Towards modeling the acquisition of intonation*

Michael Fry
University of British Columbia

Abstract: Computational models of language learning focus primarily on the emergence of segmental categories to the exclusion of intonation. This focus is understandable as the intelligibility of speech is underpinned by consistent segmental category sequences; however, there is substantial evidence that language learners rely as much, if not more, 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 employed reinforcement learning to model the development of phonation. The learning simulations use Praat's speech synthesizer to generate spontaneous utterances that are then processed into intonational phrases, analyzed as f0 tracks and amplitude envelopes. An utterance is reinforced if it is similar, as measured with distance in self-organizing maps, to a training set of infant-directed intonational phrases. Results demonstrate that, over time, the model learns to produce adult-like intonational phrases.

Keywords: Language Learning, Computational Modeling, Intonation, Sensorimotor Integration

1 Introduction

Current computational models of language learning focus primarily on the emergence of segmental categories to the exclusion of intonation – the modulations of pitch and amplitude in speech. These models learn to partition acoustic space into vowels and consonants independent of fluctuations in Fundamental Frequency (f0) or intensity (acoustic measures of pitch and amplitude, respectively). This focus is understandable for several reasons: (1) typical language development is gauged by segmental category milestones; (2) vowel-consonant sequences are intelligible without intonation; and (3) it is difficult to computationally evaluate whether an intonational pattern has been acquired because intonation is highly variable; the same vowel-consonant sequence can be realized with any number of intonational patterns depending on situation and speaker goals. Nonetheless, there is clearly early and active development of intonation in infants as they learn their native language (e.g. see Astruc, Payne, Post, del Mar Vanrell, and Prieto 2013; Esteve-Gibert and Prieto 2013; Snow and Balog 2002). The absence of intonation in current computational models, therefore, provides an intriguing research opportunity. The goal of this paper is to provide a first step towards incorporating intonation into computational models of language learning.

I begin by providing a brief overview of language development research and current computational models of language learning. This is followed by the proposed method to model the development of intonation using the Sensorimotor Integration Model (Westermann 2001) and an explicit experiment to train the model. Finally, results of the training and future directions for this research area are discussed.

* Thanks to the dinosaurs, for showing us bigger isn't always better.
Contact info: mdfry20@gmail.com

In *UBC Qualifying Papers 4 (2015–2017)*,
University of British Columbia Working Papers in Linguistics,
Emily Sadlier-Brown, Oksana Tkachman, and Avery Ozburn (eds.). Vancouver, BC: UBCWPL, 2019.

2 Background

2.1 Language Development and Intonation

Typical language development is, in general, measured by segmental category development. Examples include canonical babbling (the alternation of nonsensical CVCV syllables; MacNeilage 1998), the one-word stage (the consistent and meaningful production of sequences of segmental categories; Clark 2009) and language-specific perceptual attunement (the reduction of an infant's ability to discriminant segmental categories not in his native language; Werker and Tees 1984). These milestones are both practical in their ability to be observed and grounded in linguistic theory, which has long considered segmental categories and their distinctive features foundational (Chomsky and Halle 1968; Jakobson, Fant, and Halle 1951). These facts, coupled with the aim of computational models of speech production to synthesize segmental categories Boersma et al. (1998); Maeda (1990), lead researchers to model the emergence of segmental categories.

Several authors have noted the need to incorporate the development of intonation into current models of language learning (Bohland, Bullock, and Guenther 2010; Heintz, Beckman, Fosler-Lussier, and Ménard 2009; Westermann and Miranda 2004), but no explicit implementation has been presented. This need is, in part, motivated by findings from language acquisition research. Snow and Balog (2002) investigated whether infants learn to produce single-word utterances before or after they demonstrate *intentional* use of intonation. Intentionality is crucial in this area of research since infants modulate their pitch and amplitude in tandem with fluctuations in their affect from near birth. Snow concluded that intentional manipulation of intonation occurs concurrently with the onset of the one-word stage. In contrast, Esteve-Gibert and Prieto (2013) found that infants make use of intonation to express pragmatic meaning prior to the one-word stage, at ages as young as seven months. Regardless of the precise timing, however, it is clear that intonation is actively being developed in the early stages of language learning.

A second line of research motivating the need to incorporate intonation into models of language learning investigates when adult-like use of intonation is achieved in infants. Astruc et al. (2013) found children aged 24 months produce intonational patterns consistent with adult speech. Similarly, Prieto, Estrella, Thorson, and Vanrell (2012) found intonational patterns developed rapidly from 11 months of age to 28 months of age, with infants producing adult-like intonation (based on pragmatic meaning) before consistently producing two-word utterances. These findings align with previous claims that adult-like intonational patterns arise around 18 months (e.g. Marcos 1987) and highlight the rapidity of intonation development in infants.

In combination, these findings provide a compelling case to incorporate intonation into models of language learning.

2.2 Computational Models of Language Learning

Of the myriad computational models of language learning, only three relevant models are summarized here. All three are implemented using artificial neural networks, a computational structure used in machine learning that consists of layers of nodes (neurons) and connections between nodes of each layer (the network). Over time, the connections spread activations from one layer to another, with the activation of a node in the deeper layer calculated by a function of the sum of feeding connections multiplied by their respective source node activations. The network learns by adjusting connections, changing how activation spreads through the network, with the goal of matching a

given set of inputs to corresponding outputs. In simplest terms, a neural network implements an unknown function to best satisfy known input-output correspondences. This makes neural networks ideal for approximating solutions to problems such as acoustics-to-segmental categories, where no known mapping exists.

Using a recurrent neural network, Kanda et al. (2008; 2009) developed a model capable of learning segmental categories from continuous acoustic data. To do this, the authors paired the acoustic data with a corresponding stream of articulator movements generated using an implementation of Maeda's (1990) model of speech production. Their system predicts the dynamics of the acoustic and articulatory streams and then learns by adjusting its connections to minimize prediction error. Once a certain error threshold is reached, points of complex dynamics are identified as segmental category boundaries and chunks between boundaries are labeled according to the bias that minimized prediction error in that chunk. Bias values were shown to correspond to segmental categories. This model provides an exemplar of an engineering approach to language learning and reinforces the prevalence of segmental categories in the literature.

