MODELING THE ACQUISITION OF INTONATION: A FIRST STEP

Michael Fry

University of British Columbia
Department of Linguistics
2613 West Mall, Vancouver, BC, V6T 1Z4

ABSTRACT

Computational models of language learning focus primarily on the emergence of segmental categories to the exclusion of intonation [e.g. 1]. This runs counter to the considerable evidence that language learners rely as much on intonation as segmental categories while learning language. The current project adapts the Sensorimotor Integration Model, a popular model of language learning, to model the development of intonation. It builds on previous work that used reinforcement learning to model the development of phonation [2]. The learning simulations use a source-filter speech synthesizer to generate utterances that are then processed into intonational phrases, analyzed as f0 and amplitude. An utterance is reinforced if it is similar, as measured via distance in a self-organizing map, to a training set of infant-directed intonational phrases. Results demonstrate that, over time, the model learns to produce adult-like intonational phrases.

***Index Terms*—** Intonation, Language Acquisition, Computational Modeling, Speech Processing, Sensorimotor Integration

1. INTRODUCTION

Current computational models of language learning focus primarily on the emergence of segmental categories to the exclusion of intonation [e.g. 1, 3, 4, 5]. These models learn to produce segments and/or partition acoustic space into vowels and consonants independent of intonation-related fluctuations in pitch, amplitude and duration. This focus is understandable because typical language development is gauged by segmental category milestones [6] and monotonic vowel-consonant sequences are still intelligible, but it sidesteps the early and active development of intonation in language-acquiring infants [see e.g. 7, 8, 9]. The goal of this paper is to provide a first step towards incorporating intonation into computational models of language learning. I do so by implementing a method that builds on previous work that used reinforcement learning in the Sensorimotor Integration Model to model the development of phonation [2, 10].

This work was supported in part by the SSHRC Insight Grant 435-2014-1673 awarded to the late Dr. Eric Vatikiotis-Bateson

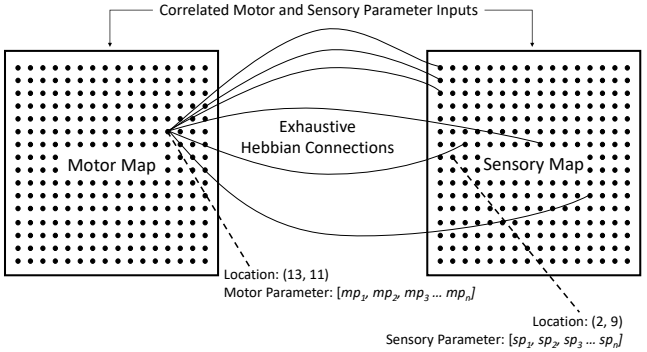
2. BACKGROUND

Typical language development is measured by segmental category development. Infants are judged according to segmental landmarks such as canonical babbling [11], the one-word stage [12] and language-specific perceptual attunement [13]. These milestones are practical in that they are observable and grounded in linguistic theory, which has long considered segmental categories and their distinctive features foundational [14, 15]. This practicality, in turn, has motivated researchers to develop computational models of speech production that focus on synthesizing segmental categories [1, 16].

Segmental categories' importance notwithstanding, several authors have noted the need to incorporate the development of intonation into current models of language learning [3, 17, 18]. The need is largely motivated by empirical findings from language acquisition research. For example, [8] and [9] (in separate studies) concluded that infants intentionally manipulate intonation either preceding the onset of the one-word stage or concurrently with it. Further, [7] found children aged 24 months produce intonational patterns consistent with those in adult speech [see also 19]. Similarly, [20] found that adult-like intonation (based on pragmatic meaning) developed rapidly in infants from 11 to 28 months of age. In combination, these studies demonstrate the prominent role played by intonation in language development that has, until now, not been implemented in a computational model of language learning.

There are many computational models of language acquisition in the literature [see 21, for a review]. With the aim of taking a first step towards modeling the acquisition of intonation, two prominent models were considered for this project: the Directions Into Velocities of Articulators (DIVA) model [1, 17] and the Sensorimotor Integration Model (SMIM) [10]. Ultimately, the SMIM was chosen because of its computational simplicity and demonstrated utility. It comprises two interconnected neural networks, corresponding to motor and sensory cortices. The SMIM learns the mapping between articulation and acoustics through two simultaneous processes: (1) the networks self-organize themselves [22], forming clusters that represent segmental categories (one in the motor domain and one in the sensory domain); and (2) connections

between the two networks update via Hebbian learning [23], associating the segmental category clusters in the two networks and allowing activation to propagate across domains. As an example of correlated input, [3]'s motor parameters were parameters for the VLAM synthesizer [24] and the sensory parameters were formant values (F1-F3) for a steady-state vowel.



