7-24-2012

Early Human Vocalization Development: A Collection of Studies Utilizing Automated Analysis of Naturalistic Recordings and Neural Network Modeling

Anne Sanda Warlaumont

EARLY HUMAN VOCALIZATION DEVELOPMENT: A COLLECTION OF STUDIES UTILIZING AUTOMATED ANALYSIS OF NATURALISTIC RECORDINGS AND NEURAL NETWORK MODELING

by

Anne S. Warlaumont

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Communication Sciences and Disorders

The University of Memphis

August 2012

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor, Kim Oller. I have benefitted from his guidance in many matters ranging from infant vocalization to speaking clearly (both of which I promise to continue to pursue). I'm also indebted to him for creating many research opportunities for me and for the freedom and encouragement he provided for my computational pursuits. Thanks also to his grant funding, including that from the NIH, which has supported the research in his lab including my own.

I also owe a big thanks to Eugene Buder for all he has taught me about speech signals and dynamical systems, for involving me in some very fun projects, and for many fun and helpful discussions, and for being a wonderful secondary advisor.

There are several other faculty I've had the privilege of collaborating with during my time at the University of Memphis. These include Robert Kozma, Gert Westermann, Linda Jarmulowicz, and Andrew Olney. All have offered very useful constructive criticism as well as encouragement. I would also like to thank them, as well as Dr. Cannito, for their time and thoughtfulness serving on my comprehensive exam and dissertation committees.

I am also thankful to the rest of the faculty, staff, and the other PhD students, past and present, in the School of Communication Sciences and Disorders, who have been so nice to work with. Likewise, I would like to thank the Institute for Intelligent Systems faculty, students, and staff, a wonderfully interdisciplinary cognitive science group that has been a source of great fun and enrichment during my graduate work. I am going to miss both communities very much and I hope that the many friendships I've made will continue into the coming years.

I owe very much to the Department of Energy Computational Science Fellowship for its generous investment in my career, funding my studies for the past four years and providing opportunities to learn from other computational scientists in a wide range of disciplines. I've very much enjoyed interacting with the folks at Krell and the other friendly fellows. As part of the fellowship, I spent a summer at Argonne National Lab,

where I was fortunate to have Mark Hereld as a very knowledgeable and friendly supervisor who pushed me to make the most of the practicum. I learned a lot as well from the other students and researchers at ANL and from the neuroscientists at the University of Chicago who were so kind and generous with their time.

Finally, thanks goes to Rick for his love, support, and good advice, to both our families for their love and encouragement, and to my mother and father for their tremendous support in all aspects of my educational journey.

ABSTRACT

Warlaumont, Anne Sanda. PhD. The University of Memphis. August 2012. Early human vocalization development: A collection of studies utilizing automated analysis of naturalistic recordings and neural network modeling. Major Professor: D. Kimbrough Oller.

Understanding early human vocalization development is a key part of understanding the origins of human communication. What are the characteristics of early human vocalizations and how do they change over time? What mechanisms underlie these changes? This dissertation is a collection of three papers that take a computational approach to addressing these questions, using neural network simulation and automated analysis of naturalistic data.

The first paper uses a self-organizing neural network to automatically derive holistic acoustic features characteristic of prelinguistic vocalizations. A supervised neural network is used to classify vocalizations into human-judged categories and to predict the age of the child vocalizing. The study represents a first step toward taking a data-driven approach to describing infant vocalizations. Its performance in classification represents progress toward developing automated analysis tools for coding infant vocalization types.

The second paper is a computational model of early vocal motor learning. It adapts a popular type of neural network, the self-organizing map, in order to control a vocal tract simulator and in order to have learning be dependent on whether the model's actions are reinforced. The model learns both to control production of sound at the larynx (phonation), an early-developing skill that is a prerequisite for speech, and to produce vowels that gravitate toward the vowels in a target language (either English or Korean) for which it is reinforced. The model provides a computationally-specified explanation for how neuromotor representations might be acquired in infancy through the combination of exploration, reinforcement, and self-organized learning.

The third paper utilizes automated analysis to uncover patterns of vocal interaction between child and caregiver that unfold over the course of day-long, totally naturalistic recordings. The participants include 16- to 48-month-old children with and without autism. Results are consistent with the idea that there is a social feedback loop wherein children produce speech-related vocalizations, these are preferentially responded to by adults, and this contingency of adult response shapes future child vocalizations. Differences in components of this feedback loop are observed in autism, as well as with different maternal education levels.

TABLE OF CONTENTS

LIST OF TABLES

Tables **Pages**

LIST OF FIGURES

Figures **Pages**

Chapter 1

General introduction

Human speech behavior is unique among primates, and unique features of human vocalization, such as contextual flexibility in the usage of vocalizations, are observed at least as early as 2 months of age [1]. Understanding early speech-related vocalization development is thus key to understanding the origins of human communication [2]. Both describing the course of vocal development and identifying the mechanisms that shape that course are of interest. In other words, what are the characteristics of early speech-related vocalizations, and how do they change with time? And what mechanisms underlie these changes?

Answering these questions requires overcoming several methodological challenges. First, one has to decide on a methodology for describing prelinguistic vocalizations. There are several options, including acoustic analysis, phonetic transcription, coding in terms of more general "protophone" categories, and, as introduced in the first paper, coding in terms of automatic, data-derived categories. The challenge is to describe infant vocalizations using constructs that map as closely as possible onto the phonetic categories or dimensions that are actually governing the child's behavior.

Second, there is a need to study infant vocal behavior naturalistically. To understand the full range of vocal behaviors infants produce, and the effects of the social and physical contexts in which they occur, one should ideally observe infants at all times of the day, in all types of activities and environments, including feeding, playing alone, interacting with different individuals and different numbers of individuals, etc. Fortunately, technology now exists that makes entirely naturalistic audio recording over long periods of time quite feasible. The challenge is that this approach generates very large quantities of data, increasing the need to make the description of infant vocalizations and the contexts in which they occur as time-efficient as possible. One solution is to develop procedures for generating these descriptions automatically.

The third challenge is the significant ethical and practical limitations on experimentation with infants and imaging of their nervous systems and vocal tracts. Additionally, human vocal development is a process for which animal modeling is likely to have limited applicability, given that speech is unique to humans and doesn't have direct analogues in other animals. Computational simulation is becoming recognized as the "third pillar of science" [3], and it has particular appeal with regard to the science of infant prelinguistic vocal development, given these limitations on experimentation, imaging, and animal modeling. Biologically-inspired models of the human infant nervous system and vocal tract can be expected to contribute toward our understanding of the processes underlying the speech-related vocal development.

This dissertation is a collection of three journal papers (one already published, one under review, and one in preparation). The first paper explores the automated derivation, using a self-organizing map neural network, of holistic spectrographic features characterizing human infant vocalizations. It uses these derived features to automatically classify vocalizations according to protophone category, age, and infant identity. Thus, it is an attempt to use artificial neural networks to address the first and second challenges listed above.

The second paper is a computational model of early vocal motor learning. It uses the same type of neural network as in the first paper, the self-organizing map, but adapts it so that it acts as a motor controller (rather than an sound perceiver) and so that learning is gated by reinforcement. The model produces sounds at random and is reinforced when it produces sounds that are phonated (as opposed to silent or just breath) and when it produces vowels that are similar to a set of targets (e.g. English or Korean vowels). All sounds are created using a realistic vocal-tract-simulating sound synthesizer. Thus, the second paper addresses the third issue above, developing a computational model to explain certain aspects infants' prespeech phonatory and vowel development.

The third paper describes a fully automated analysis of day-long (12-hour) audio recordings of 16- to 48-month-old children with and without autism. We look at the quantity of speech-related vocalization produced by the children, at the general interaction dynamics between the children and the adults in their environment, the contingency of adult response on the quantity of child speech-related vocalization produced, and the contingency of child speech-related vocalization on previous adult responses. The analyses are motivated by our hypothesis that there is a social feedback loop supporting speech development, and we explore several ways in which the proposed feedback loop may be affected in autism. Thus, the third paper exemplifies an automated analysis approach to studying infant vocalizations in large, naturalistic data.

In addition to sharing a focus on computational methodologies, the three papers explore the ideas that infant vocalizations change in meaningful ways across the first year of life and that reinforcement, be it social or intrinsic, can play a role in shaping this development.

Chapter 2

Data-driven automated acoustic analysis of human infant vocalizations using neural network tools

I. Preface

This chapter was published as an article in 2010 in the *Journal of the Acoustical Society of America*. Its authors are Anne S. Warlaumont, D. Kimbrough Oller, Eugene H. Buder, Rick Dale, and Robert Kozma.

II. Introduction

Over the course of their first year of life, human infants vocalizations become progressively more speech-like in their phonation, articulation, timing, and in other respects [4, 5, 6]. The exploration of the sound-making capability by infants, the formation of new contrastive categories of sound, and the systematic use of these categories in vocal play and in flexible expression of emotional states appear to form a critical foundation for speech [7, 8]. In fact, failure to reach milestones of vocal development is associated with hearing impairment and other medical conditions as well as with slower vocabulary development [9, 10, 11, 12]. However, in the first months of life, infant sounds bear little resemblance to speech and thus their description presents unique methodological challenges.

Acoustic analysis is central to the study of prelinguistic vocalization development. Since recordings of infant vocalizations constitute high-dimensional time series data, their acoustic analysis presents a challenge of data reduction. It is necessary to represent the signal in terms of the most significant features, the ones around which development is fundamentally organized. Some of the acoustic measures that have been applied to infant vocalizations include: duration, f_0 means, peaks, standard deviations, and contours, formant frequencies, spectral concentration/standard deviation, and degree of tremor (as measured by within-syllable f_0 and amplitude modulation) [13, 14, 15]. Such measures are inspired by a priori assumptions rooted in acoustic phonetic theory. They are usually treated as independent, with relatively limited attention paid to possible interactions. This is likely an oversimplification, since vocal categories are based on interactive, multivariate acoustic features in mature speech [16], and it seems likely that early infant sounds are also composed of acoustic features in interactive ways. Further, the traditional approach assumes that the selected a priori acoustic measures represent the fundamental dimensions of vocal development, exploration, and manipulation. There is a need for methods that address the multivariate and high-dimensional acoustic properties of infant vocalizations directly.

In addition, the need for automated analysis of infant vocal development is rapidly growing. Samples involving millions of utterances from thousands of hours of all-day audio recordings are being collected and analyzed [17]. It is important to develop a set of automated acoustic analysis tools appropriate for infant vocalization data which would be impractical to analyze manually.

Here a method is presented for reducing high-dimensional samples of infant vocalizations to a smaller set of holistic acoustic features derived directly and automatically based on the patterns exhibited by a set of infant vocalizations. The approach makes relatively few a priori assumptions and is intended to complement research using more traditional acoustic measures derived from speech science principles. It utilizes a computational algorithm that would be suitable as an automated analysis method for application to large sets of infant utterances from naturalistic recordings.

Infant vocalizations are first analyzed using a type of unsupervised artificial neural network, the self-organizing map (SOM). The SOM derives a set of sixteen holistic spectrographic features based on clusters detected in an input corpus consisting of spectrograms of infant utterances. Then a type of supervised neural network, the single-layer perceptron, is used to classify utterances on the basis of the SOMs derived acoustic features. The classification types are: (1) prelinguistic vocal categories (*squeals*, *vocants*, and *growls*), (2) when in the first year of life the utterances were produced, and (3) the identity of the individual who produced a given utterance.

The relationship between the SOMs features and vocal categorizations, age, and individual differences is explored by looking at the patterns of activations across the SOM features and through some simple acoustic measurements (spectral mean, spectral standard deviation, and duration) made on the SOM features and the perceptrons weightings of those features. The perceptrons performance is also evaluated quantitatively and is compared to performance by a prominent neural network classifier, the multilayer perceptron (MLP). Note that the SOM and perceptron neural networks can be used either (1) purely for statistical analysis purposes or (2) as models of human perception and classification. The present study falls into the first category of usage, with the second being a potential future direction.

The next sections provide background on prelinguistic vocal categories, developmental changes, and individual differences. This is followed by a brief review of previous work that has used SOMs or perceptrons to analyze vocalization data.

A. Three areas of investigation in infant pre-speech vocalization research

1. Prelinguistic phonological categories

The fact that vocalizations produced during the first year exhibit some of the characteristics of adult speech yet are still in many respects immature poses a challenge to phonological description. It is clear that phonetic transcription at the phonological segment level is not appropriate for early infant vocalizations [18]. As an alternative, some researchers have identified pre-linguistic vocal categories, termed “protophones” [2], that seem to appear relatively universally during development across the first months of life [4, 19].

Some protophone categories relate to the purposeful variation of phonatory characteristics, especially pitch and voice quality. One such category is *squeal*, which includes utterances that are high in pitch and often accompanied by pitch variation, loft (falsetto) quality, and/or harshness [20]. Another category is *growl*, which includes utterances with low pitch, harshness, creaky voice, and/or pulse register [21]. Perhaps the most frequently occurring protophone is the *vocant*, which refers to vowel-like utterances [22, 23]. *Vocants* have intermediate pitch and relatively normal phonation. Purposeful variation of pitch and vocal quality usually appears by at least 4 months of age and continues to be explored throughout the first year and beyond [4]. Although other protophone categories address maturation in the timing of syllable production (*marginal* and *canonical syllables*; [5]) and the capacity to produce multisyllabic utterances of various sorts (*reduplicated* and *variegated babbles*; [24]), the present study focuses only on the early-emerging phonatory protophones *squeal*, *growl*, and *vocant* as an illustration of how our method can be applied to the acoustic analysis of protophone categories.

Protophone categories have an inherent element of subjectivity, since they are seen as proto-phonological constructs that form the basis for interpretation of emotional states and intentions by caregivers [15, 25]. Their validity is supported by the fact that *squeals*, *growls* and *vocants* are often spontaneously reported by parents when asked to identify sounds their babies produce (*vocants* being called “vowels”; [26]). Laboratory research involving these categories primarily uses trained adult listeners perceptual judgments [19].

Little relevant acoustic data on the key categories has been published for the *squeal*, *vocant*, and *growl* protophones. However, a primary acoustic correlate has been proposed to be fundamental frequency (f_0) [20, 5, 4]. A goal of the present study is to explore the acoustic correlates of human listeners protophone judgments via inspection and visualization of neural network weights and activations. The present study also lays a foundation for the development of automatic protophone classification. This is important because protophone classification is otherwise a costly and time-consuming procedure, involving prior training of analysts and repeated careful listening to individual utterances.

2. Developmental changes across the first year

Because during most or all of the first year of life infants do not produce recognizable words, their prelinguistic vocalizations are the main means of assessing the development of speech- and language-related production capabilities. While ethologically oriented auditory studies of changes in vocalizations across the first year have been informative in determining stages of vocal development and the protophones that emerge with each stage [27, 7], developmental patterns have also been studied using acoustic phonetic methods. For example, Kent and Murray [13] tracked a number of acoustic measurements, including duration, mean f_0, f_0 intonation contours, first and second formant frequencies, and a variety of glottal and supraglottal quality characteristics such as fry, tremor, and the spectral concentration of noise, in a cross-sectional study of 3-, 6-, and 9-month-old infants vocalizations. Across age, they found changes in formant frequency values (see also [28], and [29]) as well as in amount of tremor.

Despite the important contributions of such research, it does not address the possibility that the changes in such acoustic measures across development are not independent of each other. For example, increases in duration and decreases in phonatory variability may emerge in coordination with each other, driven by common physiological and cognitive maturation that lead to increased control over the larynx. Unsupervised statistical analysis may help to address this concern, either (1) by reducing the large number of acoustic measures to a smaller number of component dimensions that are weighted on each of those acoustic parameters or (2) by deriving a limited number of new, holistic acoustic measures directly from relatively unprocessed recordings of infant vocalizations. The present study takes the second approach.

An aim of this work is to develop potential methods for automatic measurement of the acoustic maturity of infant utterances. This goal is motivated by fact that "language age" or "age-equivalence", is commonly used as an index of language development status in both research and clinical assessment of children older than one year (e.g., [30, 31]).

Automatic classification of vocalization maturity is already being pursued with some success using statistical algorithms incorporating automatic calculation of more traditional acoustic measures, such as duration, and automatic detection of phonetic features, such as bursts, glottal articulations, sonorant quality, and syllabicity [32]. The method presented here lays groundwork for the automatic measurement of the maturity of an utterance on the basis of holistic, data-driven features, which could prove a worthwhile addition to current methods for automatic detection of utterance maturity.

3. Individual differences

The ordering of phonological stages of vocal development appears to be robust across infants, even those from different language environments, with differing socioeconomic status, and in large measure with differences in hearing function [33]. However, reports of notable individual differences are also common in the literature on infant vocal development [4, 34, 35]. These individual differences appear to be associated with differences in later language styles and abilities. For example, Vihman and Greenlee [36] found that the degree of use of true consonants (consonants other than glottals and glides) in babble and words at one year of age predicted phonological skill at three years. It is important to be able to quantify individual differences in preverbal vocalizations within normally developing infants as this might be used to predict later differences in speech and language ability and usage. The study of individual differences in typical infants also sets the stage for the study of infant vocalizations across groups, e.g., across various language or dialect environments, genders, and populations with hearing, language learning, or cognitive impairments.

As with the study of age differences, the study of individual differences is likely to benefit from the introduction of data-driven acoustic measures that convert high-dimensional acoustic input to a smaller number of essential holistic features. In this study, the problem of characterizing and quantifying individual differences among infants is addressed through exploration of differences across infants in the presence of such holistic features. Automatic detection of infant identity provides groundwork for future detection of differences in the vocalization patterns across different infant populations of clinical significance.

B. Previous applications of neural networks to related problems

Neural networks are often used as tools for statistical pattern analysis and are particularly appropriate for high-dimensional data that are suspected of having nonlinear cluster or class boundaries [37]. The networks are typically trained through exposure to data exemplars. They can be used both in cases where the classes in a data set are known and used to provide explicit feedback to the network (supervised networks), or when they are unknown and discovered without explicit supervision (unsupervised networks).

The perceptron is perhaps the most commonly used supervised neural network. It consists of an input layer, an output layer, and zero or more hidden layers. Each layer except the output has a set of weights that describes the strength of the connections between its nodes and the nodes of the following layer. Activation from the input is propagated to the hidden layers (if there are any) and then to the output. The networks output activations are then compared to the known classifications for that input and the networks error is determined. Based on that error, the networks weights are adjusted, typically using the delta rule, or with backpropagation if there are any hidden layers [37].

A common unsupervised network is the self-organizing map (a.k.a. Kohonen network, SOM). SOMs are typically used for unsupervised cluster analysis and visualization of multidimensional data [38, 39, 40]. A SOM consists of an input layer and an output layer and a set of connection weights between them. The nodes of the output layer are arranged spatially, typically on a 2-D grid. When an input is presented, each of the output nodes is activated to varying extents depending on the input and its connection weights from the input layer. The output node with the highest activation is the winner. It, and to a lesser extent, its neighboring nodes, have their connection weights strengthened so that their receptive fields (i.e., their preferred inputs) more closely resemble the current input stimulus. The result after training is that the output nodes receptive fields reflect the patterns found in the input and that the receptive fields are topographically organized, i.e., nearby nodes have similar patterns of weights from the input layer.

There appear to be few, if any, previous applications of neural networks to recordings of infant pre-speech non-cry vocalizations. However, neural networks have been used to analyze recordings of vocalizations produced by songbirds, disordered and normal adult human voice, and infant crying. Many of these applications were developed in response to a need to represent high-dimensional, complex acoustic signals in a data-driven way. For example, Janata [41] used a SOM to cluster spectrographic representations of segments of juvenile zebra finch song into 200 topographically-arranged holistic spectrogram prototypes. The visualizations of the loadings of features across 30 consecutive days represented a map of the developmental pathways by which adult songs emerged. In addition to permitting data-driven detection of song features, Janata pointed out that the SOM provides automated acoustic analysis and classification of a very large set of vocalization data, permitting the study of a data set that would have been unrealistic to attempt to score manually.

In another application of neural networks to avian vocalizations, Nickerson *et al.* [42] used a single layer perceptron, a type of supervised neural network, to discover the acoustic features most relevant to the distinction between three different note types in black-capped chickadee (*Poecile atricapillus*) “chick-a-dee” calls (notes being the primary units of these calls). The network received seven frequency- and duration-related acoustic features as input and learned to predict the note type for these inputs. Testing the network with systematically modified inputs enabled them to determine which acoustic features were most important in discriminating note types.

SOMs or SOM-inspired networks have also been used in a number of studies to model the perception and classification of speech sounds of ones native language. For example, Guenther and Gjaja [43] trained an unsupervised network on formant frequency inputs. They then showed that the distribution of learned receptive fields exhibited the perceptual magnet effect humans exhibit in the perception of the vowels of their native language. Another example is a study by Gauthier *et al.* [44] that used a SOM to successfully classify Mandarin tones based on the first derivative of f_0 contours. This classification was robust in the face of the surface variability present in the multiple speakers connected speech from which the inputs were taken.

SOMs have also been applied to the study of disordered adult human voices. In one study, Leinonen *et al.* [45] trained a SOM on short-time spectra from 200 Finnish words. They then provided the network input from both normal and dysphonic speakers and tracked the trajectory of winning SOM nodes for the vowel [a:]. Normal and dysphonic voices differed in the amount of area on the SOM that was visited by these vowel trajectories. The work illustrates that a SOM tool can discriminate between normal and dysphonic voices, and that acoustic differences for these two populations can be portrayed topographically. Callan *et al.* [46] also used a SOM to study normal and dysphonic voices. However, instead of raw spectra, their inputs were scores on six acoustic measures that had previously been used in studies of dysphonia (e.g., amplitude perturbation quotient, first cepstral rahmonic amplitude, standard deviation of f_0). After training, they marked each SOM node according to which clinical group activated it the most. The SOM was able to reliably classify voices according to group. Output node activations and weights from the input (the six acoustic measures) were also visualized.

Finally, in an application of a neural network to the study of infant vocalizations, Schnweiler *et al.* [47] used a SOM to cluster recordings of cries by normal and deaf infants. The input consisted of 20-step Bark spectra. It was noted that different individuals cries mapped onto different areas of the SOM, which is in agreement with the idea that different infants produce identifiably different cries.

The results of the studies reviewed in this section suggest that neural networks, including the unsupervised SOM and the supervised perceptron networks, are appropriate and useful tools for visualization, feature-extraction, and classification purposes in the study of acoustic vocalization data. Thus, it seems fitting to explore the application of neural networks to study infant vocal development.

III. Method

A. Participants

Data from six typically developing human infant participants, three female and one male, are used in this study. Participants were recruited for a study of both interactive and spontaneously produced vocalizations and were recorded longitudinally from early in the first year until age 30 months (see Buder *et al.* [21] for additional details on participants and recording setup and procedures). The present study focuses on a subset of those recordings spanning three age intervals across the first year of life: 3;0–5;4, 6;0–8;4, and 9;0–11;4.

B. Recording

Infants were recorded for 2–3 20-minute sessions on each day of recording. For each infant, two of the 20-minute sessions were selected from each age interval for use in the present study. The selections were made from among available recordings based on there being a relatively high vocal activity level of the infant and a relative lack of crying.

Recordings took place in a minimally sound-treated room furnished with soft mats and toys while the parent was present. Siblings were sometimes present during recordings as well. Infants and their mothers interacted relatively naturalistically although some periods of time were dedicated to an oral interview between laboratory staff and the parent while the infant played nearby. Both mother and infant wore wireless microphones (Samson Airline UHF transmitter, equipped with a Countryman Associates low-profile low-friction flat frequency response MEMWF0WNC capsule, sending to Samson UHF AM1 receivers). The infants was sewn into a custom-built vest adapted from models designed by Buder and Stoel-Gammon [48]. The microphone capsule was housed within a velcro patch to locate the grill at a distance of approximately 5–10 cm from the infants mouth. Using TF32 [49] operating a DT322 acquisition card (Data Translation, Inc.), signals were digitized at 44.1–48.1 kHz after low-pass filtering at 20kHz via an AAF-3 anti-aliasing board. Microphone signals were concurrently sent to digital video recorders via separate UHF receivers to eliminate contamination to the signals that would otherwise have been transmitted from the video monitors via direct cables. The recordings for infant 1 are an exception to this procedure. This infants recordings were made according to an earlier laboratory protocol in which audio from the infants and mothers microphones were compressed in mp3 format as part of an mpeg recording file that combined audio with video. These recordings were subsequently extracted from mp3 format to wav format. Based on detailed inspection of these wav files, the only noticeable compression-based difference between the mp3-based wav file and those for infants 2-6 was that mp3 compression eliminated frequency components above about 12 kHz. To ensure signal comparability across all the recordings, only frequencies 12 kHz or lower are included in the signals processed by the neural networks in this study.

C. Utterance location and coding by human analysts

Prior to analysis by the neural networks, recordings underwent two types of processing by trained adult human analysts: 1) location of infant utterances within recording session files and 2) labeling these utterances according to protophone categories. Infants utterances were located within each recording using the spectrographic display feature of Action Analysis Coding and Training (“AACT”) software [50], marking the beginning and end of each utterance. In addition to listening to the recordings, analysts were permitted to consult spectrograms, waveform views, RMS contours, and videos for both the infant and the caregiver as they performed this localization task. An utterance was defined as a vocalization or series of vocalizations perceived as belonging to the same breath group [51]. Crying and other distress vocalizations as well as vegetative sounds were excluded. The first 49 utterances from each 20-minute session are used in this study.