A second model of segmental category learning is the Directions Into Velocities of Articulators (DIVA) model, pioneered by Guenther (1995). DIVA comprises a large array of interconnected neural networks, each of which emulates the function of an empirically established brain region important for speech production (see Guenther, Ghosh, and Tourville 2006). For example, DIVA contains networks that simulate the functions of motor, somatosensory and sensory cortices. Through adjusting the feed-forward and feedback connections between networks, DIVA learns to map motor movements in a simulated vocal tract (Maeda 1990) to speech sounds in an inventory (i.e. DIVA learns to produce segmental categories). Of note is DIVA's basis in neurophysiology and explicit computational implementation, which allows it to both make psychologically relevant, testable predictions and be modified to ensure it aligns with new findings (Lane, Denny, Guenther, Matthies, Menard, Perkell, Stockmann, Tiede, Vick, and Zandipour 2005; Tourville and Guenther 2011).

Like DIVA, Westermann's (2001) Sensorimotor Integration Model (SMIM) has the goal of learning the mapping between motor movements and acoustics. The SMIM comprises two neural networks, corresponding to motor and sensory cortices, and connections between the two networks. Learning in the SMIM happens through two simultaneous processes: (1) the networks self-organize themselves (Kohonen 1990), 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 (Hebb 2005), associating the segmental category clusters in the two networks and allowing activation to propagate across domains. In practical terms, the SMIM generates its own speech sound inventory and motor movement dictionary and learns how they correspond, which in turn allows for cross-domain effects such as motor parameters biasing categorization in the sensory domain. Because of its simplicity and utility, the SMIM has recently been adapted to model the emergence of canonical babbling (Warlaumont, Westermann, and Oller 2011) and the acquisition of vowels (Heintz et al. 2009).

While no model has been modified to incorporate intonation, both DIVA and the SMIM have been adapted in relevant ways. Bohland et al. (2010) developed an extension of DIVA, titled Gradient Order DIVA (GODIVA), that incorporates CVCV frame/content templates (MacNeilage 1998) as speech plans. These speech plans are then filled in as DIVA produces segmental categories. The authors note that speech plans could be modified to contain prosodic information through the addition of a phonology tier to the current CVCV tier. Next, to model the learning of phonation, Warlaumont, Westermann, Buder, and Oller (2013) introduced reinforcement into the self-organization

of a motor map (like that used in the SMIM). In their model, neurons in the motor map propagate activation to vocal tract parameters in Praat's simulated vocal tract (Boersma et al. 1998) to generate an utterance that is evaluated for the presence of phonation (a defined f0 value 250 ms after the start of the utterance). If phonation is present, the network self-organized, pulling motor parameters in the map towards values that generate phonation. The use of an evaluation function to gate self-organization (i.e. reinforce desirable utterances) will be termed **gated learning** in this project.

2.3 Model Selection

For the current purpose, only one model needs to be adapted to incorporate intonation. The Kanda et al. model was developed to identify segmental category-sized units. Intonation, in comparison, spans arbitrary temporal lengths. This mismatch makes the Kanda et al. model not ideal for the current purpose. Next, while the use of phonology tiers in GODIVA is possible, it is not clear how the phonology tier would be parameterized analogously to a CVCV tier. DIVA learns to produce segmental categories that slot in nicely to CVCV tiers; however, there are no such categories currently in GODIVA to slot into a phonology tier. Either pitch and amplitude would need to be incorporated into the entirety of DIVA or a hard-coded alternative would need to be provided. The latter option is clearly undesirable; the former option, while achievable, is a formidable task beyond the aim of the current work.

The SMIM, in contrast, is not constrained by fixed temporal frames (as inputs to the motor and sensory map can be of arbitrarily length) nor is it prohibitively computationally complex. What is more, the study by Warlaumont et al. (2013) that used gated learning to model the development of phonation is a natural lead-in to a study of intonation. Phonation, or vocal fold vibration, is measured acoustically as fundamental frequency (f0), and f0 modulations are perceived as pitch contours. There is, however, a crucial difference to be considered for the current project. In Warlaumont et al. (2013), utterances were evaluated for the presence or absence of phonation. Intonation, by comparison, cannot be evaluated as simply being present or not. It needs to be evaluated for goodness of fit to exemplars in a training space. Before continuing, a more detailed account of the SMIM is pertinent.

2.3.1 The Sensorimotor Integration Model

The SMIM was first proposed by Westermann (2001) to investigate cross-modal influences on speech perception. In particular, Westermann was investigating how motor neurons could shift the perception of sensory neurons. The model described here is based on Westermann and Miranda (2002).

The primary components of the SMIM, shown in Figure 1 are: (1) correlated motor and auditory inputs; (2) a motor map of neurons, with each neuron containing a location and a randomly initialized motor parameter; (3) a sensory map of neurons, with each neuron containing a location and a randomly initialized sensory parameter; and (4) connections between each neuron in the motor map to each neuron in the sensory map.