(a) The Sensorimotor Integration Model [25]

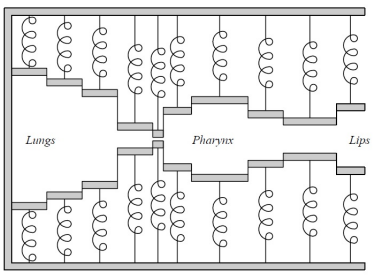
In previous research, the SMIM has been adapted to model canonical babbling [26] and the acquisition of vowels [3], and the motor map component was used to model the learning of phonation – the vibrating of the vocal folds to produce voicing [2]. The latter study is a natural precursor to the current project as phonation is measurable via f0; f0 is correlated with perceived pitch; and, pitch modulations are a crucial component of intonation. In [2], reinforcement was implemented as an on-off switch controlling self-organization of the motor map. The motor map activated vocal tract parameters in a simulated vocal tract [27] to generate an utterance that was then evaluated for the presence of phonation. If phonation were present, the network self-organized, pulling motor parameters in the map towards values that generate phonation (i.e. reinforcing desirable utterances). The current project extends this method to intonation by allowing the vocal tract parameters to vary in time and measuring intonation via f0 and RMS-amplitude. This modification necessitates a new method to evaluate when a particular motor configuration should be reinforced. To do this, self-generated utterances are compared to a set of training utterances comprised of Infant-Directed Speech, a choice meant to be analogous to an infant learning by comparing its own utterances to those of its caregiver. This comparison is done in the sensory domain, and thus the current project also extends the method of [2] to encompass the sensory map of the SMIM.

3. METHOD

3.1. Modifying the SMIM for Intonation

To incorporate intonation into the SMIM, three modifications are necessary: (i) the motor map needs to generate intonation, (ii) the sensory map needs to perceive intonation, and (iii) the perceived intonation needs to be evaluated so the model knows whether to reinforce a self-generated utterance. Intonation, herein, is defined as 'the stress, tune, phrasing...and their interactions' of spoken speech, following [28, p. 2]. In the motor domain, stress and tune are produced via modulations (through time) in energy (i.e. air pressure from the lungs) and vocal fold vibration (i.e. phonation). In the sensory domain, stress and tune are observed as modulations (through time) in pitch (measured via f0) and loudness (measured via RMS-amplitude). Of note is that duration, although important, is not explicitly considered herein because it is tied to other components of language such as syllables and segments. In both sensory and motor domains, phrasing entails that an utterance can be chunked into discrete intonational events, termed **intonational phrases** that are separated by **intonation boundaries**, defined also by Liberman as 'systematically significant pause[s]' [28, p. 286]. It is also worth stating that the standard linguistic system to annotate intonation, the Tones and Break Indices annotation system [29], similarly chunks utterances based on pauses.

To achieve (i), following the methodology of [2], motor parameters in the motor map activate muscles of the simulated vocal tract used in Praat's speech synthesizer [27]. With this synthesizer, the vocal tract is modeled as a series of ducts that approximate airflow from the lungs to the lips (each duct controlled by 'muscles'); a diagram is shown in Figure (b). Using the physics of a mass-spring system, the synthesizer estimates resonance in each duct and the cumulative effect of resonances is a speech-like vocalization. As intonation requires dynamic movement of articulators through time, muscle configurations are specified at seven time points to allow for a reasonable maximum of five inflection points per intonational phrase [28]. Also, as the current work focuses on pitch and amplitude modulations, only motor parameters for the laryngeal and lung muscles in the simulated vocal tract are set. As there are five such muscles, motor parameters in the self-organizing motor map are 35 values long (5 muscles x 7 time points).



(b) The simulated vocal tract of ducts in Praat

To achieve (ii), utterances are chunked into intonational phrases at intonation breaks (IBs) (i.e. 'significant pauses'), defined as pauses of at least 260 ms in length. This value

is based on the results of [30] and is one standard deviation shorter than the average length of phrasal pause durations of participants reading a paragraph aloud. For each intonational phrase, pitch is measured by its acoustic correlate f0 [27, 31] and amplitude is measured as RMS-amplitude with a frame length of 25ms and shift of 10ms. Sections of speech that do not have an f0 (i.e. voiceless segments) have f0 values polynomially (second degree) interpolated between the known preceding and following f0 values. While this ensures a smooth f0 signal with no null values that can serve as partial input for the self-organizing sensory map, it notedly does sidestep potential effects that voiceless segments may have on perceived pitch. Finally, the f0 signal and amplitude envelope are each normalized between [0-1] and are concatenated. The concatenation is then down-sampled to 100 data points. This 100 point vector serves as the sensory parameter (i.e. the perceived intonation) for the self-organizing sensory map.

