TR-H-309

Robustness of an Auditory-to-Articulatory Mapping for Vowel Production by the DIVA Model to Subsequent Developmental Changes in Vocal Tract Dimensions

Daniel E. CALLAN (ATR-HIP/ATR-I), Kiyoshi HONDA, Shinobu MASAKI (ATR-HIP/ATR-I), Ray D. KENT (Univ. Wisconsin-Madison), Frank H. GUENTHER (Boston Univ.) and Houri K. VORPERIAN (Univ. Wisconsin-Madison)

2001.2.7

ATR人間情報通信研究所
〒619-0288　京都府相楽郡精華町光台2-2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Telephone: +81-774-95-1011
Fax : +81-774-95-1008

Robustness of an Auditory-to-Articulatory Mapping for Vowel Production by the DIVA Model to Subsequent Developmental Changes in Vocal Tract Dimensions

Daniel E. Callan[ab], Kiyoshi Honda[a], Shinobu Masaki[ab], Ray D. Kent[c], Frank H. Guenther[d] and Houri K. Vorperian[c]

[a] ATR Human Information Processing Research Laboratories
[b] ATR-I Brain Activity Imaging Center
[c] University of Wisconsin-Madison
Department of Communicative Disorders
[d] Boston University
Department of Cognitive and Neural Systems

Address Correspondence to:
Daniel E. Callan
ATR-I Brain Activity Imaging Center
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Tel +81-774-95-1050 (desk) Fax +81-774-95-2647
e-mail: dcallan@isd.atr.co.jp

Abstract

There are considerable changes in the size, shape, and the corresponding acoustical properties of the vocal tract throughout the course of development. It is necessary for a model of speech production either to adapt to these changes or to be robust with respect to them. This study explores the robustness of an auditory-to-articulatory directional map to drive vowel production during the course of development. A robust mapping is likely to be better able to accommodate for developmental alterations in muscle and sensory responses by means of co- registration within the same reference space for planning articulation than would be a mapping that is continually updated throughout development. The robustness of an auditory-to-articulatory map is investigated by using a modified version of the DIVA neural network model [Guenther et al., Psych. Rev. (1998)] in which the dimensions of the articulatory model are modified to reflect developmental changes in the dimensions of the vocal tract. Experiments using two implementations of the DIVA model were conducted: a modifiable and a static implementation. In the modifiable implementation the weights of the auditory-to-articulatory directional map are allowed to adapt throughout development by using auditory feedback as a training signal. In the static implementation weights learned during early development are used to test production performance throughout development. The results of the simulations show that, for most vowels, formant values characteristic of child speech, are produced by both the modifiable and static implementations. The performance of the static implementation demonstrates the robustness of an auditory-to-articulatory directional map to drive vowel production throughout development. This is an important finding because it demonstrates that a robust auditory-to-articulatory mapping could plausibly be used by the neural control system for speech production. Furthermore, a robust auditory-to-articulatory mapping may be able to accommodate for developmental alterations in muscle and sensory responses by modification of articulatory maps regarding changes in muscle length and innervation patterns as well as orosensory information based on adaptive feedback of self produced speech.

Introduction

Despite considerable research concerning the acquisition and production of speech, the target reference space used by the neural control system to drive articulation is still unknown. There are two main theoretical positions concerned with the possible target reference space. One approach maintains that the goal of speech production is to move the articulators to reach gestural targets defined by degree of constriction at various locations in the vocal tract (gesture approach) (Saltzman & Munhall, 1989). Another approach maintains that the goal of speech production is to move the articulators to reach auditory/acoustic based targets (auditory approach) (Perkell, Matthies, Lane, Guenther, Wilhelms-Tricarico, Wozniak, & Guiod, 1997; and Guenther, Hampson, & Johnson, 1998).

When assessing the efficacy of various target reference spaces for speech production, it is important to determine plausible means by which the mapping from the target reference space to the reference space responsible for articulator movement (articulator reference space) can be acquired. The speech production system must establish a mapping that is able to accomplish the functional goal at hand working in concert with biophysical constraints and taking into account the contextual environment of the motor control system. The mapping between a target reference space and an articulator reference space must be learned under conditions in which the size, shape, and relationship between the various articulators is changing during the course of development. The infant vocal tract is not just a miniature version of the adult vocal tract (Kent, 1999). Differential changes in the anatomical structures that occur during development result in varying degrees of influence with respect to the acoustic properties of the vocal tract.

In current implementations of models of speech production based on the gestural approach, the mapping between the target reference space and the articulator reference space is hand set by the experimenters. As has been stated by Guenther et al., (1998), it is unclear what can serve as a plausible supervisory training signal, apart from the acoustics, that could be used to acquire a mapping between the constriction target reference space and the articulator reference space. A plausible training signal that may be used to acquire a mapping between an auditory target reference space and an articulator reference space is auditory feedback of self produced speech (Guenther et al., 1998).

One model that utilizes auditory feedback of self produced speech to acquire a mapping between an auditory target reference space and an articulator reference space is the DIVA neural network model (Guenther et al., 1998) (a detailed description of the DIVA model is given below). For other neural network models that use an auditory reference space, see Bailly (1997) and Markey (1994). The DIVA model accomplishes the goals of speech production by establishing a learned mapping between the direction in which to move the articulators and the corresponding direction in an auditory target reference space needed to accomplish some functional goal (producing a vowel). The DIVA model consists of a training (babbling) phase and a performance phase. It has been demonstrated that the DIVA neural network model is able to produce 11 English vowels with fairly good performance throughout the course of development despite changes in the size and shape of the vocal tract by means of using auditory feedback as a training signal to continuously adapt weights responsible for the mapping between the auditory target reference space and the articulator reference space (Callan, 1998; Callan et al., 2000). It has also been demonstrated that the DIVA neural network model is able to show flexible motor equivalent speech production throughout the course of development (Callan, 1998; Callan et al., 2000).

It is important that the mapping between the auditory target reference space and the articulator reference space be fairly robust with respect to morphological change during development. If the mapping between the auditory target reference space and the articulator reference space is continually updated regardless of performance (as it was for the simulations carried out in Callan et al., 2000) there will be a great deal of fluctuation in the articulatory configurations and corresponding formant values produced. The increased degree of fluctuation resulting from adaptation of the weights is caused by the system jumping in and out of different local minima depending on the nonlinear nature of the solution space. A mapping between the auditory target reference space and the articulator reference space that is robust with respect to morphological change (the weights do not need to be continuously adapted throughout the course of development) is less likely to show as great a deal of fluctuation during development. Changes in the size and shape of the articulators and associated structures results in alteration in muscle innervation patterns and sensory fields. It is unlikely that a reference space that is continuously changing could accommodate for developmental alterations in muscle and sensory responses by means of co- registration with the same reference space for planning articulation. However, a somewhat stable target reference space (auditory speech target space) allows for modification of articulatory maps regarding changes in muscle length and innervation patterns as well as orosensory information based on adaptive feedback of self produced speech.

The purpose of this article is to explore within the framework of an implementation of the DIVA speech production model the extent to which mappings learned during early development can drive speech production in later development despite the considerable degree of morphological change in the structures involved with speech production. Two implementations of the DIVA model are tested. In the first implementation (modifiable) the weights of the phoneme-to-auditory map and the auditory-to-articulatory directional map are allowed to adapt using decay and learning rate parameters that change during the course of development in correspondence with the values of neural plasticity given by Huttenlocher (1993). In the second implementation (static) the weights of the maps for a particular vowel are fixed at the value of the first time step at which the target regions in auditory space defined by the weights of the phoneme-to-auditory map are learned in the modifiable implementation. As described below the phoneme-to-auditory map for a vowel is learned

when random articulator movements produce formant ratio values that correspond to the vowel target regions defined within the speech recognition system (see below for details).

In the implementations of the DIVA model carried out in this paper, direct auditory feedback is used to determine the position of the articulators in auditory space during the performance stage (see figure 1). This is unlike the implementations of the DIVA model in Callan, et al., (2000) in which an articulatory-to-auditory map (Forward Model) is learned between the articulator position vector and the planning position vector. Which is then used to determine the position of the articulators in auditory space during the performance stage. The forward model was left out of the implementations carried out in this paper in order to focus on the issue of whether an auditory-to-articulatory directional map learned during early development is robust with respect to subsequent changes in the size and shape of the articulators that occur throughout development. The addition of a forward-model that learns an articulatory-to-auditory map throughout development would be a trivial matter and would not alter the performance of the simulations as they were tested.