In the SMIM, a learning trial occurs as follows: (i) correlated motor and auditory inputs are input into the motor and sensory maps; (ii) neurons, in each map respectively, are activated according to a Gaussian activation function whereby the neurons with the motor/sensory parameter that are the shortest Euclidean distance from the inputs are the most highly activated, and activations decrease with distance; (iii) activated neurons shift their location and motor/sensory parameter to be

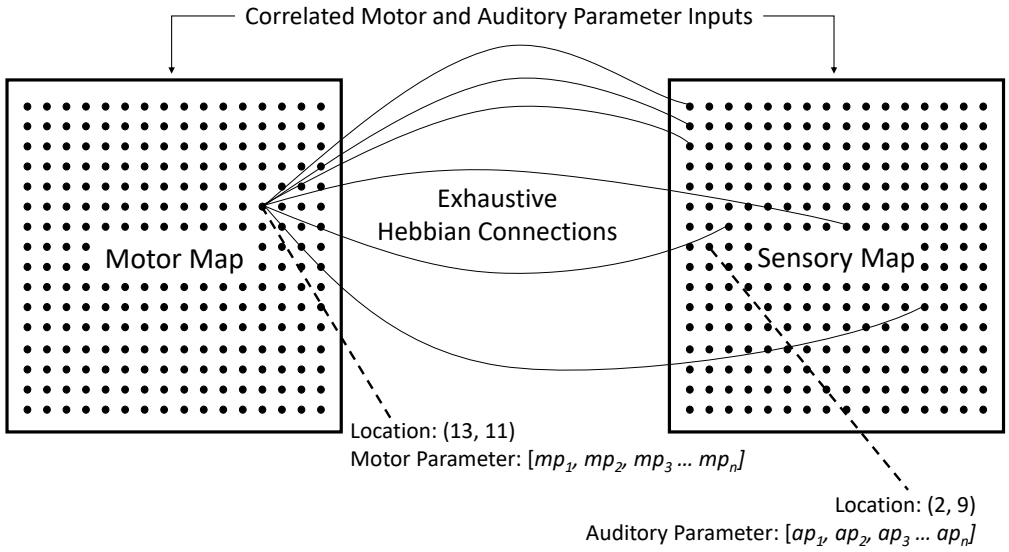

Figure 1: Schematization of the Sensorimotor Integration Model from Westermann and Miranda (2002)

closer to the most highly activated neuron; and (iv) connections between *simultaneously activated* neurons (across maps) are strengthened. Steps (ii-iii) have the practical consequence of simulating a receptive field that gradually generates higher whole-network activation (for frequent stimuli) over time (see Westermann and Miranda 2002: for discussion on the biological basis of receptive fields). Step (iv) is the implementation of Hebbian Learning, which allows activation to propagate across domains and consequently influence perception (i.e. shift which neurons have the highest activation).

In the original paper, the motor and auditory parameters were entirely artificial – two-dimensional arrays generated by Gaussian distributions. In subsequent work (e.g. Heintz et al. 2009; Warlaumont et al. 2011; Westermann and Miranda 2004), the motor parameters were articulator configurations for a speech synthesis model (either Praat's Mass-Spring model (Boersma et al. 1998) or the VLAM (Maeda 1990)) and the sensory parameters were formant values. Also, as of Heintz et al. (2009), both motor and sensory maps have been formalized as Self-Organizing Maps (SOMs) (Kohonen 1990). SOMs differ from the maps used above in that map locations do not shift, only motor/sensory parameter values. For an overview of SOMs, please refer to the Appendix A.

3 Method

3.1 Technical Challenges

In order to model the development of intonation using the SMIM, we first need a computational representation of intonation to serve as auditory parameters in the sensory map. Thereafter, we need an evaluation of intonation that will gate self-organization and, therefore, learning.

3.1.1 A Computational Representation of Intonation

In previous research, neurons in the SMIM's sensory map had formant values as auditory parameters (Heintz et al. 2009; Warlaumont et al. 2011; Westermann and Miranda 2002). Formants were chosen because they provide a static,[1] numeric representation of segmental categories in the auditory domain. These qualities are important as the sensory map self-organizes according to the Euclidean distance between parameters. To be used in the same framework, therefore, a computational representation of intonation must be *static* and *numeric*. To address how this is achieved, a general overview of intonation is necessary.

Liberman (1975) defines intonation as "the stress, tune, phrasing...and their interactions" of spoken speech. Acoustically, stress and tune are observed as modulations of pitch, amplitude and duration. Phrasing entails that intonation can be chunked into discrete events known as **intonational phrases**. An intonational phrase, then, comprises pitch, amplitude and duration modulations that occur between two **intonation boundaries** (or breaks), defined as "systematically significant pause[s]" (Liberman 1975: p. 286). This definition is consistent with the Tones and Breaks Indices (ToBI) annotation framework that is regularly used to annotate prosody in languages (Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert, and Hirschberg 1992).

If intonational phrases are taken as discrete events, then, a *static* representation of intonation is achieved. Computationally, utterances can be chunked into intonational phrases once a formalization of intonation breaks (IBs) is determined. Following Liberman's definition, a "significant pause" can be interpreted as a pause in speech of some minimal time needed to be an IB. As Liberman's second requirement, systematicity, is related to syntactic phrase breaks, it is beyond the scope of the current project and not formalized here. Based on the results of Brubaker (1972), the minimal time needed to mark an IB is 260 ms in English read-speech. This value is one standard deviation shorter than the average length of phrasal pause durations of participants reading aloud and should serve as a conservative estimate of pause length to qualify an intonation break. This is a very simple technique to approximate an intonation boundary and future work will be benefited by a further refined formalization of intonation breaks (perhaps incorporate things like pitch reset).

Once intonational phrases are isolated, they can represented *numerically* as pitch contours and amplitude envelopes. As duration is yoked to syllable and segmental categories (neither of which are considered in this project), it is not included in the computational representation of intonation presented here. Pitch contours are calculated by estimating f0 in the phrase. While there are myriad algorithms available to estimate f0, the current project uses a combination of autocorrelational analysis and sub-harmonic to harmonic frequency ratio (Sun 2002)—see Appendix B for details. Using two f0 algorithms facilitates a higher confidence in f0 accuracy. If no f0 component is present in a portion of the phrase (i.e. there is no voicing), f0 values are polynomially interpolated between the known preceding and following f0 values. The amplitude envelope is calculated by taking the root-mean-square of frames of the speech signal with a frame length of 25 ms and frame shift of 10 ms. Figure 2 shows the resulting pitch and amplitude modulations after processing an intonational phrase.

Due to the use of self-organizing maps, the current approach also requires that auditory parameters be normalized for length. This is to ensure consistency when calculating the Euclidean distance between intonational phrases. Normalizing intonational phrases for length is an oversimplification

[1] Static is used in the sense of a fixed representation of a discrete event, such as the formant values in the steady state of a vowel.