Since 49 was the minimum total number of utterances produced in a session, this ensures equal representation of recording sessions, infants, and ages (see Table 2.1).

Table 2.1: Number of vocalizations of each vocal type for each infant at each age

Infant	Age 3;0 - 5;4			Age 6;0 - 8;4			Age 9;0 - 11;4			Total
	Voc.	*Sq.*	*Gr.*	*Voc.*	*Sq.*	*Gr.*	*Voc.*	*Sq.*	*Gr.*	
1	73	2	23	53	5	40	77	6	15	294
2	72	21	5	67	15	16	70	22	6	294
3	79	14	5	72	19	7	74	23	1	294
4	68	20	10	78	7	13	80	3	15	294
5	71	0	27	66	0	32	84	1	13	294
6	71	2	25	91	1	6	75	11	12	294
Total	434	59	95	427	47	114	460	66	62	1764

After locating infants utterances, analysts then coded each utterance as one of the following protophones: *vocant*, *squeal*, *growl*, or *other*. Analysts were encouraged during training to use intuitive auditory judgments rather than strict criteria. They were told that generally *squeals* are perceived as high pitched (beyond the range of habitual pitch for the individual) and can be dysphonated as well. *Growls* were portrayed as often having low pitch (again out of the range of habitual pitch) and as often being harsh or dysphonated, but it was noted that they are sometimes within the habitual pitch range with harsh or rough voice quality. *Vocants* were portrayed as the kinds of sounds that fit within the normal pitch of the infant, with relatively little deviation from normal phonation. Analysts were encouraged to attend to the most salient aspects of utterances in making *squeal* and *growl* judgments; that is, an utterance was not required to be high pitched throughout to be categorized as *squeal* a brief but salient high pitched event could form the basis for the categorization. These instructions were designed to encourage coders to mimic the discriminatory behavior presumed to underlie the categorizations reflected in reports of caregivers regarding these kinds of sounds in their infants [4, 26]. The coding procedures are similar to those used by Nathani et al.s [20] V (*vocant*) and SQ (*squeal*) categories. The difference was that in this study there is an additional growl category (see [21]) and classifications regarding vocal type category were made independently of any syllabicity judgment. Table 2.1 provides a summary of the number of utterances in each protophone category for each infant at each age.

D. Pre-processing of utterances

Processing of utterances from this point on was done in MATLAB using the Signal Processing and Neural Networks Toolboxes [52]. Each utterance was extracted from the digital recording for the session during which it was recorded. As all inputs to a standard SOM (see following description) must be the same length, only the first second of each utterance was used (utterances were therefore aligned at the beginning). Longer utterances were truncated and shorter utterances were zero-padded. A spectrogram was obtained for each utterance using the FFT-based spectrogram function. 15 time bins were used, each with 50% overlap and a maximum frequency of 22 kHz. The frequency scale of this spectrogram was converted to a 15-bin sine-wave approximation of the Bark scale [53] and the maximum frequency was capped at 12 kHz using Elliss inverse hyperbolic sine approximation algorithm from the Rastamat toolbox [54]. For each utterance, the power spectral density values represented by this spectrogram were normalized to the maximum power spectral density magnitude within that utterance. Each utterance was thus represented as 225 spectrogram pixels corresponding to the normalized power spectral density at each frequency bin for each time bin. Figure 2.1 illustrates some examples of the spectrographic representations of infant utterances in our data set.

E. Neural network architecture

In this section, the architecture of the neural networks and the functions of each component are described. The following two sections will describe neural network

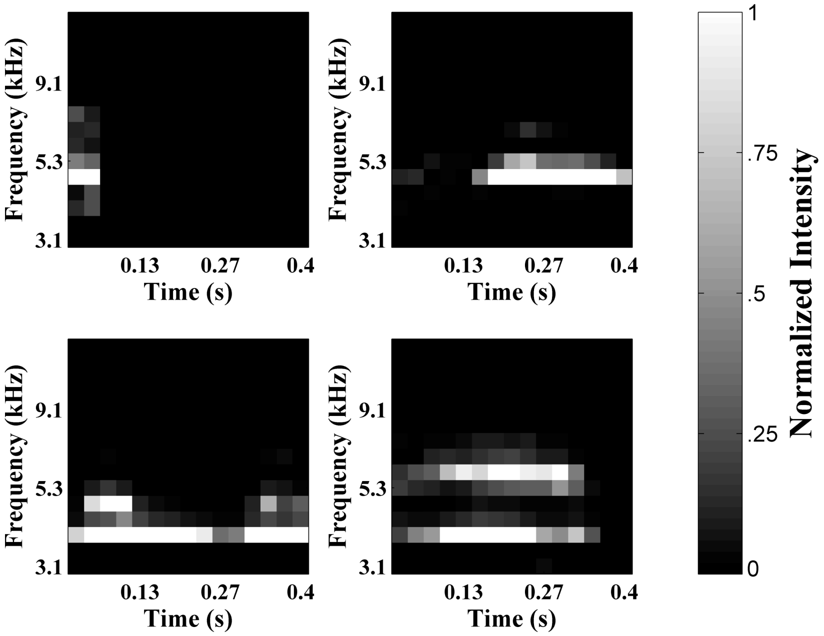

Fig. 2.1: Four examples of inputs provided to the SOM. Inputs are 225-pixel Bark-scaled spectrograms of utterances produced by infants recorded naturalistically. All inputs are one second long, with longer utterances truncated and shorter utterances zero-padded. White indicates high intensity and black indicates zero intensity. All spectrograms are normalized to the value of the highest intensity pixel. Clockwise, from top-left: a *vocant* produced by infant 1 at 3;2, a *squeal* produced by infant 2 at 4;1, a *growl* produced by infant 4 at 6;2, a *vocant* produced by infant 3 at 10;2.

training. This will be followed by a description of how the infant utterance data was divided up into a set for training and a set for testing each network.

The main type of neural network used in this study is a hybrid architecture with two components (Fig. 2.2). The first component is a self-organizing map (SOM) consisting of 16 nodes arranged on a 4×4 grid. The choice of number of nodes and their arrangement was made on the basis of pilot analyses using various configurations, considering ease of visualization and balance between specificity and over-fitting of data. The SOM receives utterance spectrograms as input, transformed into a vector with the time-slice columns of the spectrogram laid end-to-end. Note that this is a common procedure for formatting neural network input data (e.g., see [41]), and that the transformation has no effect on the function of the SOM since the SOM algorithm does not take the location of input nodes into account. The SOM categorizes these utterances according to learned holistic features extracted based on a set of training utterances, as described in the following section. Learning in the SOM is unsupervised and involves changing the weights from the input layer to each of the SOM nodes over the course of training. Eventually, these weights come to represent the nodes ideal inputs (or receptive fields), and neighboring nodes end up having similar ideal inputs (topographic organization). This SOM component of the hybrid architecture thus serves as a data-driven holistic feature detector, reducing the 225-pixel spectrographic input to 16 learned features. It also serves as a means for visualizing utterance features topographically [38, 39, 40]. The SOM component was implemented using functions custom written by the first author for this study in MATLAB.

The second component is a set of three single-layer perceptrons which are used to read the output from the SOM in order to obtain a quantitative measure of the learned SOM features relevance to various classification tasks, to actually perform those classifications, and to determine which SOM nodes best distinguish different classes of utterances from each other. The perceptron is a type of supervised classifier and in single-layer form it essentially performs a logistic linear regression [37]. Each perceptron receives activations of the SOM layer nodes (produced in response to a single utterance input to the SOM) as input (see [55] for another example of a perceptron trained on SOM activations). Based on the product of these SOM activations and the perceptrons weights (which can be either positive or negative) from the SOM layer to its output nodes, the perceptron classifies a given utterance according to its perceived protophone type as judged by trained human analysts, the age at which an utterance was produced, and the identity of the infant who produced it. Thus, the supervised perceptrons relate the features learned by the unsupervised SOM to known protophone, age, and identity classifications.

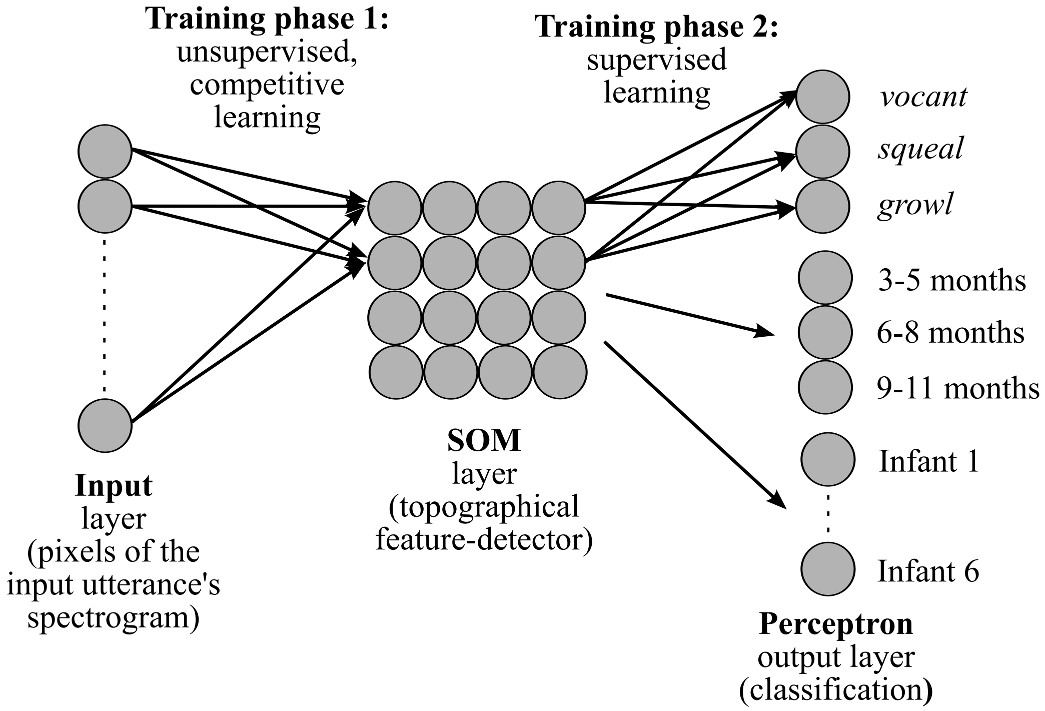

Fig. 2.2: Schematic diagram of the neural network used in the present study. The network is a hybrid of a self-organizing map (SOM) and a single-layer perceptron. Pixels of an utterance are presented first to the SOM. Activations of the SOM nodes are then sent to the perceptron output nodes for classification according to protophone, age, and infant identity. The weights from the input layer to the SOM layer are trained first. After this first phase of training, weights to the SOM are frozen and the perceptrons weights are trained.

The output layer of each of these perceptrons was constructed to have one node for each class of utterances. The vocal type protophone perceptron thus has three output nodes: one for squeals, a second for vocants, and a third for growls. The age-predicting perceptron has three output nodes: one for utterances produced at age 3;0–5;4, a second for utterances produced at age 6;0–8;4, and a third for utterances produced at age 9;0–11;4. Finally, the identity-predicting perceptron has six output nodes, one for each infant in our data set. The perceptron component was implemented using the feed-forward network functions in MATLABs Neural Network Toolbox [56]. Logistic activation functions were used for the output nodes of the perceptron classifiers and default values were used for all other parameters in initializing the network (further details can be found in [56]).

To compare the hybrid SOM-perceptron classifier to the multilayer perceptron (MLP), which is probably the most popular multilayer neural network used in supervised classification [37], we also trained a set of MLPs to perform the age and vocal type classifications using the leave-one-infant-out training data. These MLPs were run using the same procedures and parameter settings as for the single-layer perceptrons described above. The number of hidden layer nodes was set to 16, which is the same as the number of nodes in the SOM layer of our SOM-perceptron hybrid. Thus, the number of weights (i.e., free parameters that the networks adjust during training) are roughly similar. We then compared the MLPs classification performance to that of our SOM-perceptron hybrid. In addition, we trained a single-layer perceptron to predict age on the basis of a protophone-trained MLPs hidden layer activations. Likewise, we trained a single-layer perceptron to predict protophones on the basis of an age-trained MLPs hidden layer activations. Comparing classifications of these perceptrons to classifications from the SOM-perceptron hybrid assesses whether using the SOM layer is truly critical to obtaining a task-general hidden layer.

F. Neural network training

For the SOM-perceptron hybrid, training was conducted in two phases. During the first phase, only the SOM component was involved. Prior to training, its weights were set to random values with a different randomization for each of the 15 SOM runs. The SOM training algorithm was adapted from Berglund and Sittes [57] parameterless SOM algorithm. This algorithm takes three parameters (β, Θ, and ε) which determine the behavior of the SOM during training. The following parameter values were used: β = 1, their method 2 for calculating Θ, and multiplying ε by a factor of .5. The exact roles of β,

Θ, and ε are described in Berglund and Sitte [57]. In essence, training involved presenting an utterance as input (randomly chosen from the set of training utterances, discussed in the next section) to the SOM, finding the SOMs node whose weights to the input layer are the most similar to that input (as measured by the Euclidean distance between the input vector and the vector representing weights from the input to a given output node), and then updating that nodes weights and (to a lesser extent) its neighbors weights to make them even more similar to the input. This procedure was repeated 1407 times. This was the number of utterances per session times the number of sessions times the reciprocal of the scaling factor for ε in the SOM training algorithm. This amount of training was more than sufficient for the networks performance to stabilize as judged by the mean squared distances between inputs and their winning nodes weights for the testing set utterances and by visual inspection of changes in network weights across training.

After completion of this first phase of training, the weights from the input layer to the SOM nodes remained fixed during the next training phase. This second phase of training the SOM-perceptron hybrid involved only the perceptron component. Perceptrons were trained using the delta rule with regularization using Matlabs *trainbr* function. This is a variation on the traditional delta rule algorithm that balances reduction of classification error against parsimony of network weight. This method (sometimes also referred to as “learning with forgetting”) has been shown to produce good generalization of performance to previously unseen data and increases the interpretability of network weights [58, 59, 56]. In essence, this training algorithm involves presenting training set examples, which are the SOM node activations produced in response to an infant utterance, one at a time. After presentation of each example, the networks classification predictions are calculated, and then, based on the difference between these classification predictions and the correct classifications, the weights from the SOM layer to each of the perceptrons output nodes are updated so as to reduce this error (as measured by the squared error) in classifying subsequent inputs while also maintaining parsimony of network weights. All parameters other than the training method (*trainbr*) and the activation transfer function (*log-sigmoid*) were set to default values. Further details can be found in the MATLAB documentation and in [56].

The MLPs were trained in mostly the same way as the perceptron described above but with the following exceptions: The MLP was trained directly on the spectrographic input and was done in a single phase. Training was performed using the same MATLAB training method (*trainbr*), but since there were two layers instead of just one, backpropagation was involved in addition to the delta rule [37].

G. Partitioning of data into training and testing sets

In order to train the SOM, perceptron, and MLP while also allowing for testing the networks generalization abilities, the infant utterance data was divided into two subsets, one for training the network and the other for evaluating the networks classification performance. From each recording session (of which there were two for each child at each age), 37 of the 49 utterances (approximately 75%) were randomly chosen to be used in training; the remaining 12 utterances (approximately 25%) were reserved for testing the network (discussed in the following section). This random partitioning was done 15 times and the SOM-perceptron hybrid was run 15 times, each corresponding to a different random partitioning. The means and standard deviations presented in the Results section were computed over these 15 runs.

In a variation on this training procedure, an alternative leave-one-infant-out method of partitioning the data into training and testing sets was applied to a second set of 36 networks, wherein all the utterances produced by five infants were used in training and the utterances from the sixth remaining infant were reserved for use in testing only. Across these 36 networks, each infant was used as the test infant six times. As with the perceptron, means and standard deviations were computer over these 36 runs. The MLP simulations were trained and tested using the leave-one-infant-out method, although only

6 simulations (rather than 36) were run due to the long time it took for MLP runs to complete. Each infant was used exactly once in testing.

In addition, a SOM-perceptron hybrid was trained on all utterances from all recordings for the purpose of visualizing the trained network weights and activations. This network was used for generating Figures 2.3-2.5 but was not included in any of the quantifications of network performance.

H. Adjusting for unequal representation of protophone categories

When training and testing the perceptrons and MLPs responsible for predicting protophone judgments, it is a concern that *vocants* occur much more frequently than *squeals* and *growls* (see Table 2.1). This inequality inflates the percent correct that would be expected by chance, since with unequal numbers, the baseline strategy would be to give all utterances the classification corresponding to the most frequent category. With such a strategy, if 70% of the utterances were *vocants*, the baseline percent correct would be 70%. This would be very difficult for even a "smart" classifier making use of acoustic information to outperform. We thus ran the perceptron component two ways: once without any adjustment for unequal numbers of vocal types and once with an adjustment. To adjust for the frequency bias, exemplars from the *squeal* and *growl* categories were repeated as many times as was necessary for their numbers to equal the number of vocants.

I. Evaluating the network's performance

After training the hybrid network, the networks performance was assessed (1) through visualization and descriptive acoustic measurement of network weights and activations and (2) through quantitative evaluation of classification performance. The visualizations are of the weights from the spectrographic input layer to the SOM output grid and from the SOMs grid to the perceptron classifier nodes. We also visualized the winning SOM nodes (an illustration of SOM activations) for utterances with different protophone judgments, from different ages, and from different individuals.

To supplement the visualizations, we made three theoretically-derived acoustic measurements for each of the 16 SOM receptive fields. The first measure was the mean of the time-averaged spectrum and the second was the standard deviation of this spectrum, both measured in absolute frequency. These correspond to the first and second spectral moments computed by Forrest *et al.* [60]. The third measure was the median point in time of the frequency-averaged intensity contour. This should give a rough measure of the preferred duration of the receptive field. After calculating these three values for each SOM node, we calculated each perceptron output nodes preferred value for each of the three acoustic measures by finding the average SOM receptive field values weighted by the perceptron output nodes weights from the SOM layer.

Quantitative evaluation involved feeding the networks utterances from the set that were reserved for testing. The networks classifications regarding the protophone, age group, and infant identity for each of these test utterances were then obtained, and an overall mean percentage correct for each type of classification for each type of network was computed. Cohens κ reliability statistics and their corresponding probabilities were computed to evaluate the magnitude and significance of the agreement between each networks classifications and the correct classifications.

IV. Results

A. Visualization and descriptive acoustic measurement of network weights and activations

Each of the SOMs output nodes can be thought of as a holistic spectrographic feature formed by the SOM based on the training inputs. This is illustrated in Figure 2.3, where the weights from the input layer to each node of the SOM are visualized as spectrograms representing the preferred spectrographic input for that node (white indicates a high value for a given weight and black indicates a zero value). Each nodes spectrogram of weights can be thought of as a receptive field, specifying a particular preferred holistic feature derived from the input infant utterance data via the SOMs training algorithm. Note that these preferred inputs are arranged topographically; that is, neighboring nodes have similar preferred inputs. This is one of the characteristic properties of SOMs. Also note that, because the SOM nodes adjust their preferred inputs (i.e. their weights from the input layer) on the basis of exemplars from the training set of utterances, the nodes of the SOM come to represent global features of a complex nature such as would occur in an actual infant utterance. Thus, it seems that these features have a complex relationship with more basic acoustic features, such as duration and spectral compactness versus diffuseness. For example, the receptive fields for the SOM nodes pictured in the leftmost column ($x = 1$) of Figure 2.3 appear to exhibit long duration. In addition, the bottom two nodes of that leftmost column ($x = 1$, $y = 1$-2) have relatively high spectral means and spectral standard deviations. These observations are supported by measurements of the frequency-averaged intensity contours median times, the time-averaged spectral means, and the time-averaged spectral standard deviations given in Table 2.2.

Table 2.2: Acoustic properties of the SOM receptive field spectrograms

Spectral means[a]				Spectral standard deviations[a]				Temporal medians[b]			
0.7	0.7	0.7	0.6	0.7	0.7	0.7	0.6	7	5	4	4
0.7	0.8	0.7	0.6	0.8	0.8	0.7	0.7	9	4	3	3
1.3	1.1	0.7	0.5	1.3	1.8	0.9	0.8	8	4	2	3
1.7	1.3	0.6	0.3	1.6	1.3	1.1	0.9	6	4	3	5

[a] In kilohertz.
[b] In number of time bins (each bin is 66 ms).

Figure 2.4 illustrates how often each node of the SOM was the winning node (defined as the node with the highest activation) for utterances of each perceived protophone type, age group, and individual. This method of visualization provides a way of mapping from global features learned by the SOM to different utterances classes.

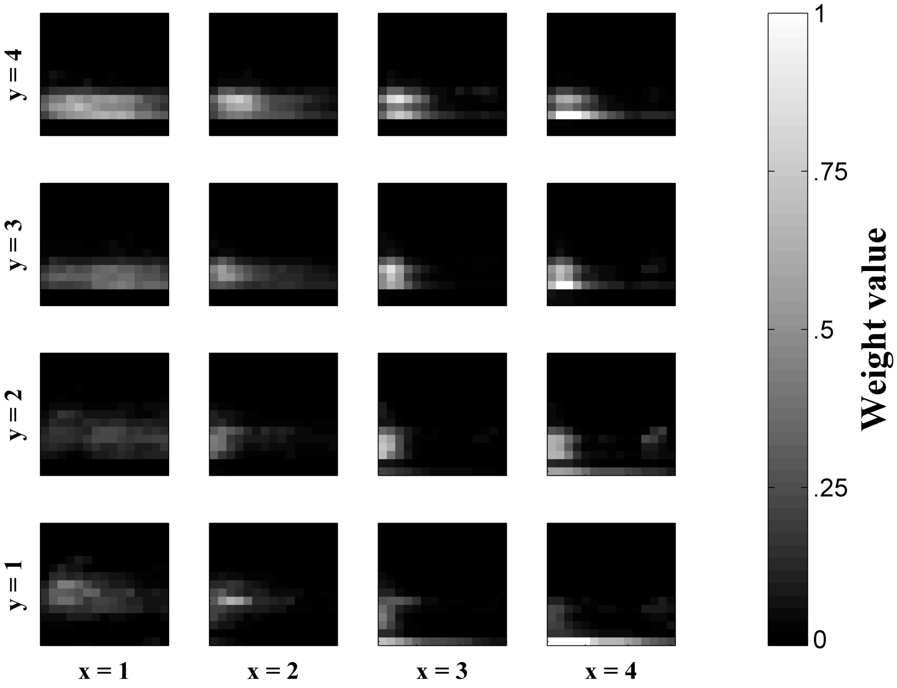

Fig. 2.3: Weights from the input layer to each SOM node for a network trained on the full set of utterances. Each spectrogram represents the input preference for that node. Note that input preferences are holistic spectrographic features and represent complex acoustic patterns. Also note the topographic organization of these inputs. White represents high weight (high intensity preference for that pixel on the input layer), and black represents zero weight.

For example, it appears in the figure that *growls* may span a broader set of global features in the acoustic space represented by the SOM, as evidenced by the large number of relatively bright squares (bright indicates high activation) for this protophone category. To quantify the diffuseness of activation across the SOM for a given utterance class, we first calculated for each node the number of inputs for which that node was the winner, divided by the total number of inputs belonging to that class. Then the median of these proportion values was computed. These medians were compared across the three protophone categories. Indeed, the median was higher for *growls* (.24) than for *vocants*

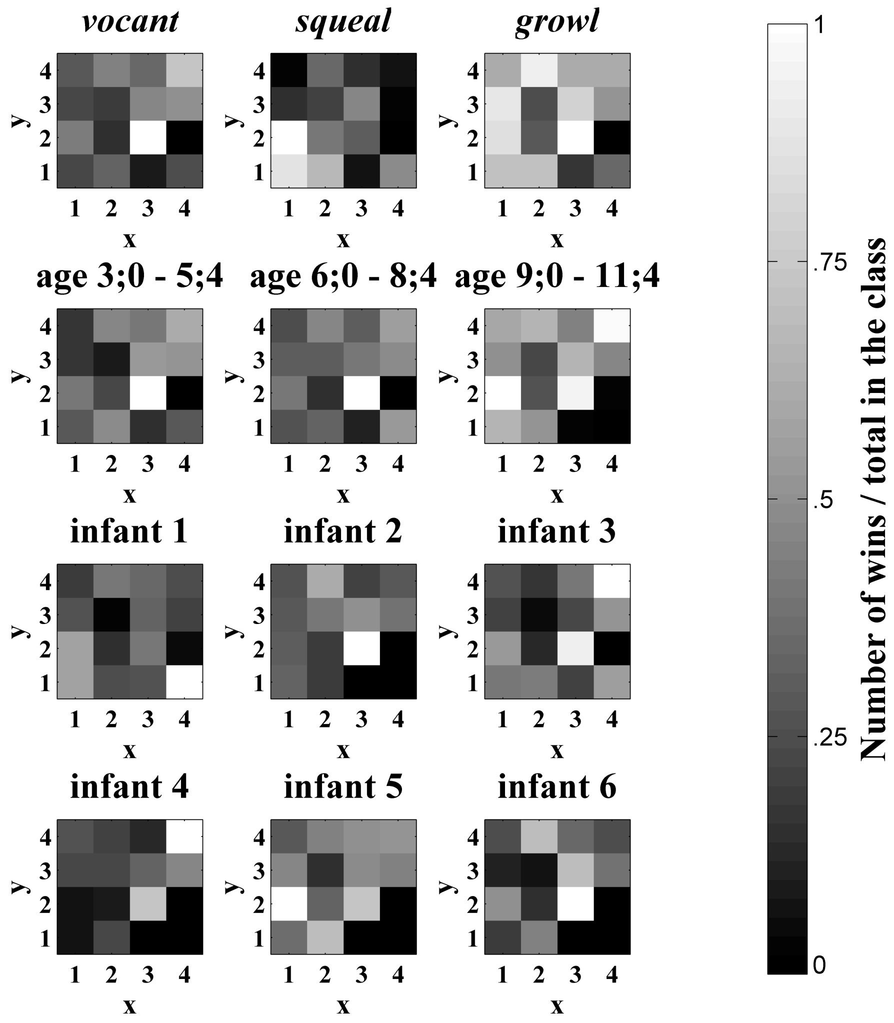

Fig. 2.4: Activations of the SOM layer by utterances with different protophone labels, produced at different ages, and produced by different infants. Bright indicates that a SOM node was often the winner and black indicates that a node was never the winner. Each 4×4 map corresponds to nodes of the SOM shown in Figure 2.3. Note that the number of utterances belonging to each protophone category was not uniform; there were more *vocants* than *squeals* and more *squeals* than *growls*.

