The Interplay of Perception and Production in Phonological Development: Beginnings of a Connectionist Model Trained on Real Speech

Christopher T. Kello† and David C. Plaut‡
† George Mason University, USA
‡ Carnegie Mellon University, USA
E-mail: ckello@gmu.edu, plaut@cmu.edu

ABSTRACT

Three forward models are presented that map articulatory positions onto acoustic outputs for a single speaker of the MOCHA speech database. Backpropagation learning was used to train the forward models on a database of 460 TIMIT sentences. Efficacy of the trained models was assessed by subjecting the model outputs to speech intelligibility tests. The results of these tests showed that enough phonetic information was captured by the models to support fairly high rates of word identification in sentences. These forward models provide the first step toward building a connectionist model of spoken word acquisition trained on real speech. The design of this model is based on a theory of phonological development in which distributed codes are learned in the service of spoken word perception, production, and comprehension.

1. INTRODUCTION

Perceptual processes exist in auditory and visual cortex to support speech perception [1], and motor processes exist in motor cortex to support speech production [2]. It is also very likely that other processes exist to serve the dual purpose of supporting both speech perception and speech production. Some of these multi-purpose processes are likely to be semantic or morphological in nature because these levels of representation are needed for both speech comprehension and speech production. Phonological representations are also needed for comprehension and production, and there are data consistent with the existence of multi-purpose phonological representations in cortex [3]. However, little is known about how such representations might be structured, or how they might emerge over the course of spoken language acquisition.

In previous work [4], we outlined a theory of spoken word acquisition in which a distributed level of representation is learned in support of three complementary functions: 1) integrate the incoming speech signal over time to activate stable representations of single words, 2) generate articulatory trajectories for single words, and 3) link the peripheral processes of spoken word perception and production with the semantic representations of words (see Figure 1). The nature of these functions has led us to following hypothesis: the structure of the distributed codes at the mediating level of representation should become correlated with the phonological structure of the native language. This hypothesis motivated the name *phonological junction* for this mediating level of representation.

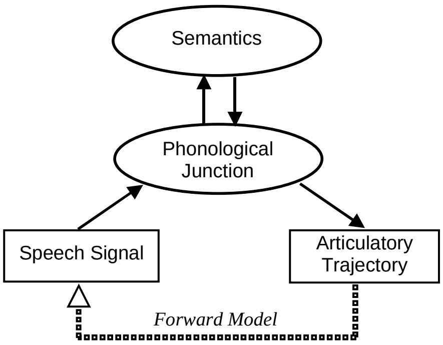

Figure 1: Junction model of spoken word acquisition

On our theory, a key to the development of the phonological junction is the relationship between articulatory gestures and their acoustic consequences. We hypothesized that a *forward model* is learned through babbling and early attempts at intentional utterances. The forward model learns to predict the acoustic consequences of any given articulatory gesture. This predictive power is used to link the learning that occurs in speech comprehension with the learning that occurs in speech production. It is by virtue of this link that the phonological junction becomes a multi-purpose level of representation.

To provide some computational support for the viability of our theory, we implemented it in a connectionist model [4]. The model was successful in that representations were learned to support the three functions just listed, and simulation results were consistent with some basic findings in the developmental literature on speech acquisition. However, the acoustic and articulatory representations used in that model were artificial, i.e., they were abstractions designed to capture at least some of major dimensions of phonetic information. The use of artificial representations makes it difficult to determine whether any shortcomings of the model are due to shortcomings in the theory, or shortcomings in the representations.

2. CURRENT STUDY

In these proceedings, we report on the initial stages of an effort to implement our theory in a model trained on real speech tokens. The use of real speech tokens will provide us with an empirical means of testing whether the model has captured at least some of the phonetic complexity of real speech. Specifically, we can test the intelligibility of speech produced by the model. Intelligible speech would help to quell concerns about the use of simplified representations. It would also enable us to directly compare model performance with empirical findings on phonological development.

A logical starting place for such an effort is to train a forward model on a database of speech tokens. In our theory, the forward model is a foundational piece of the larger model because speech production cannot develop properly without at least a partially working forward model. From a more pragmatic point of view, a working forward model would verify that enough information is present in the articulatory recordings to “regenerate” the phonetic information in the acoustic recordings. This is an essential precursor to embarking on the task building a full model of our theory.