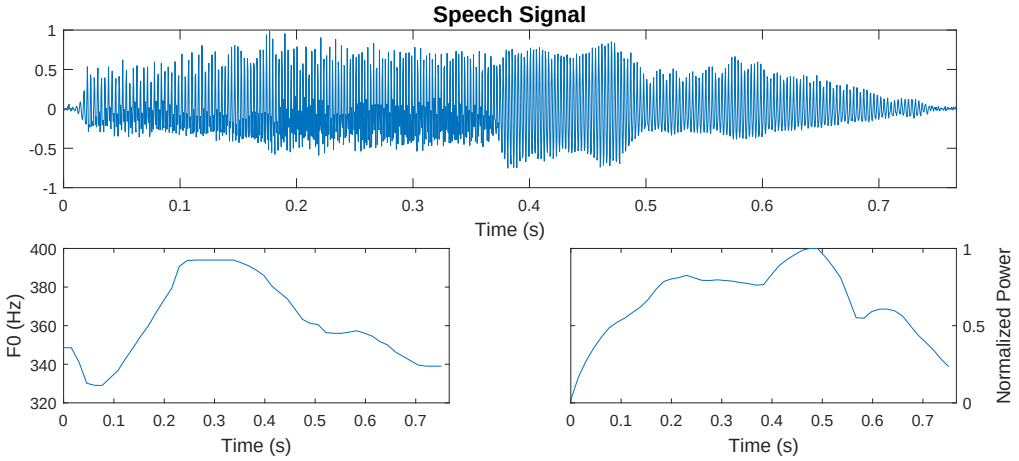

Figure 2: Top: the speech signal. Bottom left: the pitch contour of the signal. Bottom right: the amplitude envelope of the signal.

that should be addressed in follow-up studies. Pitch contours and amplitude envelopes are normalized to 50 data points each, resulting in each data point representing at most 40 ms of an utterance. This resolution provides approximately 5-6 data points per syllable and is more than sufficient to observe pitch and amplitude modulations.

3.1.2 Reinforcement in the Sensorimotor Integration Model

Following the general method of Warlaumont et al. (2013), reinforcement is implemented as an evaluation during learning that gates self-organization. This approach ensures only utterances that are positively evaluated contribute to the organization of the map, meaning that, over time, the map comes to represent primarily desirable utterances. In Warlaumont et al. (2013), the evaluation was the presence or absence of phonation in a spontaneously generated utterance. Intonational phrases, however, cannot be evaluated in the same present-or-not manner; they require an assessment of the goodness of fit to a training set.

To evaluate an intonational phrase, its computational representation is compared to a target **intonation space**. Intonation spaces, as defined here, are self-organizing maps trained on speaker-specific intonational phrases processed as in Section 3.1.1, with parameters normalized between 0 and 1. Intonation spaces are speaker-specific because the pitch range of speakers varies with the size of their vocal tract and such variation adversely affects self-organization. Once an intonation space is generated, any arbitrary, normalized intonational phrase can be compared to the intonation space by calculating the distance between it and the best-matching unit in the space. The best-matching unit in a self-organizing map is the neuron that has parameter values closest to the input values. For learning, if the distance between the intonational phrase being evaluated and the best-matching unit in the target intonation space is below a threshold, the intonational phrase is reinforced (the map self-organizes with respect to that phrase). For the current project, the threshold is calculated as the average distance between *all* training data and their corresponding best-matching unit in the target intonation space. This value approximates how well the training data is represented in the intonation space and provides a reasonable means to evaluated the quality of an intonational phrase. For the

current project, the threshold is 1.57 units of distance.

Figure 3 demonstrates how learning occurs in the model. To differentiate this model from the original SMIM, it will be referred to as the Sensorimotor Integration Model with Gated Learning (SMIMGL).

