INTERSPEECH 2009 BRIGHTON

Evaluating parameters for mapping adult vowels to imitative babbling

Ilana Heintz[1], *Mary Beckman*[1], *Eric Fosler-Lussier*[2], *Lucie Ménard*[3]

[1]Dept. of Linguistics, [2]Dept. of Computer Science & Engineering,
The Ohio State University, Columbus, OH
[3]Dept. de linguistique et de didactique des langues,
Université de Québec à Montréal, Montréal, Québec
{heintz.38,beckman.2,fosler-lussier.1}@osu.edu, lucie.menard@uqam.ca

10.21437/Interspeech.2009-238

Abstract

We design a neural network model of first language acquisition to explore the relationship between child and adult speech sounds. The model learns simple vowel categories using a produce-and-perceive babbling algorithm in addition to listening to ambient speech. The model is similar to that of Westermann & Miranda (2004), but adds a dynamic aspect in that it adapts in both the articulatory and acoustic domains to changes in the child's speech patterns. The training data is designed to replicate infant speech sounds and articulatory configurations. By exploring a range of articulatory and acoustic dimensions, we see how the child might learn to draw correspondences between his or her own speech and that of a caretaker, whose productions are quite different from the child's. We also design an imitation evaluation paradigm that gives insight into the strengths and weaknesses of the model.

Index Terms: language acquisition, neural networks, self-organizing maps, language development

1. Introduction

We present a computational model of speech sound acquisition by a child learning his or her first language. The model is designed to capture three basic abilities of a normally developing child: the ability to produce sounds via the speech organs, to hear those self-produced sounds, and to hear speech sounds from adults in the environment. We assume that the child can learn the associations between his or her own gestures and the sounds produced, and that the child can associate the adults' productions with his or her own. These two assumptions lead us to a model that allows for articulatory imitation of adult acoustic input. We describe a method for using this imitation as an evaluative measure of models of babbling.

The first goal of this study is to learn more about how the child forms a mapping between his own speech and his parents' despite inherent obstacles. Due to differences in vocal tract length and configuration, the formant frequencies of an infant and an adult rarely overlap for vowels of the same category. To resolve this, we test and discuss the use of formant differences and formant ratios as data structures that could be used in forming the child-to-adult mapping. We also present a new algorithm for testing the type of neural network based acoustic-articulatory model used in this study, and based on those in [1, 2], and others. Producing a child-like imitation of an adult input and testing that imitation's accuracy provides a clear evaluation of the model's strength.

The model takes advantage of the dynamic nature of self-organizing maps to capture the changing nature of the child's speech; in particular, we model the finding that an infant begins life vocalizing in only a small range of the possible vowel space, and increases exploration of that space over the first year [3]. We use a Hebbian update algorithm to promote learning between the auditory and articulatory domains. As discussed in [1], the use of self-organizing maps and Hebbian updates offers a plausible way of modeling actual neural activity, including interactions between parts of the brain responsible for audio perceptions, proprioception, and even visual perception. Like the DIVA model of [2] and the auditory-motor model of [1], learning takes place due to both the influence of regular ambient speech input and the self-produced speech of the child's babbling. Unlike these previous studies, however, part of our focus is how the child might overcome the obstacles that are the many differences between his or her own speech and that of the caretaker. Since a baby's /i/ has a different first and second formant than his mother's /i/, how is it that he is able to learn the correspondence between them? We train our models with both adult- and child-based data and look for the correspondences between them that might lead to a vowel-recognition mapping. Here we focus on how different data representations, in particular the use of different articulatory and acoustic dimensions for the child and adult data, affect the accuracy of the model in attempting to imitate adult input.

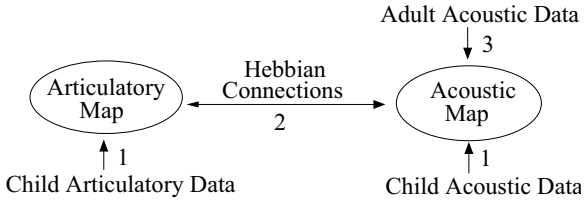

Figure 1: Model Components. We present associated child data points to the articulatory and acoustic maps (1). We then update the Hebbian connections between the two maps (2). We alternate babbling training with listening training, (3), in which adult acoustic data is presented to the acoustic map only.