To achieve (iii), reinforcement is implemented as an evaluation during learning that gates self-organization of the motor and sensory maps [cf. 2]. This approach ensures only utterances that are positively evaluated contribute to the organization of the map; meaning, over time the map comes to represent primarily desirable utterances. In [2], the evaluation was the presence or absence of phonation, defined as an f0 value 250ms after the start of the utterance. In the current implementation, intonational phrases are evaluated relative to a training set which is comprised of thirteen hours of Infant-Directed Speech (IDS). These data were chosen based on the assumption that a learner is motivated, at least in part, to produce utterances similar to his caregiver. A total of 7623 IDS utterances/intonational phrases were used, collected by [32] and retrieved from the CHILDES database [33]. All IDS tokens were from the same caregiver speaking to her infant in English. The training data were processed into intonational phrases (as described above) and used to train a self-organizing map [22] of the same type as the sensory map component of the SMIM. This map serves as the **target map** with which an utterance generated by the model is compared. If the utterance is sufficiently close (Euclidean Distance) to the Best-Matching Unit in the target map, it is used to self-organize the SMIM (i.e. it is reinforced).

3.2. Learning Simulations

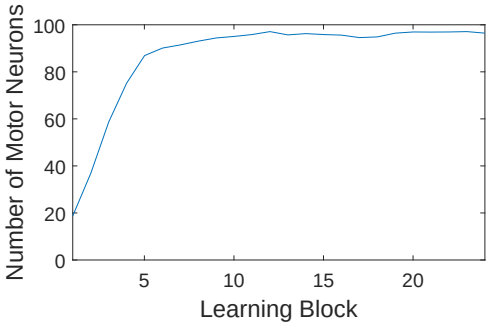
Before learning, the SMIM's motor and sensory maps are initialized with random values. Each map consists of 100 neurons, structured as 10x10 matrices. Each neuron in each map has a location and a motor/sensory parameter as in standard implementations of self-organizing maps. As previously stated, motor parameters are of length 35, feeding five muscles in the Praat Synthesizer with seven values for the seven points in time. Sensory parameters are of length 100, corresponding to the down-sampled concatenation of the f0 signal and amplitude envelope of an intonational phrase. Finally, Hebbian connections are initialized to zero.

After initialization, the SMIM begins a learning block. In a learning block, each motor parameter is fed to Praat and is used to generate an utterance. The utterance is then processed into its corresponding sensory parameter. This parameter is then matched to its Best-Matching Unit (BMU) in the target map (i.e. the map trained on IDS). If the distance between the sensory parameter and the BMU is below a threshold, those motor parameters and the corresponding sensory parameters are used to update their respective maps in the SMIM. Hebbian weights are strengthed for the pair of inputs following the method used by [3]. For the current project, the threshold was set as the average distance between each IDS utterance and its Best-Matching Unit in the target map, a value of 1.57 units.

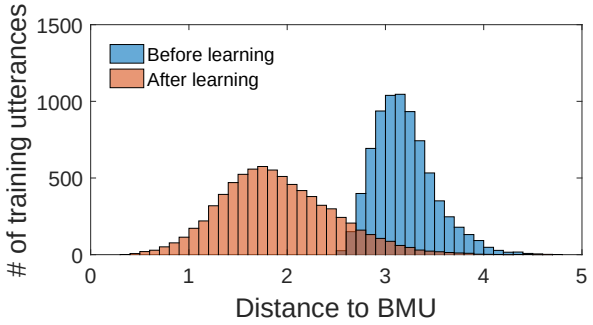
4. RESULTS

As the SMIM is randomly initialized, the end state of each learning simulation varies; however, the general performance of each simulation is similar. The results reported here are averages over 10 learning simulations. Two metrics to evaluate learning are presented: (1) the number of reinforced utterances from the SMIM's motor map per learning block; and (2) histograms of distances between all IDS utterances and the BMU of the SMIM's sensory map before and after learning. Following this, an exemplar that demonstrates the change of a single sensory neuron throughout learning is presented.

The total number of motor neurons with parameters that produce reinforced utterances demonstrates learning as more such neurons means the network produces more adult-like intonational utterances. Figure (c) shows that the number of randomly initialized motor parameters that produce reinforced utterances is initially around 20. Over time, as the model self-organizes, the number of motor parameters that produce reinforced utterances nears 100, encompassing all motor neurons in the motor map.