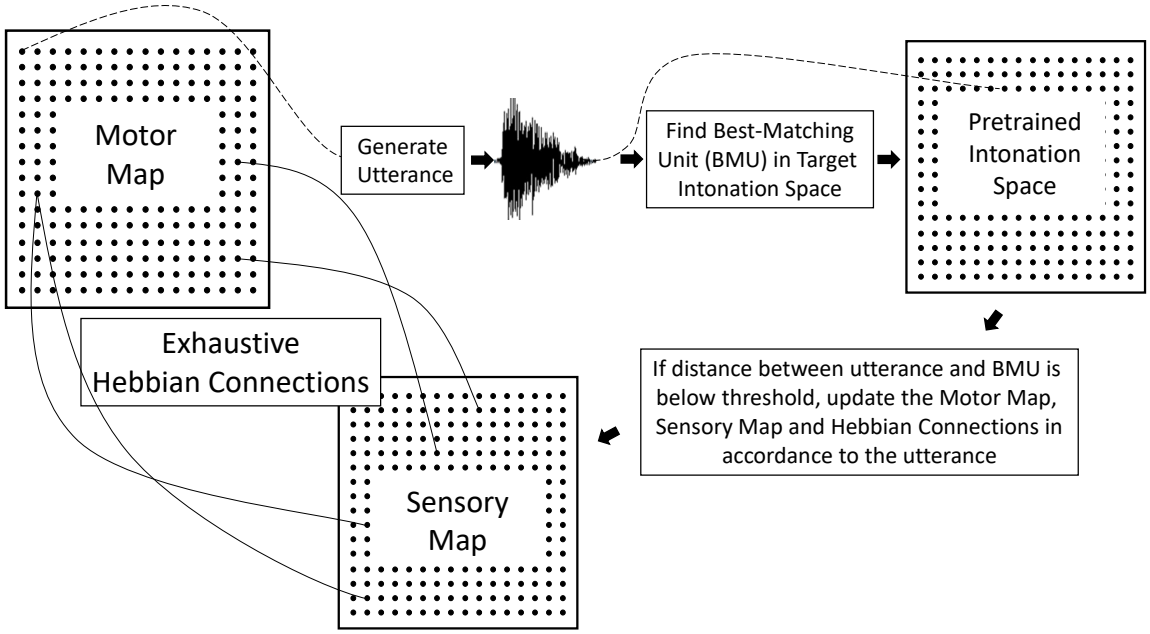

Figure 3: The learning mechanism for the SMIMGL

3.2 Learning Intonation with the SMIMGL

3.2.1 Initialization of SMIMGL

Before training, the SMIMGL's motor and sensory maps are initialized with a location and a random motor/sensory parameter. Each map consists of 100 neurons, structured as 10x10 matrices. One hundred neurons is fewer than would be expected for a human brain, but the amount is sufficient for the current experiment and is consistent with previous work (e.g. Warlaumont et al. 2011). Motor parameters, as in Warlaumont et al. (2013), correspond to muscle activations of the simulated vocal tract used in Praat's speech synthesizer (Boersma et al. 1998). In Praat, the vocal tract is modeled as a series of ducts that approximate airflow from the lungs to the lips; a diagram is shown in Figure 4. 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 the current work focuses on pitch and amplitude, motor parameters set only the laryngeal and lung muscles in the simulated vocal tract. Further, as intonation requires dynamic movement of articulators through time, each muscle is specified for its configuration at 7 time points. This is to ensure at most 5 inflection points in an intonational phrase, a reasonable expectation for a complex intonational phrase (Liberman

1975). For the sake of continuity, the first motor parameter for each muscle is randomized and the subsequent ones vary in place by a maximum of 10%.

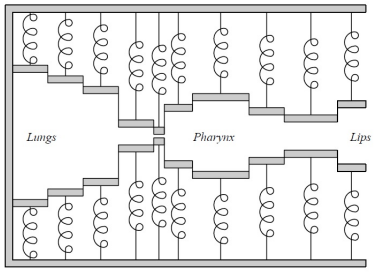

Figure 4: The simulated vocal tract of ducts, implemented as masses and springs, in Praat.

The sensory parameters are randomly initialized with a length of 100, the length of the combined pitch (50 data points) and amplitude (50 data points) modulations processed in Section 3.1.1. Finally, Hebbian connections are initialized to zero.

3.2.2 Generating the Target Intonation Space

As this project is an initial attempt to model the development of intonation for a language learner, the target intonation space is generated from Infant-Directed Speech (IDS). This follows from the assumption that a learner is motivated, at least in part, to produce utterances similar to his caregiver. As the target intonation space remains static throughout learning, a secondary assumption is that the learner has an internal representation of his caregiver's intonation space before producing utterances himself. In other words, the learner has *two* intonation spaces, one representing his caregiver's intonation space and one representing his own. The former is taken to be innate and unchanging, the latter changes through learning.

The data for the target intonation space consists of thirteen hours of caregiver-infant interactions. A total of 7623 IDS utterances/intonational phrases were identified and processed into their respective computational representations of intonation as described in Section 3.1.1. The data were collected by Brent and Siskind (2001) and were retrieved from the CHILDES database MacWhinney (2000). Audio of the interactions was recorded on a portable digital audio tape (DAT) recorder, worn by the caregiver, using a lavaliere microphone. All IDS tokens were from the same caregiver speaking to her infant in English.

After processing (Section 3.1.1), all IDS utterances were used to train the SOM target intonation space (i.e. the learners representation of the caregiver's intonation space).

3.2.3 Learning

After initialization (Section 3.2.1), the SMIMGL begins a learning block. In a learning block, each motor parameter (muscle activations) is used to generate an utterance that is then processed into its computational representation of intonation (i.e. the corresponding sensory parameters). The distance between each sensory parameter and the best-matching unit in the target sensory map is then calculated. If the distance is less than the threshold (1.57 units), both the motor parameter and

sensory parameter are added to a list of data points to be used in self-organization at the end of that learning block. Once all motor parameters have been processed, the list of motor/sensory parameter pairs is used to self-organize each map respectively. Simultaneously, Hebbian connections are strengthened for corresponding pairs following Heintz et al. (2009)—see Appendix C for details.

4 Results

As the SMIMGL is randomly initialized, the end state of each learning simulation varies slightly; however, the general performance of each simulation is similar. The results reported here are averages over 10 learning simulations. Three metrics to evaluate learning are presented: (1) the topographic correlation between the SMIMGL's sensory map and target intonation space; (2) the number of reinforced utterances from the SMIMGL's motor map per learning block; and (3) histograms of distances between all training utterances and the best-matching unit (BMU) of the SMIMGL's sensory map before and after learning. Following this, an exemplar that demonstrates the change of a single sensory neuron throughout learning is presented.

Topographic correlation measures the similarity between two SOMs. While it is unique to this project, it is related to previous work by Kirt, Vainik, and Võhandu (2007) in which visual inspection of neighborhoods and correlation analysis were used for the same purpose. The measurement is termed topographic as it is concerned with the *distance* between neighboring neuron parameters in a SOM and not the parameters of the neurons themselves. As a SOM learns to represent a space, clusters of similarly parameterized neurons form and separate from other clusters (where separation is in terms of their parameter values, not their map location); the pattern of distances within and between clusters represents a SOM's topography and allows it to be compared to other SOMs. This comparison, however, must be done for one SOM to *every rotation* of the other as SOMs variably expand depending on initial parameter values. Algorithmically, the topographic correlation between two SOMs (A and B) is calculated by: (1) measuring the Euclidean distance between neighboring neurons' parameters to generate a matrix of distances for each SOM; (2) calculating the correlation between the matrix of distances of SOM A and the matrix of distances of SOM B for every rotation around its middle; (3) selecting the highest correlation from the rotations.

