; Ghazanfar, Takahashi, Mathur, & Fitch, 2012). The work shows that chewing in macaques is a stereotyped movement that has a relatively slow frequency even in adult monkeys. In contrast, lip smacking is a stereotyped communicative behavior that in adult monkeys has a faster frequency more similar to that of syllabic movements in speech. Furthermore, juvenile macaques execute the lip smacking behavior more slowly, paralleling the trajectory of development of speech movements in humans, which tend to be slower for children than adults (Morrill, Paukner, Ferrari, & Ghazanfar, 2012). This possible connection between lip smacking and speech is an important insight, but it still leaves open the question of what immediate developmental processes underlie human infants' and juvenile macaques' development of rhythmic jaw movements.

One possibility is that rhythmic vocal movement is a result of entrainment to rhythmic behavior of other body parts, cortical maturation, or some combination of the two. Rhythmic vocal babbling and rhythmic arm movement tend to emerge at around the same age (Ejiri & Masataka, 2001; Iverson, Hall, Nickel, & Wozniak, 2007; Locke, Bekken, McMinn-Larson, & Wein, 1995), suggesting possible entrainment of rhythmic vocalization to rhythmic limb movement. Additionally, the brain undergoes major reorganization over the course of the first year of life (Huttenlocher, 1990), and neural network simulations suggest that after the first few months of life human brain networks shift into a state of criticality, generating changes in overall neural activity patterns (Kozma, Puljic, & Freeman, 2012). Changes in brain dynamics could potentially underlie the changes in the dynamics of behavior seen across motor systems. Thus, it is conceivable that rhythmic jaw and other articulator oscillations originate from oscillations that emerge in other motor systems and/or in brain dynamics (see also Ghazanfar & Katz, 1998).

Another possibility is that that jaw, lip, and other articulator oscillation emerges in human infancy through learning. Rhythmic babbling emerges and gradually increases over the course of human infancy (Koopmans-van Beinum & van der Stelt, 1986; Nathani Iyer & Oller, 2008; Oller, 1980, 1997; Stark, 1980). Prior to producing canonical babbling, infants progress through stages of gooing and marginal babbling, where syllables are less frequent and sloppier. Even after progressing into the canonical stage, infants continue over a period of months to increase their rates of production of canonical syllables. Furthermore, canonical babbling is delayed in infants with severe or profound hearing impairment, indicating that audition plays a role in the development of syllabic vocalizations (Nathani Iyer & Oller, 2008; Oller, 1988). These findings suggest environmental influences on the development of rhythmic speech articulator movement and a role for learning in the generation of oscillatory jaw movement itself. Although the rest of this paper will focus on this role of learning, note that there being a role of learning does not exclude the possibility that rhythmic entrainment, neural maturation, or anatomical maturation are also important factors. Canonical babbling is likely an emergent phenomenon, supported by a number of underlying processes (Davis & Bedore, 2013; Oller, 2000).

The next section will provide some brief background on other computational models besides MDSB's that also address how consonant–vowel combinations are learned while assuming pre-existing syllabic structure. This will be followed by discussion of two computational models that have addressed the question of how infants might *learn* to produce syllabic utterances.

3. Approaches to modeling syllable generation

MDSB are not alone in assuming pre-existing syllabic structure when modeling the development of articulatory skills. Guenther (1994), Kröger, Kannampuzha, and Neuschaefer-Rube (2009), Howard and Messum (2011), and Kröger, Kannampuzha, and Kaufmann (2014) have assumed a CV or VC form and then focused on how the specific combinations of articulators and their

degrees of closure could be learned. These modeling approaches have proven very useful for studying a number of interesting aspects of vocal learning, such as how infants form correspondences between articulatory motor movements and auditory percepts, and how they form correspondences between these and phoneme representations. A slight variation is to allow the timings of the consonants and vowels within a syllable to vary. This is the approach taken by Howard and Messum (2011, 2014) and Philippsen, Reinhart, and Wrede (2014). The coupled oscillator model of Nam, Goldstein, Giulivi, Levitt, and Whalen (2013) assumes a syllabic unit but allows the timing of single cycles of oscillation of specific articulators to dynamically reorganize in a coarticulatory fashion. The model has shown good performance in predicting the frequencies of various CV combinations by calculating the stability of the coupled articulatory oscillators involved (see also Lindblom, MacNeilage, & Studdert-Kennedy, 1983).