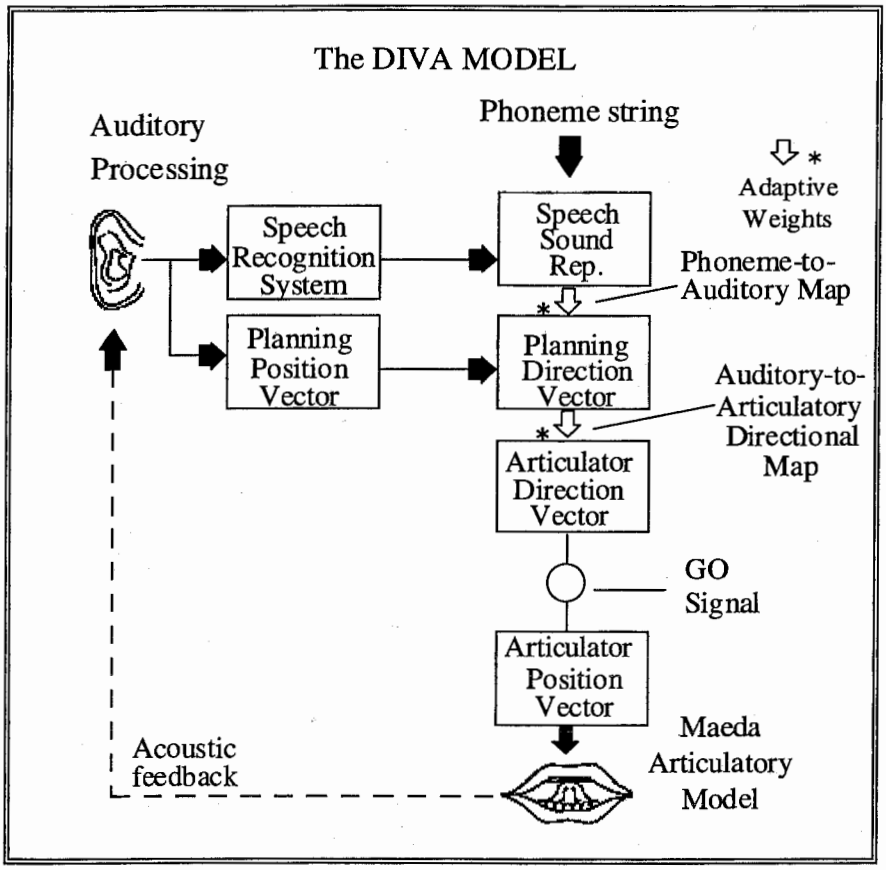


Figure 1: Overview of the DIVA model. Boxes are input or output representations and arrows are weights. Solid arrows represent the passage of the representation with a weight of one. The unfilled arrows represent learned maps, the value of the weights, for which, are determined during training. The articulatory-to-auditory map (forward model) was left out of the simulations in this study (see text for details). Rep. = Representation.

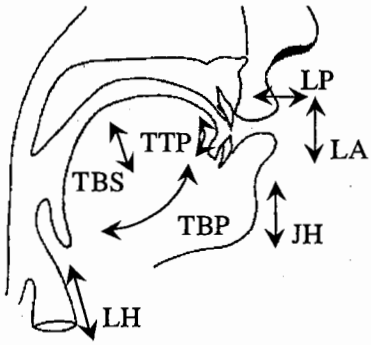
The rest of the paper focuses on the modifiable and static implementations of the DIVA model. First a description of the components of the DIVA model used in these implementations is given. This is followed by a description of the training and performance phases of the DIVA model. A description of the developmental measures used to modify the dimensions of the articulatory system of the DIVA model is then given. Next, the training of the DIVA model is described. This is followed by the results and discussion section in which the performance of the modifiable and static implementations are evaluated with respect to

vowel formant values, articulator configuration patterns, and vocal tract constriction patterns produced throughout development. The paper concludes with a discussion of how further work can include an investigation of trajectories for the two implementations, as well as a discussion of how the model could be extended to include consonants as well as vowels.

Components of the DIVA Model

Maeda Articulatory Model

The Maeda (1990) model of speech production is a shape factor model based on cineradiographic and labiofilm data of French adult speakers. The seven parameters of the Maeda (1990) model control the movement directions of the various articulators (see figure 2). They can be individually shifted between –3 and +3 standard deviations to derive different vocal tract shapes. It is important to note that the Maeda (1990) model is a two-dimensional model working in the midsagital plane. The cross sectional area is determined using a scaling factor (Maeda, 1990). The area function of the vocal tract shape is used to determine the acoustic output of the model (this is the source of auditory feedback that is used to determine formant values for the speech recognition system of the DIVA model). Formant values are determined by a peak-picking algorithm working on the area transfer function in the frequency domain. As will be discussed further below, the vocal tract dimensions of the Maeda (1990) model were altered during the course of training to simulate the developmental restructuring of the speech production system.



Maeda Articulation Parameters	
JH	Jaw Height
TBP	Tongue Body Position
TBS	Tongue Body Shape
TTP	Tongue Tip Position
LA	Lip Aperture
LP	Lip Protrusion
LH	Larynx Height

Figure 2: Direction of movement for the seven Maeda (1990) articulation parameters. The seven articulatory parameters can be individually shifted between –3 and +3 standard deviations to derive different vocal tract shapes. Taken from Callan et al., (2000).

Speech Recognition System

The purpose of the speech recognition system is to determine if acoustic signals produced by the vocal tract are speech and if so determine the signals phonological content. The Miller (1989) auditory perceptual model is used in the current implementation. The Miller (1989) model uses formant ratios taking into account fundamental frequency. It is hoped that this will allow for sufficient normalization between the acoustic properties of child and adult speech. The vowel target regions are defined by the first three dimensions *R1*, *R2*, and *R3* of the Miller (1989) model (see figure 5b for calculation of values). This speech recognition system does not take into account the acquisition of the perceptual targets for each of the vowels. The target regions for each vowel are handset by the experimenter. Vowel target regions are based on formant values taken from Peterson and Barney (1952) with the exception of [e] and [o] taken from Hillenbrand, Getty, Clark, & Wheeler (1995). The DIVA model uses auditory feedback from the Maeda (1990) articulation model and determines whether the ratio of the formant values, in Miller (1989) space, fall within one of the vowel target regions. The output of the speech recognition system identifies which of the 11 vowels is recognized or is zero corresponding to no vowel recognized.

Speech Sound Representation

The speech sound representation consists of 11 nodes each one corresponding to one of the 11 target English vowels (i = 1 to 11) to be learned by the model. A node can be activated by the output of the speech recognition system or directly by the experimenter during the performance phase.

S_i = {1 : if recognition system hears ith vowel or if directly set by experimenter
0 : otherwise

The activated node has an output value of one and all other nodes are set to zero. Each node has associated with it antagonistically paired sets of weights for each of the dimensions in auditory space (in this case the dimensions are *R1*, *R2*, and *R3*).

Planning Position Vector

The planning position vector represents the present state of the vocal tract within the scaled auditory perceptual space (reference space for movement planning). Determining the present state of the vocal tract in auditory space is accomplished by means of auditory feedback (figure 1). In this study the auditory space consists of formant ratios *R1*, *R2*, and *R3* (Miller, 1989). The vector nodes are composed of antagonistically paired variables (r_{j+} and r_{j-}) for each of the formant ratios (R_j; where j = 1 to 3) in Miller (1989) auditory space normalized such that they fall within the interval [0,1] and their sum is equal to 1 (equation 1).

$$r_{j+} = \frac{\log(R_j) - \log(R_{j\,min})}{\log(R_{j\,max}) - \log(R_{j\,min})} \qquad (1)$$

$$r_{j-} = 1.0 - r_{j+}$$

The $R_{j\,min}$ and $R_{j\,max}$ correspond to the minimum and maximum values for the jth dimension in auditory space that can be encountered during the babbling phase of learning.

Planning Direction Vector

The planning direction vector represents the movement direction in auditory space needed to achieve the current vowel target (see figure 1). The planning direction vector nodes are composed of antagonistically paired variables (d_{j+} and d_{j-}) that are determined by subtracting the planning position vector nodes (r_{j+} and r_{j-}) from the learned weights (z_{ij+} and z_{ij-}) that make up the phoneme-to-auditory map (see below) representing the vowel (s_i) target regions in auditory space (equation 2).

$$d_{j+} = \sum_i s_i z_{ij+} - r_{j+} \qquad (2)$$

$$d_{j-} = \sum_i s_i z_{ij-} - r_{j-}$$

Articulator Direction Vector

The articulator direction vector represents the movement direction of the seven Maeda (1990) parameters in articulation space corresponding to the direction in auditory space needed to reach the vowel target region (see figure 1). The articulator direction vector nodes are composed of antagonistically paired variables (a_{k+} and a_{k-}) for each of the seven (k = 1 to 7) Maeda articulator parameters. The articulator direction vector node values (a_{k+} and a_{k-}) are determined by multiplying the planning direction vector nodes (d_{j+} and d_{j-}) by the learned weights (w_{j+k+}, w_{j+k-}, w_{j-k+}, w_{j-k-}) that make up the auditory-to-articulatory directional map (equation 3).

$$a_{k+} = \sum_j [d_{j+}]^+ w_{j+k+} + \sum_j [d_{j-}]^+ w_{j-k+} \qquad (3)$$

$$a_{k-} = \sum_j [d_{j+}]^+ w_{j+k-} + \sum_j [d_{j-}]^+ w_{j-k-}$$

w_{j+k+} is the weight projecting from the j+th planning direction vector node to the k+th articulator direction vector node (with analogous definitions for the various +, – combinations) and $[x]^+$ is a rectification function such that $[x]^+ = 0$ for $x < 0$ and $[x]^+ = x$ for $x >= 0$ (Guenther, 1995; Johnson, 1998). During the training (babbling) phase the articulator direction vector nodes are randomly activated to produce movements of the articulators (see below).