Figure 5 is a plot of the topographic correlation of the SMIMGL's sensory map to the target intonation space throughout learning averaged across all learning simulations. When the SMIMGL is randomly initialized, its topographic correlation is ≈ 0.3. This raises to an average of ≈ 0.7 after 15 learning blocks. The plateauing seen after 15 learning blocks indicates stabilization of the SMIMGL's sensory map. The significance of ≈ 0.7 is unclear currently, but it is expected that the topographic correlation would not reach 1 as the learning model only generates its own utterances and therefore will never be able to replicate a target intonation space perfectly. What is more, a SOM's final state is dependent on its random initialization as best-matching units (and their neighbors) will be different. For reference, two randomly initialized SOMs trained on the same data (the training data in this work) achieved a correlation of ≈ 0.8

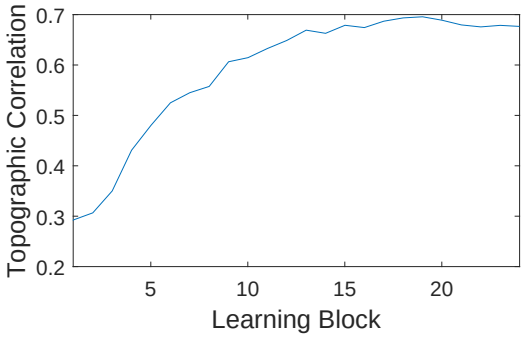

Figure 5: The topographic correlation between the target intonation space and the model's sensory map over time

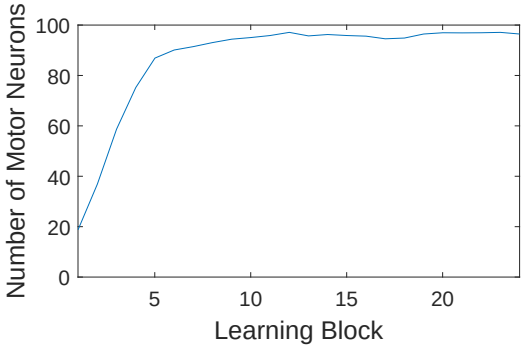

Figure 6: The number of reinforced motor parameters per learning block

Counting the total number of motor neurons with parameters that produce reinforced utterances is a second metric of the quality of learning. An utterance that is reinforced is one that has a distance less than the 1.57 unit threshold between its sensory parameterization and the best-matching unit in the target intonation space. Figure 6 shows that the number of motor parameters that produce reinforced utterances is initially around 20. Over time, as the SMIMGL self-organizes, the number of motor parameters that produce reinforced utterances nears 100, encompassing all motor neurons in the motor map. This is due to the fact that self-organizing makes neighbors more similar to each other, meaning one motor neuron that produces reinforced utterances pulls its neighbors, making them likely to do the same.

A final measure of the SMIMGL's performance is shown as the average distance between all 7623 training utterances and their respective best-matching units in the model's sensory map before and after learning. As the SMIMGL learns to produce desirable utterances, the distances between training data and their respective best-matching units should decrease. Figure 7 shows the histograms of distances before and after learning. The mean distance before learning is ≈ 3.1 units of distance; the mean after learning is ≈ 1.6 units of distance. This indicates that the SMIMGL's sensory map better represents the training data after learning. A Student's T-Test performed on the distribution confirms a significant difference with $p < 1e-10$.

Finally, as a concrete example of learning, Figure 8 shows the parameterization of the SMIMGL's best-matching unit to the training utterance in Figure 2 before and after learning. The random initial-

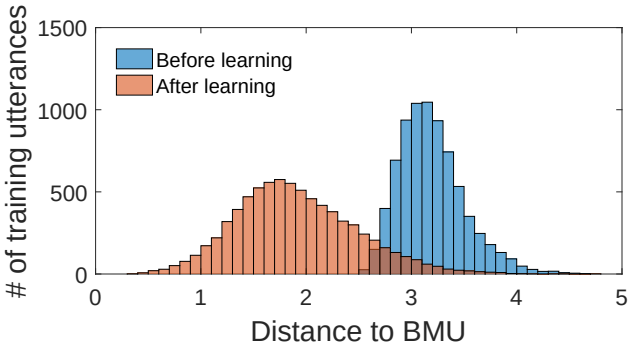

Figure 7: Histogram of distances between training data and each respective best-matching unit in the SMIMGL's sensory map before and after learning.

ization of the SMIMGL is seen in the jagged pattern before learning; the smooth sensory parameter seen after learning is a direct result of desirable utterances being reinforced.

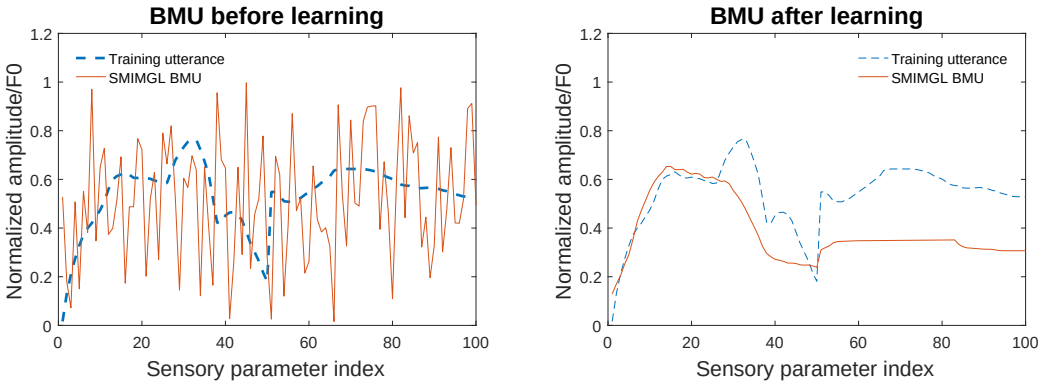

Figure 8: The best-matching unit from the SMIMGL to a training utterance before and after learning. The training utterance has been processed into its computational representation of intonation (3.1.1)—the first 50 values are the normalized amplitude, the latter 50 values are normalized f0.