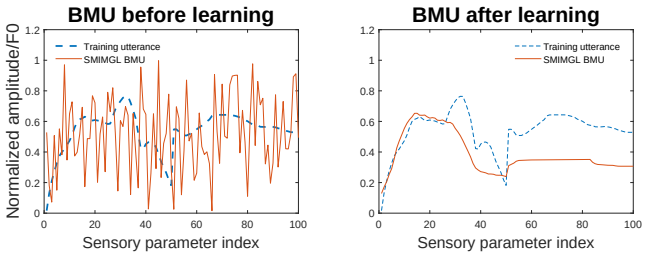
(c) Motor neurons that produce reinforced utterances per learning block



(d) Histogram of distances between IDS utterances and their BMUs in the SMIM before and after learning

The average distance between IDS training utterances and their respective BMU in the SMIM's sensory map before and after learning is shown in Figure (d). As the SMIM learns to produce desirable utterances, the distances between IDS utterances and their respective BMUs should decrease. This indicates that the learned sensory map better matches utterances that are known to be desirable (i.e. adult speech). The mean distance before learning is ≈ 3.1 units of distance; the mean after learning is ≈ 1.6. This confirms that the SMIM's sensory map better represents the IDS data after learning. A Student's T-Test performed on the distributions confirms a significant difference with t(7622) ≈ 172, p<$1e-10$, and a Cohen's D ≈ 1.62.

Finally, Figure (e) shows an explicit example of learning; it compares the SMIM's BMU unit to a single IDS utterance before and after learning. The random initialization of the SMIM is seen in the jagged pattern before learning; the smooth sensory parameter seen after learning is a direct result of desirable utterances being reinforced.



(e) The SMIM's BMU compared to a training utterance before and after learning

5. DISCUSSION

The results confirm that the SMIM is able to learn to produce adult-like intonational phrases. The number of motor neurons that produce desirable utterances reaches ceiling after ≈ 13 learning blocks, showing that each neuron in the motor map has learned to produce an adult-like intonational phrases. Also, the distribution of distances to the BMU of IDS utterances shows that the sensory map after learning more closely represents the caregiver's utterances. This is of particular note as the learning model has only modified its own sensory map from its own self-generated utterances and not explicitly from IDS utterances. That said, the model does guide its own learning through comparison to the IDS utterances. This method is consistent with learning by imitation, but it is unclear that a human infant would actually learn in the similar way. It is an open question whether an infant generates a perceptual map of its own utterances via comparison to an already known map of its caregiver's utterances, or, if there is a single sensory map that the learner is continually updating from multiple inputs. In fact, evidence from profoundly deaf infants that phonate and marginally babble up until ≈ 6 months of age [34] suggests that infants do not require a target sensory map in order to begin developing their ability to manipulate intonation. This may result from learning to intonate in two (likely overlapping) stages: an autonomous, exploratory learning stage and then a stage of imitation.

There are also clear improvements to be made following the first steps taken here. For one, the treatment of intonation as solely a pitch contour and amplitude envelope is an oversimplification. For instance, fundamental frequency (the measurement of the pitch contour) may vary because of inherent vowel quality or aerodynamic consequences from a particular articulation and not because of intentional manipulation. Also, modeling the development of intonation without the use of segmental categories is incomplete as intonation and segmental categories develop simultaneously. Finally, intonation interacts with lexical stress in non-trivial ways. Future work should address these complicated issues.

In conclusion, this project set out to take a first step towards incorporating intonation into current models of language learning. This was achieved by modifying the motor and sensory parameters in Sensorimotor Integration Model to allow for the processing of intonation.

References

[1] Frank H Guenther, "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production.," *Psychological review*, vol. 102, no. 3, pp. 594, 1995.

[2] Anne S Warlaumont, Gert Westermann, Eugene H Buder, and D Kimbrough Oller, "Prespeech motor learning in a neural network using reinforcement," *Neural Networks*, vol. 38, pp. 64–75, 2013.

[3] Ilana Heintz, Mary E Beckman, Eric Fosler-Lussier, and Lucie Ménard, "Evaluating parameters for mapping adult vowels to imitative babbling.," in *INTERSPEECH*, 2009, vol. 9, pp. 688–691.

[4] Hisashi Kanda, Tetsuya Ogata, Kazunori Komatani, and Hiroshi G Okuno, "Segmenting acoustic signal with articulatory movement using recurrent neural network for phoneme acquisition," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 1712–1717.

[5] Ian Spencer Howard and Piers Ruston Messum, "A computational model of infant speech development," in *XII International Conference" Speech and Computer"(SPECOM'2007)*, 2007, pp. 756–765.