2. Model Components and Data Set

The model has three main parts. A self-organizing map [4] represents the child's knowledge of his or her own speech articulators, the articulatory map. A second self-organizing map is excited by both the babbled speech of the child and the ambient speech of the adults in the child's surroundings, the acoustic map. Third, a set of weights describes how the neurons in the

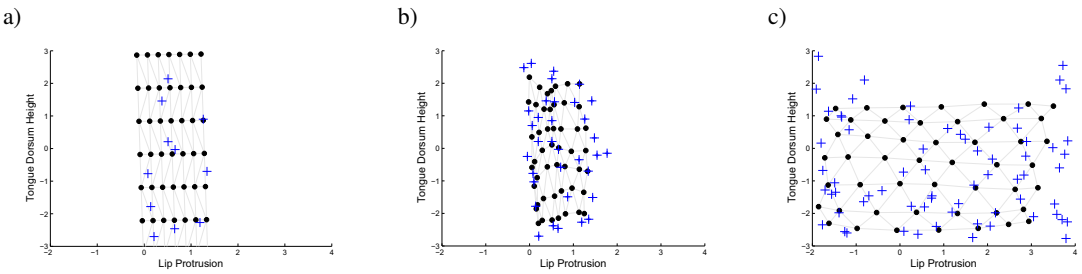

Figure 2: Training the Articulatory Map. In part a) we have initialized the map to cover a small range of articulatory data. The neurons (black circles) are arranged linearly. The data points (blue crosses) are all within the boundaries of the initialized map. Halfway through training, in part b), the map has shrunk to accommodate the centralized babbling data. Part c) shows the fully-trained map, which accommodates both central and the more extreme data points to which it has now been exposed. Higher values on the x axis indicate more lip protrusion, and therefore backer vowels. Higher values on the y axis indicate that the tongue is closer to the palate, producing a higher vowel.

two maps are related. Called Hebbian weights, these describe which pairs of articulatory and acoustic neurons tend to fire together in response to a single instance of babbling. These three components are shown in Fig. 1.

The data set, a subset of the Variable Linear Articulatory Model (VLAM, [5]), also has three parts: a set of articulatory configurations designed to mimic the vocalic speech capabilities of an infant; the acoustic output that corresponds to each of those articulatory configurations; and a set of acoustic data points representing the range of vowels that an adult woman might produce. The full data set comprises 12 articulatory and 10 acoustic dimensions, where an algorithm described in [5] derives the acoustic dimensions from each setting of the 12 articulatory parameters. The mapping between spaces for different areas is complex because the overall size of the vocal tract and the proportional size of the pharyngeal cavity both increase as the individual ages. In this study, we model the child using data designed to emulate a 1-year-old, and the adult data is that of the data designed to emulate an adult woman. We use a subset of the articulatory and acoustic dimensions to model production and perception. We convert all of the acoustic values into the Bark scale to approximate their psychoacoustic value [6].

We also extract from the adult acoustic data a number of examples that we label *point vowels*. These examples represent the vowels at the edge of the adult vowel space, which have been shown to be exaggerated in motherese, as described in [7]. The point vowels are chosen as follows: /i/ examples have low F1 and high F2, /a/ examples have high F1 and low F2, and /u/ examples have low F1 and low F2. We use different proportions of point vowels to all other vowels as a feedback variable. This allows us to test the model's sensitivity to the adult input. The varying proportion of point vowels in the data could represent the child's varying attentiveness to such cues, or caretakers' variable tendency to use child-directed speech containing such exaggerated vowels.

2.1. Self-Organizing Maps and Hebbian Weights

We use the SOM Toolbox for Matlab [8] to train and test the self-organizing maps. We begin by initializing a set of neurons to cover the range of a data set. For instance, if we use two articulatory variables, each with a range of [-3,3], then each neuron in the articulatory map is two-dimensional with values in the same range. The neurons are initialized linearly. Fig. 2a shows the initialized neurons of an articulatory map. We initialize the articulatory map with data that represent very central productions. We assume that the child is at first incapable of producing extreme or particularly distinct vowel sounds [3], and therefore we limit the initial babbling data set to central vowels. The acoustic map is initialized with the corresponding child acoustic data and a subset of the adult data.

The map is then trained to accommodate and model subsequent data sets. First we find the best matching unit (x_{BMU}): the neuron x in the map for which the distance between that neuron and the data point d is the smallest. Then, the value of the BMU is changed:

$$x_{BMU} = \underset{x}{\operatorname{argmin}}(||x - d||) \quad (1)$$

$$x_{BMU}(t+1) = x_{BMU}(t) + \alpha(t)||x_{BMU} - d|| \quad (2)$$

The value of neuron x at time $t+1$ is calculated according to the distance between that neuron and the data point at time t multiplied by some learning rate α. Some of the neighbors of x may also be changed to better model the data. Figures 2b and 2c show how the articulatory map changes.