Here we present the results of building a forward model based on an articulatory and acoustic database of real speech tokens. The model was built as an artificial neural network, and the model parameters (i.e., connection weights) were learned through the back-propagation of error signals from acoustics. The model was tested by presenting the acoustic outputs to listeners who transcribed what they heard. The percentage of words transcribed correctly served as a measure of speech intelligibility. The amount of phonetic information captured by the model was estimated by comparing intelligibility of the model tokens with intelligibility of the database tokens (i.e., the targets of the model). Model generalization was assessed by training separate models on half of the tokens in the database, and measuring the intelligibility of the untrained tokens.

3. METHODS

Database. Speech tokens were drawn from one female speaker of British English (subject ID “fsew”, southern dialect) in the multi-channel articulatory (MOCHA) speech database [5] recorded at the Edinburgh speech production recording facility. The database is being developed primarily for purposes of automatic speech recognition (acoustic-to-articulatory mapping, i.e., an *inverse* model), but it is well-suited for building a forward model.

Speech tokens. The database consisted of one token each of 460 sentences based on the TIMIT database. All 460 tokens were used for training and testing.

Articulatory recordings. The MOCHA database consists of three different types of articulatory recordings (see Figure 2): electromagnetic articulograph (EMA), laryngograph, and electropalatograph (EPG). All three types were used for the forward model as follows.

- EMA sensors captured XY positions in the midsagittal plane for the upper and lower lips and incisors, the soft palate (velar port opening), and three positions on the tongue (tip, blade, and dorsum). On the basis of these sensor positions, 18 EMA dimensions were derived for input to the forward model. Each dimension was an X or Y coordinate of one of the sensors, set relative to the incisor positions or the centroid of the three tongue positions.

- EPG sensors captured tongue contact with the hard palate. Forty-eight binary sensors (contact or not) were recorded for the database, but only the front 24 sensors were used as input to the forward model (the back 24 carried little or no information for speaker FSEW).

- Laryngograph recordings captured source information in the form of vocal fold vibrations. Fourier (FFT) analysis was performed on Hamming windows 64 ms wide, taken at 32 ms intervals. Given the recording sample rate of 16 KHz, this procedure resulted in 512 frequency bins of log magnitude per window. To capture voicing and pitch information, only the lower 25 bins (up to 400 Hz) were used as input to the forward model.

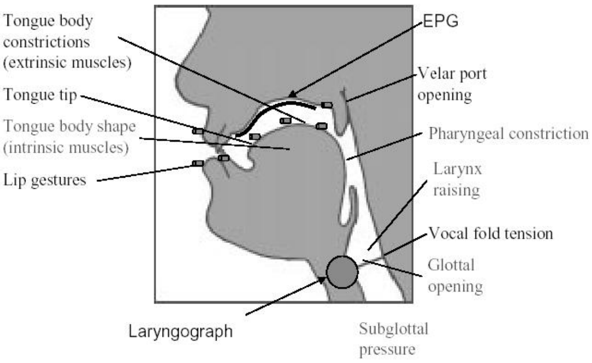

Figure 2: Placements of articulatory sensors (from [5])

Acoustic recordings. In the MOCHA database, the speech signal was recorded with 16 bit sampling at a rate of 16 KHz. To create acoustic targets for the forward model, the same FFT analysis as for the laryngograph recordings was performed on the acoustic wave forms. All but one (DC offset) of the FFT bins were used in the models (511 total).

Model representations. Each articulatory and acoustic dimension was standardized to real values between 0 and 1 (except for the EPG dimensions, which were binary). To help spread out values along the scale, extremely low and high measurements were truncated prior to standardization. For acoustic, laryngograph and EPG measurements, each dimension was assigned one unit in the neural network model (see below). For EMA measurements, each dimension was assigned two units, one to represent the lower range of values (0 to 0.5), the other to represent the upper range (0.5 to 1).

In the MOCHA database, the data streams were sampled at various rates, but for the model, all dimensions were sampled at 32 ms intervals (31.25 Hz), aligned with the acoustic and laryngograph FFT windows. Over the 460 speech tokens, there was a total of 41,791 samples.