GO Signal

The value of the articulator direction vector is passed through a multiplicative gating function that controls movement speed (equation 4), the GO Signal (G, varying between 0 for minimum speaking rate and 1 for maximum speaking rate). In the simulations conducted here, a GO signal of 1.0 is used during both the training and testing phase.

$$V_k = G\,[a_{k+} - a_{k-}] \tag{4}$$

Articulator Position Vector

The articulator position vector represents the position of the seven articulation parameters determined by integrating the activity of the articulator direction vector after it has been passed through the GO signal. The articulator position vector is used to set the position of each of the Maeda articulator parameters between –3 and +3 standard deviations.

Training (Babbling) Phase of the DIVA Model

During the babbling phase, the activity of the nodes composing the articulator direction vector are randomly set thus producing random positions for each of the Maeda (1990) articulator parameters. The acoustic consequence of the resulting vocal tract shape is determined and used in this study as feedback to train the two learned maps (the auditory-to-articulatory directional map and the phoneme-to-auditory map) of the DIVA model. Training the maps takes place during a two-stage babbling phase. During the first stage the auditory-to-articulatory directional map is learned by using auditory feedback as a training signal utilizing hyperplane radial basis function neural networks (HRBF). In the second stage the phoneme-to-auditory map is learned by correlating activation of the speech sound map with movement direction in auditory planning space.

Auditory-to-Articulatory Directional Map

The training of the auditory-to-articulatory directional map between movement directions in the three-dimensional auditory planning space (planning direction vector) and the corresponding movement directions in the seven-dimensional articulator space (articulator direction vector) occurs during the first stage of babbling. The weights (w_{j+k+}, w_{j+k-}, w_{j-k+}, w_{j-k-}) that compose this map are learned by HRBFs using the difference (articulator direction vector) between the babbled movement and the predicted movement (based on the auditory signal) as an error signal. Learning occurs regardless of whether the babbled movement falls within one of the phoneme target regions. Training continues for a pre-specified number of iterations given by the experimenter. Learning (ε_1) and decay (α_1) rate parameters control the degree to which the weights can be changed on each iteration during training (see Guenther, 1998, for derivation of RBF learning rules). With this kind of mapping, the configuration used to produce a desired set of formants will depend on factors such as starting configuration and externally imposed constraints on the articulators.

Phoneme-to-Auditory Map

The training of the phoneme-to-auditory map between the speech sound representation and the planning direction vector occurs during the second stage of babbling. The mapping represents the learned target regions in auditory space for each of the vowels. During the babbling phase under conditions in which a random articulator position produces

an auditory signal that falls within the range of one of the vowel target regions (as identified by the speech recognition system) the antagonistically paired weights (z_{ij+}, z_{ij-}) between the node (s_i) in the speech sound representation (corresponding to the vowel recognized) and the nodes (d_{j+}, d_{j-}) in the planning direction vector are adapted in accordance with equation 5. Training continues for an experimenter determined pre-specified number of iterations.

$$d/dt\ z_{ij+} = \varepsilon_2 s_i(\alpha_2 z_{ij+} - [d_{j+}]^+) \qquad (5)$$
$$d/dt\ z_{ij-} = \varepsilon_2 s_i(\alpha_2 z_{ij-} - [d_{j-}]^+)$$

Where ε_2 is a learning rate parameter, α_2 is a learning decay parameter, and $[x]^+$ is a rectification function the same as described above (Guenther, 1995; Johnson, 1998).

The Performance Phase of the DIVA Model

During the performance phase, production of one of the 11 English vowel targets specified by the user, is accomplished in the following manner: First, the node in the speech sound representation corresponding to the vowel target is activated (see figure 2). This in turn activates the learned weights for the vowel target between the speech sound representation and the planning direction vector. These weights are subtracted from the values of the current state of the vocal tract in auditory reference space (given by the planning position vector) to give the value of the planning direction vector (which represents the desired movement direction in auditory planning space needed to reach the vowel target). The planning direction vector in auditory reference space is then transformed into a set of articulator velocities (articulator direction vector) by means of multiplying its node values by the learned auditory-to-articulatory directional map to obtain the values of the nodes in the articulator direction vector. The node values of the articulator direction vector are then passed through a multiplicative gating function controlling movement speed (GO signal) and then integrated to produce the position values of the seven Maeda (1990) articulator parameters (articulator position vector). The position of the seven Maeda (1990) parameters is used to determine the area function of the vocal tract. The area function is used to determine the acoustic output of the model (this is the source of auditory feedback that is used to determine formant values for the speech recognition system of the DIVA model). The model iterates through this process moving the articulators closer and closer to the vowel target in auditory reference space. As the production of the model gets closer to the vowel target, the magnitude of the planning direction vector becomes smaller leading to a slowing down of articulator movement and a halting when the vowel target is reached or a certain number of iterations have passed. The trajectory of movement is carried out automatically based on the temporal dynamics of the model. There is no internal executive agency 'homunculus' that has a predetermined plan for producing the vowel target. For a more extensive discussion of the performance phase of the DIVA model see Guenther, 1995; 1998; Johnson, 1998).

Developmental Measures used to Modify the Dimensions of the Maeda Model

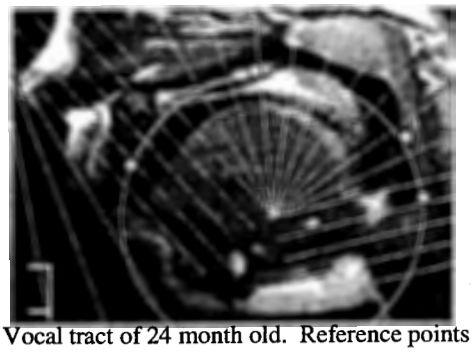
This Section is Adapted from Callan et al., (2000)

In order to simulate the developmental restructuring of the speech production system, the vocal tract dimensions of the Maeda (1990) model (see figure 2) were altered during the course of training. Measures of vocal tract dimensions were approximated from mid-sagittal MRI slices of four males at ages 3, 7, 15, 24, 36, and 45 months; 216 months represents the adult Maeda articulation dimensions. Most of the measurements were from a single individual (15, 24, and 36 months). The other three ages (3, 7, and 45 months) were from three different individuals that were normalized by differences in the length of the vocal tract based on images collected at one of the ages (15, 24, or 36 months) of the above mentioned individual. The MRI scans of the children were acquired as part of a medical examination at the University of Wisconsin–Madison Hospital. The children were sedated during the scans.

The ailment from which each of the children suffered was considered not to influence morphological development of the speech production system.

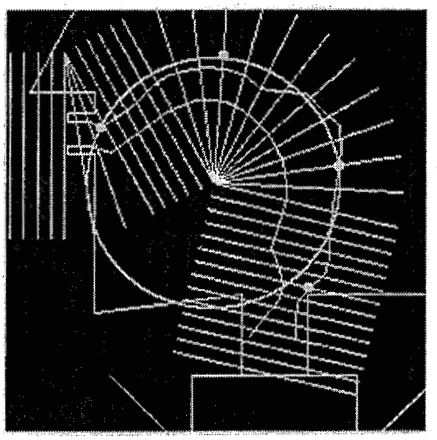
The semipolar coordinate gridlines (shown in figure 3) were made using programs developed by Mark Tiede at ATR that run on the public domain image analysis software Scion Image (http://www.scioncorp.com/). Coordinate grids are spaced by 0.5 cm in the two linear dimensions and by 11.2 degrees in the polar region. It can be seen by comparison of figure 3a and 3b that there are considerable differences in the geometrical configuration of the four reference points that are used to define the placement of the semipolar grid. This results in a different number and respective ratio of gridlines within each section during the course of development (see figure 4a-c).

Simipolar Gridlines of
24 Month Old



Vocal tract of 24 month old. Reference points and semipolar grid are shown in white.