(.18) or for *squeals* (.15). This indicates that the winning nodes for this category are distributed more evenly across the map than for the other categories.

Another observation that is evident in Figure jasa:fig4 regards the overlap between utterance classes. While there is some distinctness across protophones, as indicated, for example, by there being different most highly activated nodes for *squeals* (the node at $x = 1, y = 2$) than for *vocants* and *growls* (the node at $x = 3, y = 2$), there is also a high degree of overlap in the SOM node activations, as indicated by numerous nodes that show gray activation for all three protophone types.

Figure 2.5 illustrates the weights from the SOM to the perceptron output nodes for each age, infant, and protophone prediction. Recall that the goal of the perceptron is differentiation of categories (protophone type, age, infant) via positive and negative weights. Thus for Figure 2.5 the scaling is different from that of Figure 2.4. In Figure 2.5, white indicates high positive weight, black indicates high negative weight, and mid-gray indicates near-zero weight. The weights indicate which of the SOMs holistic features are informative for classification purposes, highlighting the differences between utterance classes and ignoring features that are common to all classes.

The visualizations in Figures 2.4 and 2.5 exhibit both similarities and differences. This is evident in the correlation coefficients between a given classs SOM activations (Fig. 2.4) and the weights from the SOM to its perceptron node (Fig. 2.5). The mean, across all class types, of these correlation coefficients is r = .31 where r was always positive, ranging from .03 to .58. As an example of a specific similarity between activation and weight patterns, the SOM nodes located at $(x = 1, y = 4)$, $(x = 4, y = 4)$, and $(x = 4, y = 3)$ are very dark for *squeals* in both figures. This indicates that these SOM nodes are both infrequent (Fig. 2.4) and negatively associated characteristics (Fig. 2.5) of *squeals*. An example of a difference between the two figures is that, while the SOM node located at (x=4,y=4) is the second highest activated for *vocants* as shown in Figure 2.4, it does not have a very large positive weight to the perceptron *vocant* node, as indicated by Figure 2.5. Differences between Figures 2.4 and 2.5 are due to the fact that Figure 2.4 indicates the frequency with which features were observed whereas Figure 2.5

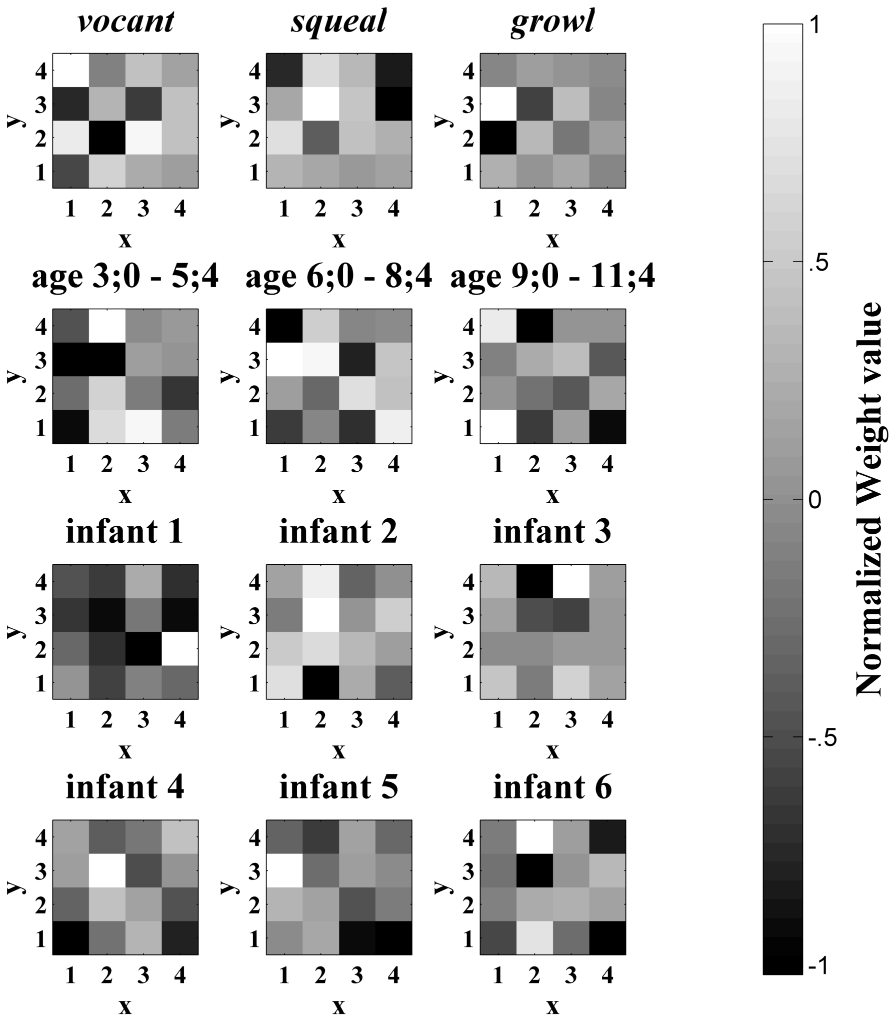

Fig. 2.5: Weights from the SOM layer to each perceptron output node. Bright indicates a large positive weight from the SOM node to that perceptron output node, black indicates a large negative weight, and gray indicates a near-zero weight. Each 4×4 map corresponds to the SOM nodes shown in Figure 2.3. Note that for protophones, the weights are based on training the perceptron on a set of utterances that was adjusted to be balanced across *vocant*, *squeal*, and *growl* protophones by randomly repeating exemplars from the less frequent categories.

highlights the particular SOM nodes that, when activated at least partially by an utterance, distinguish utterances of one class (e.g., *vocants*) from utterances of other classes (e.g., *squeals* and *growls*).

Recall the discussion of duration, spectral mean, and spectral standard deviation from the discussion of the SOM receptive field spectrograms (Fig. 2.3). It was observed that the leftmost column was associated with long duration and that the bottom two nodes of that column also had high spectral means and standard deviations. Interestingly, this leftmost column appears both in Figure 2.4 and in Figure 2.5 to be associated more (as evidenced by light-colored pixels in this column) with the older two age groups than with the younger age group. This suggests that increase in duration is associated with increase in age. In addition, the bottom two nodes of that leftmost column are associated with the oldest age group. This suggests that the oldest age group is associated also with increase in spectral mean and standard deviation. Combining information about the acoustic properties of SOM weights and the values of the weights from the SOM layer to each of the perceptron output nodes, it is possible explore whether these acoustic features are present in the nodes that distinguish between different ages. Table 2.3 shows the spectral mean, spectral standard deviation, and temporal duration properties for each age, protophone type, and infant. Indeed, the spectral duration of perceptron weights appears to increase across the three age groups and the spectral mean and standard deviation are highest for the oldest age group.

Table 2.3 also reveals interesting patterns with respect to the three protophones acoustic properties. Squeals have the highest spectral mean and spectral standard deviation. This is in accordance with previous descriptions of this category as high pitch often accompanied by harshness and/or pitch variation. However, growls do not differ from vocants in either mean or spectral standard deviation. Perhaps the high harshness/pulse/creaky-voice combine with the low pitch of growls to yield moderate spectral mean values. Thus, although the differentiating acoustic properties of squeals fit

Table 2.3: Acoustic properties of the perceptron weights from the SOM layer, given the acoustic features of the SOM nodes

		Spectral mean[a]	Spectral SD[a]	Temporal median[b]
Age	3;0–5;4	0.80	0.91	4.11
	6;0–8;4	0.75	0.89	4.67
	9;0–11;4	0.87	0.95	4.84
Protophone	Vocant	0.78	0.89	4.49
	Squeal	0.85	0.96	4.69
	Growl	0.80	0.91	4.68
Infant	Infant 1	0.81	0.94	4.55
	Infant 2	0.83	0.93	4.62
	Infant 3	0.82	0.94	4.69
	Infant 4	0.77	0.87	4.47
	Infant 5	0.89	0.95	5.12
	Infant 6	0.84	0.92	4.41

[a] In kilohertz.
[b] In number of time bins (each bin is 66 ms).

with their previous perceptual descriptions, the differentiating acoustic properties of growls may be less straightforwardly defined in this neural network.

B. Classification performance

1. Protophone classification performance

When predicting human-judged protophone categories after equated-frequency training, the 15 hybrid networks had a mean percent correct on the previously unseen test utterances (selected randomly at the recording session level) of 54.4% (see the first column of Table 2.4). Since there were three protophone types, each of which was equally represented in both the training and the testing utterance sets, the classification performance that would be expected for a classifier performing at chance is 33.3%. The vocal-type-predicting networks 54.4% correct performance was significantly better than chance, κ = .316, $p < .001$. When no adjustment was made for the inequality in the number of exemplars in each protophone category, the percentage correct was 73.4% where the baseline percent correct for an algorithm that always guessed *vocant* would be 74.9%.

Table 2.4: Classification task performance of the SOM-perceptron hybrid neural network

	Protophones		Ages		Identities
Type of test set	25% per recording	100% of one infant[a]	25% per recording	100% of one infant	25% per recording
Mean % correct	54.4 (chance = 33.3)	55.0 (chance = 33.3)	42.8 (chance = 33.3)	35.6 (chance = 33.3)	32.4 (chance = 16.7)
Standard deviation	3.2	6.5	1.4	4.7	1.9
Mean Cohen's κ	.316	.325	.142	.034	.189
Mean p	<.001	<.001	<.001	.146	<.001

[a] With adjustment for unequal catcategory sizes. When there is no adjustment and one infant is reserved for testing, the mean percent correct is 73.4 (chance = 74.9) with a standard deviation of 5.4.

Recall that 36 additional hybrid networks were trained on utterances from five infants and tested on the sixth remaining infants utterances. Each infant was used for testing for exactly 6 of the 36 networks. The purpose of this variation on the method for partitioning utterances into training and testing sets was to see if classification of protophones would generalize across infants. Mean classification performance for these networks was 55.0% correct, where chance level performance would have been 33.3% (see the second column of Table 2.4). This was statistically better than chance, κ = .325, p $<$ .001. This shows that for protophone prediction, performance did not differ from when the session-level train-test partition method was used to when the leave-one-infant-out method was used. Thus, it appears that the networks protophone-classification capabilities are based on features of utterances that are generalizable even to infants the network has never previously encountered.

For the six MLPs that were trained using a leave-one-infant-out data partition to predict protophones (where the numbers of protophones were adjusted to give equal representation of all categories), the mean percent correct was 45.9% (see the first column of Table 2.5). This was not quite as high as performance of the SOM-perceptron hybrid, although across runs, this performance was within a standard deviation of the SOM-perceptron hybrids. When no adjustment was made for the inequality in the number of exemplars for each protophone category, the percentage correct was 65.3% where the baseline percent correct for an algorithm that always guessed *vocant* would be 74.9%. When 6 MLPs were trained using the same leave-one-infant-out method to predict age and then a single-layer perceptron layer was trained to take those MLPs hidden layer activations as input and produce protophone classifications as output, performance was 46.6% correct (see the second column of Table 2.5). This was lower than the SOM-perceptron hybrid by more than eight percentage points. These combined results of the MLP networks suggest that while a MLP trained to perform protophone prediction may perform similarly to the SOM-perceptron hybrid, the hidden layer of other another MLP trained on a different classification task (age-prediction) is not as good as the general-purpose unsupervised SOM layer. Furthermore, the MLP did not fare any better than the SOM when there was no adjustment for the overrepresentation of *vocants*.

Table 2.5: Classification task performance of the MLP neural network. All data are for leave-one-infant-out partitioning of utterances.

	Protophones		Ages	
Type of hidden layer	Protophone-predicting[a]	Age-predicting	Age-predicting	Protophone-predicting
Mean % correct	45.9 (chance = 33.3)	46.6 (chance = 33.3)	35.1 (chance = 33.3)	36.1 (chance = 33.3)
Standard deviation	10.3	5.7	3.6	3.0
Mean Cohen's κ	.191	.200	.026	.041
Mean p	<.001	<.001	.118	.018

[a] With adjustment for unequal category sizes. When there is no adjustment and one infant is reserved for testing, the mean percent correct is 65.3 (chance = 74.9) with a standard deviation of 6.9.

2. Age classification performance

For the 15 hybrid networks trained to predict infant age with a session-level training-test data partition, the mean percent correct was 42.8% (see the third column of Table 2.4). This was significantly better than the 33.3% that would have been expected by chance, κ = .142, $p < .001$. Mean classification performance for the 36 additional hybrid networks that were trained on utterances from five infants and tested on the sixth was approximately 35.6% correct, where chance level performance would have again been 33.3% (see the fourth column of Table 2.4). This did not reach statistical significance, κ = .034, p = .146.

The 6 MLPs that were trained using a leave-one-infant-out data partition to predict age had a mean percent correct was 35.1% (see the third column of Table 2.5). This was very similar to the performance of the SOM-perceptron hybrid. When 6 MLPs were trained using the same leave-one-infant-out method to predict protophones (numbers adjusted for equal representation of protophone categories) and then a single-layer perceptron layer was trained to take those MLPs hidden layer activations as input and produce age classifications as output, performance was 36.1% correct (see the fourth column of Table 2.5). This was again very similar to the performance of the SOM-perceptron hybrid. These combined results of the two MLP variations suggest that both a MLP and the SOM-perceptron hybrid are approximately equally suited to the task of predicting age, though neither does very well when forced to generalize to an infant it has never previously encountered before.

3. Infant identity classification performance

For the 15 hybrid networks trained to predict the identity of the infant who produced an utterance (with session-level training-test data partition), the mean percent correct was approximately 32.4% correct (see the fifth column of Table 2.4). Compared to the 16.7% correct that would be expected had the networks been performing at chance, this performance was statistically significant, $\kappa = .189$, $p < .001$.

V. Discussion

A. Visualization of network weights and activations

One of the main advantages of the SOM-perceptron hybrid is its usefulness for data visualization purposes. By plotting the weights from the input layer to the SOM (Fig. 2.3), it is possible to visualize the range of holistic spectrographic features exhibited by the vocalizations in the present data set. These holistic features are extremely complex, which can be seen as both an advantage, in that they retain the complexity of prototypical utterances, and as a disadvantage, in that they are difficult to interpret. By plotting the activations of each SOM node according to protophone, age, and identity, and by plotting the weights from each SOM node to each perceptron node, it is also possible to explore the relationship between the holistic spectrographic features learned by the SOM and different categories of utterances.

One method that was used to quantitatively interpret the trends observed in the figures was to get the median number of wins per SOM node for a specific utterance type (e.g., for each of the protophone types) to see which tended to occupy more of the SOMs representational space. Using this method, it was found that *growls* had more diffuse activation of the SOM than *squeals* or *vocants*, suggesting that *growls* have a larger range of acoustic variability.

In another approach to interpreting the trained network we showed that since the SOMs receptive fields take the same form as their inputs, which in this case are coarse-grained spectrograms, more traditional acoustic descriptions, such as spectral mean, spectral standard deviation, and temporal median (related to duration) can be gotten. As observed in the Results section, the leftmost column of SOM nodes in Figure 2.3 had long durations and the bottom two nodes of that column had high spectral mean and standard deviation. These nodes also had a tendency to be activated more by utterances from the older two age groups (6;0–11;4) than by utterances from the oldest age group (3;0–5;4), as evidenced by their lighter colorings in Figures 2.4 and 2.5. Thus, a hypothesis for future investigation might be that utterances produced at older ages are not only longer in duration but also higher in spectral mean and variance.

B. Classification performance

The hybrid neural network, consisting of a perceptron classifier operating on the SOMs holistic spectrographic features, is able to reliably classify one-second-long utterance samples according to vocal type protophones, ages at which they were produced, and the identities of the individuals who produced them. Reliable performance on these classification tasks provides support for the validity of the SOMs learned utterance features, suggesting that they reflect meaningful acoustic variation in infants vocal productions. One of the most important possible applications of the work represented here may be in contributing to the rapidly growing field of automated analysis of vocalization. MLPs trained on the same classification tasks also performed well, so when the goal is purely classification, and comparison of holistic features across different classifications is not important, MLPs may also be a good choice of tool.

It should be emphasized that the critical issue for the future of automated vocal analysis is that reliability be significant, not necessarily that it be high. With very large datasets, relatively low kappa values should present no important problem. If a signal is consistently (even though at low levels) detectable, it can become highly discernible at high Ns. This principle is widely recognized for example in the field of averaged evoked potentials [61]. It should also be noted that, although the methods used in the present study did involve some processing by human analysts, this was only in order to perform utterance extraction and protophone labeling. An automated infant utterance extraction method has already been developed for very large vocalization data sets taken from day-long recordings (see [17]), and such a method could be substituted for the manual utterance extraction performed here. As for protophone labeling, for model training and evaluation, the use of human judgments in establishing gold-standard classifications is unavoidable. However, for automated analysis of large data sets, training and evaluation of a network using manually labeled utterances need not be done on the entire large data set, but only on a sample of data large enough to ensure satisfactory network performance and generalization.

The ability to reliably classify utterances according to protophone is of considerable interest. At present, protophone categories are widely used in studies of infant speech development in both typically and atypically developing children (e.g., [62, 63, 64, 65]) as well as in tools that use infants vocalizations in their assessment of infants communicative function (e.g., [66]). The ability to predict trained analysts perceptual judgments suggests that neural networks and other data-driven statistical tools have the potential to be used for automatic classification of protophone categories (although a workaround for the issue of frequency imbalance across categories would have to be devised). This would be useful in many research and clinical settings where coding by trained analysts is costly or difficult. In the future, it would be interesting to apply either the SOM-perceptron-hybrid or a MLP to the classification of other protophones, such as marginal syllable and canonical syllable categories related to the articulation timing of syllables that have been shown to be of particular importance as indicators of normal development [2].

The ability to identify age, combined with the networks ability to identify the individual who produced a given utterance suggests that neural networks and related approaches have the potential for future use in classifying utterances produced by delayed or disordered individuals. Prediction of infant identity also lays groundwork for future work that might attempt to classify utterances produced by infants from different cultural or linguistic backgrounds and by female versus male infants.

C. Future development

1. Manipulating the network's input

The SOM-perceptron architecture is highly flexible with regard to the type of information it can be given as input. Although 225-pixel spectrograms of one-second-long utterance samples were used in this study, such an input representation was chosen primarily for its computational efficiency and because it involved relatively little pre-processing of data. It is possible that other formats of input would yield better performance or additional insights. Future studies might compare features learned by SOMs trained on different types of input, be they relatively raw input (e.g., raw waveforms, spectrograms of various frequency and time resolutions), more traditional acoustic measures (e.g., f_0 means, formant frequency means, amplitude means, durations), or measures that represent intermediate amounts of pre-processing (e.g., f_0 contours, formant frequency contours, amplitude contours).

In discussing the visualizations afforded by the SOM-perceptron hybrid, reference was made to how these visualizations might be related to acoustic patterns described in more traditional terms. For example, it was noted that the SOM features duration preferences appear to increase with increasing age. Although beyond the scope of the present study, this hypothesis could be tested by comparing the present SOM-perceptron hybrid (trained on raw spectrographic input) with a SOM-perceptron hybrid network trained on duration alone. That is, rather than providing the network with pixels of spectrograms as input, one could provide the network with a single value representing an utterances duration. If such a network trained on utterance durations also performs significantly well, this would indicate that changes in utterance duration are indeed associated with increasing age. One could then train a SOM-perceptron network on input consisting of a spectrogram plus an additional feature representing the utterances duration. If this network performed better than the network trained only on spectrograms (e.g., as measured by a hierarchical regression), this would imply that duration changes systematically with development but was not adequately represented by the SOM trained only on spectrograms. On the other hand, if the two networks perform equally well, this might suggest that the SOM had already encoded the relevant duration information in its features. This type of approach could provide a means for parsing out the role of various acoustic measures in how well they predict the age (or identity, protophone category etc.) of infant utterances.

Finally, it would be highly desirable to explore input representations that deal more flexibly with temporal aspects of vocalizations. Infant utterances vary in length and often have prosodic and syllabic components that vary in their timing. The current static spectrograms used as input do not adequately deal with this fact. A better solution might be one in which small time segments of spectrograms (or other acoustic features) of infant utterances are presented in sequence. The network would then make classifications at each time point or at the conclusion of the entire sequence. A change of this sort in the temporal nature of the input would, however require changes in the network architecture. Some possibilities are proposed as part of the next section.

2. *Alternative architectures and algorithms*

The choice of a SOM-perceptron hybrid architecture was motivated by the fact that these architectures had been previously applied to related problems involving the visualization and classification of acoustic vocalization data, including avian song, disordered adult voice, and infant crying. The choice of a SOM as the first element of this architecture was also motivated by studies suggesting that SOMs can produce results that are comparable to other statistical clustering and visualization methods [67, 68]. Choosing a SOM for the first component of the two-component hybrid network also has the advantage that the same first component is used regardless of the classification task performed by the subsequent perceptron component. Thus, the middle layer activations and weights can be compared across different classification tasks (e.g., the SOM node activations and weights for younger utterances can be compared to the SOM node activations and weights for vocants, squeals, and growls). Finally, the biologically-inspired features of the SOM, notably its topographical self-organization and incremental learning algorithm are also seen as advantages (see the section below on future modeling directions; [69, 70, 39].

Nevertheless, exploration of other architectures could yield better performance or additional information. For example, a two-layer perceptron may be worth using for situations where classification performance and differentiation between classes is the primary goal. Furthermore, non-neural-network statistical models, such as mixtures of Gaussians, k-nearest-neighbors analysis [40], and possibly even linear discriminant analysis and regression techniques could potentially yield as good or better clustering and classification performance, respectively. Future work could compare such methods on their performance on a specific visualization or classification task.

In addition, recurrent neural networks are often considered better for temporal sequence processing than networks that take static input [71]. Thus, given that infant vocalizations are temporal patterns occurring in temporal sequences, it would be worthwhile to explore recurrent versions of the SOM (e.g., [72]) when unsupervised analysis is desired, or the simple recurrent network (SRN; [71]), when classification or prediction is the primary goal. Perhaps even a hybrid of the recurrent SOM and the SRN could be used, which would be analogous to the static SOM-perceptron hybrid explored in the present study. Moving to such temporal architectures would involve changing the nature of the networks input representation as discussed in the previous section. Instead, a fixed moving window of spectral input would be appropriate.

Finally, variations on the SOM that allow for uncertainty in the number of features/categories or that allow for hierarchical organization of features/categories [73, 74] might also prove useful and informative. The SOM-perceptron hybrid presented in this early study is thus only one of a number of statistical and neural network options.

3. Modeling the perception and production of infant vocalizations

The SOM is a neural network inspired in large part by biological considerations, namely the self-organizing topographic nature of its feature representations and unsupervised learning in response to stimulus exposure [69, 70, 39]. Although the present study focuses solely on acoustic analysis and classification applications, this work provides a potential foundation for future modeling of the perception of infant vocalizations by humans, including learning through exposure to such vocalizations.

Caregivers are commonly infants primary communication partners, responding and providing feedback to infants. Furthermore, much of the current research on infant vocal development relies critically on naturalistic judgments by laboratory personnel. It is therefore important to understand how adults perceive infant vocalizations and to understand what acoustic features are relevant to adult communication partners. There are several ways in which the ability of the SOM to model adult humans perceptions of infant utterances might be assessed. One way would be to have human participants perform tasks directly matched to those the SOM-perceptron hybrid performed. Another possibility would be to compare the topography of features on the SOM to listeners similarity judgments.

VI. Acknowledgments

This study was supported by a Department of Energy Computational Science Graduate Fellowship (A. S. Warlaumont), by NIH Grant R01 DC006099-04 (D. K. Oller, PI, and E. H. Buder, Co-PI), and by the Plough Foundation. The penultimate author was supported by NSF grants BCS-0720322 and BCS-0826825 (Dale, PI). The authors wish to thank Lesya Chorna, Kyounghwa Kwon, Elisabeth Callihan, Courtney Jacks, the University of Memphis Infant Vocalization Lab analysts, and the infant participants and their caregivers.