Model architecture. A forward model must be able to produce an acoustic trajectory through time, given an articulatory trajectory as input. However, the physical relationship between the vocal tract and the resulting acoustics can, at least for practical purposes, be described instantaneously. Therefore, the forward model was built to generate an acoustic output at a given moment in time, i.e., for a single FFT window 64 ms wide. To capture articulatory positions and movements over this 64 ms window, the model was built to generate an acoustic output on the basis of the previous, current, and next articulatory states (each 32 ms apart). Model trajectories were generated by inputting successive articulatory states to generate successive acoustic states.

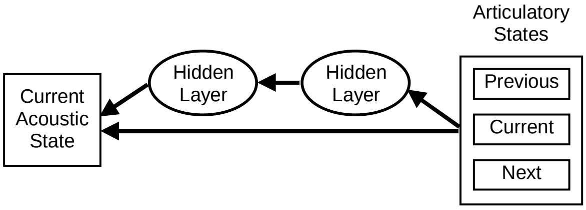

Figure 3: Architecture of the forward models

The model inputs were mapped onto the outputs via weights on direct connections, and weights mediated by two layers of hidden units (see Figure 3). This architecture allowed first order relationships between the inputs and outputs to be coded separately from higher order relationships. The number of hidden units (50 per group) was chosen to be relatively small to avoid overfitting of the training data.

Articulatory values were clamped to the input units, and activations of the hidden and output units were calculated by the logistic function of their net inputs. Net inputs were calculated as the dot product of the incoming weight vector with the vector of presynaptic activations.

Model training. Weights were adjusted to minimize squared error between acoustic targets and outputs (i.e., supervised learning). Weight derivatives were calculated by the back-propagation of error signals (i.e., gradient descent learning). Derivatives were accumulated over batches of 5000 samples drawn at random from the training sets (see below). After each batch, weight changes were made according to

$$\Delta w_{ij}^{[b]} = \eta_N \eta_{ij} \frac{\partial E}{\partial w_{ij}} + \alpha\left(\Delta w_{ij}^{[b-1]}\right), \qquad (1)$$

where η_N was the overall network learning rate (decreased from 5e-4 to 5e-5 over the course of training), η_{ij} was a weight-specific learning rate, α was a momentum term (fixed at 0.8), and *b* was the Nth batch over the course of training. Weight-specific learning rates were adjusted on the basis of the consistency of weight derivatives across batches.

Three models were trained with three different sets of training samples. One set consisted of all samples from all 460 speech tokens (*full*), one set was trained on the odd-numbered tokens (*odd*), and one on the even-numbered tokens (*even*). The arbitrary odd/even split was made to test generalization to untrained tokens. Each model was trained on 30,000 batches of samples.

4. RESULTS

At the end of training, the average amount of squared error per unit per training example was 0.023 for the full set, 0.028 for the odd set, and 0.028 for the even set (the maximum possible squared error was 1). In Figure 4, the error for each model is plotted as a function of frequency. The figure shows that there was a general trend for the models to perform more poorly as frequency increased. This trend was probably due, in part, to the fact that frication causes unpredictable fluctuations in energy at the higher frequencies. Another result was that there was a dip in error at the low end of the frequency range. This dip was likely due to the fact that the laryngograph inputs provide voicing and pitch information that exists at these lower frequencies (i.e., up to 400 Hz).

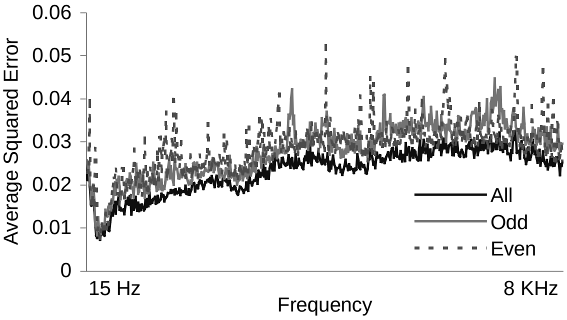

Figure 4: Error as a function of frequency for each model

Intelligibility tests. Four types of tokens were subjected to listening tests: target tokens, and output tokens from the full model, the even model, and the odd model. The model tokens were generated by inputting the articulatory states for a given sentence to the models, and inverting the FFT outputs to create corresponding wave forms.