Semipolar Gridlines of Maeda
Articulation Model



Vocal tract of of 216 month old. Based on coordinates given in the Maeda (1990) articulation model. Reference points and semipolar grid are shown in white.

Reference Points (shown in white) used to define the coordinates of the semipolar grid were placed at:
- The Bottom of the Alveolar Ridge.
- The Maxilla above the hard palate.
- The Rear of the Pharyngeal Wall.
- The Rear of the Pharyngeal Wall above the Aryepiglottic tissue.

Three regions are defined by the semipolar coordinates:
- Palatal-Dental Region (Linear)
- Velar Region (Polar)
- Pharynx Region (Linear)

Figure 3: Semipolar coordinate grids used to make measurements of mid-sagittal MRI slices of four males at ages 3, 7, 15, 24, 36, and 45 months. These measurements were later used to construct developmental curves for various features of the vocal tract. Taken from Callan et al., (2000).

Developmental curves used to modify the dimensions of the Maeda articulation model were approximated from measures taken from the MR images as well as data given in Kent & Vorperian (1995) (see figure 4a-c). Spline interpolation was used to generate data from 3 to 216 months of age at one-month intervals (216 months represents the adult Maeda articulator model dimensions). Points were added to produce smooth curves. It is important to note that the same ratio is used to calculate the cross-sectional diameter within each respective region (Palatal-Dental, Velar, and Pharynx). From the various developmental curves, the coordinate system defining the midsagittal outline of the vocal tract was modified to reflect the corresponding dimensions for each age. It is recognized that these developmental changes used to alter the dimensions of the Maeda (1990) articulatory model

are only gross approximations of actual developmental changes that occur in the structures involved with speech production. However, with respect to the objectives of this study (see above), the changes made in the dimensions of the Maeda (1990) articulatory model are believed to be adequate.

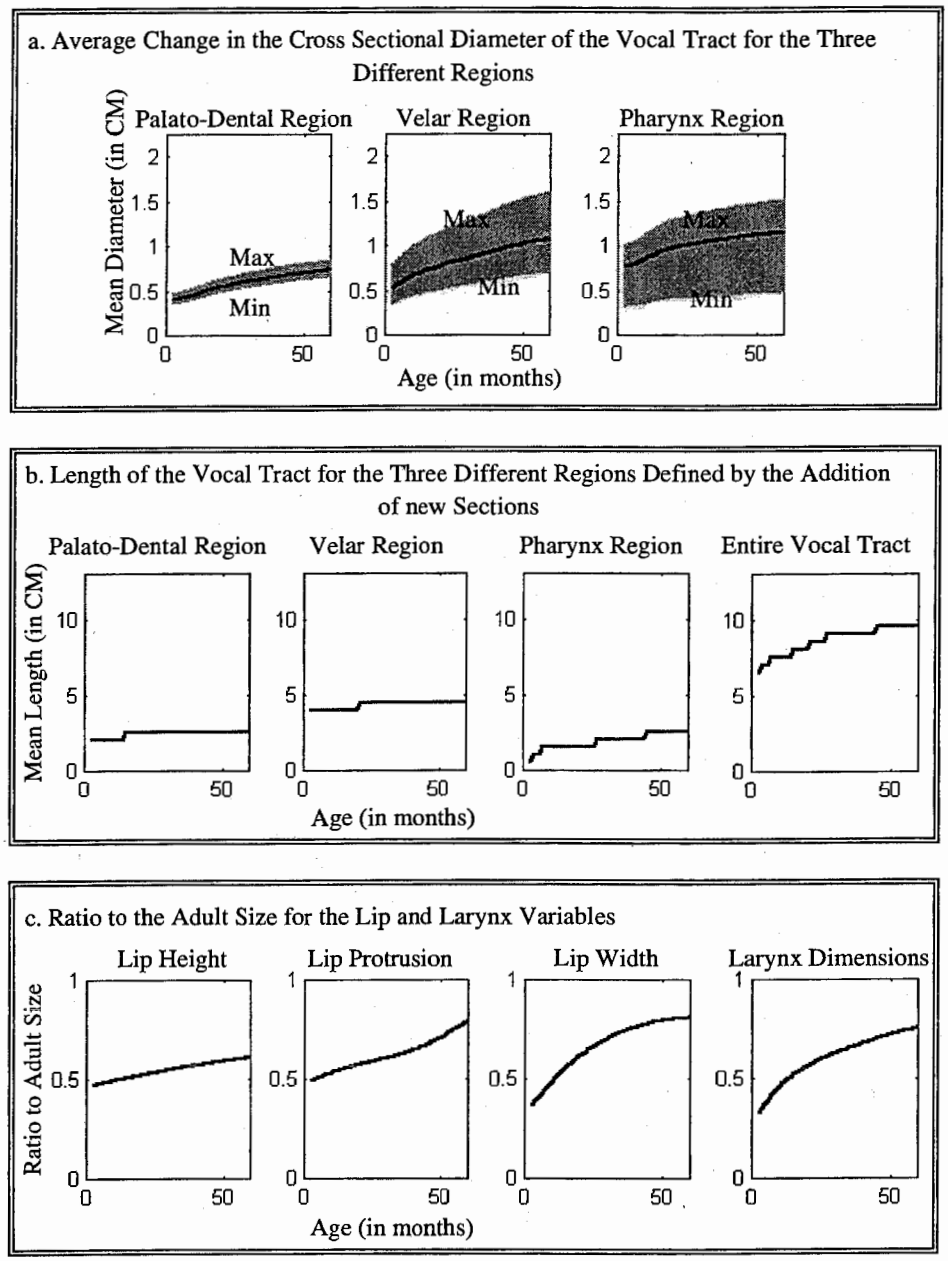


Figure 4a-c: Developmental curves used to modify the dimensions of the Maeda (1990) model. See text for details. Taken from Callan et al., (2000).

The factor patterns (control parameters) determining the respective shape resulting from manipulation of the seven Maeda (1990) articulatory parameters were not altered during the course of development. The factor patterns of the Maeda (1990) model are the principle components obtained from a statistical analysis of cineradiographic and labiofilm data, and as such they relate only indirectly to the positions of the articulators. It is unclear to what extent the control parameters (factor patterns) of the Maeda (1990) model agree with control parameters used by developing children. Despite this limitation, the original factor patterns of the Maeda (1990) articulation model were used because data to produce fully-fledged child vocal tract models was not available. Given changes in muscle innervation patterns of the

various structures involved with speech production that are known to occur during development (Kent, 1999) it is likely that the use of adult control parameters will limit the accuracy of the model when comparing it to articulatory patterns used by developing children. However, even though adult control parameters are used, it is still possible to demonstrate that an auditory-to-articulatory map learned early in development (static condition) is capable of producing vowels throughout subsequent development despite changes that occur in the size and shape of the vocal tract. It should be noted that although adult control parameters are used there are far fewer vocal tract sections (reflecting vocal tract length) in the child models developed here. A reduction in the number of sections limits the degree to which adult like movements can be made.

Training of the Neural Networks

The DIVA model was trained to produce 11 English vowels using vocal tracts from 12 to 60 months of age in steps of three months. In the modifiable implementation, the model was allowed to adapt its weights at each of the age steps. For each age step, 500 iterations were used to train the auditory-to-articulatory directional map and 1000 iterations were used to train the phoneme-to-auditory map (see above for details). In the static implementation, the model was tested at different age steps using weights that were learned during early development. The weights used for each vowel in the static implementation correspond to the time step of those of the modifiable implementation for which the phoneme-to-auditory map was learned. Learning occurs when the vowel produced falls within the target region defined in the speech recognition system. The implementations were trained using adult target formant ratio values, using the Miller (1989) transform, based on a fundamental frequency of 100 Hz. It is believed that children learn perceptual speech targets based on predominantly adult speech and are able to normalize their own speech to fall within the same perceptual space. There is substantial evidence that infants less than 6 months of age can normalize between child and adult speech (Kuhl, 1979; Kuhl, 1983; Kuhl & Meltzoff, 1996). In this model the Miller (1989) transform serves to normalize between the productions made by the model and the target regions based on the transform of adult formant values. Age appropriate fundamental frequency (as given by Kent, 1997; see figure 8 bottom) was used at each age step to evaluate the performance of the model. Only one initial random weight scheme was used to train the implementations. To ensure that the results are not spurious, it may have been better to compare results using several different initial random weight schemes. However, given the similarity to the results of Callan et al., (2000), in which a different random weight initialization was used, it is unlikely that the results are spurious. It should also be noted that in the Callan et al., (2000) study an articulatory-to-auditory map (forward model) was used, whereas in the study presented here direct auditory feedback is used to determine the position of the articulators in auditory space during the performance stage.