Chapter 3

Prespeech motor learning in a neural network using reinforcement

1. Preface

This chapter was submitted for journal publication in January 2012. It is currently under review. Its authors are Anne S. Warlaumont, Gert Westermann, Eugene H. Buder, and D. Kimbrough Oller. It is formatted for submission to *Neural Networks*.

1.1. Human infant vocal development

During the first year of life, human infants make considerable progress in learning to produce speech-like sounds. One of the first achievements in prelinguistic vocal development is acquiring the ability to control phonation, producing voiced sounds at will [2]. Basic modal phonation is so readily produced by a healthy adult that its complexities may easily be overlooked. In fact, phonation involves active settings of a number of muscles that contribute to the positions, compressions, stresses in the tissues of the larynx [75]. To further complicate things, it has recently become clear that the larynx and the upper vocal tract interact nonlinearly [76]. How infants learn to control this system in order to support phonation is an open question.

Soon a number of other milestones are achieved, such as expansion of the range of pitches, durations, and vocal qualities, and the emergence of syllabic consonant-vowel timing [2, 4, 7, 51]. Toward the end of the first year of life, infant vocalizations have been reported to begin to show adaptation to the phonetic characteristics of their particular language environment as opposed to those of other languages [77, 78]. For example, a study by de Boysson-Bardies *et al.* [78] of 10-month-old infants from monolingual French, English, Cantonese, and Arabic speaking households compared the vowel sounds produced during canonical babbling by each infant to the vowels and their frequencies in adult speech in the household language. The study found that mean first and second formant frequencies of vowels produced by infants were significantly different across language backgrounds and that the patterns of differences matched those estimated for adult speech for the four languages. The results were taken as evidence that a child's language environment influences the range of movements of the infant's articulators, particularly the tongue and lips, supporting the development of the vowel system of the target language.

1.2. Reinforcement in early vocal development

The human infant develops within a social environment of interaction with parents and other adults and children. For this reason, the developing social brain has recently become a focus in infancy research (e.g., [79, 80]). Speech production development is one of the many behaviors that develops in the context of and is shaped by social interaction. Caregivers direct vocalizations (such as acknowledgments, imitations, playful vocalizations, and object labels) toward their infants as well as smiling at, looking at, and touching their infants. These caregiver behaviors, particularly the vocal ones, are modulated in response to the infants' vocalization behaviors [81, 82] and they serve as reinforcers to the infant: experimental manipulation of caregiver responses has been shown to selectively increase infant production of specific sound characteristics. For example, when infants are reinforced for producing fully resonant vowels, the infants later produce more fully resonant vowels; when they are reinforced for producing well-timed consonant-vowel syllable transitions, infants later produce more well-timed syllables [65].

In addition to social sources, reinforcement may also come from internal sources. For example, the high auditory salience of a self-produced sound or its matching to the infant's auditory preferences may function as reinforcers. It is likely that auditory salience and preference are influenced both by innate factors and by exposure to ambient language input. Salience-based reinforcement-learning, though it has not been addressed in research on development of vocalization abilities, has been shown to be feasible in a computational model of the development of eye movements for joint attention [83]. Whether it originates from social sources or from internal preferences, the idea is that positive reinforcement for producing speech-like vocalizations facilitates the development and increased usage of the reinforced vocalizations, consistent with the principles of operant conditioning [84].

Functionally, positive reinforcement provides an agent with feedback that its vocalization was on the right track, without directly indicating what the motoric target is. It is useful to compare reinforcement-based learning to two other types of learning, unsupervised self-organization (e.g., in learning by Kohonen maps and Hebbian networks) and supervised learning (e.g., utilized by feed-forward and simple recurrent networks that learn via the delta rule and backpropagation). On the one hand learning from reinforcement does, unlike unsupervised learning, rely on the model's receiving feedback about how well it performed. However, this feedback is not as targeted as in supervised learning in that the exact desired modeled behaviors are not assumed to be known by the entity providing the feedback.

Reinforcement-based learning is suitable for situations where the optimal behavioral or motoric output is unknown, as in the case of a modeler or roboticist who wishes to make a realistic synthesizer produce certain types of sounds. Infants as well may not have direct access to the correct motor configurations for producing target vocalizations, so reinforcement from caregivers or the infants' own learned or innate auditory preferences may serve as useful guides in the infants' learning to produce vocalizations of a given type.

1.3. Previous vocal development models

Additional mechanisms likely also play important roles in learning to produce speech-like sounds. One proposal is that adaptations of infant vocalizations to the ambient language result from self-organized perceptual and perceptual-motor learning. For example, it has been argued that by monitoring their own vocalizations, infants learn sensorimotor mappings that enable them to reproduce sounds heard from others [85, 29]. Most computational neural network modeling work to date has focused on this mechanism

and not on reinforcement (note that the two are not mutually exclusive)
[86, 87, 88, 89, 90, 91].

The DIVA (Directions Into Velocities of Articulators) model [92, 86] focuses on self-organizing synaptic mappings between primary auditory, higher-level auditory, somatosensory, and motor brain regions. The DIVA model is assumed to have knowledge about which specific vowels and consonants exist in its language and their acoustic properties (for example, the first three formant frequencies). During a “babbling” phase, the model randomly moves its articulators, i.e., its tongue, jaw, and lips. The model learns by updating the synaptic mappings between its motor and sensory cortices to reflect the associations between articulatory motor commands and their somatosensory and auditory consequences discovered during the babbling experience. When the model’s random movements happen to produce a synthesized sound that corresponds acoustically to a sound in its language, the synaptic mappings from a premotor speech sound layer to motor cortex and to sensory cortices are also updated. The effect is that future activation of the speech sound simultaneously activates the appropriate motor commands and inhibits the appropriate auditory and somatosensory expectations. The inhibition of auditory and somatosensory regions enables the model to detect if there is any error in its production of the sound and if so, to make appropriate motor corrections.

The DIVA model is the most comprehensive and well-tested model of human speech sound learning to date. It has been compared to adult fMRI data and has been used to model normal adult performance under various experimental manipulations, differences in hearing impairment and stuttering, and robustness in the face of developmental changes across childhood in the size and shape of the vocal tract [86, 93, 94, 95]. However, there are a number of aspects of early vocal learning that it has not yet addressed. For one, it has not yet been used to model self-initiated behavior; instead speech sounds are activated directly by the modeler [86]. Relatedly, it does not address the role that reinforcement might play in shaping spontaneous vocal behavior. Finally, it does not address phonatory learning, i.e. learning to produce voicing and learning to control the pitch, amplitude, etc. of vocalizations, despite this being a major aspect of early speech development.

Several other models, narrower in scope than the DIVA model, aim to explain how infants might learn to imitate vocalizations produced by others via Hebbian learning of perceptual-motor connections [87, 89, 90]. These models each consist of two layers of neurons, one auditory and one motor with weighted connections between the two layers. As in the DIVA model, learning in these models involves having the model produce random motor outputs, determining vocal tract configurations, which in turn determine the acoustics of synthesized vocalizations. In Yoshikawa *et al.* [87], each model production is then imitated by a human adult, and sensorimotor connections are updated in a Hebbian fashion so as to link the acoustics of the adult imitation to the motor outputs of the model. After training, adult vowels can be input and the model produces correct vowel imitations. In Heintz et al. [89] and Warlaumont *et al.* [90], learning from model productions is based on Hebbian associative learning between the acoustics of the model's own vocalization and its motor outputs. In addition to learning based on self-production, these models include passive listening events, in which the model receives external auditory input, as if from a caregiver, and the model self-organizes its perceptual receptive fields and/or its Hebbian perceptual-motor connections as a result. However, in these models, the utility of such passive learning from adult input for improving imitation accuracy has not been established, although in a similar model by Westermann and Miranda [88] it has been shown that such adult input does produce ambient language effects on perceptual representations. Presumably even if passive perceptual input does not produce improvements in imitation accuracy, it is possible that were the post-learning spontaneous vocalizations of these models to be explored, ambient language effects of the sort shown in the literature on human infants might be observed. This possibility has not yet been examined.

Kanda et al. [96] have also addressed learning to produce the vowels of a given language. The model is a recurrent neural network with parametric bias (RNNPB). In a first phase of learning, inputs are sequences of adult vowels. The model is trained to predict, on the basis of the acoustics and corresponding motor parameters at the current and previous time steps, the acoustics and corresponding motor parameters that will be input at the next time step. After this first phase of training, the model is able to segment sequences of vowels based on where prediction errors are highest. In a second phase of learning, the model learns to represent segmented vowels as constant values of two ”parametric bias” neurons. After this second phase of learning, the parametric bias neurons can be activated by the modeler and the network accurately produces the correct vowels. Although the model performs well on segmentation, recognition, and production tasks, its plausibility is questionable. It is assumed that during training the model knows, for each adult vowel, both its acoustic parameters and the precise articulatory motor parameters that generated the vowel. Such an assumption is consistent with Liberman and Mattingly’s motor theory of speech perception [97], which posits that from birth, infants’ perception of speech sounds is innately linked with the articulatory gestures that produce those speech sounds. However, whether infants innately, without any prior learning, possess direct access to the precise motor commands they would need in order to produce a sound that they hear someone else in their environment has produced is a strong assumption, especially given the fact that infants do not at birth or even within the first few months of life produce vocalizations that sound like speech, except perhaps accidentally [2].

Other work by Oudeyer [91] is unique in that it does explicitly address ambient language effects on spontaneous vocalizations. The model consists of multiple agents, each with a layer of auditory neurons connected to a layer of motor neurons that in turn connect to three articulatory parameters: lip rounding, tongue height, and tongue position. At each iteration, an agent is randomly chosen and its motor neurons are randomly activated. The agent adjusts, in a self-organizing manner, its neuro-articulator weights as well as the connection weights between the two layers. The geographically closest neighbor hears the first agent's vocalization, has activation propagated from its auditory to its motor layer and then also updates its neuro-articulator weights. In this way, the second agent becomes more likely to spontaneously produce sounds similar to those of the first agent. The model provides an impressive demonstration of how self-organized learning and interaction among agents can affect clustering of the vowel space as well as adaptation of vocal productions to others in the environment. However, by design it does not include modeling of either social or intrinsic reinforcement effects. Also, like the other models, it does not address phonatory learning.

Thus, despite the insights obtained from previous work, many aspects of early vocal motor learning in human infancy remain to be modeled. For one, reinforcement has not been incorporated, despite its important role in the empirical human infancy literature. Second, the focus has been on responses to caregiver vocalizations or production of given sequences of phones and has rarely (an exception being [91]) addressed spontaneous productions. Third, previous work has focused heavily on learning vowel sounds and has not addressed development of control over phonation, which is also an important aspect of speech production. In the present study, we introduce a new neural network architecture that addresses each of these three aspects of early vocal motor learning.

1.4. Our model

Our model consists of a topographically organized layer of neurons that control a physiologically realistic vocalization synthesizer [98, 99] via neuromotor connections. During learning, the model explores its vocalization capabilities. If and only if it produces a vocalization that is reinforced, its neuromotor connections are updated to reflect its current neuronal and muscle activations. This dependence of learning on reinforcement is consistent with neurophysiological findings that learning in motor cortex is modulated by dopamine, a neurotransmitter strongly associated with reinforcement [100]. Updating of neuromuscular weights follows the learning procedure for the self-organizing map [101], a popular type of neural network consisting of a layer of neurons with topographically-organized receptive fields that adapt to the environment. The topographic organization corresponds to the topographic organization observed throughout the brain.

The combination of self-organizing topographic map learning and reinforcement gating represents a novel neural network modeling approach. Note that the approach has a different emphasis from most computational reinforcement learning work such as those focusing on temporal difference learning and related methods [102]. For example, we do not consider reinforcement that is significantly delayed. Another difference is that in our model the primary function of reinforcement is to gate the learning of neuromotor connections. While reinforcement learning systems have been developed that use neural networks for processing sensory inputs, only a few have incorporated neural networks for producing behavioral outputs [103, 104]. Those that have have shown promising results, but have not to our knowledge used the self-organizing map network or addressed problems of learning to produce complex output patterns.

2. Method

2.1. Vocalization synthesis and analysis

All of the simulations in the present study used Boersma's articulatory speech synthesizer, implemented in Praat, a free speech analysis and synthesis software [98, 99]. The synthesizer consists of a model of the human vocal tract, including the lungs, trachea, larynx, pharynx, oral cavity, and nasal cavity. The walls of the vocal tract are modeled as coupled mass-spring systems. The synthesizer includes several options for the number of masses used in modeling the vocal folds; for the present study, we used the default two-mass option. The synthesizer also offers three sizes of vocal tract: adult female, adult male, and child; we used the default adult female version, since our target vowel acoustic measurements came from a study of adult female speakers. Based on the volume of air in the lungs and the activation of laryngeal and upper vocal tract (i.e., pharynx, oral cavity, and nasal cavity) muscles, specified by the user, the synthesizer calculates the positions and mechanical parameters of the vocal tract walls and the air pressures at each section of the vocal tract over time. The fluctuating air pressure at the mouth determines the synthesized sound. An advantage of using this synthesizer over the synthesizers used in most previous models of infant vocal development is that it allows for motor control of the larynx to be modeled, which is necessary for phonatory development to be addressed.

For this study, all synthesized sounds lasted .5 s. Similar to the example given by Boersma [99], the Lungs parameter, which represents the speaker's lung volume, was set to .2 at time 0 s and to 0 at time .1 s (-.5 corresponds to maximum exhalation and 1.5 corresponds to maximum inhalation). The activations of twenty muscle parameters, listed in Table 3.1, varied across vocalization events according to the procedures described below. Within a vocalization event muscle activations were static, i.e. there was no intra-vocalization variation. How each muscle's activation for a given vocalization event was determined is described below in Section *2.3.*.

The synthesized sounds were analyzed automatically, also in Praat, to get estimated measures of fundamental frequency (f_0) and first and second formant frequencies (F1 and F2) at 250 ms after the start of vocalization synthesis. When Praat could not identify an f_0 at this time in the sound, which tends to happen for example when the synthesized sound is silent or breathy but lacking phonation, then the f_0 was considered undefined. Ellis's RASTAMAT toolbox [54] was used to convert frequencies from Hz to mel, as the nonlinear mel scale better reflects the frequency scaling of the human auditory system. f_0, F1, and F2 were the quantities that determined whether or not a given vocalization was reinforced, as described in Section *2.5.* below.

2.2. Neural network architecture

The neural network contained 25 neurons arranged on a 5×5 grid. Each neuron had a spatial location defined by (x, y) coordinates (see Fig. 3.1) and each neuron had modifiable connection weights to each of the twenty muscles. The connection weights

Table 3.1: The vocal tract synthesizer muscles controlled by the neural network. Laryngeal muscles are those mainly involved in phonation and articulatory muscles are those mainly involved in controlling the shape of the upper vocal tract.

Muscle number	Name	Grouping
1	Interarytenoid	Laryngeal
2	Cricothyroid	Laryngeal
3	Vocalis	Laryngeal
4	Thyroarytenoid	Laryngeal
5	Posterior Cricoarytenoid	Laryngeal
6	Lateral Cricoarytenoid	Laryngeal
7	Styloglossus	Articulatory
8	Masseter	Articulatory
9	Upper Tongue	Articulatory
10	Lower Tongue	Articulatory
11	Orbicularis Oris	Articulatory
12	Vertical Tongue	Articulatory
13	Transverse Tongue	Articulatory
14	Levator Palatini	Articulatory
15	Risorius	Articulatory
16	Genioglossus	Articulatory
17	Hyoglossus	Articulatory
18	Mylohyoid	Articulatory
19	Lateral Pterygoid	Articulatory
20	Buccinator	Articulatory

from each neuron to the set of all muscles determined a specific state of the synthesizer's vocal tract. In turn, each vocal tract state was associated with a synthesized vocalization for which f_0, F1, and F2 traces could be automatically estimated (although they could be undefined at the measured point in time).

2.3. Learning

Prior to learning, the neurons' connection weights to the vocal tract muscles were chosen from a uniform random distribution. Each simulation had 1,000 learning events, each of which corresponded to a discrete time step. A learning event began by randomly activating the motor neurons in an exploratory fashion. The extent of this random

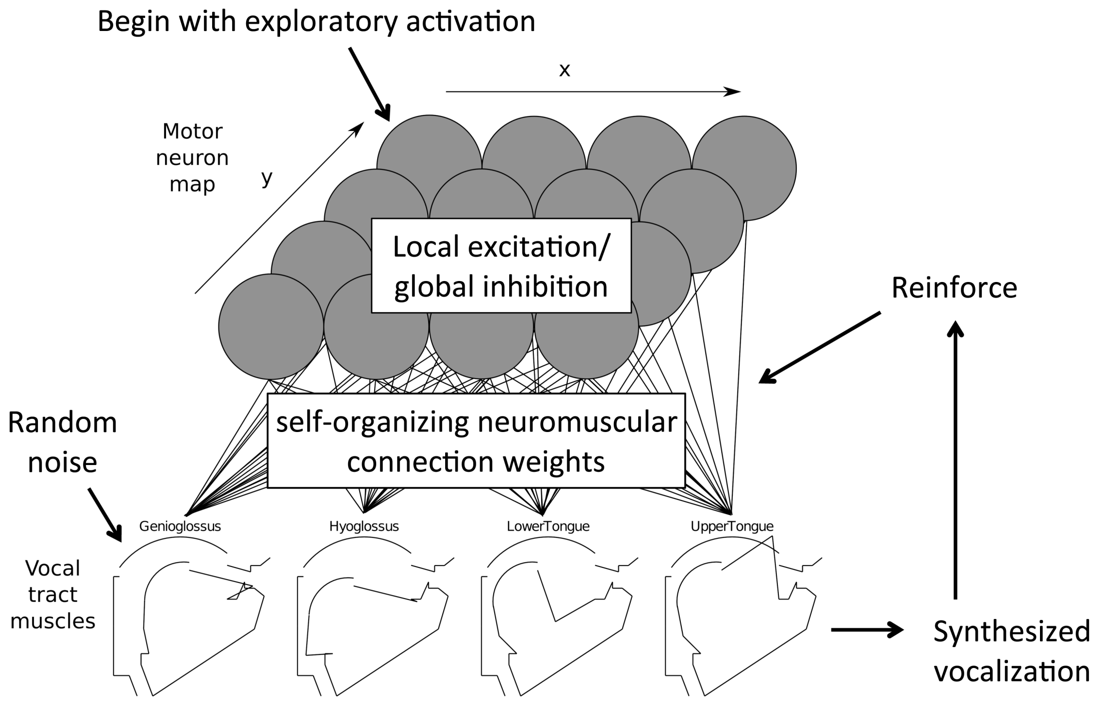

Fig. 3.1: Schematic diagram of the neural network model.

exploration depended on whether the previous vocalization event had been reinforced. If the model had not received reinforcement on the previous event, its activation was drawn from a uniform random distribution ranging from zero to one. Alternatively, if the model had indeed received reinforcement for its previous vocalization, instead of resetting the neuronal activations, a small amount of noise, ranging from -.25 to .25 was added to the previous learning event's neuron activations, subject to the constraint that the resulting activations had to remain between 0 and 1. Thus, if the previous vocalization had been reinforced, exploration was more precisely targeted. At this point, the most active node and its closest neighbors were identified and local excitation and lateral inhibition was effected. Each neuron had 2, 3, or 4 closest neighbors depending on whether it was located on a corner, on an edge, or on the interior of the motor neuron grid, respectively. All other neurons besides those five were inhibited by setting their activations to zero.

Activation was then propagated from the neurons to the muscles. Muscle activations were given by

$$\overline{m} = \frac{\overline{a}}{\sum \overline{a}} W + \overline{n}$$

where $\overline{m}$ is a row vector representing the activation level of each vocal tract muscle, $\overline{a}$ is a column vector representing the activation of each neuron, and W is a matrix giving the connection weights from each neuron (in rows) to each muscle (in columns). Thus, muscle activations were a function of the normalized neuron activations propagated through the weighted neuromuscular connections. The $\overline{n}$ is Gaussian noise added at the muscular level, intended to model incidental variation in the shape of the vocal tract. Such variation would correspond to changes in infants' vocal tract positioning due to feeding, mouthing of objects [105], or postural stabilization. Vocal-tract-level "noise" facilitated broad exploration by the model of its full range of vocal capabilities. As with the exploratory activation at the neuron level, noise at the muscular level was dependent on whether the previous vocalization had been reinforced. Muscular noise was more restricted if the model had previously been reinforced, having a standard deviation of 1 if the previous vocalization had not been rewarded and a standard deviation of .25 if it had. After the muscle activations for the current event were determined, a vocalization corresponding to those activations was synthesized, the vocalization's f_0, F1, and F2 were estimated, and based on these values it was determined whether the network would receive reinforcement for the vocalization event. The specific acoustic criteria for reinforcement are described in more detail in Section *2.5.*.

If no reinforcement was given, the event concluded without any changes to the network weights. However, if reinforcement was in fact given, the weights from the motor neurons to the vocal tract muscles were modified according to a self-organizing map algorithm [101],

$$W_{p,t+1} = \begin{cases} W_{p,t} + \alpha(\overline{m} - W_{p,t}) & \text{if } \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} \leq \theta \\ W_{p,t} & \text{otherwise,} \end{cases}$$

where $W_{p,t}$ gives the connection weights from neuron p to the vocal tract muscles at the time of the current event, (x, y) are the coordinates on the motor neuron map for a given neuron, q is the most active motor neuron, α is the learning rate and was set to .8 for the simulations presented here, and θ is the size of the learning neighborhood and was set to 1. In other words, the neuromotor connection weights were adjusted so that muscle activations similar to those just produced would be more likely to be produced on subsequent events. At this point, the learning event was complete.

2.4. Performance evaluation

At the beginning and end of each simulation, we tested the network to see what kinds of vocalizations it would spontaneously produce. Each simulated network was made to vocalize 25 times in the same manner as in training except that no reinforcement was ever provided and there was no noise added at the muscular level. The muscular-level noise was left out in order to provide a clear view on what the network learned at the neural level.

2.5. Reinforcement criteria

Six different reinforcement conditions were evaluated, with the goal being to compare the sounds produced and the neural representations developed across the different conditions. We ran 50 simulations for each reinforcement condition.

In the first condition, reinforcement was always given, no matter what the network produced. In the second condition, reinforcement was given if the sound produced by the model had a defined f_0 at time .25 s which had the effect of reinforcing voiced (i.e. phonated) but not unvoiced (e.g., silent or breath-only) sounds. Although the reinforcement criterion is quite simple, the act of phonation involves coordination of a number of muscles (see Table 3.1) in order to cause vibration in a nonlinear system of laryngeal tissues [21, 76].

In the third condition, in order to be reinforced the model's vocalization had to not only be phonated (operationalized as having defined f_0) but also had to be similar to one of thirteen American English vowels. Similarity to a vowel was operationalized as Euclidean distance in the two-dimensional space defined by F1-f_0 and F2-F1). The model had to become increasingly similar to one of the vowels, or else fall within a threshold degree of similarity, in order to be reinforced. The threshold degree of similarity was 3 mels (in other words, the target region around a vowel was a circle with a 3 mel radius). Throughout training, a record was kept, for each American English vowel, of the top ten model vocalizations that were closest to that American English vowel. The increasingly similar criterion for reinforcement was defined such that on a given trial, the model's production, if it did not fall within the 3-mel radius of an American English vowel, had to at least be closer to one of the American English vowels than one of the top ten previous model vocalizations.

The fourth condition was the same as the third except that ten Korean (instead of English) vowel targets were used. The American English and Korean vowel targets were taken from a prior study of vowels produced by adult female native speakers of the two languages [106].

The fifth reinforcement condition was the same as the third except that instead of all American English vowels being targeted, only the vowel /a/ was reinforced. The sixth and seventh conditions were the same as the fourth except that the individual target vowels were /e/ and /u/, respectively. Focusing on single vowel targets allowed us to see how well the model can learn to produce specific vowels, and to clearly visualize the effects of reinforcement.

We tested both phonatory and articulatory performance after learning. Greater post-learning tendency both to produce voiced sounds and to produce sounds resembling the target language would show generalizability of the reinforcement-gated self-organized learning approach. In particular, it would show that the model can, using a single learning mechanism, simultaneously learn two foundational speech skills, phonation and vowel articulation.

3. Results

3.1. Phonation before and after learning

As shown in Fig. 3.2, before any learning, across simulations the mean number of vocalizations that had identifiable f_0 was approximately 5 (out of a possible 25). When the model was reinforced at every trial, regardless of phonation, the mean number of vocalizations with identifiable f_0 after learning was only 2.28. For the various reinforcement conditions where reinforcement was contingent on phonation, the mean number of vocalizations with identifiable f_0 after learning ranged between 20.2 and 24.5. For each reinforcement condition, the difference between the number of sounds with f_0 before versus after training was highly significant, $p < .001$. This indicates that when reinforcement was contingent on voicing (i.e. phonation), the model learned to reliably produce sounds that were clearly voiced (not silent or purely breathy). When reinforcement was given all the time, without regard to voicing, the model's production of sounds that were voiced actually decreased after learning.