Thus, assuming some sort of syllabic structure or frame has allowed computational modelers to provide mechanistic explanations for a number of different phenomena regarding the development of speech production capabilities. However, in assuming that infants have a pre-existing propensity to produce syllabic vocalizations (i.e, vocalizations that have alternations between the vocal tract being relatively open and being relatively closed), these models necessarily only address aspects of speech production skill development from the canonical babbling period (beginning at about 7 months of age) and beyond. Furthermore, by ignoring the ontogenetic origins of the propensity to babble syllabically, the models may miss out on key features of how syllabicity is represented in the infant's motor control system. For example, are the relative frequencies of different consonant–vowel combinations in human infant babble and speech best accounted for by an oscillatory frame that purely concerns the jaw, or by an oscillatory frame that may be biased toward jaw movement but also includes oscillations of other articulators? Having an understanding of how syllabic structure emerges in the first 7 months of life (and why it continues increasing in frequency for many months afterward) should help inform this question. A closely related question is how syllabic utterances relate to non-syllabic utterances, which continue to be frequently produced even well after canonical babbling appears.

At least two recent models, one by Moulin-Frier, Nguyen, and Oudeyer (2014) and one by myself (Warlaumont, 2012, 2013), have attempted to totally do away with the assumption of a pre-existing syllabic frame and instead explain how syllables containing both consonants and vowels could emerge through learning. Moulin-Frier et al.'s (2014) model controls vocalization by setting the locations and heights of the centers of five Gaussian functions, each defined in a multidimensional space the size of the number of articulators in the model's vocalization synthesizer. The model is driven by an intrinsic motivation to maximize its learning about its own sensorimotor mappings. By “intrinsic motivation”, the authors refer to the tendencies for children (and for their model) to be curious about the way the actions they do affect the world around them and to desire to make progress in learning these relationships. Intrinsic motivation contrasts with motivation that is driven by objective rewards or punishments. The model keeps track of how much progress it is making with regard to understanding which actions it takes to achieve various sensorimotor outcomes. It then chooses to aim for specific sensory goals based on which will lead to the greatest sensorimotor learning. Early in development, the model does not know anything about the relationship between its actions and the sensory consequences, so it learns the most from setting goals to do relatively simple things. As it masters these simpler goals, it progresses to targeting more complex behaviors. Thus, at first, Moulin-Frier et al.'s model tends to produce mostly unphonated sounds. After a short time, it begins to preferentially explore its unarticulated (i.e. vowel only) sound space. Finally, the model begins to prefer to explore articulated sounds (i.e. syllabic sounds containing at least one vowel and at least one consonant). By taking a very flexible articulatory control approach and focusing on maximizing learning progress, the model exhibits a transition from producing mostly unphonated sounds to producing mostly vowel sounds to producing mostly syllabic babbling, thus helping explain how syllables could emerge over the course of an individual's development. Learning to produce vocalizations with syllabic structure and learning to produce specific consonant–vowel combinations are accomplished through a single learning mechanism.

My own recent modeling efforts have also targeted the question of how syllabicity might be learned. I have utilized a recurrent spiking neural network to control movement of the masseter (which promotes jaw closure) and orbicularis oris (which promotes lip closure) during vocalization (Warlaumont, 2012, 2013). The model contains a network of 1000 spiking cortical neurons. Due to the recurrent connections between neurons as well as the nonlinear functions that determine the neurons' responses to input current, the overall activity of the network exhibits complex oscillatory dynamics. These oscillations are harnessed for the dynamic control of speech articulators. Sounds synthesized using these vocal tract muscle activations are then selected for reinforcement either by a human listener or by an algorithm that estimates the overall auditory salience of the sound. An assumption, based on previous research with human infants, is that both social responses and auditory interest are possible sources of intrinsic rewards for infants. The neural network updates its connection weights via spike timing dependent plasticity, a form of Hebbian learning, but only when reinforced. Reinforcement, whether by human or by auditory salience, tends to be more likely the more syllabic the model's vocalizations are. Initially the model produces vocalizations that only rarely include consonant elements and mostly are pure vowel sounds. After a period of exploration and selective reinforcement, however, the model comes to more frequently produce vocalizations that contain consonant elements. The spiking neural network's natural oscillatory activity when translated into muscle activity will occasionally rise above threshold so that the lips meet or nearly meet and a consonant or semivowel sound is generated. Through reinforcement-modulated plasticity, the model modifies this oscillatory activity so that it that rises above this threshold more often, thus generating consonant and semivowel sounds more often. The learning mechanism contrast somewhat with Moulin-Frier's approach, which relies on intrinsic motivation to learn new sensorimotor mappings; here no sensorimotor mappings are learned, and instead patterns of neural activity leading to receiving a reward are learned.