Results and Discussion

Testing the Performance of the DIVA Implementations

The production performance for both the modifiable and static implementations was evaluated for each of the vowels learned from 12 to 60 months of age at three-month intervals. Measures of performance include formant and ratio values, articulator configuration patterns, and vocal tract constriction patterns (area functions). The starting position for each of the vowels tested during the performance stage was defined by a neutral articulatory configuration pattern (the SD of all 7 Maeda articulatory parameters were set to 0.0 [see figure 2]). The corresponding neutral articulatory configuration pattern in auditory space is denoted by asterisks in formant and ratio space in figures 5 a-b. As one would expect, it can be seen that there is a general decrease in formant frequencies as the vocal tract increases in size with age (figure 6).

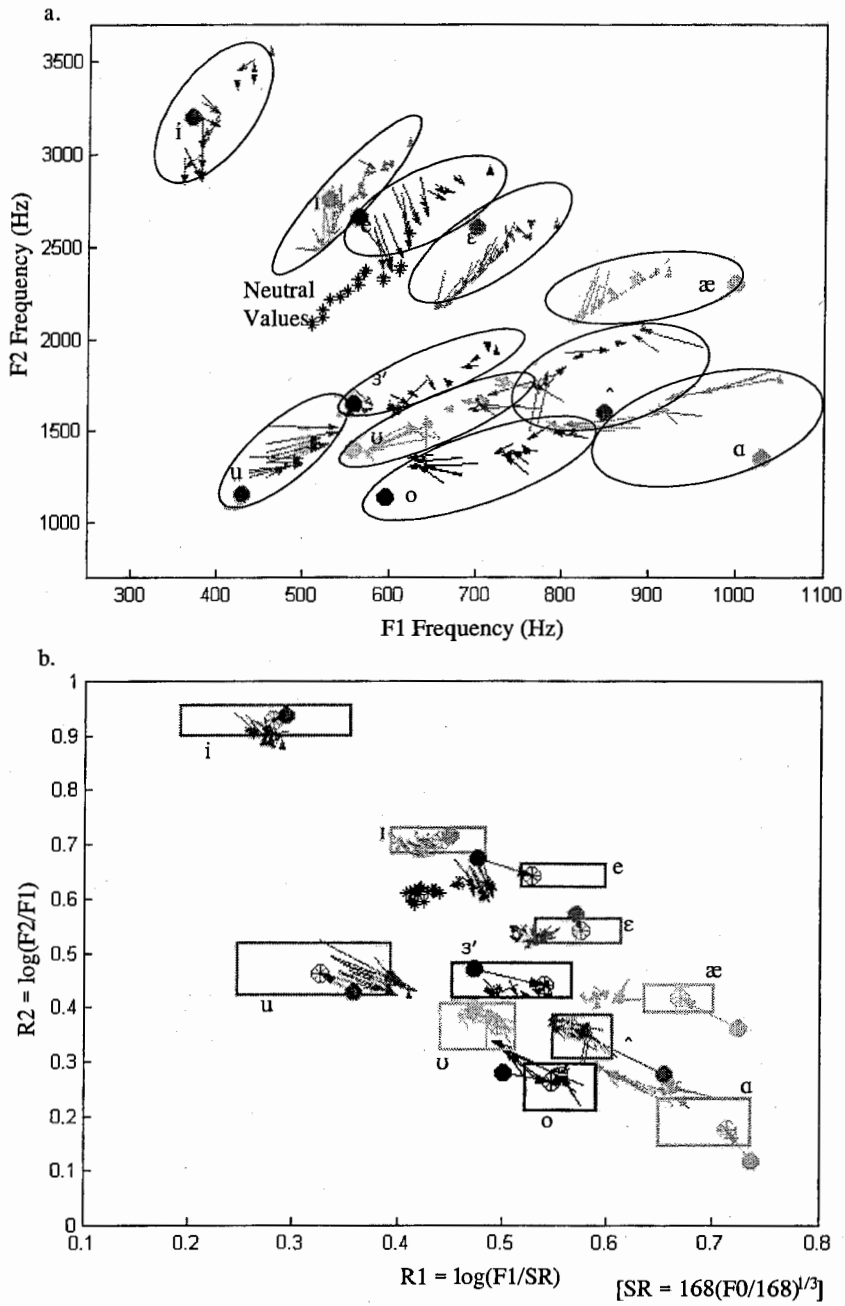


Figure 5a-b: a. Difference in the performance between the modifiable and static implementations of the DIVA model for each of the 11 vowels in F1 by F2 formant space for each of the stages of development, 3 to 60 months of age (the arrows point from the modifiable to the static implementation). The formant values corresponding to the neutral articulatory configuration are displayed as small black asterisks. The big circles represent mean child formant values taken from Peterson and Barney (1952) with the exception of [e] and [o] taken from Hillenbrand, Getty, Clark, & Wheeler (1995). Ellipses circle the main clustering for each vowel produced by the network as well as the child formant values (large circles). b. Difference in the performance between the modifiable and static implementations of the DIVA model for each of the 11 vowels in R1 by R2 ratio auditory target space from 3 to 60 months of age. The ratio values corresponding to the neutral articulatory configuration are displayed as small black asterisks. The rectangles represent the hyperplane target regions derived from using the Miller (1989) transform of formant values. The arrows projecting from the filled large circles to the empty large circles denote the difference between ratio values calculated from child formant values and adult formant values upon which the target regions are based. F0 = Fundamental Frequency.

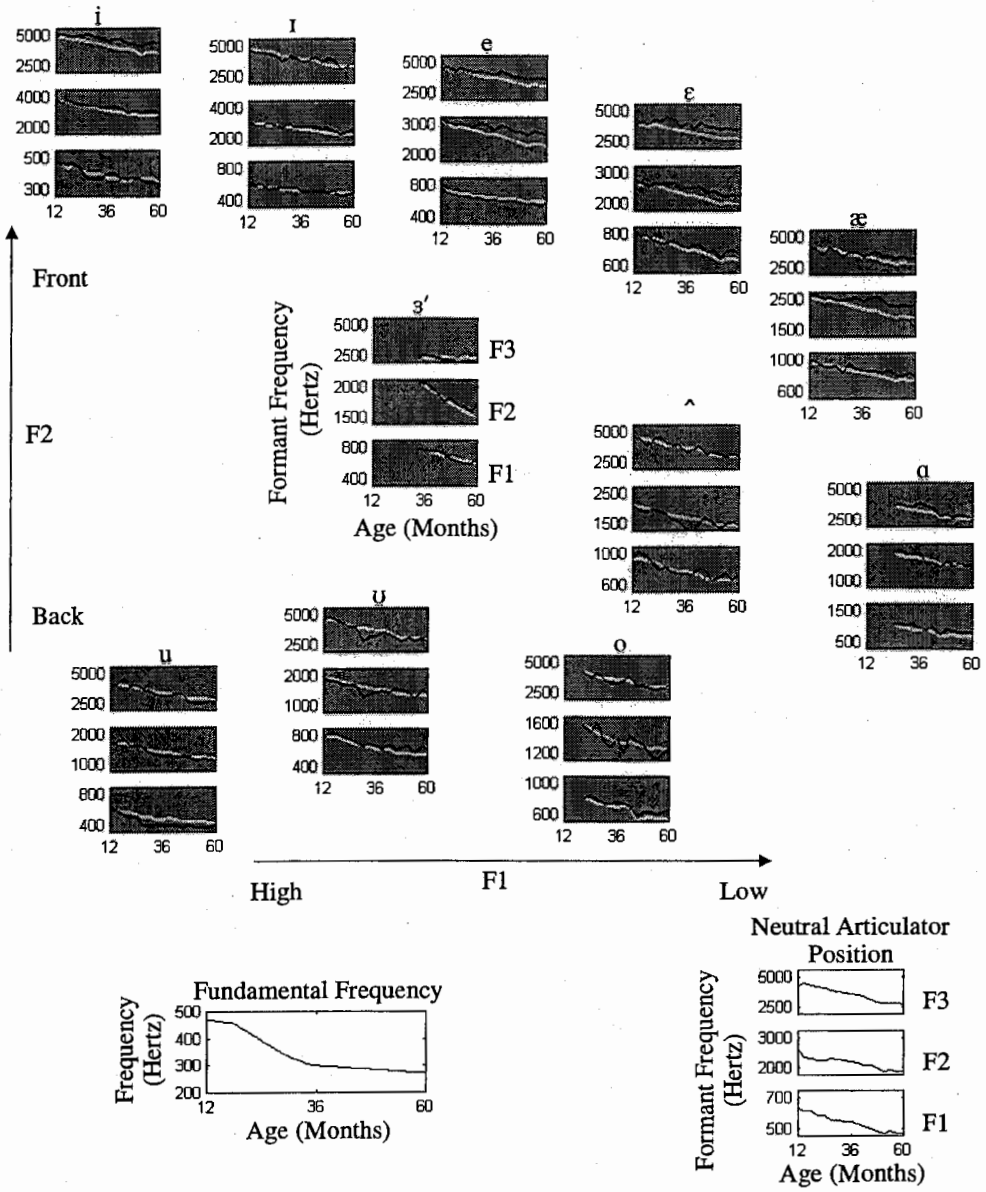


Figure 6: Formant frequencies in Hertz produced by the network for each of the 11 vowels from 3 to 60 months of age (the modifiable implementation is displayed in black and the static implementation is displayed in white). Also shown is a plot of the formant frequencies for the neutral articulator position from 3 to 60 months of age as well as the fundamental frequency value used during the performance phase for each step of development. Empty regions at the initial part of the plots indicate that the phoneme-to-auditory map for that vowel has not been acquired yet. Plot of vowels is made in F1 by F2 space.