Figure 8 demonstrates a bias in learning towards the first 50 values in the training data (i.e. the first 50 values are better matched than the latter 50). These values are the normalized amplitude component of the computational representation of intonation used in this project. The bias likely results from the fact that amplitude envelopes across intonational phrases are more similar than f0 contours (the latter 50 values are normalized f0), and thus generated utterances are more easily matched for amplitude than f0.

5 Discussion

All the results confirm that the SMIMGL is able to learn to produce adult-like intonational phrases in accordance with the assumptions and simplifications laid out here. The raw count of motor neurons that produce desirable utterances reaches ceiling after approximately 13 learning blocks, showing that the motor map has learned to produce adult-like intonational phrases. Next, the topographic

correlation of the SMIMGL's sensory map becomes more similar to the target intonation space over time, showing that the learner's utterances better match utterances heard from its caregiver. Finally, the distribution of distances to the BMU of training data shows that the learned sensory map better represents the caregiver's utterances; this is particularly impressive as the learner only modified its own sensory map from self-generated utterances and had never 'heard' the training data.

Moving towards theory, the above results highlight a crucial aspect of the current modeling approach, namely, that the learner is evaluated relative to a secondary intonation space (or sensory map). In fact, all three metrics depend on comparing the learners self-generated utterances to a previously learned intonation space of its caregiver. While this may be consistent with some notion of learning by imitation, it is unclear that a human infant would actually learn in the same way as the SMIMGL. That is, does an infant generate a perceptual map of his own utterances and then try to match those to his caretaker's utterances, or is the perceptual mapping done simultaneously on the same map? If the latter, the current projects use of evaluation should be reworked. Regardless, the map of caretaker utterances is undoubtedly dynamic and is not held constant like was done in this implementation.

Despite its successes, the current project does leave much to be desired. For one, its treatment of intonation as solely a pitch contour and amplitude envelope is a drastic oversimplification. Much like the criticism of current models of language learning that was shared at the beginning of this paper, modeling the development of intonation without any use of segmental categories is incomplete. Intonation and segmental categories develop simultaneously during language learning. Further, intonation interacts with lexical stress and with the inherent f0, amplitude and duration qualities of vowels. Future work would be improved by dealing with some of these interactions.

Another simplification used herein is the normalization of the length of intonational phrases. Intonational phrases span arbitrary temporal lengths and it is not clear that simply collapsing all phrases to the same length captures general patterns appropriately. For example, the interactions of intonation with lexical stress could distort intonational phrases in ways that are artificially and incorrectly avoided in the current length normalization.

A third simplification is the absence of an exploratory learning stage. Even profoundly deaf infants phonate and marginally babble up until ≈ 6 months of age (Oller and Eilers 1988). This is evidence that infants do not require a target intonation space in order to begin developing their ability to manipulate intonation. Future work would benefit from investigating the development of intonation in at least two overlapping stages, one in which the learner 'practices' on its own and another in which the learner imitates its caregiver.

A final, yet crucial, issue that needs to be addressed in future work is the use of Praat's speech synthesis. While an impressive mathematical model, Praat's synthesis is based on models of adult vocal tracts. Infant vocal tracts change substantially from birth through 6 months (including the lowering of the larynx between 2 and 3 months), and then change further from 6 months through adolescents (Crelin 1987). Nonetheless, this mismatch is possibly less severe for the current project than it would have been for a project on the production of segments. That is, any segmental category articulated using Praat would not match the configuration of an infant's vocal tract; this comparisons is less clear when only laryngeal muscles are considered. Regardless, for future work to better capture how infants learn to produce adult-like intonation, new synthesis models need to be developed.

In addition to the results and limitations discussed, the current project also required the development of several tools that may find application elsewhere. The computational representation of intonation may provide an interface for computational modelers and speech scientist. The cre-

ation of speaker-specific intonation space could find use in studies that match intonational phrases with pragmatic meanings. Also, intonation spaces could be used to compare intonational tendencies across speakers or even perhaps within the same speaker in different social situations or when speaking different languages. Finally, the modification of the SMIM to the SMIMGL is an advancement that could be used to model the learning of phenomenon other than intonation.

Finally, if the same research program is to be pursued in the future, it may also be beneficial to do a detailed analysis of the learning of the SMIMGL over time. There may be testable predictions concerning the order in which intonational phrases are acquired. Similarly, the time-course of acquisition is of interest to many language acquisition researchers.

6 Conclusion

The current work presents a first attempt towards incorporating intonation into computational models of language learning. The work includes the development of a computational representation of intonation, a computational comparable intonation space using self-organizing maps and an evaluation metric to compare such maps.

References

Astruc, Lluïsa, Elinor Payne, Brechtje Post, Maria del Mar Vanrell, and Pilar Prieto. 2013. Tonal targets in early child English, Spanish, and Catalan. *Language and speech* 56:229–253.

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

Bohland, Jason W, Daniel Bullock, and Frank H Guenther. 2010. Neural representations and mechanisms for the performance of simple speech sequences. *Journal of cognitive neuroscience* 22:1504–1529.

Brent, Michael R, and Jeffrey Mark Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition* 81:B33–B44.

Brubaker, Robert S. 1972. Rate and pause characteristics of oral reading. *Journal of Psycholinguistic Research* 1:141–147.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English.*. New York: Harper & Rowe.

Clark, Eve V. 2009. *First language acquisition*. Cambridge, UK: Cambridge University Press.

Crelin, Edmund S. 1987. *The human vocal tract: Anatomy, function, development, and evolution*. Vantage Press.

Esteve-Gibert, Nuria, and PILAR Prieto. 2013. Prosody signals the emergence of intentional communication in the first year of life: Evidence from Catalan-babbling infants. *Journal of child language* 40:919–44.

Guenther, Frank H. 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological review* 102:594.

Guenther, Frank H, Satrajit S Ghosh, and Jason A Tourville. 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language* 96:280–301.

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

Heintz, Ilana, Mary E Beckman, Eric Fosler-Lussier, and Lucie Ménard. 2009. Evaluating parameters for mapping adult vowels to imitative babbling. In *INTERSPEECH*, volume 9, 688–691.