The advantage of this model is that it relates the production of oscillatory neural activity to oscillatory vocal activity and provides an account for how syllabic vocalizations might emerge through a biologically plausible learning mechanism. A current shortcoming is that, for the sake of simplicity as an initial version of a new model, it controls only one articulatory degree of freedom. Work still needs

to be done to test whether such a model will scale up to the control of multiple articulators. Nevertheless, it supports the idea that cortical dynamics could generate the oscillatory motor activity suitable for syllable generation, and that learning could help harness this oscillatory cortical activity to generate consonants at increasing rates (see also Ghazanfar & Katz, 1998). With the jaw being one of the articulators controlled, the work could be viewed as providing a potential mechanistic theory for how the jaw oscillation invoked in FCT might emerge during the first 7 months or so of life.

Taken together, these two models show that it is possible to build computational models that learn to produce increasingly syllabic structures, providing mechanistic accounts for the process that occurs in human infancy. This is a significant step forward, as most previous models addressing consonant and vowel learning have assumed that the infant already produces syllabically structured vocalizations, and that the problem is primarily one of filling in which specific movements should fill the consonant role(s) and which should fill the vowel role within a syllable. The fact that the two models have substantial differences in their learning mechanism (intrinsically motivated sensorimotor learning vs. social or salience-based reward-driven learning) suggests the possibility that multiple learning mechanisms could support the development of canonical, i.e. syllabic, babbling, an idea consistent with the overall robustness of the development of canonical babbling observed across many different groups of infants (Oller, 2000).

4. Conclusion

Assuming a preexisting syllabic frame, or some sort of syllabic template more generally, in computational modeling of the emergence of speech sound systems has its advantages. For one, it has allowed modelers to temporarily side-step the issue of how syllabicity might have emerged in the first place and instead focus on important questions such as how humans learn to associate speech motor acts with sensory inputs, how they learn to associate both of these with phonemic representations, which specific consonant–vowel combinations are most stable, and more.

However, considering the fact that syllabic babbling is not consistently produced until around 7 months of age, and that it appears to develop somewhat gradually during the preceding time period, it is worth seriously studying how syllabicity emerges in development and might have emerged as such a stable phenomenon across the world's languages. Some work has already begun to provide computational accounts for how syllabicity might emerge in development. A greater effort toward modeling not only segment-level learning but also the emergence of syllabic structure, exploring different learning mechanisms (including intrinsically motivated sensorimotor learning, reward-based learning, and perhaps others), assessing the role of anatomical and neural maturation, and exploring different levels of biological detail (biologically plausible neural learning and more abstract models that provide a more high-level explanation of the learning process) can be expected to lead to a more complete understanding of the development of speech production skills in human infancy.

Another important step is to begin to address the emergence of syllabicity in agent-based models of the emergence of vocal communication systems. Considering that syllabic vocalization is unique to humans among the primates, and that it appears to precede the emergence of control over specific segment types, this will also allow agent based models to address an even more fundamental characteristic of human speech.

References

Davis, B. L., & Bedore, L. M. (2013). *An emergence approach to speech acquisition*. New York: Psychology Press.

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech and Hearing Research*, *38*, 1199–1211.

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

de Boer, B. (2001). *The origins of vowel systems*. Oxford: Oxford University Press.

Ejiri, K., & Masataka, N. (2001). *Developmental Science*, *4*, 40–48.

Ghazanfar, A. A., & Katz, D. B. (1998). *Behavioral and Brain Sciences*, *21*, 516–517.