Vowel Formant and Ratio Values Produced during the Course of Development

Figure 5 a-b display the difference (arrows) in formant and ratio values between the modifiable and the static implementations throughout the course of development in corresponding formant and ratio space. At 12 months of age (the first training step), 8 of the 11 vowels were learned ([i], [ɪ], [e], [ɛ], [æ], [ʊ], [u], [ʌ]). The remaining three vowels were learned at 21 months for the vowel [o], 24 months for the vowel [ɑ], and 33 months for the vowel [ɜ′]. This type of mastery is not usually seen in children until 36 months of age (Kent, 1992). As stated above the age step for which a vowel is learned is based on when the phoneme-to-auditory map for a particular vowel is learned. In total there were four static simulations, corresponding to freezing the weights at 12 months (during which 8 of the 11 vowels were learned), 21 months ([o]), 24 months ([ɑ]), and 33 months ([ɜ′]). The reason that

some vowels are learned earlier or later than others is dependent on the extent to which random articulator movements produces formant ratio values that fall within one of the target regions defined in the speech recognition system. Vowels that have a large range of articulator configurations that correspond to the target regions of the speech recognition system are learned early and those that don't are learned later. It is important to note that training for ten times the number of iterations did not increase the number of vowels learned at the initial 12-month age step. This suggests that there is a restriction as to what auditory-to-articulatory relationships can be learned based on differences in the size and shape of the articulatory system during development. One possible reason why some vowels are learned at such an early age by the model compared to children is that the targets are defined in the speech recognition system apriori, they do not have to be learned.

Formant values characteristic of child speech (Hillenbrand et al. 1995; Peterson & Barney, 1952; Lee, Potamianos, & Narayanan, 1999) are produced by both the modifiable and static implementations of the DIVA model for most of the vowels. In addition, both implementations also show a tight clustering of vowels in formant and ratio space with only minor overlap (see figure 5). The model shows some difficulty in reaching ratio targets for the vowels [e], [æ], [ɑ], and [ɝ] (the error for [ɝ] is mainly for R3, not shown in figure 5b). This is an interesting finding in that the vowels [e], [æ], and [ɝ] are acquired later in life by children (Kent, 1992). The vowel [ɑ] has been shown to be difficult to produce by the DIVA model even with adult vocal tract dimensions (Guenther et al., 1998) perhaps resulting from restrictions of the Maeda articulation model. One can see that for the vowels [e], [æ], [ɑ], and [ɝ] the Miller (1989) transform of child formant values are outside the vowel target regions based on adult formant values (see arrow between large filled and empty circles in figure 5b). This suggests that error in reaching ratio targets of some vowels may result from inaccuracies in normalizing between child and adult formant values. It should be noted that even though ratio targets were not met for some of the vowels, the formants produced by the model are characteristic of child speech (Hillenbrand et a., 1995; Peterson & Barney, 1952; Lee et al., 1999). It can be seen in figure 6 that the static implementation shows a much smoother decrease in formant frequencies with age than does the modifiable implementation. It should be noted that the performance of the modifiable implementation using auditory feedback to determine the planning position vector is similar to the implementation, as reported in Callan et al., (2000), in which a forward model is used to determine the planning position vector. Although the synthesized vowels produced by the model can be distinguished and identified by a human listener the quality is far from real speech produced by children.

Developmental Articulator Configuration Patterns and Vocal Tract Constriction Patterns

It has been noted that there is a fair degree of variability in speech production in children during the course of development (Green, 1998; Sharkey & Folkins, 1985). In the study reported here the extent to which these developmental changes can be accounted for by a static auditory-to-articulatory map formed early in development is explored. The articulatory configuration patterns of the seven Maeda (1990) parameters for the modifiable and the static implementations are displayed in figure 7 (also see figure 2 for the definition and movement pattern of the 7 Maeda (1990) parameters). Although the articulatory configuration patterns of the static implementation are far less dramatic than for the modifiable implementation, there is still a fair degree of change that occurs throughout the course of development (see figure 7).

Both the modifiable and the static implementations are able to alter their pattern of articulation during development to compensate for changes in the acoustical properties of the vocal tract. Compensation is made possible by means of the auditory-to-articulatory directional map and the phoneme-to-auditory map. Instead of mapping positions in auditory space onto positions of the articulators, the DIVA model maps directions in auditory space onto directions in which to move the articulators in order to reduce the distance to the nearest region of the target phoneme in the speech sound representation (see figure 1). Whereas the

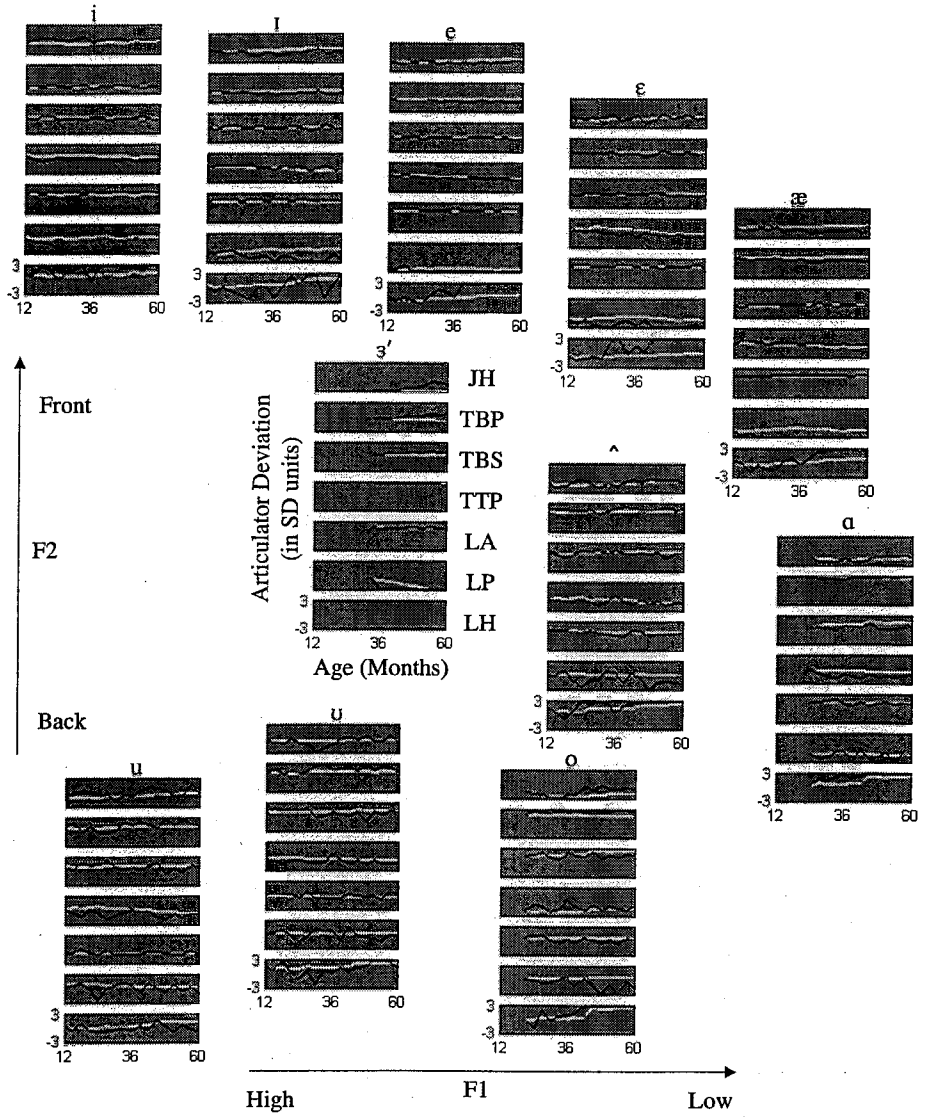


Figure 7: Articulator configuration patterns in standard deviation units produced by the network for each of the 11 vowels from 3 to 60 months of age (the modifiable implementation is displayed in black and the static implementation is displayed in white). Empty regions at the initial part of the plots indicate that the phoneme-to-auditory map for that vowel has not been acquired yet. A negative standard deviation (SD) value of jaw height (JH) corresponds to lowering of the jaw, a negative SD value of tongue body position (TBP) corresponds to a anterior position, a negative SD value of tongue body shape (TBS) corresponds to a flat tongue, a negative SD value of tongue tip position (TTP) corresponds to a posterior tongue tip position, a negative SD value of lip aperture (LA) corresponds to a closure of the lips, a negative SD value of lip protrusion (LP) corresponds to a retraction of the lips, and a negative SD value of larynx height (LH) corresponds to a lengthening of the larynx. Plot of vowels is made in F1 by F2 space.