Two undergraduates and one graduate student listened to all 460 sentence tokens. For each listener, one quarter of the tokens were the targets, one quarter were generated by the full model, one quarter by the odd model, and one quarter by the even model. Tokens were presented in random order over Sennheiser mh80 headphones in a quiet room at a comfortable volume. Listeners heard each token exactly three times before typing in their responses. Listeners were instructed that all the stimuli would be legal English sentences spoken by a female with a British accent (the listeners spoke American English), and that they should type in every word that they heard, in the order that they heard the words. Guessing was encouraged, but incomplete responses were also allowed.

Responses were corrected for spelling errors, and the percentage of words transcribed correctly was calculated as a measure of intelligibility. Results are shown in Table 1.

Token Type	Listener 1	Listener 2	Listener 3	Average
Target	93.1	88.2	94.7	92.0
Full	82.8	67.2	74.2	74.8
Odd	75.8	67.7	75.9	73.2
Even	77.1	67.2	71.0	71.8
Trained	80.5	73.4	81.5	78.5
Untrained	72.4	61.5	65.5	66.5

Table 1: Percent words correct on intelligibility tests

Intelligibility of the targets was fairly high, and intelligibility of the models could not be expected to exceed the target intelligibility. Less than perfect performance for the targets may have been due to the British accent of the speaker, or to the FFT processing of the original wave forms (which removes and distorts some information in the signal). Also, many of the sentences were long (causing memory to be a limiting factor on performance), and they often contained unusual words or phrases.

Model intelligibility was also fairly high (average of 73.2% words correct), although it was 18.8% worse than performance on the targets (more listeners are necessary to assess the statistical reliability of any observed differences in intelligibility). Intelligibility for the full, odd, and even models was about the same on average. For the odd and even models, intelligibility was 12% worse on the untrained tokens compared with the trained tokens. This relatively small difference in performance indicates that the mappings learned by the models were highly, but not perfectly, generalizable to novel inputs.

5. CONCLUSIONS

The results of this study demonstrate that the mapping from articulation to acoustics can be learned for at least one of the speakers in the MOCHA database. The intelligibility tests showed that enough phonetic information is captured by this learned mapping to support fairly high rates of word identification in sentences. The comparisons between trained and untrained tokens showed that the learned mappings had captured general relationships between articulation and acoustics, rather than specific I/O pairings.

Further work is necessary to determine why model intelligibility fell short of the targets. One potential factor is that the articulatory inputs provided only a crude sampling of the vocal tract. A second potential factor is that the model representations, computational capacity, or learning mechanisms may have been deficient in some way. Exploration of the model parameters may reveal some of these potential deficiencies. However, if model performance never reaches the targets, it will be difficult to ultimately determine the causes for this shortcoming.

A more tractable problem is to determine the kinds of phonetic information that were and were not captured by the models. This problem can be addressed in the current study by analyzing word omissions and substitutions to estimate which phonetic features were more or less likely to be preserved in the models. This work is currently underway. A more direct method would be to conduct phoneme identification tests with simple CVC or VCV speech tokens. Phoneme confusion matrices could then be generated to more clearly reveal the transmission of phonetic feature information. Unfortunately, the MOCHA database does not contain such speech tokens.

In conclusion, the relative success of the forward models in this study is a first step towards building a model of spoken word acquisition based on real speech. The biggest and most daunting steps lie ahead. Most notably, a level of representation must be shaped to integrate over incoming words, generate articulatory trajectories to produce words, and connect these input and output processes with the semantics of words. If successful, we will be able to listen to the model over its course of learning, and compare it against the empirical data on spoken word acquisition. Such a model would be a valuable tool for making progress towards an understanding of speech development.

REFERENCES

[1] P. Belin, R. Zatorre and P. Ahad, "Human temporal-lobe response to vocal sounds," *Cognitive Brain Research*, vol. 13, pp. 17-26, 2002.

[2] J. Fiez., "Neuroimaging studies of speech: An overview of techniques and methodological approaches," *Journal of Communication Disorders*, vol. 16, pp. 445-454, 2001.

[3] G. Hickok, "Functional anatomy of speech perception and speech production: Psycholinguistic implications," *Journal of Psycholinguistic Research*, vol. 30, pp. 225-235, 2001.

[4] D. Plaut and C. Kello, "The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach," in The Emergence of Language, B. MacWhinney, Ed., pp. 381-415. Maweh, NJ: Erlbaum, 1999.

[5] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," *Proc. 5th seminar on speech production: models and data*, pp. 305-308, 2000.