Fig. 3.3 illustrates, for one of the networks that was reinforced for any sound with identifiable f_0, the sounds that were produced before and after learning when each of the 25 neurons was activated in isolation. As can be seen in Fig. 3.3, most of the spectrograms of sounds produced by the neurons before training showed little acoustic energy and were

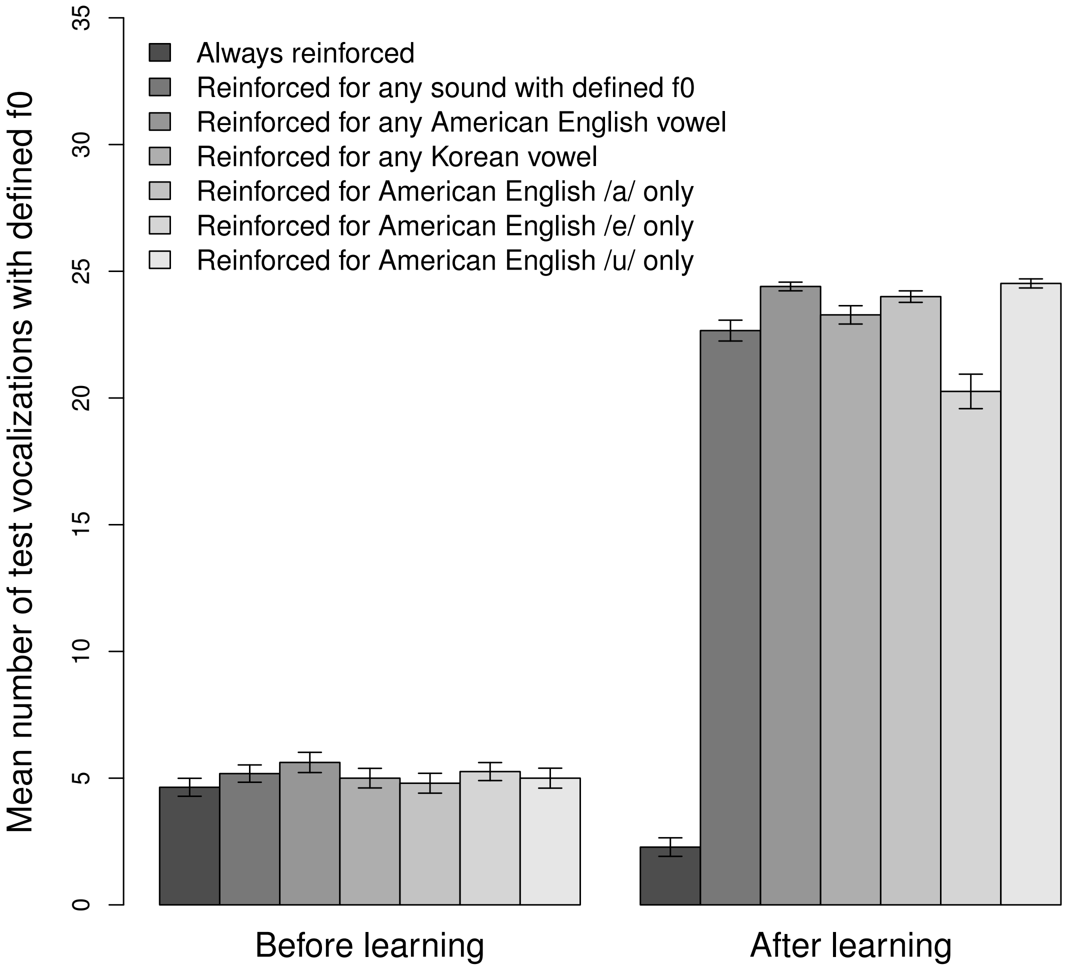

Fig. 3.2: Mean numbers of vocalizations with identifiable fundamental frequency before and after learning. Means are over the 50 simulations within a given reinforcement condition. Error bars indicate standard errors.

essentially silent. In contrast, after learning, almost all of the neurons produced sounds with high acoustic energy, indicating that the neuron has learned to produce audibly voiced sounds, that is, to phonate. It can also be seen that there was a range of duration, spectral quality, and amplitude: apparently, the simple requirement of defined f_0 at time

.25 left opportunity for the neural network to develop representations for motor control of sounds with a variety of different phonatory characteristics. Finally, note that after learning had taken place, the network exhibited topographic organization—neurons located near each other tended to produce sounds with similar-looking spectrograms.

Fig. 3.4 shows the laryngeal muscle activations responsible for producing the vocalizations spectrograms in Fig. 3.3. The figure shows consistencies with what is known about roles of the various laryngeal muscles in phonation. In particular, muscle number 4, the thyroarytenoid, a muscle that courses beside each vocal fold and promotes phonation by adducting the vocal folds (it also relaxes and shortens them), is highly activated, as would be expected. Additionally, muscle number 6, the lateral cricoarytenoid, shows greater activation than muscle number 5, the posterior cricoarytenoid; this corresponds to the fact that the lateral cricoarytenoid is a vocal fold adductor and therefore promotes phonation whereas the posterior cricarytenoid is a vocal fold abductor, inhibiting phonation.

3.2. Vowel types produced before and after learning

To investigate the types of vowels produced under the various vowel reinforcement conditions, we used the same set of test vocalizations as was used for the phonation evaluations. We compared the simulations in which any sound with identifiable f_0 was reinforced, in which vocalizations resembling any of the American English vowels were reinforced, and in which vocalizations resembling any of the Korean vowels were reinforced. The dependent variables were the number of sounds falling within 3 mel of the American English vowels and the number of sounds falling within 3 mel of the Korean vowels.

As can be seen in in Fig. 3.5, all of the networks produced fewer vowels resembling the vowel targets before learning than after learning. This pattern may be driven in part by the fact that before learning all networks produced fewer sounds with

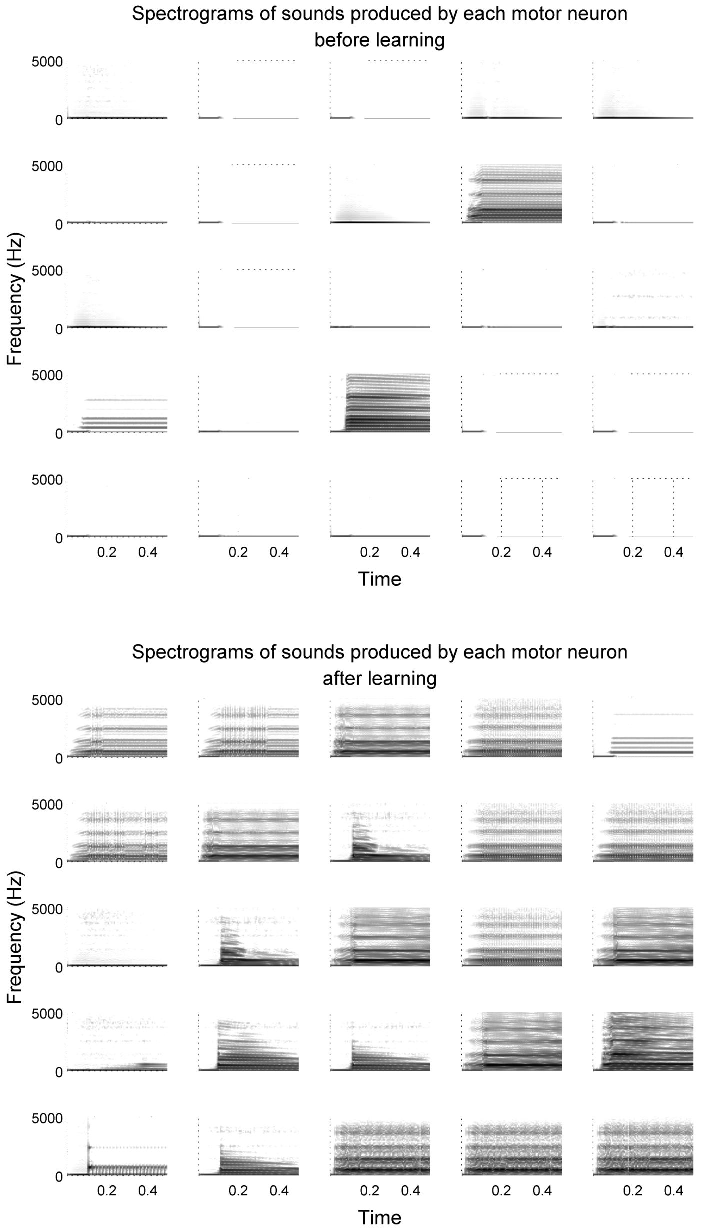

Fig. 3.3: Spectrograms of sounds produced by individually activating each neuron in one of the simulations from the second reinforcement condition, in which any sound with identifiable f_0 was reinforced. Before learning (pictured at top), three neurons' productions were judged as being voiced: these are located at row 2, column 4; row 4, column 1; and row 4, column 3. After learning (pictured at bottom), all neurons' productions were judged as being voiced except for one, located at row 3, column 1.

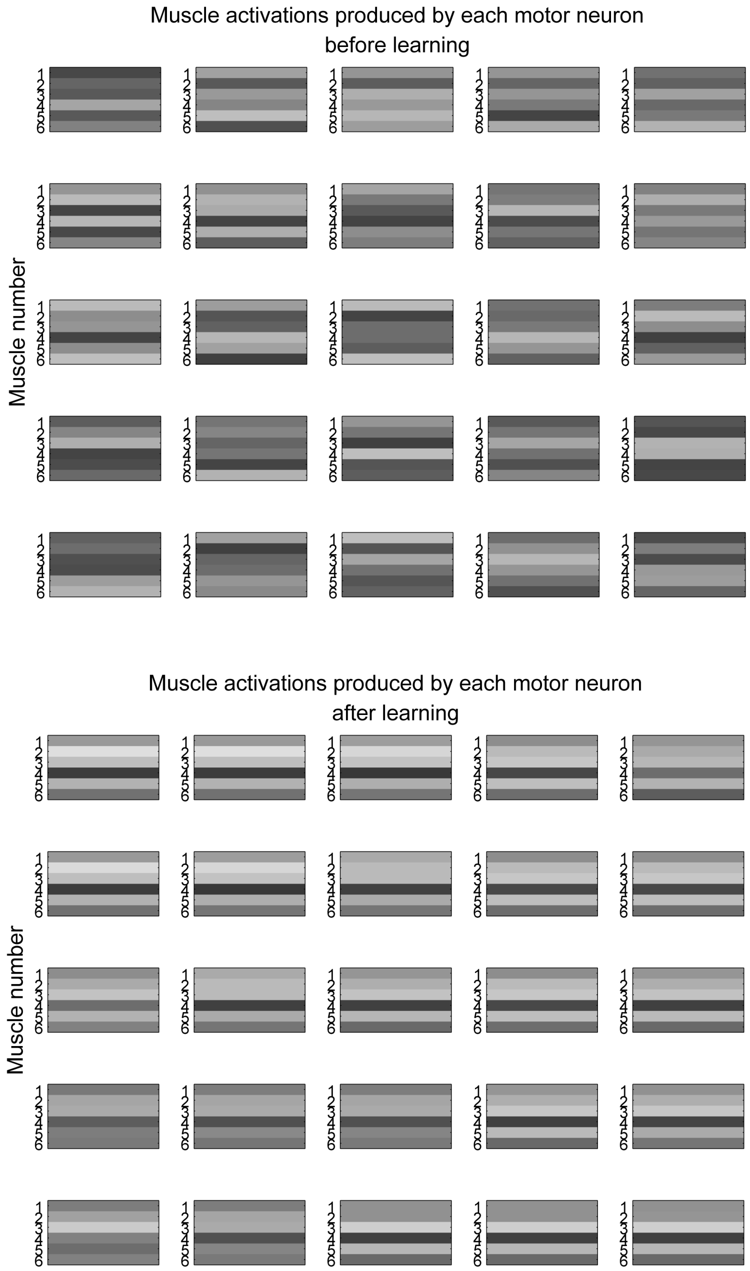

Fig. 3.4: Connection weights from each neuron in one of the simulations from the second reinforcement condition (the same simulation as in Fig. 3.3) to each of the laryngeal vocal tract muscles (see Table 3.1). Darker colors indicate higher weights and thus higher muscle activations.

defined f_0, and if a sound did not have defined f_0, it was automatically considered not similar to any of the target vowels.

After learning, the American-English-reinforced model produced the most sounds falling within the 3 mel target range of the American English vowels. A mixed-model regression with vowel and simulation as random effects, reinforcement for American English versus Korean as a fixed effect, and number of vowels resembling American English targets as the dependent variable showed that the difference between reinforcement conditions in the number of American-English-like productions after learning was statistically significant, $\beta = .23$, $p < .001$. While the number of vowels falling within 3 mel of the Korean targets was overall lower for all reinforcement conditions, which would be expected since there were thirteen English vowels but only ten Korean vowels, the Korean-reinforced model was the best-performing. A mixed model regression with number or vowels resembling the Korean targets as the dependent variable revealed the effect of the reinforced language to again be statistically significant, $\beta = .28$, $p < .001$. Note that β, the standardized regression coefficient, is comparable across the two target languages, indicating that the effect of reinforced language was similar in maginitude for both. Fig. 3.6 shows the relative formants of the productions from each version of the model after training. In sum, the model learned to produce more of the vowels from the language for which it was reinforced.

For a closer look at the model's learning of specific vowels, we compared the simulations in which only the American English /a/ was reinforced, in which only the American English /e/ was reinforced, and in which only the American English /u/ was reinforced. The dependent variables were the number of sounds falling within 3 mel of /a/, /e/, and /u/. As can be seen in Fig. 3.7, it was the simulations that were reinforced for /a/ that produced the most vowels resembling /a/ after learning. The differences between /a/-reinforcement and /e/-reinforcement and between /a/-reinforcement and /u/-reinforcement were both statistically significant with $p < .001$ in both cases. Similarly,

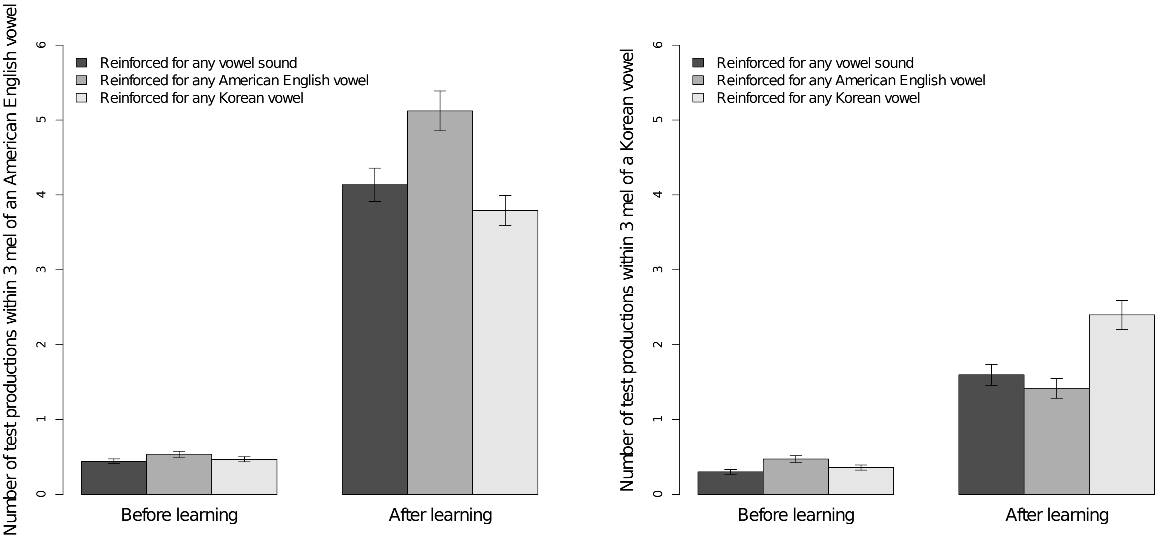

Fig. 3.5: Mean numbers of vowels within 3 mel of American English (left) and Korean (right) vowels for models in three different reinforcement conditions before and after learning. Means are over the 50 simulations within a given reinforcement condition. Error bars indicate standard errors.

the simulations reinforced for /u/ produced more vowels resembling /u/ than the simulations reinforced for /a/ and /e/, $p < .001$ in both cases. The simulations reinforced for /e/ produced more vowels resembling /e/ than the simulations reinforced for /a/, $p = .02$, and marginally more than the simulations reinforced for /u/, $p = .10$. Overall, fewer productions were close to the /e/ target than were close to the /a/ or /u/ targets. Fig. 3.8 shows the relative formants of each model's productions after training. The plots confirm that while the model readily learned to produce precise /a/'s and /u/'s, it had more difficulty learning to produce /e/ as evidenced by the broad distribution of vocalizations produced in simulations where /e/ was the target vowel.

4. Discussion

We have presented a new neural network model wherein exploration and reinforcement are integrated with topographic self-organized learning. A layer of neurons is connected to the muscle inputs of a realistic human vocal tract synthesizer. The model

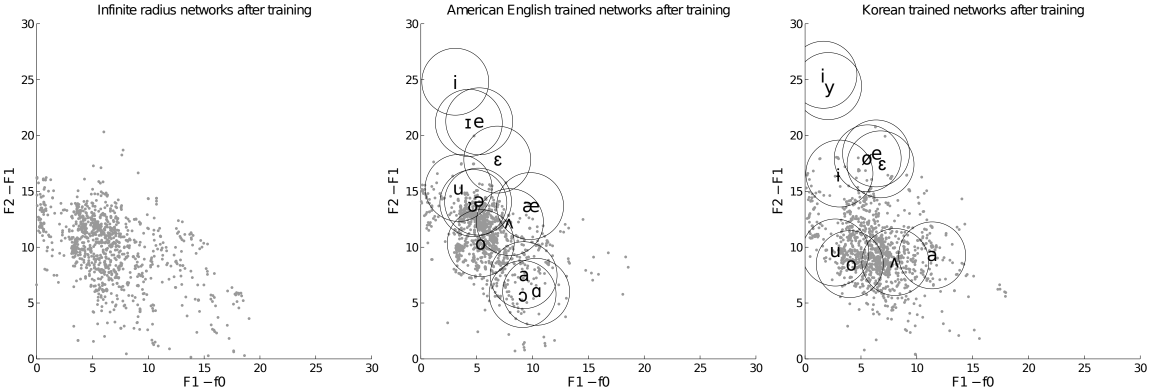

Fig. 3.6: Relative vowel formants of the vocalizations produced by individually activating each motor neuron from all the simulations in each of three different reinforcement conditions. Left: reinforced for any sound with defined f_0. Middle: reinforced for any American English vowel. Right: reinforced for any Korean vowel. Each gray dot represents one neuron's vocalization. Vocalizations from the 50 simulations in the same condition are superimposed. For each reinforcement condition, the targets of training are shown in black characters with circles delineating the 3 mel radius around each target.

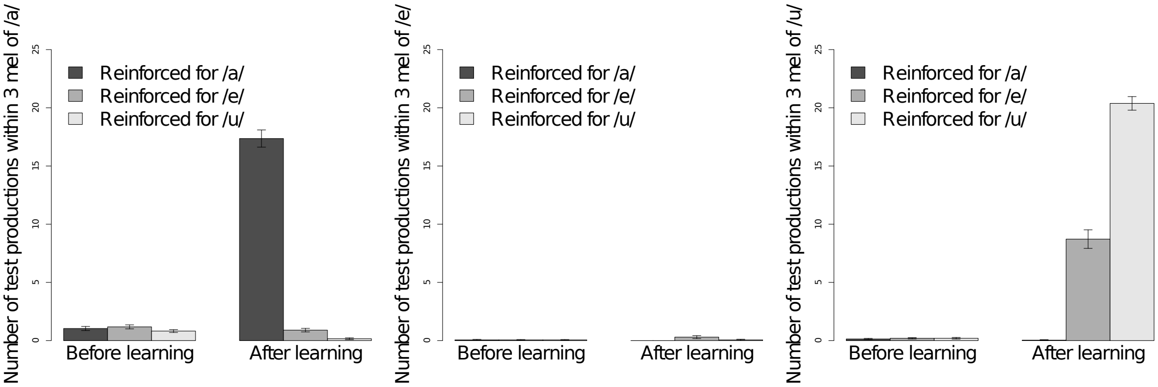

Fig. 3.7: Mean numbers of vowels within 3 mel of /a/, /e/, and /u/ for models trained on /a/, /e/, and /u/, before and after learning. Error bars indicate standard errors.

explores its vocalization abilities by randomly activating neurons, with some noise added at the muscular level. When it receives reinforcement for a vocalization, it updates its neuromuscular connection weights so that similar motor commands become more likely

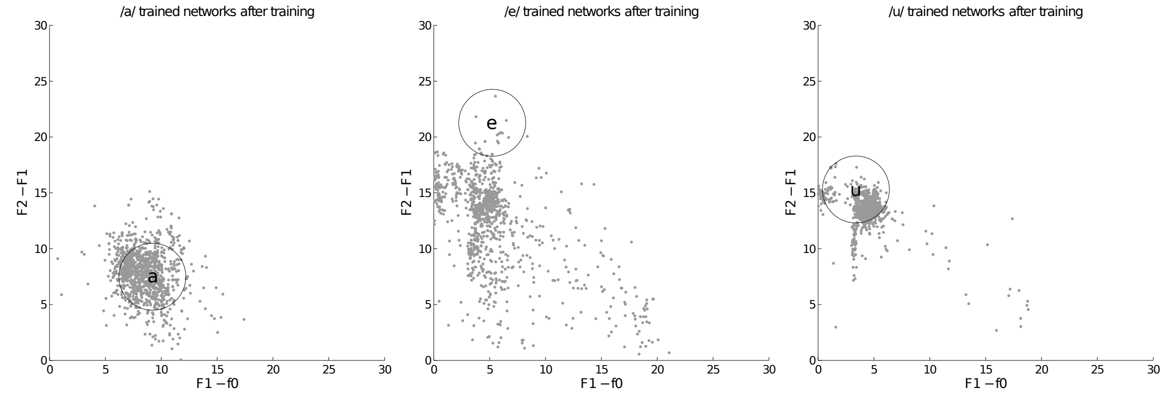

Fig. 3.8: Relative vowel formants produced when activating neurons in isolation from networks in trained on /a/, /e/, or /u/. Neurons from different simulations in the same reinforcement condition are superimposed. For each condition, the targets of training are shown in black characters with circles delineating the 3 mel radius around each target.

to be produced in the future. We show that the model can learn at least two foundational speech-related skills: production of phonated sounds and production of specific vowel types.

One of the contributions of this work is that it specifies at a mechanistic level how reinforcement, which is known to play a role in speech development ([81, 65, 107], may be used by human infants as they develop sounds with speech-like characteristics. For example, it has been shown that when caregivers' reinforcement is contingent on infants' production of resonant phonation, the frequency of resonant phonation increases [65]. In our model, reinforcement that is contingent on phonation (as measured by the sound having an identifiable f_0) signals the model to modify the connection weights from its motor neurons to its vocal tract muscles so that future neuronal activity will be more likely to result in phonated sounds. As discussed in Results, the laryngeal muscle activation levels produced after learning correspond to what would be expected based on previous physiological studies of speech production. Note, however, that our model is agnostic regarding the source of reinforcement. Reinforcement could come directly from social sources, such as a mother vocalizing toward her infant. It is also possible for an infant to be reinforced intrinsically, for example by producing an appealing sound, where the appeal is based on auditory salience or similarity to sounds that the infant has previously heard other individuals produce. In future work, it would be worthwhile to model these distinct possible sources of reinforcement in more detail. Regardless of whether the reinforcement is social or intrinsic, ours appears to be the first to address the role of reinforcement in infant vocalization development and also appears to be the first to model the development of phonatory control.

In addition to learning to phonate, the model also develops a propensity toward producing vowels like those for which it has been reinforced, whether that be the whole set of American English or Korean vowels or a single isolated vowel. A process of reinforcement-gated learning may be one of the mechanisms underlying babbling drift findings, i.e. shifting of vowels toward those that are most frequent in the infant's language environment [78]. Previous neural network models of speech production learning have all depended critically on learning sensorimotor correspondences in order to achieve ambient-language effects [87, 92, 89, 90, 96, 88] (note that none of the studies report data on spontaneous vocal productions, although those that include learning of connections between motor neurons and the vocal tract would be expected to exhibit ambient language effects on spontaneous productions). Our model, in contrast, requires no learning of sensorimotor correspondences, relying instead on reinforcement-gated learning of neuromotor connections and therefore illuminating an additional pathway through which the ambient language environment may shape spontaneous productions.

The model appears to exhibit not only learning effects but also biases with regard to the sounds the realistic vocal tract simulator can learn to reliably produce. In particular, we observed that learning /e/ proved more challenging to the model than learning /a/ or /u/. On the one hand, as discussed by Oudeyer [91], physiological vocal tract constraints also play a role in his model of speech sound learning and evolution and presumably play a role in the human system as well. In support of this, it is observed that /e/, /i/, and /u/ are less frequent in infants' vocalizations than /a/ [78, 108]. Thus, the model's weakness on /e/ and /i/ corresponds to relatively low production frequencies by human infants. However, our model's strong performance on /u/ does not correspond to the pattern from human data. Furthermore, since the synthesizer used in the present study models an adult female vocal tract and the acoustic vowel targets are based on average adult female productions from the literature, the particular pattern of difficulty on mid and high front vowels such as /e/ in the present study is surprising. We suspect the difficulty with /i/ and /e/ reflects issues with our acoustic measure for evaluating vowel similarity. First and second formant frequencies, despite being the most popular metric for quantifying vowel acoustics, do not completely account for listener perceptions of vowel type [109, 110]. Future research should explore other acoustic correlates of vowel productions as well as human listener judgments to see if better results on /i/ and /e/ can be obtained.