[6] Mildred C Templin, "Certain language skills in children; their development and interrelationships.," 1957.

[7] Lluïsa Astruc, Elinor Payne, Brechtje Post, Maria del Mar Vanrell, and Pilar Prieto, "Tonal targets in early child english, spanish, and catalan," *Language and speech*, vol. 56, no. 2, pp. 229–253, 2013.

[8] David Snow and Heather L Balog, "Do children produce the melody before the words? a review of developmental intonation research," *Lingua*, vol. 112, no. 12, pp. 1025–1058, 2002.

[9] Nuria Esteve-Gibert and PILAR Prieto, "Prosody signals the emergence of intentional communication in the first year of life: Evidence from catalan-babbling infants," *Journal of child language*, vol. 40, no. 5, pp. 919–44, 2013.

[10] Gert Westermann, "A model of perceptual change by domain integration," in *Proceedings of the 23rd annual conference of the cognitive science society*, 2001, pp. 1100–1105.

[11] Peter F MacNeilage, "The frame/content theory of evolution of speech production," *Behavioral and brain sciences*, vol. 21, no. 04, pp. 499–511, 1998.

[12] Eve V Clark, *First language acquisition*, Cambridge University Press, 2009.

[13] Janet F Werker and Richard C Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant behavior and development*, vol. 7, no. 1, pp. 49–63, 1984.

[14] Roman Jakobson, Gunnar Fant, and Morris Halle, "Preliminaries to speech analysis. the distinctive features and their correlates," 1951.

[15] Noam Chomsky and Morris Halle, "The sound pattern of english.," 1968.

[16] Shinji Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*, pp. 131–149. Springer, 1990.

[17] Jason W Bohland, Daniel Bullock, and Frank H Guenther, "Neural representations and mechanisms for the performance of simple speech sequences," *Journal of cognitive neuroscience*, vol. 22, no. 7, pp. 1504–1529, 2010.

[18] Gert Westermann and Eduardo Reck Miranda, "A new model of sensorimotor coupling in the development of speech," *Brain and language*, vol. 89, no. 2, pp. 393–400, 2004.

[19] Haydée Marcos, "Communicative functions of pitch range and pitch direction in infants," *Journal of Child Language*, vol. 14, no. 02, pp. 255–268, 1987.

[20] Pilar Prieto, Ana Estrella, Jill Thorson, and Maria del Mar Vanrell, "Is prosodic development correlated with grammatical and lexical development? evidence from emerging intonation in catalan and spanish," *Journal of Child Language*, vol. 39, no. 02, pp. 221–257, 2012.

[21] Michael Frank, "Computational models of early language acquisition," unpublished manuscript.

[22] Teuvo Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[23] Donald Olding Hebb, *The organization of behavior: A neuropsychological theory*, Psychology Press, 2005.

[24] Lucie Ménard, Jean-Luc Schwartz, Louis-Jean Boë, Sonia Kandel, and Nathalie Vallée, "Auditory normalization of french vowels synthesized by an articulatory model simulating growth from birth to adulthood," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1892–1905, 2002.

[25] Gert Westermann and Eduardo Reck Miranda, "Modelling the development of mirror neurons for auditory-motor integration," *Journal of new music research*, vol. 31, no. 4, pp. 367–375, 2002.

[26] Anne Warlaumont, Gert Westermann, and D Kimbrough Oller, "Self-production facilitates and adult input interferes in a neural network model of infant vowel imitation.," 2011.

[27] Paul Boersma et al., *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*, Holland Academic Graphics/IFOTT, 1998.

[28] Mark Yoffe Liberman, *The intonational system of English.*, Ph.D. thesis, Massachusetts Institute of Technology, 1975.

[29] Kim EA Silverman, Mary E Beckman, John F Pitrelli, Mari Ostendorf, Colin W Wightman, Patti Price, Janet B Pierrehumbert, and Julia Hirschberg, "Tobi: a standard for labeling english prosody.," in *ICSLP*, 1992, vol. 2, pp. 867–870.

[30] Robert S Brubaker, "Rate and pause characteristics of oral reading," *Journal of Psycholinguistic Research*, vol. 1, no. 2, pp. 141–147, 1972.

[31] Xuejing Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–333.

[32] Michael R Brent and Jeffrey Mark Siskind, "The role of exposure to isolated words in early vocabulary development," *Cognition*, vol. 81, no. 2, pp. B33–B44, 2001.

[33] Brian MacWhinney, *The CHILDES project: The database*, vol. 2, Psychology Press, 2000.

[34] D Kimbrough Oller and Rebecca E Eilers, "The role of audition in infant babbling," *Child development*, pp. 441–449, 1988.