We allow the articulators to reach a larger range of values at each iteration, mimicking the growth described in [3]. The child acoustic input accordingly becomes more varied. At each data presentation, the neurons in the self-organizing map change position to better match the variation in the data.

The acoustic map is trained by presenting data points chosen in correspondence with the articulatory data, according to VLAM. This is called babbling. We alternate the babbling with additional presentations to the acoustic map of adult acoustic data points. We choose acoustic dimensions that provide considerable overlap between adult and child data, specifically, the difference between F2 and F1, and the difference between F3 and F2. These intervals are similar in both infants and adults, even though the absolute formant frequencies are dissimilar. This causes the same neurons to react to both child and adult data; they are superimposed on the same space.

2.2. Training the Hebbian Connections

We use Hebbian updating to connect the articulatory and acoustic maps. We begin by initializing to zero a weight between each articulatory and acoustic neuron pair. For each presentation of babbling data, we find the best matching unit in both the articulatory and acoustic maps. We calculate the error, or distance, between the BMU and the input. The weight connecting the two activated neurons is increased by the inverse of the sum of the errors between those points and the input data points:

$$W(a,b)(t+1) = W(a,b)(t) + \frac{1}{(err_a + err_b)} \quad (3)$$

Where a and b represent neurons in the articulatory and acoustic maps, err_a is the distance between the BMU on the articulatory map and the input, and err_b is the distance between the BMU on the acoustic map and its input. We also diminish some of the weights: the weights connecting a and all *non*-BMU acoustic neurons (and vice versa) are lowered by a constant fraction, experimentally tuned to .001. Thus, pairs of neurons that are simultaneously activated many times develop strong weights, while non-coordinating pairs develop weak or negative weights.

The Hebbian weights describe the relationship between the articulatory and acoustic maps; they show how the child learns to relate his or her own articulatory gestures to the sounds that they produce. They also allow the child to imitate the adult: when an adult input activates the acoustic map, the Hebbian weights can be used to project that activation to a place on the articulatory map, spurring a configuration that may or may not emulate that of the adult. The accuracy of the imitation will vary depending on how well the adult and child acoustic data correspond on the acoustic map, and how strong the connections are to the articulatory map. This imitation only works if the child and adult acoustic input is overlapping, otherwise the adult input activates areas of the map that do not correspond to the right articulatory configurations. We test the accuracy of our maps and connections by performing an imitation test.

3. Training and Testing the Model

We begin **training** by choosing the parameters to model the articulatory and acoustic data. In all models, we use two articulatory and two acoustic parameters. As described above, we initialize both maps on a centralized subset of the child data, and the acoustic map is also initialized with a subset of adult data. In each babbling iteration, we present to both maps a corresponding subset of child articulatory and acoustic data with the desired amount of variation. We update the neurons in both maps to account for the new data, according to Equations 1 and 2 above. Then, we update the Hebbian connections between the acoustic and articulatory maps according to these same input pairs, as described in Equation 3. Because we choose a slightly more varied data set at each iteration, the maps grow as the child matures. In listening iterations, we choose a random subset of the adult acoustic data, assuring that the desired percentage of point vowel data is included in that subset. We update the values of the acoustic map by presenting these data to the acoustic map only. Areas of the acoustic map corresponding to often-repeated sounds develop close clusters of neurons. The Hebbian weights do not change on listening iterations.

We repeat the babbling and listening iterations until the coverage of the child's articulatory data has reached 100%. We continue training for several iterations to strengthen the connections between neurons after their rate of change has slowed. The Hebbian connections are built as the neurons move around, so the associations between some close neurons will vary more than is desirable. After coverage has reached its maximum, that variation is diminished, so we allow the weights to strengthen at this point before testing.