Jakobson, Roman, Gunnar Fant, and Morris Halle. 1951. *Preliminaries to speech analysis. the distinctive features and their correlates*. Cambridge, MA: MIT Press.

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

Kanda, Hisashi, Tetsuya Ogata, Toru Takahashi, Kazunori Komatani, and Hiroshi G Okuno. 2009. Continuous vocal imitation with self-organized vowel spaces in recurrent neural network. In *IEEE International Conference on Robotics and Automation (ICRA 2009)*, 4438–4443. IEEE.

Kirt, Toomas, Ene Vainik, and Leo Võhandu. 2007. A method for comparing self-organizing maps: Case studies of banking and linguistic data. In *Local Proceedings of ADBIS*, ed. B. Novikov Y. Ioannidis and B. Rachev, 107–115.

Kohonen, Teuvo. 1990. The self-organizing map. *Proceedings of the IEEE* 78:1464–1480.

Lane, Harlan, Margaret Denny, Frank H Guenther, Melanie L Matthies, Lucie Menard, Joseph S Perkell, Ellen Stockmann, Mark Tiede, Jennell Vick, and Majid Zandipour. 2005. Effects of bite blocks and hearing status on vowel production. *The Journal of the Acoustical Society of America* 118:1636–1646.

Liberman, Mark Yoffe. 1975. The intonational system of English. Doctoral Dissertation, Massachusetts Institute of Technology.

MacNeilage, Peter F. 1998. The frame/content theory of evolution of speech production. *Behavioral and brain sciences* 21:499–511.

MacWhinney, Brian. 2000. *The childes project: The database*, volume 2. Psychology Press.

Maeda, Shinji. 1990. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling*, ed. William J. Hardcastle and Alain Marchal, 131–149. Springer.

Marcos, Haydée. 1987. Communicative functions of pitch range and pitch direction in infants. *Journal of Child Language* 14:255–268.

Oller, D Kimbrough, and Rebecca E Eilers. 1988. The role of audition in infant babbling. *Child development* 59:441–449.

Prieto, Pilar, Ana Estrella, Jill Thorson, and Maria del Mar Vanrell. 2012. Is prosodic development correlated with grammatical and lexical development? Evidence from emerging intonation in Catalan and Spanish. *Journal of Child Language* 39:221–257.

Silverman, Kim EA, Mary E Beckman, John F Pitrelli, Mari Ostendorf, Colin W Wightman, Patti

Price, Janet B Pierrehumbert, and Julia Hirschberg. 1992. Tobi: A standard for labeling English prosody. In *ICSLP*, volume 2, 867–870.

Snow, David, and Heather L Balog. 2002. Do children produce the melody before the words? A review of developmental intonation research. *Lingua* 112:1025–1058.

Sun, Xuejing. 2002. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, I–333–I–336. IEEE.

Tourville, Jason A, and Frank H Guenther. 2011. The diva model: A neural theory of speech acquisition and production. *Language and Cognitive Processes* 26:952–981.

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

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

Werker, Janet F, and Richard C Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development* 7:49–63.

Westermann, Gert. 2001. A model of perceptual change by domain integration. In *Proceedings of the 23rd annual conference of the cognitive science society*, 1100–1105.

Westermann, Gert, and Eduardo Reck Miranda. 2002. Modelling the development of mirror neurons for auditory-motor integration. *Journal of new music research* 31:367–375.

Westermann, Gert, and Eduardo Reck Miranda. 2004. A new model of sensorimotor coupling in the development of speech. *Brain and language* 89:393–400.

7 Appendix

A Self-Organizing Maps

SOMs learn the topological representation of a data space in that the distance between high-dimensional data points is preserved in a two-dimensional space. SOMs consist of neurons that are initialized with a location C and a randomly generated parameter vector W (of the same dimensionality as the data points). When presented with a new data point V, the **Best Matching Unit** in the map is identified, defined as the neuron with W most similar to V. Once identified, a neighbourhood of the BMU is selected based on its distance from other neurons in the map. A learning trial concludes by shifting the parameter vectors (Ws) of the BMU and neurons in its neighbourhood to be more similar to the data point (V). The shift decreases according to a Gaussian Decay Function based on the distance of a neuron from the BMU. Equation 1 denotes what has just been read – t is time and η is a pre-set learning rate.

$$W(t+1) = W(t) + \eta \cdot e^{\frac{dist(C_{BMU}, C_{NEURON})^2}{2\sigma^2}} \cdot (V - W_{NEURON}) \quad (1)$$

B f0 Estimation

Two algorithnms were used to estimate f0 for the current project: (1) autocorrelation; and (2) sub-harmonic to harmonic ratio (Sun 2002). The former was coded by myself using MATLAB's built in Butterworth filter, cross-correlation and Savitzky-Golay FIR smoothing filter. The latter, which is publicly available, estimates f0 through calculating the ratio of energy between whole number factors of a frequency and the frequencies half-way between the whole number factors – the frequency with the largest ratio is the f0 estimate.

C Hebbian Learning

Hebbian Learning in the current project is implemented following the work of Heintz et al. (2009). Corresponding motor parameter/sensor-parameter pairs in learning increase the Hebbian connection between their respective best-matching units. This increase is proportional to the errors between the given motor parameter and its best-matching unit, and, the error between the given sensory parameter and its best-matching unit. Explicitly: take H to be a set of Hebbian connections indexed with (i, j) for motor-neuron i in the motor map and sensory neuron j in the sensory map; take motor parameter a and sensory parameter b to be corresponding occurrences in learning; and, take $BMU(a)$ and $BMU(b)$ to be the best-matching units respectively to a and b, then:

$$H(loc(BMU(a)), loc(BMU(b)))(t+1) = H(loc(BMU(a)), loc(BMU(b)))(t) + \frac{1}{(\sqrt{(BMU(a)-a)^2} + \sqrt{(BMU(b)-b)^2})} \quad (2)$$