Previous studies involving neural network models of infant vocal development have not reported quantitative results regarding ambient-language effects on spontaneous vocal productions and have not addressed the development of phonation. In the future, doing so would permit direct comparison of our results to those of the previous models discussed in the Introduction. Additionally, more detailed comparison of the behavior of this and other models to the behavior of human infants and their caregivers will be helpful in further developing the work. Increased efforts to tie neural network modeling directly to neurophysiological findings, to anatomical changes across the lifespan, and to patterns of difference observed in clinically relevant groups, such as those with hearing impairment or those with autism, would also be expected to improve the models and therefore increase their scientific and clinical value.

Our mechanism and those of previous models are not mutually exclusive. Reinforcement-gated motor learning, perceptual learning, and sensorimotor associative learning are likely all involved in infant vocal development. The various mechanisms also likely interact with each other. For instance, changes in perceptual representations as a result of exposure to sounds from an ambient language may affect how the infant perceives sounds to be salient or otherwise intrinsically rewarding. In the future, a more comprehensive model of vocalization development that combines these various mechanisms should be developed and evaluated. Additionally, all existing models of vocal development must be extended in the future to problems of the development of fine-grained dynamic sequences, such as those required for the precise syllable timing that also emerges in the first year of life and is a critical pre-speech skill. Finally, It is worth exploring the possibility that the same principles exemplified by our model may generalize to domains such as in the development of gestures and reaching skills.

5. Conclusions

We have presented the first neural network model to address the role of reinforcement in human vocalization development. It introduces a new approach that combines self-organization with selective reinforcement. The model exhibits several general characteristics of human infant vocal development, including sensitivity of vocal productions to reinforcement, development of phonatory skill, and development of a tendency of vowel production acoustics to be more consistent with the vowels in the ambient language than with vowels from other languages. These successes warrant the further development of the model and others that address the role of reinforcement in vocal motor learning.

6. Acknowlegements

This work was supported by a U.S. Dept. of Energy Computational Science Graduate Fellowship (DE-FG02-97ER25308), by the Plough Foundation, and by NIH DC011027.

Chapter 4

A social feedback loop for speech development is diminished in autism

I. Preface

This paper is being prepared for journal submission. Its authors are Anne S. Warlaumont, Jeffrey A. Richards, Jill Gilkerson, and D. Kimbrough Oller. It is formatted for submission to *PNAS*. Note that the Methods section therefore appears after the Results section and there is a Supplementary Information (SI) section following the main text.

II. Introduction

We propose a social feedback loop supporting speech-language acquisition for both typically developing children and children with autism spectrum disorders. This feedback loop provides an explanation for how language development and social interaction, two areas where differences are seen in autism [111], may influence each other.

Development of the ability to produce vocalizations with speech-like acoustics is an essential component of early language learning [2]. It is known that children with autism produce vocalizations with atypical acoustics. For example, Sheinkopf et al. [112] have found that preverbal children with autism have different vocal quality (increased rates of squealing, growling, and yelling as opposed to normal speech-like phonation) than preverbal children with developmental delay but not autism. More recently, Oller et al. [113] examined acoustic characteristics of child vocalizations that were automatically measured from day-long recordings of children with autism, children with developmental delay but not autism, and typically developing children. They too found differences in the autism group's vocalization acoustics, specifically in parameters associated with syllable form and duration, spectral tilt, and pitch. These acoustic differences reliably discriminated the autism group from the typically developing and developmentally delayed group. Others have found differences in the prosody of children with autism compared to typically developing children [114, 115]. In addition to these differences in vocalization acoustics, children with autism are slower and less likely to produce words and sentences than typically developing children [116]. Such expressive language differences are already apparent by two years and increase with age.

Differences in social interactions are also central to autism spectrum disorders (ASDs). For instance, individuals with ASD show reduced initiation of interactions and atypical patterns of turn-taking [111, 115]. Children with autism have also been found to both initiate joint attention and respond to joint attention bids less often than typically developing and developmentally delayed children [117, 112, 116] and to initiate social interactions less frequently [112]. And in automated analyses of day-long audio recordings, Warren et al. [118] found that for preschool-age children with autism there were fewer child-adult conversational turns.

We propose that as children learn to produce speech, a positive social feedback loop is in place to help support the process (Fig. 4.1). When a child produces sounds that are relatively mature, understandable, or otherwise appealing to another individual in their presence, such as a parent, they are likely to receive immediate, positive responses from those individuals. Receiving such a response provides an opportunity for learning, signaling to the child that the sound they just produced was something worth exploring or repeating again in the future, thus encouraging the development of mature, understandable utterances. Individual interactions of this sort accumulate over time, promoting speech development over the course of days, months, and years. The social feedback loop proposal is in keeping with constructivist theories of cognitive development [119], in that the child’s behavior affects the environmental input they receive and in that any atypicalities in the feedback loop would be expected to magnify group differences in speech skill across time.

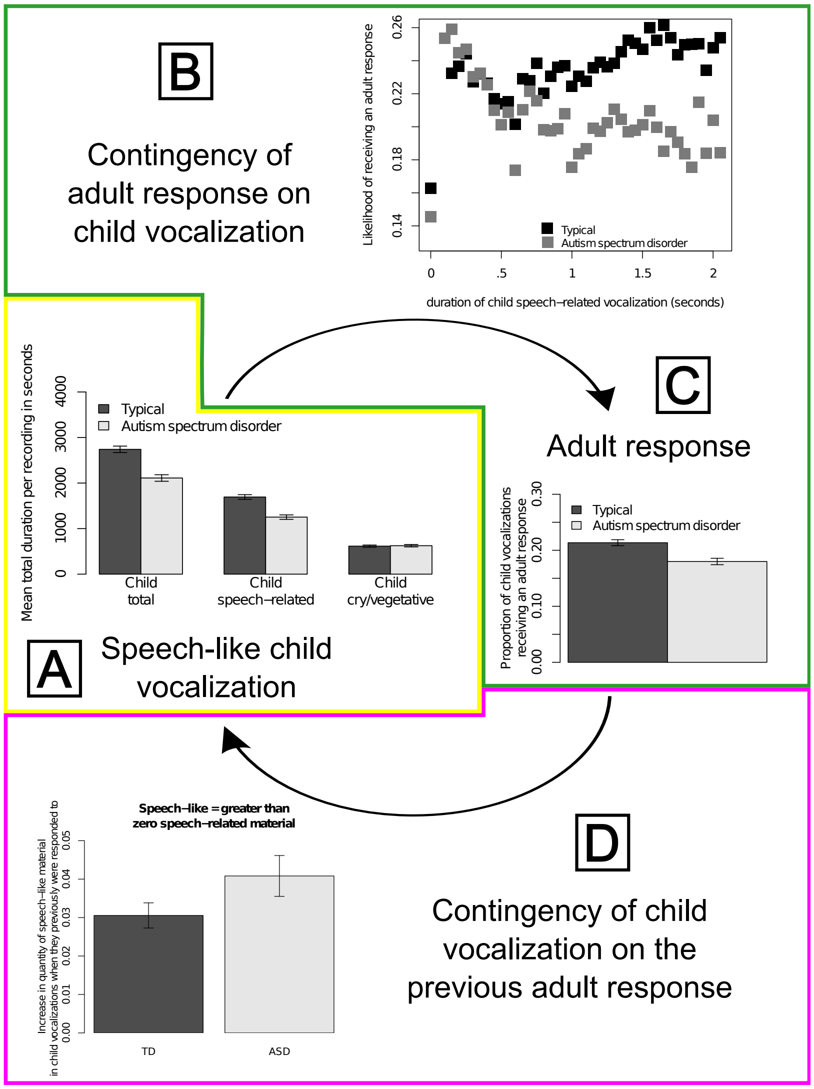
B
Contingency of adult response on child vocalization
Likelihood of receiving an adult response
0.14
0.18
0.22
0.26
0
.5
1
1.5
2
duration of child speech–related vocalization (seconds)
■ Typical
■ Autism spectrum disorder
C
Adult response
Proportion of child vocalizations receiving an adult response
0.00
0.10
0.20
0.30
■ Typical
□ Autism spectrum disorder
Mean total duration per recording in seconds
0
1000
2000
3000
4000
■ Typical
□ Autism spectrum disorder
Child total
Child speech–related
Child cry/vegetative
A
Speech-like child vocalization
Speech–like = greater than zero speech–related material
Increase in quantity of speech–like material in child vocalizations when they previously were responded to
0.00
0.01
0.02
0.03
0.04
0.05
TD
ASD
D
Contingency of child vocalization on the previous adult response

Fig. 4.1: The proposed social feedback loop, along with results supporting its existence in typical development and its existence at a diminished level in autism spectrum disorder. A: At the left side of the feedback loop in yellow are the child's speech-related vocalizations. The quantity of child vocalization, child speech-related vocalization, and child cry/vegetative vocalization are shown in the bar plot, separately for the TD and ASD groups. The ASD group produced less speech-related vocalization, which is reflected in their total vocalization quantity. B and C: When the child produces a mature, speech-related vocalization it is expected that this will often lead to an adult response, with the rate of adult responding contingent on the quantity or quality of speech-related material the child produced. The adult response component of the feedback loop is shown in green. Plot C illustrates the difference in rate of adult responding, regardless of the nature of the child vocalization, for the ASD vs. TD groups. Plot B illustrates the difference between diagnostic groups in the contingency of adult response on quantity of speech-related material in a child vocalization: In ASD, adult response is less contingent on the quantity of speech-related material within a child vocalization than in TD. D: Adult responses to child speech-related vocalizations would in turn be expected to increase the quantity and quality of child vocalizations, since the adult response, according to this feedback loop hypothesis, serves as a reward and an opportunity for the child to learn that vocalizations of the type they just produced elicits responses from other individuals. This dependency of a child's subsequent vocal behavior on adult responses to the child's prior behaviors is illustrated in magenta. The associated bar plot illustrates that in both groups, child quantity of speech-related vocalization is greater when the child's previous speech-related vocalization received an adult response. Error bars indicate standard errors of the mean across recordings. Note that the statistical effects of ASD were not based exactly on these error bars, but on mixed effects regressions that simultaneously took into account a variety of individual and demographic variables. Overall, while the social feedback loop is supported by the data for both groups, it is affected in two ways by autism: the quantity of child speech-related vocalization on which the feedback loop can operate is less and there is reduced contingency of adult response on a child's vocalization being speech-related.

Such a feedback loop is supported by experimental studies of typically-developing children finding that when caregivers' responses are contingent on certain characteristics of their infants' vocalizations, those responses affect the characteristics of the children's subsequent vocalizations [107, 65]. Additionally, the degree to which a mother is responsive to her child's communicative behaviors predicts language performance at a later age for both typically developing infants [120] and children with developmental disabilities [121, 122]. With regard to interaction more generally, greater coordination between infant and adult vocalization timing at 4 months of age predicts higher scores at 12 months on the Bayley MDI, a test that assesses language, cognitive, and perceptual abilities [123], and greater vocal interactivity during late infancy and early childhood is also associated with higher Bayley and Preschool Language Scale (PLS) scores [124].

Supporting the idea that a social feedback loop of this sort may also apply to children with autism spectrum disorders, it has also been shown that language development gains in young children with autism are predicted by caregivers' responsiveness, both verbal and nonverbal, to their children's activities and attentional focuses as observed during short play sessions [125]. In the same sample, it was also observed that children's tendencies to initiate and respond to adult bids for joint attention predicted rates of language development (see also [116]). The child and adult behaviors had independent contributions in the prediction of language gain, and both were more predictive than various demographic and individual performance variables, indicating that both children's and caregivers' social behaviors contribute substantially to language development. Similarly, it has been proposed that feedback to children with high functioning autism on their attempts to produce functional prosodic patterns facilitates the acquisition of distinctions between prosodic functions [114].

There are a number of ways in which the proposed feedback loop could be dampened for children with autism compared to children who are typically developing. First, motor or other impairments may lead to reduced ability to produce mature speech-like vocalizations [126], which would provide less material for other individuals to respond to [121]. A reduced number of contingent responses would then provide a child with autism with less information about what kinds of vocalizations are of high social value, further reducing the pace of speech vocalization development compared to that of a child without such impairments.

Second, it is possible that caregivers of a child with autism might exhibit reduced rates of responding to their children's vocalizations. When a child is not as communicative to begin with or is not as responsive to their caregivers' social responses, this could conceivably lead to changes in caregivers' overall responsiveness [121, 127]. Autism spectrum disorders may also be associated with additional demands on the caregiver's attention in general, which may reduce parents' ability to respond to their children's vocal behaviors. Reduced caregiver responsiveness could also be due to differences in the conversation-related behaviors of parents of children with autism [128], which can be expected due to sharing of autism-related genetic traits between parents and their children.

Third, Mundy and Neal [129] have proposed that children with autism's social orienting and joint attention deficits, originally arising from a variety of possible neural and psychological mechanisms and including orienting to speech sounds [130], cause them to have trouble extracting information from social situations. This in in turn impacts their cognitive, social, and language development, further slowing the children's development of social orienting and joint attention abilities. Reduced social orienting thus diminishes an important social feedback loop. This social feedback loop is somewhat different from the vocalization social feedback loop we propose here, though the two are not mutually exclusive and they may very well overlap. In our feedback loop, social orienting impairments [117, 129] could lead to impairments in processing caregivers' contingent social responses to the child's vocalizations, which could in turn stunt the development of mature, speech-like vocalizations.

In the present study, we harness the power of day-long naturalistic recordings and automated labeling and analysis in order to test the plausibility of our proposed speech feedback loop for vocalization development and its diminishment in autism. We first compare the rates of speech-related vocalizations produced by preschool-age child vocalization across typical and autism groups. Second we compare patterns of child-caregiver vocal interaction across the two groups. We then test the hypotheses that are specific to our feedback loop hypothesis by asking (1) whether the speech-likeness of a child vocalization predicts whether or not it will receive an immediate adult response and (2) whether such a contingent adult response predicts how speech-like the child's next vocalization will be. We test whether this holds in typical and autistic groups of children and also test for any differences in these contingencies across groups.

Although the focus of the present study is on differences between children with and without autism, we also test to see if there are differences across age, maternal education level, and gender. Such demographic factors likely also affect the functioning of the social feedback loop. For example, we expect that as maternal education increase, child speech-related vocalization rate and child-adult interactivity will increase, given that parents of higher socioeconomic status have been observed to on average spend more time talking with their children [131]. We also expect similar increases as age increases, given previous findings that child vocalization and conversational turn rates increase with age [131, 124]. Going beyond vocalization rate and general interactivity, we ask whether as SES and age increase there are differences across in contingency of adult response and in sensitivity of the child to adult response contingencies.

III. Results

A. Quantity and type of vocalizations

Fig. 4.2 gives an example of a series of labels for a small portion of one recording. Fig. 4.1A shows the quantities of child vocalization types in the day-long recordings. The automated labeling confirms, in line with previous research, that children with autism exhibit differences in the kinds of vocalizations they produce. In particular, children with autism produced less speech-related vocalization within the 12-hour recordings than children who were typically developing, $\beta = -.455, p < .001$. There was a statistically significant interaction between age and diagnostic status, $\beta = -.104, p = .024$ such that the rate of increase in quantity of speech-related vocalization as age increased was faster for the typically developing group than the group with ASD—the two groups tended to diverge with time. Quantity of adult vocalization did not differ significantly between the ASD and TD groups. The SI (VIII.) reports on the frequency of additional vocal types.

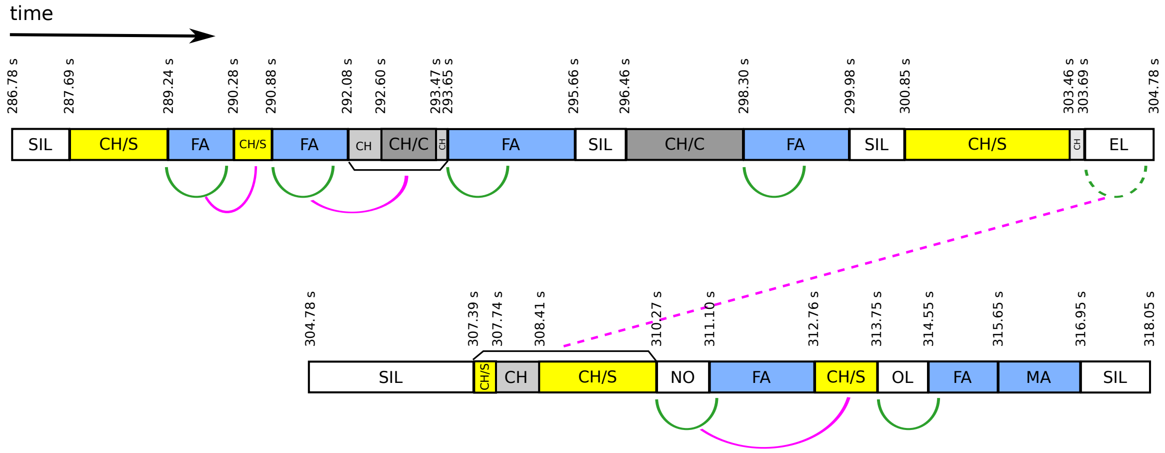

Fig. 4.2: Example, from a subsection of one recording, of sound source labels and how adult responses and child dependence on adult responses were measured. In yellow are the child speech-related vocalizations (CH/S), in dark gray are the child cry/vegetative vocalizations (CH/C), and in light gray are unspecified child vocalization segments (CH). In blue are female adult (FA) and male adult vocalization segments (MA). In white are silence (SIL), electronic sounds (EL), noise (NO), and overlap (OL). Solid green curves illustrate cases where an adult vocalization followed a child vocalization within one second and was thus considered a response. The dashed green curve illustrates a case where a child vocalization did not receive an adult response within one second. Solid magenta curves show the association between an adult response to a child vocalization containing speech-related material and the subsequent child vocalization whose duration of speech-related material is measured to determine the dependence of child speech-related vocalization on previous contingent adult responses. The dashed magenta line illustrates the association between lack of adult response to a child vocalization containing speech-related material and the following child vocalization. Supplementary section IX. provides the raw text data on which this example is based.

B. Interaction dynamics

In addition to differences in the kinds of vocalizations produced by children with autism, we also found differences in the dynamics of child-adult vocal interactions in the ASD group compared to the TD group. Interaction dynamics were investigated using two different approaches: cross-recurrence, for investigating global interaction patterns, and response frequency, for investigating local response patterns.

C. Cross recurrence

A diagonal cross recurrence profile (DCRP), was used to visualize the temporal relationship between all possible pairings (within a 30 s sliding window) of vocalizations from the child and vocalizations from the adult. Average DCRPs for the TD and ASD groups are shown in Figure 4.3. The left side of each DCRP plot indicates pairings where the adult vocalization preceded the child vocalization, i.e., when the adult was leading. The right side indicates pairings where the child vocalization led the adult. Moving from the center of the plot outward the lag between two paired vocalizations increases. The overall height of the DCRP gives a measure of the amount of child-adult interactivity. The ratio of the height of the right side of the DCRP to the height of the left side gives a measure of the extent to which the child tended to initiate as opposed to follow. The overall height of the DCRP was lower for the ASD group than the TD group, $\beta = -.218$, $p < .001$, indicating reduced levels of child-caregiver interaction in ASD

The ratio of the child leading side of the DCRP to the caregiver leading side was lower for the ASD group than the TD group, $\beta = -.386$, $p < .001$. Thus, ASD appears to be associated with reduced child leading relative to adult leading. Put another way, the children with ASD exhibiting a pattern of following the caregiver rather than leading, with the caregivers exhibiting a greater tendency to lead the interactions, to a greater extent than for the TD group.

D. Immediate responses

We also measured the proportion of child vocalization segments that received an immediate adult response, which we defined as one occurring within one second or less after the end of a child vocalization. This window was chosen to correspond to that used by Keller et al. [132], based on the fact that a 1 s window has been shown to be short enough for even young infants to detect temporal contingencies. When an adult response that occurs within this time window, a child should therefore be able to detect the relationship between the type of vocalization they themselves produced and the adult

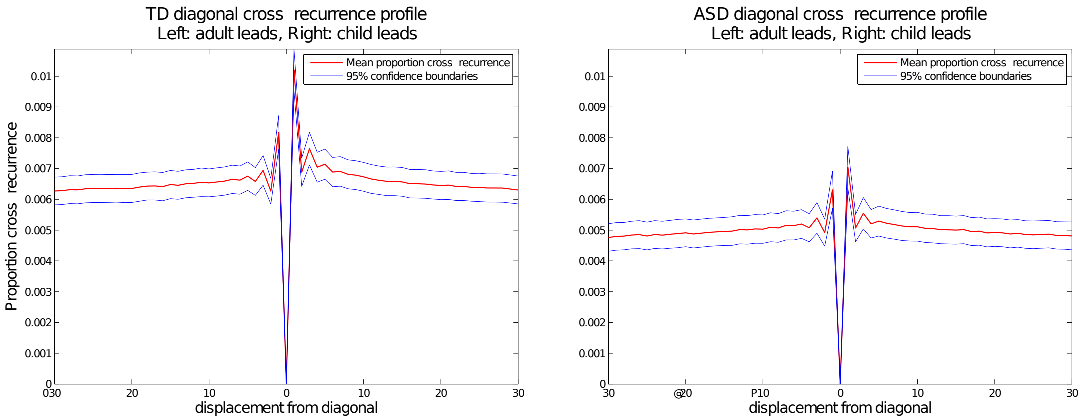

Fig. 4.3: Diagonal cross recurrence profiles, averaged across TD recordings (left) and ASD recordings (right). Red are the group means and blue are their 95% confidence intervals. Overall height of the plot indicates the overall level of child-adult interactivity. Displacement along the diagonal refers to the diagonal of the recurrence plots, not shown here, on which this profile is based (see section X. for further details). Displacement along the diagonal measures the difference between pairings of child and adult vocalizations in seconds. Here there is always a dip to zero cross-recurrence at lag 0 since adult and child speaker labels never overlap each other. The typically developing children's recordings show overall higher levels of interaction and a larger child leading to adult leading (right side over left side) ratio compared to the recordings for the children with autism.

response that followed. Note that the one-second interval also corresponds to peaks in interactive timing observed in the DCRPs discussed in the previous section. Figure 4.2 illustrates how whether a child vocalization received an adult response was determined. There was a significant effect of ASD such that children with ASD to receive proportionally fewer adult responses, $\beta = -.183$, $p = .002$, as shown in Fig. 4.1C. Similarly, proportions of adult speaker segments receiving child responses were lower for the ASD group than the TD group.

These immediate response results and the results from the cross recurrence analysis confirm and extend previous findings that vocal interaction dynamics differ substantially in autism. More specifically, children with autism are involved in less child-caregiver interaction at a range of time lags and including responses both of the adult to the child and of the child to the adult. Children with autism also tend have a lower ratio of leading to following compared to typically developing children. In the next section we explore the possible relationship between ASD-related differences in the types of vocalizations children produce and differences in child-adult interaction patterns.

E. Testing the social feedback hypothesis

Figure 4.2 illustrates how we tested to see whether the social feedback for speech development hypothesis was supported by the day-long recording data. For a given recording we first tested whether the speech-related utterance duration in a child vocalization segment predicted whether or not that child vocalization received a response (green arcs). We then tested whether the child having received an adult response for his/her most recent previous speech-like vocalization predicted whether his/her next vocalization would be speech-related (pink arcs and line).

F. Contingency of adult responses on content of child vocalizations

In the recordings of typically developing children, presence of speech-related vocalization within a child segment did indeed correspond to increased likelihood of that child vocalization receiving an immediate adult response: the difference between the proportion of speech-related child vocalizations receiving an adult response and the proportion of non-speech-related child vocalizations receiving an adult response was .054, $p < .001$ (see the scatterplot showing probability of response given duration of speech-related vocalization in Fig. 4.1B). The same pattern held for the recordings of children with autism, with the difference between the two proportions being .040, $p < .001$. The strength of this effect of child speech-related vocalization on adult response differed across the two groups, being weaker in ASD than in TD.

G. The effect of contingent adult responses on subsequent child vocalizations

In the recordings of typically developing children, a child vocalization was more likely to be speech-related if the child's previous speech-related vocalization had received an immediate adult response. The average difference between proportion of child vocalizations that were speech-like when the preceding speech-related vocalization was responded to and the proportion when the preceding speech-related vocalization was not responded to was .031, which was a statistically significant difference, $p < .001$ (see the bar plot in Fig. 4.1D). The same pattern again held for the children with autism as well, with the differences in proportions being .041, $p < .001$. There were no significant differences corresponding to ASD diagnosis.