For the **imitation test**, we are most interested in whether the model has learned enough to respond to an adult point vowel by producing a vowel of the same quality with the articulatory map. We label the neurons in the acoustic map as *i, a, u*, or *other*, according to their response to the training stimuli (Step 1 in Fig. 3). The point vowel labels form clusters. These labels allow us to categorize each test adult acoustic stimulus and the subsequent child production response as belonging to one of

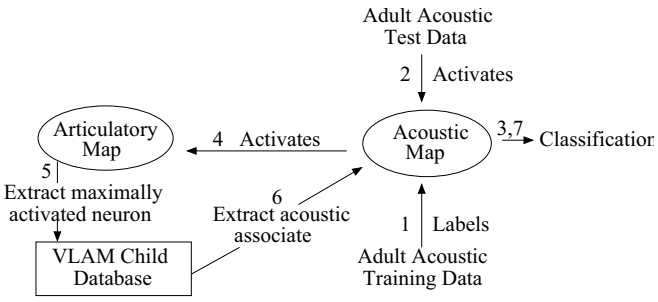

Figure 3: Imitation: testing whether the child can produce an accurate response to an adult stimulus.

four categories according to the trained acoustic map.

We choose a data point from a held-out set of adult acoustic data and present it to the acoustic map (2), which causes an external activation of the acoustic map:

$$act_{i\_ext} = e^{\frac{-\|d-x\|}{\sigma^2}} \quad (4)$$

where d is the value of the data point, x is the value of neuron x, and σ^2 is the width of the Gaussian activation in the model. In other words, σ^2 defines how far the activation of a single data point spreads across the map. The result of this equation is a description of how much one neuron is activated by the presentation of one data point. We calculate this external activation for every neuron on the map to find the map's overall activation.

We determine the best matching unit of the acoustic map, and classify the data point according to that neuron's label (3).

We multiply the overall acoustic map activation and the Hebbian weights to derive the articulatory activation (4). This is the level of activation of the articulatory map according to its learned associations with the activated acoustic map.

The next task is to associate the articulatory map activation with an acoustic production, in order to complete the child's imitation of the adult stimulus. We find the neuron in the articulatory map that has the maximum activation. We assume the value of this neuron to be the articulatory configuration chosen by the child. We use Euclidean distance to find the nearest matches in the database of VLAM child training data (5). Because we are currently limited to only two dimensions, we usually find several near-matches. The acoustic values associated with these sets of near-matches often diverge considerably, because the acoustic values are derived from all of the articulatory settings. If we randomly choose one articulatory sample from the five closest matches, and then extract the acoustic value associated with that sample, the results are no better than chance in matching to the adult category. Instead, we choose from among the near-matches the sample with the acoustic value that is closest to the adult acoustic stimulus (6). We expect that when we increase the articulatory map's dimensionality and develop a more sophisticated forward model, we will be able to choose the correct sample without such an oracle.

Next we test the match between stimulus and response. We classify the acoustic response according to the best matching unit on the acoustic map (7). We record whether the same category as the adult stimulus is chosen. This is a broad test, but one that is useful in learning about the effectiveness of the model.

In our experiments, we vary the features we use to train the maps. We also vary the proportion of point vowel data to other adult data presented at each listening iteration, as mentioned in Section 2. We leave constant all other variables. The experimental variables and results are shown in Fig. 4.

Exp	Artic1	Artic2	Acoustic1	Acoustic2
1	Lip Pro.	Tng Dorsum	F2-F1	F3-F2
2	Lip Pro.	Tng Dorsum	F2-F1	F3-F1
3	Lip Pro.	Tng Dorsum	F2/F1	F3/F2
4	Lip Pro.	Tng Dorsum	F1	F2
5	Tng Apex	Jaw Height	F2-F1	F3-F2

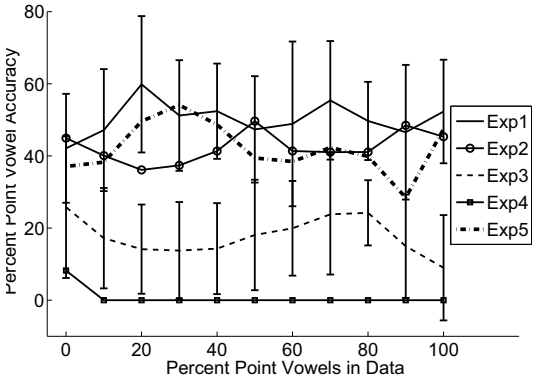

Figure 4: Results: Percentage of correct classifications of child responses when adult stimulus was classified as a point vowel, against the percentage of point vowel data in the adult training stimuli. Error bars shown for Exp. 1 & 3, they are similar for Exp. 2 & 4. Zero on the x axis represents a babbling-only trial.