modifiable implementation is able to adapt the weights of the auditory-to-articulatory directional map and the phoneme-to-auditory map, the static implementation relies on the robustness of the maps learned early in development to compensate for changes in the acoustical properties of the vocal tract. Given the changing formant values corresponding to the neutral position of the vocal tract that occurs during the course of development (see figure 5 and 6 bottom), the DIVA model may use differing patterns of articulation in order to produce the target phoneme. A different starting position in auditory space results in a different direction in auditory space to the closest region of the phoneme target. Since the direction in auditory space differs during development, one would expect the auditory-to-articulatory directional map to produce differing directions in which to move the articulators resulting in differing articulatory configurations during the course of development. For most

vowels tested in this study the auditory-to-articulatory directional map learned during early development was robust to changes in the acoustical properties of the vocal tract. Even though the acoustical properties of the vocal tract change throughout development, the directions in which to move the articulators to meet a vowel target in auditory space (auditory-to-articulatory directional map) for the static implementation appear to be fairly robust with respect to this change. It can be seen that the endpoint positions of the seven articulators (figure 7) as well as the corresponding formant values (figure 6) of the produced vowel targets do not show a great deal of fluctuation across development for the static implementation. The larger degree of fluctuation across development occurring in the modifiable implementation (figure 6 and 7) results from updating the weights of the auditory-to-articulatory directional map and the phoneme-to-auditory map in response to evolving tracts. It is unlikely that these results are merely due to the discontinuity (three-month intervals) of the evolving tracts used in this model because one would expect the articulatory configuration patterns of the static implementation to be equally affected. It is important to note that an auditory-to-articulatory positional map would use the same articulatory configuration throughout development. A positional map would probably be unable to compensate for changes in the size and shape of the speech production system resulting in the production of vowels with somewhat inconsistent formant frequencies and ratio values.

As one would expect, both the static and modifiable implementations show a general increase in the area of the constriction patterns as the dimensions of the vocal tract increase throughout development. Many aspects of the vocal tract constriction patterns produced by the model are consistent with the radiographic analysis of the constriction patterns for vowels of adults (Wood, 1979). Consistent with the study conducted by Wood (1979) it can be seen in figure 8 that front vowels are characterized by more anterior constrictions of the vocal tract than back vowels. In addition, large areas are seen in posterior regions of the vocal tract for front high vowels that progressively diminish in the direction of front low vowels. There are also, however, inconsistencies in the constriction patterns produced by the model and those of adults as reported by Wood (1979). In the model back high vowels have narrow constrictions which progressively become less narrow in the direction of back low vowels (figure 8). In contrast, the data reported in Wood (1979) show a narrow constriction in the pharynx for the vowel [ɑ], as well as a less narrow velar constriction for the vowel [u] than for the vowel [o]. These inconsistencies may be a result of inadequacies of the model or may reflect true differences between adult and child constriction patterns used to produce the same vowels. Experiments need to be conducted in order to determine whether children produce different constriction patterns than adults.

The constriction patterns for the modifiable implementation are more variable than for the static implementation (figure 8). Front vowels show much more similarity between the modifiable and static implementations than do the back vowels. The static implementation shows constriction patterns with maximum areas somewhat more anterior for front vowels and smaller maximum areas for back vowels (with the exception of [o]) as compared to the modifiable implementation. One interesting finding shown in figure 8 is that rounded vowel ([u], [o], and [ʊ]) tract shapes are more variable than unrounded vowel ([i], [ɪ], [ɛ], etc...) tract shapes for the modifiable implementation than for the static implementation. This finding is consistent with several studies demonstrating that for adults rounded vowels show more articulation variability than unrounded vowels (Hashi, Westbury, & Honda, 1998; Perkell, 1996; Wood, 1986). A possible explanation for why the model demonstrates this behavior is based on the few-to-many characteristic of the auditory-to-articulatory directional map which allows for many tract shapes to agree with a given formant pattern. The formant patterns of rounded vowels are the outcome of two constrictions in the vocal tract that entertain trading relations, which increase variability (Perkell, Mathies, Svirsky, & Jordan, 1995). Repeated updating of the weights in the modifiable implementation to developing vocal tracts would likely result in variable constriction patterns that can achieve the same auditory goal. Although one might expect this form of variability for any model that employs a few-to-many map, it is expected that models that employ

directional maps will show more variability because the formant trajectories are dependent on initial conditions.

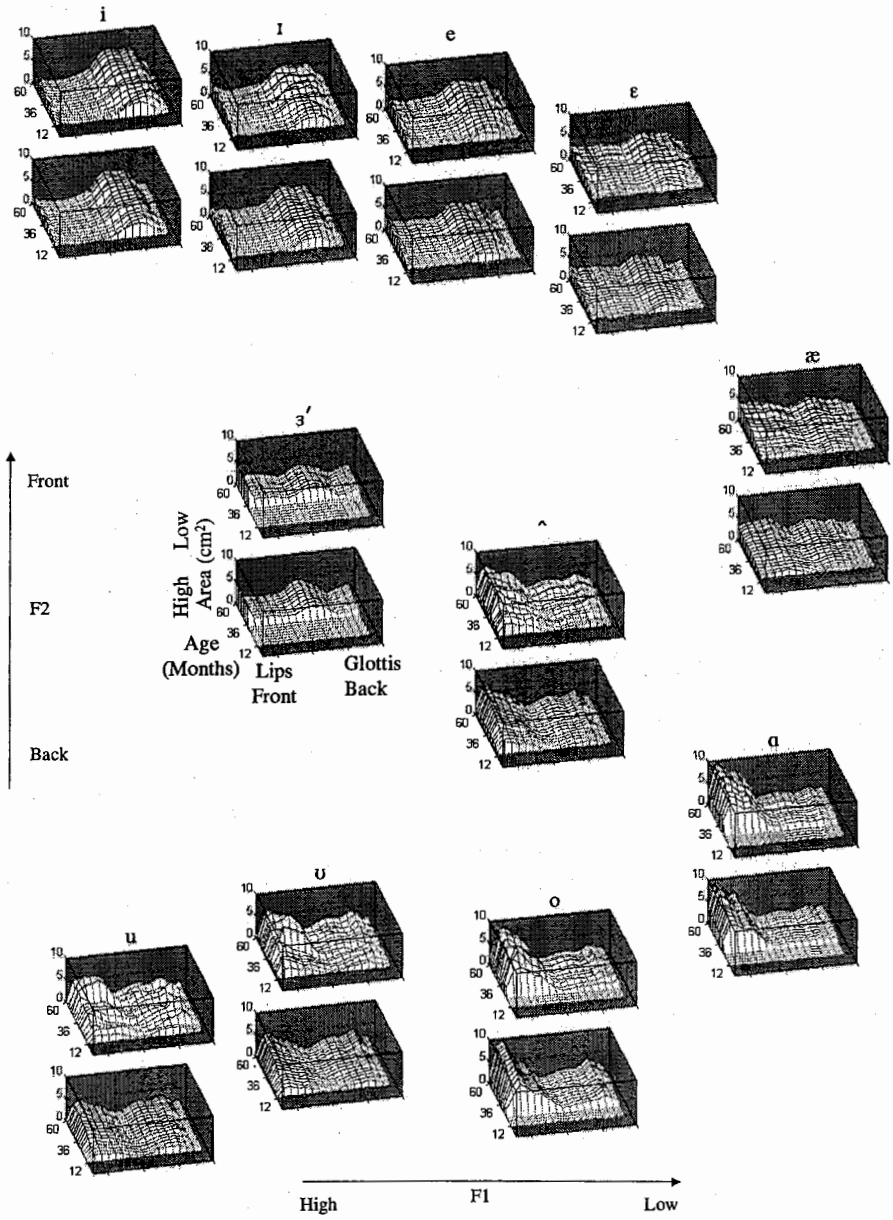


Figure 8: Vocal tract constriction pattern (area function) produced by the network for each of the 11 vowels from 3 to 60 months of age, modifiable implementation (top plot) and static implementation (bottom plot). Empty regions at the initial part of the plots indicate that the phoneme-to-auditory map for that vowel has not been acquired yet. Plot of vowels is made in F1 by F2 space. The graphs displaying the vocal tract constriction pattern are constructed by interpolating the vocal tract area function into 17 divisions. Therefore, the x-axis only represents the relative position along the vocal tract not the actual length of the vocal tract. By using this method the relative location of constriction and location of the maximum area can be compared throughout development.