H. Age, maternal education, and gender

The quantity of child speech-related vocalization also increased with age, $\beta = .309$, $p < .001$, with maternal education, $\beta = .183$, $p = .001$, and was higher for females than males, $\beta = .133$, $p = .014$. There was no statistically significant interaction between age and maternal education, $\beta = .008$, $p = .692$. Quantity of adult vocalization increased significantly with maternal education, $\beta = .285$, $p < .001$, but did not change significantly with age, or gender.

The diagonal cross recurrence profile was higher as age increased, $\beta = .1551$, $p = .011$, and was greater as maternal education increased, $\beta = .257$, $p < .001$. There was no significant effect of gender. These results indicated that age and maternal education are both associated with increased quantities of child-caregiver interaction. There were no significant effects of age, maternal education, or gender on the ratio of the child leading side of the DCRP to the adult leading side.

As age increased, the proportion of child vocalizations receiving an adult response within 1 s decreased, $\beta = -.120$, $p = .043$. As maternal education increased, this proportion increased, $\beta = .258$, $p < .001$. There were no significant trends with regard to gender. With regard to children's immediate responses to adults, they were more likely for females than for males, $\beta = -.313$, $p < .001$, and were higher for females, but there were no significant trends with regard to age or maternal education.

As for the test of the top arrow of the feedback loop depicted in Fig. 4.1B, the contingency of adult response on there being speech-related material in a given child vocalization was stronger as maternal education increased, $\beta = .263$, $p < .001$, and stronger for males, $\beta = .125$, $p = .027$. There was no statistically significant effect of age. The contingency of child speech-related vocalization on whether the child's previous speech-related vocalization received a response did not differ with regard to age, maternal education, or gender.

IV. Discussion

We found that, compared to typically developing children, children with autism produced less speech-like vocalization and exhibited differences in the quantity and temporal patterning of vocal interaction with adults. These results support the patterns uncovered by previous research, using day-long naturalistic recordings, lag-1 analysis of immediate responses, and cross-recurrence, a method for analyzing dyadic interaction that is new to this domain.

Our subsequent analyses explored the interrelationship between speech-language development and social interaction. We found that in both typical development and autism spectrum disorder, adult responding was contingent on the duration of speech-related material within a child vocalization: more child speech-related vocalization increased the likelihood of adult response. We also found that the likelihood of there being speech-related material within a child vocalization was contingent on whether or not the child's previous speech-related vocalization did receive an adult response: having received a response previously was associated with a increased likelihood of speech-related material in the very next child vocalization. These findings taken together provide support, based on microscopic analysis of naturalistic data, for the idea that there is a social feedback loop between child and adult that facilitates speech development in both typically developing children and children with autism spectrum disorders. Added up across the huge number of child vocalizations that take place within a day, let alone a month or year, these contingencies could be expected to compound, contributing substantially to speech development.

Although this social feedback loop is operational in autism, it appears to be diminished in several ways. First, children with autism produce fewer speech-related vocalizations (Fig. 4.1A). If part of this reduction in speech-related vocalization stems from endogenous factors, such as differences in motor skills or innate propensity to vocalize, then this would inherently lead to fewer iterations of the social feedback loop, reducing the child's opportunities to learn from contingent social feedback.

Second, adults interacting with children with autism exhibit lower rates of responding to children's speech-related utterances, as measured both by proportion of child speech-related vocalizations receiving a response (where a marginally significant difference between ASD and TD recordings was found) and by the dependence of response likelihood on the speech-relatedness of the child vocalization. These reductions in contingent responding by adults to children with ASD could be in part due to a variety of factors, such as genetic differences between parents of children with ASD and parents of TD children, the demands of parenting a child with special needs affecting a parent's ability to attend and respond at high rates, differences in the speech and social behaviors of children with ASD affecting parents' expectation that the child will produce mature speech-like vocalizations or that their response will lead to rewarding social interaction. Additionally, the reduced contingency of adult responses in the ASD group may be due to lower quality of the speech-related vocalizations of the children in that group. In this study we only took into account the duration of speech-related material in a child vocalization, but the quality of the vocalization, such as its voice quality, the words and phrases in contains, its prosody, etc. likely play a role in determining the likelihood of adult response. Regardless of the reason, the reduction in adults' contingent responding again provides children with fewer opportunities to learn about the social effects of their vocal behavior.

A third way the social feedback loop could be diminished in autism would be if children with autism have reduced ability to make use of contingent social responses as a cue indicating how they should direct future behaviors. In the present study we did not find any statistically significant differences across groups in the degree to which children's future vocalization speech-relatedness was positively affected by having received a response for previous speech-related vocalizations. Since this lack of effect may reflect a true lack of difference across groups or a lack of sensitivity of the design of the present study, whether or not this is a source of reduction in the social feedback loop remains an open question for future research, using larger numbers of samples or perhaps looking in greater detail at the characteristics of the children's vocalizations.

Our social feedback loop hypothesis predicts an interaction between age and ASD such that group differences in speech skills may magnify over time, as any differences in the operation of the feedback loop or the frequency with which it is experienced would magnify over the course of months and years. In keeping with this prediction, we found an interaction between age and ASD in their effects on quantity of speech related vocalization such that group differences magnified with age. Additionally, previous research also using day-long naturalistic recordings has also found slower maturation of acoustic-phonetic features of child vocalizations for children with ASD compared to children who were TD or who had language day but not ASD [113]. Interestingly, the language-delay children in that study did not have slower maturation of vocalizations compared to TD children—they seemed to have a similar rate of maturation, just delayed in time. Perhaps in the case of those language-delayed children, the social feedback loop was not as badly affected by their impairment, or perhaps the interventions they received compensated for the negative effects on the feedback loop—these factors may affect whether differences across groups compound over time. The success of interventions with children with autism and developmental delays that have targeted adult responsivity to child communication acts [133, 134] can be interpreted as affecting the contingent adult response component of the feedback loop proposed here.

In addition to finding differences with regard to autism spectrum disorder, we also found differences with regard to age and maternal education. Overall levels of interaction increased with age and maternal education. Additionally, adult responsiveness and the contingency of adult responses on the speech-relatedness of child vocalizations increased with maternal education. We found no statistically significant effects of gender. These results suggest that in addition to child differences, such as having autism spectrum disorder, other factors such as socioeconomic status may also affect aspects of the social feedback loop.

In summary, we propose that there is a social feedback loop that supports speech development, wherein adults respond contingently to speech-related child vocalizations and children's vocalization characteristics are in turn contingent on previous adult responses. We explicitly tested each component of this social feedback loop, using automated labeling and analysis of the microstructure of interaction in day-long naturalistic recordings. The results support the social feedback loop hypothesis for speech development in children both with and without autism spectrum disorder. The results also reveal two ways in which autism appears to reduce the effectiveness of the social feedback loop in supporting development: (1) the disorder reduces the child vocal material that is available for adults to respond to and (2) the disorder is associated with decreases in adult responsiveness and its contingency on the child's behavior. Given the constructivist nature of the social feedback loop, these group differences are expected to have accumulating, cascading effects across development, providing a possible account for the magnification of group differences across time due to sensitivity of the system to initial conditions.

V. Materials and Methods

We used the same database of recordings as a previous study by Warren and colleagues [118]. The autism participants were 26 children between 16–48 mo. of age who were diagnosed with the Autism subtype of ASD, except for two of the youngest children who fell into the Pervasive Developmental Disorder–Not Otherwise Specified subtype. None of the children was reported in their diagnosis to exhibit echolalia. The typically developing participants were 78 children who were matched 3-to-1 to the ASD children on gender and maternal education and, collectively across the 3 TD children, on age. More information about participant recruitment, recording, and demographics can be found in [118] and about the automated speaker and utterance type labeling can be found in [113]. Informed consent was obtained from all participants.

Recordings were made and pre-processed using the LENA system. A small device fit into custom clothing records the child's voice as well as other sounds in the child's environment. Parents were instructed to begin recording when the child awoke in the morning. Some of the TD group recordings were made using earlier versions of the recorder. All recordings lasted 12 hours or longer. For those recordings longer than 12 hours, only data from the first 12 hours was used in the present study so as to equate for recording length. Recordings took place in varied settings including in the home, car, preschool, and, for the ASD group, speech-language therapy. In total, 438 recordings and 5,256 hours of recording were included. All procedures were approved by the Essex Institutional Review Board.

The LENA system software was used to analyze each recording, automatically segmenting it according to various types of sound sources [113]. For the purposes of the present study, we looked only at the following categories of sound source: child wearing the recorder and adult. Within child segments, the system also identifies sub-segments of vocalization and labels them as speech-like utterances or as non-speech like vocalizations, which include crying and vegetative sounds such as burping, coughing, etc. Reliability of the automated labeling compared to human listener labeling has been computed for recordings of typically developing children, yielding the following percentages correct: 82 for adult speaker, 76 for child wearing the recorder, 75 for child speech-like utterances, and 84 for child non-speech-like vocalizations. The start and end times of these labels were extracted from the LENA system's ".its" files using custom-written perl scripts. These start and end times served as the input to our analyses.

Cross recurrence analysis starts with the formation of a cross recurrence plot [135]. The plot is essentially a matrix that shows the temporal relationship between all possible combinations of child vocalizations with adult vocalizations occurring anywhere within a recording. Each diagonal of this plot corresponds to a particular temporal relationship between the pair of child and adult vocalizations. By finding the proportion of points along each diagonal that are occupied, we derive a diagonal cross recurrence profile [136]. See the SI (section X.) for more detail on the diagonal cross recurrence plot analysis.

For each vocalization and interaction variable, statistical tests for differences across groups used a linear mixed effects model with participant ID as a random effect and ASD, age, mother's level of education, and gender as fixed effects. For the model where quantity of speech-related vocalization was the dependent variable, two interaction terms, age by ASD and age by maternal education, were also included, since this interaction was of particular interest for this variable. Markov chain Monte Carlo was used to determine the p-values. All variables were standardized prior to inclusion in the linear mixed effects models.

In testing the two directions of influence in our social feedback loop hypothesis, significance of the predictions was based on bootstrapped null distributions. The relationship between whether a child vocalization contained child speech-related material and the likelihood that it received an adult response variable was tested. To obtain the p-value for the effect of child speech-related material, all recordings for the group in question had the adult response variable shuffled relative to the child speech-related utterance duration. Shuffling was done at the level of the recording. This was repeated 1000 times. The proportion of speech-related child vocalizations receiving adult responses minus the proportion of non-speech-related child vocalizations receiving adult responsesfor the original (not reshuffled) data was then compared to the difference between the two proportions for the 1000 shuffled datasets. The p-value was then determined based on the proportion of reshuffled proportion differences that were greater than the real proportion differences. An analogous procedure was done in order to test the significance of the hypothesis that speech-related material in a child vocalization would be predicted by whether the previous child segment containing some speech-related vocalization had received an adult response.

VI. acknowledgments

This work was supported by a Department of Energy Computational Science Graduate Fellowship, the NIDCD (R01 DC011027), the Plough Foundation, and the LENA Foundation. We thank Rick Dale and Dongxin Xu for their helpful advice and contributions to earlier presentations of portions of the work and Gert Westermann, Eugene Buder, and Robert Kozma for their comments on a draft of this manuscript. We also thank the families who participated in this study and the LENA Foundation for the use of their databases.

VII. Supplementary information

VIII. Prevalence of sound types

Figure 4.4 shows the prevalence of each of nine sound source types, separately for TD and ASD recordings. A mixed effects regression was run for each sound type, with participant ID as a random effect and ASD, age, maternal education, gender, recorder version, interaction between ASD and age, and interaction between maternal education and age as fixed effects. The results of this analysis are given in Table 4.1. Note that although Other Child vocalization segments were more frequent in the TD group's recordings, the number of siblings was not very different across groups, with the mean number being .95 ($SD = .86$) for the TD groups's recordings and .93 ($SD = 1.04$ for the ASD group's recordings.

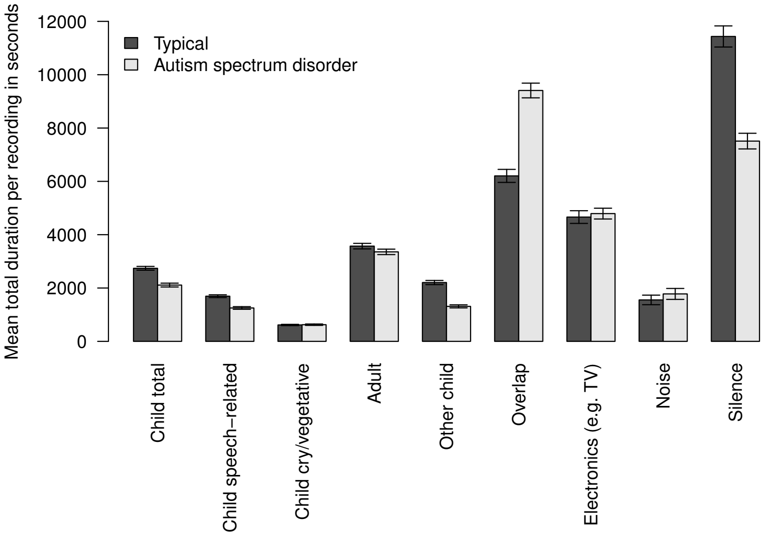

Fig. 4.4: Prevalence of sound source types in TD and ASD recordings. Error bars indicate standard error of the means across all recordings.

Table 4.1: Effects of ASD, age, maternal education, gender (female), and recorder version on prevalence of sound types

Sound source	ASD	Age	MomEd	Gender	Recorder	Age*ASD	Age*MomEd
Child total	$\beta = -.377$ $p < .001$	$\beta = .241$ $p < .001$	$\beta = .165$ $p = .005$	$\beta = .125$ $p = .030$	$\beta = .075$ $p = .164$	$\beta = -.080$ $p = .107$	$\beta = .021$ $p = .685$
Child speech-related	$\beta = -.455$ $p < .001$	$\beta = .309$ $p < .001$	$\beta = .183$ $p = .001$	$\beta = .133$ $p = .014$	$\beta = .185$ $p = .001$	$\beta = -.100$ $p = .026$	$\beta = -.100$ $p = .549$
Child cry and vegetative	$\beta = .079$ $p = .353$	$\beta = -.025$ $p = .693$	$\beta = .001$ $p = .871$	$\beta = .040$ $p = .574$	$\beta = -.106$ $p = .176$	$\beta = .024$ $p = .745$	$\beta = -.007$ $p = .847$
Adult	$\beta = -.110$ $p = .132$	$\beta = .093$ $p = .141$	$\beta = .288$ $p < .001$	$\beta = -.040$ $p = .447$	$\beta = .036$ $p = .542$	$\beta = .075$ $p = .387$	$\beta = -.086$ $p = .167$
Other child	$\beta = -.324$ $p < .001$	$\beta = .306$ $p < .001$	$\beta = .209$ $p < .001$	$\beta = -.097$ $p = .109$	$\beta = -.223$ $p < .001$	$\beta = .024$ $p = .633$	$\beta = .081$ $p = .114$
Overlap	$\beta = .156$ $p = .021$	$\beta = .142$ $p = .008$	$\beta = .087$ $p = .113$	$\beta = -.061$ $p = .208$	$\beta = .334$ $p < .001$	$\beta = .037$ $p = .461$	$\beta = -.024$ $p = .694$
Electronic	$\beta = -.127$ $p = .062$	$\beta = .069$ $p = .359$	$\beta = -.134$ $p = .028$	$\beta = -.064$ $p = .307$	$\beta = .226$ $p < .001$	$\beta = -.035$ $p = .498$	$\beta = .054$ $p = .333$
Noise	$\beta = .108$ $p = .169$	$\beta = -.186$ $p = .004$	$\beta = -.124$ $p = .083$	$\beta = .185$ $p = .002$	$\beta = -.063$ $p = .429$	$\beta = -.050$ $p = .454$	$\beta = .061$ $p = .557$
Silence	$\beta = -.047$ $p = .521$	$\beta = -.250$ $p < .001$	$\beta = .021$ $p = .727$	$\beta = -.056$ $p = .243$	$\beta = -.420$ $p < .001$	$\beta = .006$ $p = .790$	$\beta = -.025$ $p = .639$

IX. Extraction of vocalization times

The LENA “.its” file is an xml file that provides information about the recorder version, software version, information about the automatically-produced sound type labels, and much other information. Here is an example of a portion of one of the its files used in this study, corresponding to Figure 4.2 in the main article:

```
<Segment spkr="SIL" average_dB="-59.47" peak_dB="-53.02"
   startTime="PT286.78S" endTime="PT287.69S" />
<Segment spkr="CHN" average_dB="-26.01" peak_dB="-18.84"
   conversationInfo="|RC|10|9|3|AICF|NT|FH|" childUttCnt="1"
   childUttLen="P1.55S" startUtt1="PT287.69S" endUtt1="PT289.24
   S" childCryVfxLen="P0.00S" startTime="PT287.69S" endTime="
   PT289.24S" />
<Segment spkr="FAN" average_dB="-43.54" peak_dB="-35.31"
   conversationInfo="|RC|10|9|3|AICF|TIFI|FI|"
   femaleAdultWordCnt="5.28" femaleAdultNonSpeechLen="P0.00S"
   femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="
   PT289.24S" endTime="PT290.28S" />
<Segment spkr="CHN" average_dB="-45.82" peak_dB="-36.08"
   conversationInfo="|RC|10|10|4|AICF|TIFR|FI|" childUttCnt="1"
    childUttLen="P0.60S" startUtt1="PT290.28S" endUtt1="PT290
   .88S" childCryVfxLen="P0.00S" startTime="PT290.28S" endTime
   ="PT290.88S" />
<Segment spkr="FAN" average_dB="-40.63" peak_dB="-31.29"
   conversationInfo="|RC|10|10|4|AICF|TIFE|FI|"
   femaleAdultWordCnt="3.49" femaleAdultNonSpeechLen="P0.00S"
   femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="
   PT290.88S" endTime="PT292.08S" />
<Segment spkr="CHN" average_dB="-10.14" peak_dB="-5.01"
   childUttCnt="0" childUttLen="P0.00S" startCry1="PT292.60S"
   endCry1="PT293.47S" childCryVfxLen="P0.87S" startTime="PT292
   .08S" endTime="PT293.65S" />
```

```
<Segment spkr="FAN" average_dB="-39.64" peak_dB="-31.73"
   conversationInfo="|RC|10|10|4|AICF|NT|FH|"
   femaleAdultWordCnt="8.86" femaleAdultNonSpeechLen="P0.00S"
   femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="
   PT293.65S" endTime="PT295.66S" />
<Segment spkr="SIL" average_dB="-66.53" peak_dB="-58.48"
   startTime="PT295.66S" endTime="PT296.46S" />
<Segment spkr="CHN" average_dB="-11.01" peak_dB="-6.58"
   childUttCnt="0" childUttLen="P0.00S" startCry1="PT296.46S"
   endCry1="PT298.31S" childCryVfxLen="P1.85S" startTime="PT296
   .46S" endTime="PT298.31S" />
<Segment spkr="FAN" average_dB="-39.58" peak_dB="-31.36"
   conversationInfo="|RC|10|10|4|AICF|TIFI|FH|"
   femaleAdultWordCnt="7.87" femaleAdultNonSpeechLen="P0.00S"
   femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="
   PT298.31S" endTime="PT299.98S" />
<Segment spkr="SIL" average_dB="-52.40" peak_dB="-40.56"
   startTime="PT299.98S" endTime="PT300.85S" />
<Segment spkr="CHN" average_dB="-16.19" peak_dB="-6.59"
   conversationInfo="|RC|10|11|5|AICF|TIFR|FI|" childUttCnt="1"
    childUttLen="P2.61S" startUtt1="PT300.85S" endUtt1="PT303
   .46S" childCryVfxLen="P0.00S" startTime="PT300.85S" endTime
   ="PT303.69S" />
<Segment spkr="TVF" average_dB="-57.17" peak_dB="-45.99"
   startTime="PT303.69S" endTime="PT304.78S" />
<Segment spkr="SIL" average_dB="-57.76" peak_dB="-49.54"
   startTime="PT304.78S" endTime="PT307.39S" />
<Segment spkr="CHN" average_dB="-25.74" peak_dB="-17.37"
   conversationInfo="|RC|10|11|5|AICF|NT|FH|" childUttCnt="2"
   childUttLen="P2.21S" startUtt1="PT307.39S" endUtt1="PT307.74
   S" startUtt2="PT308.41S" endUtt2="PT310.27S" childCryVfxLen
   ="P0.00S" startTime="PT307.39S" endTime="PT310.27S" />
```

```
        <Segment spkr="NOF" average_dB="-50.92" peak_dB="-44.79"
           startTime="PT310.27S" endTime="PT311.10S" />
        <Segment spkr="FAN" average_dB="-43.20" peak_dB="-31.10"
           conversationInfo="|RC|10|11|5|AICF|TIFI|FI|"
           femaleAdultWordCnt="5.68" femaleAdultNonSpeechLen="P0.00S"
           femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="
           PT311.10S" endTime="PT312.76S" />
        <Segment spkr="CHN" average_dB="-28.98" peak_dB="-18.43"
           conversationInfo="|RC|10|12|6|AICF|TIFR|FI|" childUttCnt="1"
            childUttLen="P0.99S" startUtt1="PT312.76S" endUtt1="PT313
           .75S" childCryVfxLen="P0.00S" startTime="PT312.76S" endTime
           ="PT313.75S" />
        <Segment spkr="OLN" average_dB="-26.88" peak_dB="-23.08"
           startTime="PT313.75S" endTime="PT314.55S" />
        <Segment spkr="FAN" average_dB="-36.17" peak_dB="-29.04"
           conversationInfo="|RC|10|12|6|AICF|TIFE|FI|"
           femaleAdultWordCnt="6.74" femaleAdultNonSpeechLen="P0.00S"
           femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="
           PT314.55S" endTime="PT315.65S" />
        <Segment spkr="MAN" average_dB="-46.61" peak_dB="-39.19"
           conversationInfo="|EC|10|12|6|AICF|NT|FI|" maleAdultWordCnt
           ="4.82" maleAdultNonSpeechLen="P0.00S" maleAdultUttCnt="0"
           maleAdultUttLen="P0.00S" startTime="PT315.65S" endTime="
           PT316.95S" />

</Conversation>


<Pause num="11" average_dB="-45.86" peak_dB="-28.31" childCryVfxLen="P0
   .00S" femaleAdultNonSpeechLen="P0.00S" maleAdultNonSpeechLen="P0.00S
   " TVF="P0.00S" FAN="P0.00S" OLN="P0.00S" SIL="P2.97S" NOF="P16.24S"
   CXF="P0.00S" OLF="P0.00S" CHF="P0.00S" MAF="P1.00S" TVN="P0.00S" NON
```

```
="P0.00S" CXN="P0.00S" CHN="P0.00S" MAN="P0.00S" FAF="P0.00S"
startTime="PT316.95S" endTime="PT337.16S" >

    <Segment spkr="SIL" average_dB="-57.86" peak_dB="-52.58"
       startTime="PT316.95S" endTime="PT318.05S" />
```

We parsed this text line-by-line in Perl, using regular expressions to search for sound types (e.g. `Segment spkr="CHN"` indicates a segment where the sound source is the child wearing the recorder). We treated sound types ending in `F` as if they were silence, since an `F` indicates that the labeling algorithm found the segment to have a high level of similarity to its Gaussian mixture model for silence. The bottom portion of the its file provides information about the bar plot data in the LENA software GUI; we ignored this portion of the its file.

X. Cross-recurrence analysis

Before performing the cross recurrence analyses, we divided the recording up into segments corresponding to any child or adult speaker segment or into 1 s bins if the corresponded to times in the recording when neither the child nor the adult was speaking. This step ensured that interactivity as measured using cross recurrence was not affected by the durations of the vocalizations themselves, only by the timing between child and adult vocalizations.

Figure 4.5 gives an example of a cross-recurrence plot and illustrates how the cross recurrence diagonal profile (DCRP) is computed. A cell in the cross recurrence plot is considered occupied (marked black in the example shown in Fig. 4.5) if there is a child segment at the time corresponding to the cell's y-coordinate and an adult segment at the time corresponding to the cell's x-coordinate. Otherwise the cell is considered unoccupied (marked white in the figure).