The results obtained from Experiments 1, 2, and 5 are not significantly different. We can use either F3-F2 or F3-F1 to model vowel height (Exp. 1 vs. Exp 2), and either lip protrusion and tongue dorsum height (Exp. 1) or tongue apex position and jaw height (Exp. 5) as articulatory features, without a great difference in the imitation task. In each case, we have for both articulatory and acoustic features one variable that we expect to correlate with vowel backness, and another with vowel height. The use of formant ratios F2/F1 and F3/F2 in Exp. 3 produced worse accuracy than formant differences in Exp. 1 ($p \leq 0.0001$), against our expectations. As expected, the use of absolute formant frequencies was not at all successful due to the non-overlapping nature of the child and adult formant frequencies (Exp. 4). We find no reliable correspondence between category accuracy and percentage of point vowels in the adult data.

4. Discussion

The imitation evaluation method allows us to explore the usefulness of the articulatory and acoustic parameters. We find that formant differences form stronger connections than formant ratios. Both are a measure of the relationship between the vocal resonances, and both produce overlapping values for the child and adult data. However, the distribution of values of the formant ratios is very peaky. This results in more of the acoustic data activating fewer neurons and less reliable Hebbian weights.

In Exp. 1, the /i/ samples constituted the bulk of the correct classifications. In studying the correspondences between the maps, we found a strong backness correspondence between articulatory and acoustic maps, but a weak height correspondence. The /a/ and /u/ neurons tended to group near each other, causing confusability, and were also confusable with the adjacent central vowel neurons. In Exp. 2 we replaced F3-F2 with F3-F1, attempting to improve the accuracy of the model by changing the vowel height definition. Using F3-F1 as an alternate to F3-F2 did not reduce the confusability between /a/ and /u/ or between these and the non-point vowels. Changing the articulatory variable did not help either, nor did adding dimensions to either map. This is an area of further research as we continue to improve the model.

The correct responses in the babbling-only trial of Exp. 4 are unreliable. Since the two data sets do not overlap in this experiment, the labels given by the adult data are essentially random. In the other trials, the neurons form two clusters to accommodate the two data sets. Only the adult-oriented cluster gets point vowel labels, and these neurons have only weak or negative connections to the articulatory map.

The evaluation also shows us that our model is not using the adult acoustic data effectively. We can likely solve this problem by updating Hebbian weights on listening iterations as in [1]. Although this finding limits the predictive power of our current model, we note that the problem may have remained hidden without such an evaluation. We therefore believe that the imitation evaluation method is a useful addition to the literature on computational modeling of babbling.

In subsequent models, we will add fundamental frequencies or formant frequencies as additional dimensions. Such a model will be able to discriminate between speakers as well as vowel categories. As we continue to make the model more robust and complex, we will discuss the role of feedback in a developing child's changing ability to perceive native and non-native contrasts [9]. We will enhance the acoustic map by introducing perception and production data from psycholinguistic experiments. We will test our model's predictions of how cross-linguistic and other environmental variables change the course of language learning. We will also use the model to make predictions about how feedback affects the learning process.

5. Acknowledgments

This work is funded in part by NSF grant 0739206.

6. References

[1] G. Westermann and E. R. Miranda, "A new model of sensorimotor coupling in the development of speech," *Brain and Language*, vol. 89, pp. 393–400, 2004.

[2] D. E. Callan, R. Kent, F. Guenther, and H. Vorperian, "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 and Hearing Research*, vol. 43, no. 3, pp. 721–736, 2000.

[3] J. Serkhane, J. Schwartz, L. Boë, B. Davis, and C. Matyear, "Infants' vocalizations analyzed with an articulatory model: A preliminary report," *Journal of Phonetics*, vol. 35, pp. 321–340, 2007.

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

[5] L. Ménard, J.-L. Schwartz, and L.-J. Boë, "Auditory normalization of french vowels synthesized by an articulatory model simulating growth from birth to adulthood," *Journal Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1892–1905, 2002.

[6] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, p. 248, 1961.

[7] P. K. Kuhl, "A new view of language acquisition," *PNAS*, vol. 97, no. 22, pp. 11 850–11 857, 2000.

[8] J. n. J. H. Vesanto, E. Alhoniemi, and J. Parhankangas, "Self-organizing map in matlab: the som toolbox," Helsinki University of Technology, Tech. Rep., 2000. [Online]. Available: http://www.cis.hut.fi/projects/somtoolbox

[9] P. K. Kuhl, E. Stevens, A. Hayashi, T. Deguchi, S. Kiritani, and P. Iverson, "Infants show a facilitation effect for native language phonetic perception between 6 and 12 months," *Developmental Science*, vol. 9, no. 2, pp. F13–F21, 2006.