Conclusion

In this study it has been demonstrated that a model using an auditory-to-articulatory directional map and a phoneme-to-auditory map learned during early development (static implementation) does fairly well in accomplishing the goals of speech production throughout development. This is quite remarkable given the changes in the size and shape of the vocal tract that occur during the course of development. The results of these simulations demonstrate the robustness of the auditory-to-articulatory directional map and suggest that an

auditory reference space can serve as a plausible target signal used by the neural control system to drive speech production for vowels. It is important to note that these findings are not constrained to the DIVA model but extend to any model that uses an auditory-to-articulatory directional map to drive speech production. During development the nervous system must adapt its sensory and motor maps to accommodate for morphological change. A reference space that is robust with respect to developmental change may be better able to accommodate for developmental alterations in muscle and sensory responses by co-registering change with the same reference space used by the neural control system for speech production.

It should be recognized that a more sensitive test of the model would be to see whether the trajectories of the articulators in reaching the vowel targets are similar to those produced by developing children for both the static and modifiable implementations. One should expect that the static implementation would produce different formant movements than the modifiable implementation in reaching the same auditory targets, and that the differences in trajectories could help us discern which model is more accurate. Given that some of the perceptual cues for vowels and consonants may be centered on the transients rather than the targets future modeling work should investigate the trajectories of the articulators in reaching speech targets.

However, even examining trajectories may not be sufficient to validate the production model. Certainly the representation of the acoustics will have a large effect on the trajectories: moving straight toward a target in formant space will produce a different formant trajectory than moving straight toward a target in Miller's transformed acoustic space or than moving toward a vocal tract constriction target. Thus, to get the correct trajectories it will be important to get the details of the target representation correct. But consider that it is possible to map acoustic signals to positions in a continuity map using an unsupervised learning technique (Hogden & Valdez, 2000), and that positions in the continuity map are correlated with articulator positions. In fact, it may yet be shown that the continuity map positions can be thought of as representing vocal tract constrictions parameters (Hogden & Valdez, 2000). If a model maps acoustics to something like positions in a continuity map, and also represents targets in a continuity map, then model trajectories could look very similar to trajectories that would be obtained if the model were using vocal tract constriction targets. This is so even though the positions in the continuity map are learned in an unsupervised manner from properties of the acoustic speech signals. Thus, many aspects of speech production will need to be considered, in addition to trajectories, to get to an accurate model.

In order for the DIVA model to be extended to consonants the targets have to be defined. The model presented here uses an auditory target reference space. Although invariant acoustic targets are relatively easy to define for vowels they are much harder to describe for consonants (Liberman, 1996). However, it has been suggested that invariants for consonants do exist and can be determined by a better understanding of how the auditory system processes the dynamic aspects of the acoustic speech signal (Diehl & Kluender, 1989; Kluender, 1994). If this is true then it should be possible to define targets for the production of consonants in auditory reference space. However, more research needs to be conducted to determine the auditory processes involved with speech perception. An alternative approach would be to use the aforementioned continuity map or perhaps a multimodal reference space that defines orosensory as well as auditory targets for the production of speech. In order to more accurately model the course of speech production acquisition in children the DIVA model needs to incorporate a target reference space that can accommodate consonants. In addition, changes in biophysical constraints and motor control that occur throughout development need to be incorporated.

Acknowledgements

Work was supported in part by NIDCD grant 1R29 DC02852. I would like to thank Shinji Maeda and Mark Tiede for the use of their code. Address all correspondence to Daniel E. Callan, ATR Brain Activity Imaging Center, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Email: dcallan@hip.atr.co.jp.

References

Bailly, G., 1997. Learning to speak, Sensori-motor control of speech movements. Speech Communication, 22, 251-267.

Callan, D. E., 1998. An Auditory-Feedback Based Model of Speech Production in the Developing Child. Dissertation. University of Wisconsin – Madison.

Callan, D. E, Kent, R. D., Guenther, F., & Vorperian, H., 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. Journal of Speech, Language, and Hearing Research, 43 (3), 721-736.

Diehl. R. L., & Kluender, K. R., 1989. On the objects of speech perception. Ecological Psychology, 1, 121-144.

Green, J. R., 1998. Physiologic Development of Speech Motor Control: Articulatory Coordination of Lips and Jaw. Dissertation at the University of Washington.

Guenther, F. H., 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychological Review, 102 (3), 594-621.

Guenther, F. H., Hampson, M., and Johnson, D., 1998. A theoretical investigation of reference frames for the planning of speech movements. Psychological Review. 105, 611-633.

Hashi, M., Westbury, J. R., & Honda, K., 1998. Vowel posture normalization. Journal of the Acoustical Society of America. 104 (4), 2426-2437.

Hillenbrand, J., Getty, L., Clark, M. & Wheeler, K., 1995. Acoustic characteristics of American English vowels. Journal of the Acoustical Society of America. 97, 3099-3111.

Hogden, J. & Valdez, 2000. Bridging the gap between speech production and speech recognition. Proceedings of the 5th Seminar on Speech Production: Models and Data. 241-244.

Huttenlocher, P., 1993. Morphometric study of human cerebral cortex development. In Johnson, M. (Ed.) Brain Development and Cognition: A Reader. Cambridge MA, Blackwell.

Johnson, C. D., 1998. Investigations of Formant and Wavelet Representations for Speech Movement Planning. Dissertation. Boston University.

Kent, R. D., 1999. Motor control: Neurophysiology and functional development. In Caruso A. J. and Strand, E. A. (eds.), Clinical Management of Motor Speech Disorders in Children (pp. 29-71). Thieme Medical and Scientific Publishers.

Kent, R. D., 1997. The Speech Sciences. San Diego, CA: Singular Publishing Group.

Kent, R. D. & Vorperian H. K., 1995. Anatomic development of the craniogacial-oral-laryngeal systems: A review. Journal of Medical Speech-language Pathology, 3, 145-190.

Kluender, K. R., 1994. Speech perception as a tractable problem in cognitive science. In Gersbacher M. (Ed.) Handbook of Psycholinguistics. San Diego: Academic Press, 173-217.

Kuhl, P., 1979. Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. Journal of the Acoustic Society of American, 66 (6), 1668-1679.

Kuhl, P., 1983. Perception of auditory equivalence classes for speech in early infancy. Infant Behavior and Development, 6, 263-285.

Kuhl., P. & Meltzoff, A., 1996. Infant vocalizations in response to speech: Vocal imitation and developmental change. Journal of the Acoustical Society of America, 100 (4), Pt. 1, 2425-2438.

Lee, S., Potamianos, A., &Narayanan, S. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. Journal of the Acoustical Society of America, 105 (3), 1455-1468.

Liberman, A. M., 1996. Speech. Cambridge, Massachusetts: MIT Press.

Maeda, S. 1990. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle & Marchal (Eds.), Speech Production and Speech Modeling, pp. 131-149. Kluwer Academic Publishers, The Netherlands.

Markey, K. L., 1994. Acoustic-based syllabic representation and articulatory gesture detection: prerequisites for early childhood phonetic and articulatory development. In: Ram, A., Eiselt, K. (Eds.), Proceedings of the 16th Annual Conf. Of the Cognitive Science Science Society. Hillsdale, NJ: Lawrence Earlbaum Associates, 595-600.

Miller, J., 1989. Auditory-perceptual interpretation of the vowel. Journal of the Acoustical Society of America, 85, 2114-2134.

Perkell, J., 1996. Properties of the tongue help to define vowel categories: hypotheses based on physiologically-oriented modeling. Journal of Phonetics, 24, 3-22.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., Guiod, P., 1997. Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. Speech Communication, 22, 227-250.

Perkell, J., Matthies, M., Svirsky, & Jordan, M., 1995. Goal-based speech motor control: a theoretical framework and some preliminary data. Journal of Phonetics, 23, 23-35.

Peterson, G. & Barney, H., 1952. Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24 (2), 175-184.

Saltzman, E. L., & Munhall, K. G., 1989. A dynamical approach to gestural patterning in speech production. Ecological Psychology, 1, 333-382.

Sharkey, S. G. & Folkins, J. W., 1985. Variability of lip and jaw movements in children and adults: Implications for the development of speech motor control. Journal of Speech and Hearing Research, 28, 8-15.

Wood, S., 1979. A radiographic analysis of constriction locations for vowels. Journal of Phonetics, 7, 25-43.

Wood, S., 1986. The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels. Journal of the Acoustical Society of America, 80 (2), 391-401.