Filled cells on the diagonal of this plot that runs from the origin to the final event represent pairs of child and adult vocalizations that occurred at the same time. Since the

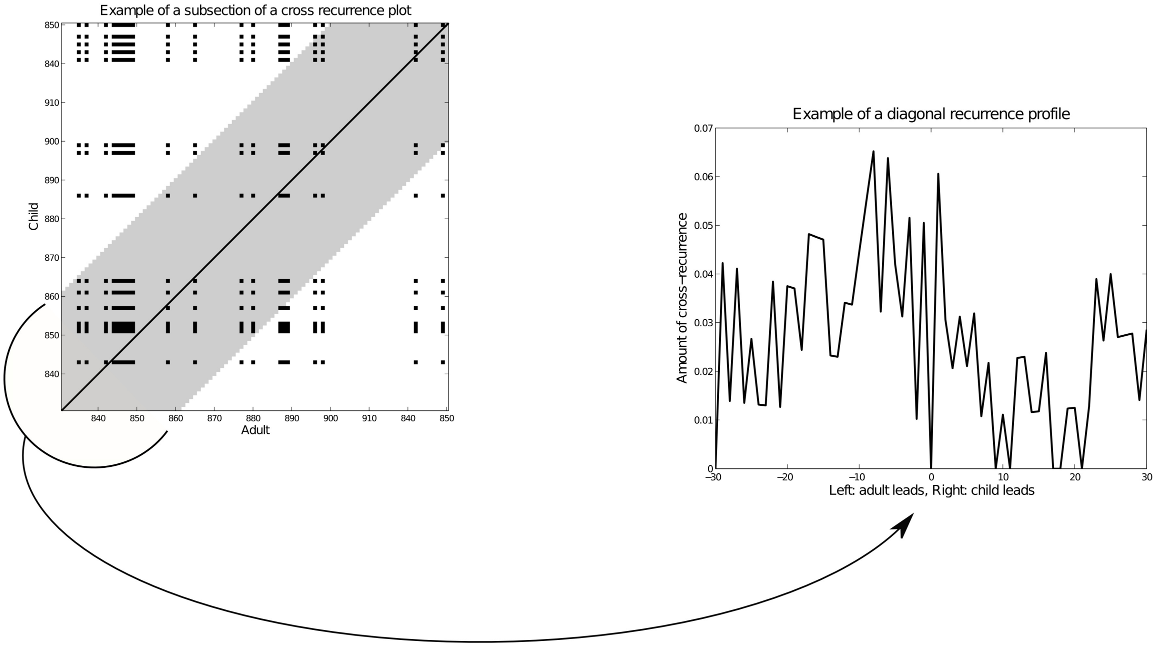

Fig. 4.5: Recurrence plot for a 200 s section of a recording (left) and the diagonal recurrence profile for the recording (right).

speaker labels in the present study are mutually exclusive, there were never any points on this central diagonal. Filled cells that are below and right of the central diagonal represent pairs of child and adult vocalizations where the child's vocalization preceded the adult's. Filled cells above and left of the central diagonal represent pairs of child and adult vocalizations where the child's vocalization followed the adult's, whereas filled cells in the bottom right represent pairings where the child led the adult. Each diagonal running from the bottom left to the top right represents pairings with the same lag relationship. Adding up the filled points along each diagonal and dividing by that diagonal's length gives the height (the y-axis value) for the lag (the x-axis value) in the diagonal cross recurrence profile (DCRP) [136]. In the physical sciences, this diagonal cross recurrence profile has been referred to as the recurrence probability or the recurrence spectrum [135].

The overall height of this profile indicates the level of interactivity within the maximum and minimum lags on the plot. The height of the right side of the plot indicates the amount of child leading and the height of the left side indicates the amount of adult leading; the ratio of the right side to the left side provides a measure of the degree to which the child lead vs. the adult leading.

Chapter 5

General conclusion

The three papers included in this dissertation focused on the nature of early speech-related vocalizations and how they change across the first few years of life. Methodologically, the approach was focused on using computational tools, namely automated analysis and neural network modeling.

In paper 1, we found that through automated analysis using a combination of self-organizing and supervised neural networks, it was possible to obtain and visualize holistic, data-derived acoustic features that characterize early speech-related vocalizations. We compared these features and noted how they change across the first year of life and how they map onto categories from a different, protophone-based system for vocalization coding. Supervised classification according to phonatory protophone category (*squeal*, *vocant*, or *growl*) and age based on these data-derived features was above chance levels, suggesting that this type of approach might be further developed into a useful automated analysis system.

In paper 2, we reported on a neural network model of early infant vocalizations. The model modified a self-organizing neural network architecture so that its learning was dependent on whether or not it received reinforcement for a particular action. We connected the model to an existing realistic vocal tract simulator in order to produce synthesized vowel sounds. The same model learned both to reliably produce phonation and to produce vowels that were acoustically similar to target vowels for which it had been reinforced. Differences were observed when the model was reinforced for producing English vowels versus when it was reinforced for producing Korean vowels. Thus, the paper introduced a computationally-specified mechanistic model for how production of speech sounds might be acquired in infancy.

In paper 3, we analyzed the interaction dynamics present in day-long automated recordings of preschoolers that had already been automatically labeled according to who was vocalizing at what time (and if it was the child, when the child's vocalizations were speech-related). We found that when children produced speech-related vocalizations, they were more likely to receive adult responses and that in turn, when children's speech-related vocalizations were responded to, the children's subsequent vocalizations were more likely to contain speech-related material. Thus, the results were consistent with our proposed social feedback loop playing a positive role in speech development. A number of differences were found between children who were typically developing and children with autism, suggesting that this feedback loop is diminished in autism. Differences with regard to maternal education and age were also found.

Taken together, the three papers provide support for the value of computationally-oriented approaches to studying human infant vocalization and its development. They also provide support for the idea that there are differentiable types of speech-related vocalizations produced by young infants, that they change with age, and that these changes can be shaped through contingent reinforcement. Future work should continue to explore automated analysis approaches and neural network modeling, seeking to improve the fit of the automated analyses and of the models to human data.

I was the primary contributor in all of the papers, designing the studies, developing the automated analysis and computational model code, running the statistical analyses, interpreting the results, and writing the papers. My co-authors provided guidance and feedback in most cases at each of these stages. The child audio data in paper 1 was from an existing dataset; I was not personally involved in the recordings and I played only a relatively minor role in the human listener judgments. In all other cases where existing datasets or software were used, this is made clear in the papers.

REFERENCES

[1] Oller DK, Griebel U (2008) in *Evolution of Communicative Flexibility: Complexity, Creativity, and Adaptability in Human and Animal Communication*, eds Oller DK, Griebel U (MIT Press, Cambridge, MA) 141–168.

[2] Oller DK *The Emergence of the Speech Capacity* (Lawrence Erlbaum Associates, Mahwah, NJ).

[3] President's Information Technology Advisory Committee (2005) *Computational Science: Ensuring Americas Competitiveness* (Executive Office of the President of the United States, Washington, DC).

[4] Stark RE (1980) in *Child Phonology, Vol. 1: Production*, eds Yeni-Komshian GH, Kavanagh JF, Ferguson CA (Academic Press, New York) 73–92.

[5] Oller DK (1980) in *Child Phonology, Vol 1: Production*, eds Yeni-Komshian GH, Kavanagh JF, Ferguson CA (Academic Press, New York) 93–112.

[6] van der Stelt JM (1993) *Finally a Word: A Sensori-motor Approach of the Mother-infant System in its Development toward Speech* (Uitgave IFOTT, Amsterdam).

[7] Koopmans-van Beinum FJ, van der Stelt JM (1986) in *Precursors of Early Speech*, eds. Lindblom B, Zetterström R (Stockton Press, New York) 37–50.

[8] Vihman M, Macken M, Miller R, Simmons H, Miller J (1985) From babbling to speech: a re-assessment of the continuity issue. *Lang* 61:397–445.

[9] Roe KV (1975) Amount of infant vocalization as a function of age: some cognitive implications. *Child Dev* 46:936–941.

[10] Stoel-Gammon C (1989) Prespeech and early speech development of two late talkers. *First Lang* 9:207–223.

[11] Eilers RE, Oller DK (1994) Infant vocalizations and the early diagnosis of severe hearing impairment. *J Pediatr* 124:199–203.

[12] Oller DK, Eilers R, Neal R, Schwartz H (1999) Precursors to speech in infancy: the prediction of speech and language disorders. *J Commun Disord* 32:223–245.

[13] Kent R, Murray A (1982) Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *J Acoust Soc Am* 72:353–365.

[14] Robb MP, Saxman JH (1988) Acoustic observations in young childrens non-cry vocalizations. *J Acoust Soc Am* 83:1876–1882.

[15] Papaeliou C, Minadakis G, Cavouras D (2002) Acoustic patterns of infant vocalizations expressing emotions and communicative functions. *J Speech Lang Hear Res* 45:311–317.

[16] Repp B (1982) Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Psychol Bull* 92:81–110.

[17] Zimmerman FJ, et al. (2009) Teaching by listening: the importance of adult-child conversations to language development. *Pediatr* 124:342–349.

[18] Lynip AW (1951) The use of magnetic devices in the collection and analysis of the preverbal utterances of an infant. *Genet Psychol Monogr* 44:221–262.

[19] Nathani S, Oller DK (2001) Beyond ba-ba and gu-gu: challenges and strategies in coding infant vocalizations. *Behav Res Methods Instrum Comput* :3:321–330.

[20] Nathani S, Ertmer D, Stark R (2006) Assessing vocal development in infants and toddlers. *Clin Linguist Phon* 20:351–369.

[21] Buder E, Chorna L, Oller DK, Robinson R (2008) Vibratory regime classification of infant phonation. *J Voice* 22:553–564.

[22] Martin JAM (1981) *Voice, Speech, and Language in the Child: Development and Disorder* (Springer-Verlag, New York).

[23] Kent R, Bauer H (1985) Vocalizations of one-year-olds. *J of Child Lang* 12:491–526.

[24] Smith B, Brown-Sweeney S, Stoel-Gammon C (1989) A quantitative analysis of reduplicated and variegated babbling. *First Lang* 9:175–189.

[25] Scheiner E, Hammerschmidt K, Jürgens U, Zwirner P (2002) Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants. *J Voice* 16:509–529.

[26] Oller DK, Eilers R, Basinger D (2001) Intuitive identification of infant vocal sounds by parents. *Dev Sci* 4:49–60.

[27] Holmgren K, Lindblom B, Aurelius G, Jalling B, Zetterström R (1986) in *Precursors of Early Speech*, eds. Lindblom B, Zetterström R (Stockton Press, New York) 51–63.

[28] in *Child Phonology, Vol. 1: Production* (Academic Press, New York) 113–142.

[29] Kuhl PK, Meltzoff AN (1996) Infant vocalizations in response to speech: vocal imitation and developmental change. *J Acoust Soc Am* 100:2425–2438.

[30] Stevenson J, Richman N (1976) The prevalence of language delay in a population of three-year-old children and its association with general retardation. *Dev Med Child Neurol* 18:431–441.

[31] Thal D, Desjardin J, Eisenberg L (2007) Validity of the MacArthur-Bates communicative development inventories for measuring language abilities in children with cochlear implants. *Am J Speech Lang Pathol* 16:54–64.

[32] Fell HJ, MacAuslan J, Ferrier LJ, Worst SG, Chenausky K (2002) in *Proceedings of the 7th International Conference on Spoken Language Processing* 2345–2348.

[33] Oller DK, Eilers RE (1988) The role of audition in infant babbling. *Child Dev* 59:441–449.

[34] Vihman M (1986) Phonological development from babbling to speech: common tendencies and individual differences. *Appl Psycholinguist* 7:3–40.

[35] Nathani Iyer S, Oller DK (2008) Fundamental frequency development in typically developing infants and infants with severe to profound hearing loss. *Clin Linguist Phon* 22:917–936.

[36] Vihman M, Greenlee M (1987) Individual differences in phonological development: ages one and three years. *J Speech Hear Res* 30:503–521.

[37] Bishop C (1995) *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford).

[38] Kohonen T (2001) *Self-organizing Maps* (Springer, New York).

[39] Ritter H (2003) in *The Handbook of Brain Theory and Neural Networks* (MIT Press, Cambridge, Massachusetts) 1005–1010.

[40] Xu R, Wunsch D (2005) Survey of clustering algorithms. *IEEE Trans Neural Netw* 16:645–678.

[41] Janata P (2001) Quantitative assessment of vocal development in the zebra finch using self-organizing neural networks. *J Acoust Soc Am* 110:2593–2603.

[42] Nickerson C, Bloomfield L, Dawson M, Charrier I, Sturdy C (2007) Feature weighting in “chick-a-dee” call notes of poecile atricapillus. *J Acoust Soc Am* 122:2451–2458.

[43] Guenther F, Gjaja M (1996) The perceptual magnet effect as an emergent property of neural map formation. *J Acoust Soc Am* 100:1111–1121.

[44] Gauthier B, Shi R, Xu Y (2007) Learning phonetic categories by tracking movements. *Cogn* 103:80–106.

[45] Leinonen L, Kangas J, Torkkola K, Juvas A (1992) Dysphonia detected by pattern recognition of spectral composition. *J of Speech Hear Res* 35:287–295.

[46] Callan DE, Kent RD, Roy N, Tasko SM (1999) Self-organizing map for the classification of normal and disordered female voices. *J Speech Lang Hear Res* 42:355–366.

[47] Schönweiler R, Kaese S, Möller S, Rinscheid A, Ptok M (1996) Neuronal networks and self-organizing maps: new computer techniques in the acoustic evaluation of the infant cry. *Int J Pediatr Otorhinolaryngol* 38:1–11.

[48] Buder E, Stoel-Gammon C (2002) American and swedish childrens acquisition of vowel duration: effects of vowel identity and final stop voicing. *J Acoust Soc Am* 111:1854–1864.

[49] Milenkovic P (2001) *TF32*.

[50] Delgado RE (2008) *Action Analysis Coding and Training Software* (AACT).

[51] Oller DK, Lynch MP (1992) in *Phonological Development: Models, Research, Implications*, eds Ferguson CA, Menn L, Stoel-Gammon C (York Press, Timonium, MD) 509–536.

[52] MathWorks (2008). *MATLAB*.

[53] Zwicker E (1961) Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *J Acoust Soc Am* 33:248–248.

[54] Ellis DPW (2007) *PLP and RASTA (and MFCC, and inversion) in MATLAB* (http://www.ee.columbia.edu/?dpwe/resources/matlab/rastamat/).

[55] Kuang Z, Kuh A (1992) A combined self-organizing feature map and multilayer perceptron for isolated word recognition. *IEEE Trans Signal Process* 40:2651–2657.

[56] Demuth H, Beale M, Hagan M (2006) *Neural network toolbox for use with MATLAB*.

[57] Berglund E, Sitte J (2006) The parameterless self-organizing map algorithm. *IEEE Trans Neural Netw* 17:305–316.

[58] Foresee D, Hagan M (1997) Gauss-Newton approximation to bayesian learning. *Proceedings of the IEEE International Conference on Neural Networks* 10:1930–1935.

[59] Kasabov NK, Kozma R, Watts M (1998) Phoneme-based speech recognition via fuzzy neural networks modeling and learning. *Inf Sci* 110:61–79.

[60] Forrest K, Weismer G, Milenkovic P, Dougall R (1988) Statistical analysis of word-initial voiceless obstruents: preliminary data. *J Acoust Soc America* 84:115–123.

[61] Burkard RF, Secor C (2002) in *Handbook of Clinical Audiology*, eds Katz J, Burkard RF, Medwetsky L (Lippincott Williams & Wilkins, Baltimore) 233–248.

[62] Hardin-Jones M, Chapman K, Schulte J (2003) The impact of cleft type on early vocal development in babies with cleft palate. *Cleft Palate Craniofac J* 40:453–459.

[63] Salas-Provance M, Kuehn D, Marsh J (2003) Phonetic repertoire and syllable characteristics of 15-month-old babies with cleft palate. *J Phon* 31:23–38.

[64] Iverson JM, Wozniak RH (2007) Variation in vocal-motor development in infant siblings of children with autism. *J Autism Dev Disord* 37:158–170.

[65] Goldstein MH, Schwade JA (2008) Social feedback to infants babbling facilitates rapid phonological learning. *Psychol Sci* 19:515–523.

[66] Bryson S, Zwaigenbaum L, Mcdermott C, Rombough V, Brian J (2008) The autism observation scale for infants: scale development and reliability data. *J Autism Dev Disord*, 4:731–738.

[67] Flexer A (2001) On the use of self-organizing maps for clustering and visualization. *Intelligent Data Analysis* 5:373–384.

[68] de Bodt E, Cottrell M, Verleysen M (2002) Statistical tools to assess the reliability of self-organizing maps. *Neural Netw* 15:967–978.

[69] Miikkulainen R, Kohonen T, Mäkisara K, Simula O, Kangas J (1991) in *Proceedings of ICANN91, International Conference on Artificial Neural Networks, volume I* (North-Holland, Amsterdam) 415–420.

[70] Kohonen T, Hari R (1999) Where the abstract feature maps of the brain might come from. *Trends Neurosci* 22:135–139.

[71] Elman J (1990) Finding structure in time. *Cogn Sci* 14:179–211.

[72] Euliano N, Principe J (1996) Spatio-temporal self-organizing feature maps. *IEEE International Conference on Neural Networks* 4:1900–1905.

[73] Carpinteiro OAs (1999) A hierarchical self-organizing map model for sequence recognition. *Neural Processing Letters* 9:209–220.

[74] Rauber A, Merkl D, Dittenbach M (2002) The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. *IEEE Trans Neural Netw* 13:1331–1341.

[75] Titze IR (1994) *Principles of Voice Production* (Prentice Hall, Englewood Cliffs, NJ).

[76] Titze IR (2008) Nonlinear sourcefilter coupling in phonation: theory. *J Acoust Soc Am* 123:2733–2749.

[77] de Boysson Bardies B, Vihman M (1991) Adaptation to language: evidence from babbling and first words in four languages. *Lang* 67:297–319.

[78] de Boysson-Bardies B, Halle P, Sagart L, Durand C (1989) A crosslinguistic investigation of vowel formants in babbling. *J Child Lang* 16:1–17.

[79] Blakemore S (2010) The developing social brain: implications for education. *Neuron* 65:744–747.

[80] Grossmann T, Johnson MH (2007) The development of the social brain in human infancy. *Eur J Neurosci* 25:909–919.

[81] Gros-Louis J, West MJ, Goldstein MH, King AP (2006) Mothers provide differential feedback to infants prelinguistic sounds. *Int J Behav Dev* 30:509–516.

[82] Papoušek M, Papoušek H (1989) Forms and functions of vocal matching in interactions between mothers and their precanonical infants. *First Lang* 9:137–157.

[83] Lewis JM, Deák GO, Jasso H, Triesch J (2010) in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, eds Ohlsson S, Catrambone R (Cognitive Science Society, Austin, TX) 278–283.

[84] Domjan M (2010) *The principles of learning and behavior* (Wadsworth, Belmont, CA).

[85] Vihman MM (1993) Variable paths to early word production. *J Phon* 21:61–82.

[86] Guenther FH, Ghosh SS, Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang* 96:280–301.

[87] Yoshikawa Y, Asada M, Hosoda K, Koga J (2003) A constructivist approach to infants vowel acquisition through mother-infant interaction. *Conn Sci* 15:245–258.

[88] Westermann G, Miranda ER (2004) A new model of sensorimotor coupling in the development of speech. *Brain Lang* 89:393–400.

[89] Heintz I, Beckman M, Fosler-Lussier E, Ménard L (2009) in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH).*

[90] Warlaumont AS, Westermann G, Oller DK (2011) in *AISB 2011 Computational Models of Cognitive Development*, eds Kazakov D, Tsoulas G (Society for the Study of Artificial Intelligence and the Simulation of Behaviour, York) 8–12.

[91] Oudeyer P (2005) The self-organization of speech sounds. *J Theor Biol* 233:435–449.

[92] Guenther FH, Hampson M, Johnson D (1998) A theoretical investigation of reference frames for the planning of speech movements. *Psychol Rev* 105:611–633.

[93] Perkell JS, et al. (2007) A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *J Phon* 28:233–272.

[94] Max L, Guenther F, Gracco V, Ghosh S, Wallace M (2004) Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: a theoretical model of stuttering. *Contemp Issues Commun Sci Disord* 31:105–122.

[95] Callan DE, Kent RD, Guenther FH, Vorperian HK (2000) An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *J Speech Lang Hear Res* 43:721–736.

[96] Kanda H, Ogata T, Takahashi T, Komatani K, Okuno H (2009) in *2009 IEEE International Conference on Robotics and Automation (Kobe)* 4438–4443.

[97] Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. *Cogn* 21:1–36.

[98] Boersma P, Wennink D (2010). *Praat: Doing phonetics by computer (Version 5.1.31)* (http://www.praat.org).

[99] Boersma P (1998) *Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives* (Holland Academic Graphics, The Hague).

[100] Molina-Luna K, et al. (2009) Dopamine in motor cortex is necessary for skill learning and synaptic plasticity. *PLoS One* 4:e7082.

[101] Kohonen T (1990) The self-organizing map. *Proc IEEE* 78:1464–1480.

[102] Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

[103] Barto AG (1995) in *Models of Iinformation Processing in the Basal Ganglia*, eds Houk JC, Davis J, Beiser D (MIT Press, Cambridge, MA) 215–232.

[104] Izhikevich EM (2006) Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb Cortex* 17:2443–2452.

[105] Fagan MK, Iverson JM (2007) The influence of mouthing on infant vocalization. *Infancy* 11:191–202.

[106] Yang B (1996) A comparative study of American English and Korean vowels produced by male and female speakers. *J Phon* 24:245–261.

[107] Goldstein MH, King AP, West MJ (2003) Social interaction shapes babbling: testing parallels between birdsong and speech. *Proc Natl Acad Sci USA* 100:8030–8035.

[108] Ishizuka K, Mugitani R, Kato H, Amano S (2007) Longitudinal developmental changes in spectral peaks of vowels produced by japanese infants. *J Acoust Soc Am* 121:2272–2282.

[109] Zahorian SA, Jagharghi AJ (1993) Spectral-shape features versus formants as acoustic correlates for vowels. *J Acoust Soc Am* 94:1966–1982.

[110] Ito M, Tsuchida J, Yano M (2001) On the effectiveness of whole spectral shape for vowel perception. *J Acoust Soc Am* 110:1141–1149.

[111] American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR* (American Psychiatric Association, Arlington, VA).

[112] Sheinkopf SJ, Mundy P, Oller DK, Steffens M (2000) Vocal atypicalities of preverbal autistic children. *J Autism Dev Disord* 30:345–354.

[113] Oller DK, et al. (2010) Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proc Natl Acad Sci USA* 107:13354–13359.

[114] Peppé S, McCann J, Gibbon F, OHare A, Rutherford M (2007) Receptive and expressive prosodic ability in children with high-functioning autism. *J Speech Lang Hear Res* 50:1015–1028.

[115] Paul R, Orlovski SM, Marcinko HC, Volkmar F (2009) Conversational behaviors in youth with high-functioning ASD and asperger syndrome. *J Autism Dev Disord* 39:115–125.

[116] Anderson DK, et al. (2007) Patterns of growth in verbal abilities among children with autism spectrum disorder. *J Consult Clin Psychol* 75:594–604.

[117] Dawson G, et al. (2004) Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Dev Psychol* 40:271–283.

[118] Warren S, et al. (2010) What automated vocal analysis reveals about the language learning environment of young children with autism. *J Autism Dev Disord* 40:555–569.

[119] Karmiloff-Smith A (1998) Development itself is the key to understanding developmental disorders. *Trends Cogn Sci* 2:389–398.

[120] Tamis-LeMonda CS, Bornstein MH, Baumwell L (2001) Maternal responsiveness and childrens achievement of language milestones. *Child Dev* 72:748–767.

[121] Yoder PJ, Warren SF (1999) Maternal responsivity mediates the relationship between prelinguistic intentional communication and later language. *J Early Interv* 22:126–136.

[122] Girolametto LE (1988) Improving the social-conversational skills of developmentally delayed children: an intervention study. *J Speech Hear Disord* 53:156–167.

[123] Jaffe J, Beebe B, Feldstein S, Crown C, Jasnow M (2001) Rhythms of dialogue in infancy: coordinated timing in development. *Monogr Soc Res Child Dev* 66:vii–viii,1–132.

[124] Greenwood CR, Thiemann-Bourque K, Walker D, Buzhardt J, Gilkerson J (2010) Assessing childrens home language environments using automatic speech recognition technology. *Commun Disord Q* 32:83–92.

[125] Siller M, Sigman M (2008) Modeling longitudinal change in the language abilities of children with autism: parent behaviors and child characteristics as predictors of change. *Dev Psychol* 44:1691–1704.

[126] Amorosa H (1992) in *Nonverbal Vocal Communication: Comparative and Developmental Approaches*, eds, Papoušek H, Jürgens U, Papoušek M (Cambridge University Press, Cambridge).

[127] Markus J, Mundy P, Morales M, Delgado CEF, Yale M (2000) Individual differences in infant skills as predictors of child-caregiver joint attention and language. *Soc Dev* 9:302–315.

[128] Landa R, et al. (1992) Social language use in parents of autistic individuals. *Psychol Med* 22:245–254.

[129] Mundy P, Neal AR (2001) Neural plasticity, joint attention, and a transactional social-orienting model of autism. *Int Rev Res Ment Retard* 23:139–168.

[130] Klin A (1991) Young autistic childrens listening preferences in regard to speech: a possible characterization of the symptom of social withdrawal. *J Autism Dev Disord* 21:29–42.

[131] Hart B, Risley TR (1995) *Meaningful Differences in the Everyday Experience of Young American Children* (Brooks, Baltimore).

[132] Keller H, Lohaus A, Völker S, Cappenberg M, Chasiotis A (1999) Temporal contingency as an independent component of parenting behavior. *Child Dev* 70:474–485.

[133] Yoder P, Stone WL (2006) A randomized comparison of the effect of two prelinguistic communication interventions on the acquisition of spoken communication in preschoolers with ASD. *J Speech Lang Hear Res* 49:698–711.

[134] Council NR (2001) *Educating Children with Autism* (National Academy Press, Washington, DC).

[135] Marwan N, Romano MC, Thiel M, Kurths J (2007) Recurrence plots for the analysis of complex systems. *Phys Rep* 438:237–329.

[136] Dale R, Warlaumont AS, Richardson DC (2011) Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *Int J Bifurcat Chaos* 21:1153–1161.