Ghazanfar, A. A., & Takahashi, D. Y. (2014). *Trends in Cognitive Sciences*, *18*, 543–553.

Ghazanfar, A. A., Takahashi, D. Y., Mathur, N., & Fitch, W. T. (2012). *Current Biology*, *22*, 1176–1182.

Giulivi, S., Whalen, D. H., Goldstein, L. M., Nam, H., & Levitt, A. G. (2011). *Language Learning and Development*, *7*, 202–225.

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

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

Howard, I. S., & Messum, P. (2014). Learning to pronounce words in three languages: *An investigation of caregiver and infant behavior using a computational model of an infant*. *PLOS ONE*, *9*, e110334.

Huttenlocher, P. R. (1990). Morphometric study of human cerebral cortex development. *Neuropsychologia*, *28*, 517–527.

Iverson, J. M., Hall, A. J., Nickel, L., & Wozniak, R. H. (2007). *Brain and Language*, *101*, 198–207.

Koopmans-van Beinum, F. J., & van der Stelt, J. M. (1986). Early stages in the development of speech movements. In B. Lindblom, & R. Zetterström (Eds.), *Precursors of early speech*. New York: Stockton Press.

Kozma, R., Puljic, M., & Freeman, W. J. (2012). Thermodynamic model of criticality in the cortex based on EEG/ECOG data. In D. Plenz, & E. Niebur (Eds.), *Criticality in neural systems*. Weinheim: John Wiley & Sons.

Kröger, B. J., Kannampuzha, J., & Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics*, *2*, 2.

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

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. *Linguistics*, *21*, 181–203.

Locke, J. L., Bekken, K. E., McMinn-Larson, L., & Wein, D. (1995). *Brain and Language*, *51*, 498–508.

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*, 499–546.

MacNeilage, P. F., Davis, B. L., Kinney, A., & Matyear, C. L. (2000). The motor core of speech: *A comparison of serial organization patterns in infants and languages*. *Child Development*, *71*, 151–163.

Morrill, R. J., Paukner, A., Ferrari, P. F., & Ghazanfar, A. A. (2012). Monkey lipsmacking develops like the human speech rhythm. *Developmental Science*, *15*, 557–568.

Moulin-Frier, C., Nguyen, S. M., & Oudeyer, P.-Y. (2014). Self-organization of arly vocal development in infants and machines: *The role of intrinsic motivation*. *Frontiers in Cognitive Science*, *4*, 1006.

Nam, H., Goldstein, L. M., Giulivi, S., Levitt, A. G., & Whalen, D. H. (2013). Computational simulation of CV combination preferences in babbling. *Journal of Phonetics*, *41*, 63–77.

Nathani Iyer, S., & Oller, D. K. (2008). Prelinguistic vocal development in infants with typical hearing and infants with severe-to-profound hearing loss. *The Volta Review*, *108*, 115–138.
Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. H. Yeni-Komshian, J. F., & C. A. Ferguson, (Eds.), *Child phonology, Vol. 1: Production*. Academic Press, New York.
Oller, D. K. (1988). The role of audition in infant babbling. *Child Development*, *59*, 441–449.
Oller, D. K. (1997). Development of precursors to speech in infants exposed to two languages. *Journal of Child Language*, *24*, 407–425.
Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates.
Oudeyer, P.-Y. (2005). The self-organization of combinatoriality and phonotactics in vocalization systems. *Connection Science*, *17*, 325–341.
Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech*. Oxford University Press.
Philippsen, A. K., Reinhart, R. F., & Wrede, B. (2014). Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In *Proceedings of the 2012 IEEE International conference on development and learning and epigenetic robotics*.
Stark, R. E. (1980). Stages of speech development in the first year. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson, (Eds.), *Child phonology, Vol. 1: Production*. Academic Press, New York.
Warlaumont, A. S. (2012). A spiking neural network model of canonical babbling development. In *Proceedings of the 2012 IEEE international conference on development and learning and epigenetic robotics*.
Warlaumont, A. S. (2013). Salience-based reinforcement of a spiking neural network leads to increased syllable production. In *Proceedings of the 2013 IEEE international conference on development and learning and epigenetic robotics*.