Using Contrastive Hebbian Learning to Model Early Auditory Processing

David A. Medler
Department of Neurology, Medical College of Wisconsin
Milwaukee, WI 53226

Abstract

We present a model of early auditory processing using the Symmetric Diffusion Network (SDN) architecture, a class of multi-layer, parallel distributed processing model based on the principles of continuous, stochastic, adaptive, and interactive processing [Movellan & McClelland, 1993]. From a computational perspective, a SDN can be viewed as a continuous version of the Boltzmann machine; that is, time is intrinsic to the dynamics of the network. Furthermore, SDNs embody Bayesian principles in that they develop internal representations based on the statistics of the environment. One of the main advantages of SDNs is that they are able to learn probabilistic mappings (i.e., mapping from m→n, where m<<n) for a single input pattern, a task impossible for many other classes of neural networks. SDNs are trained using the Contrastive Hebbian Learning (CHL) algorithm which is based on positive and negative learning phases. The basic model has been trained on two separate tasks: (i) a signal detection task, and (ii) a phonetic/nonphonetic discrimination task. In the signal detection task, the model was able to capture the accuracy data of human participants, but only grossly approximated participants' reaction time data. Reanalysis of the human data, however, showed that the network correctly predicted the reaction times in the early phases of the experiment. In the phonetic/nonphonetic discrimination task, the network was able to show both categorical and continuous perception of the stimuli. Importantly, the model predicted learning curves for categorical perception of nonphonetic stimuli that was subsequently confirmed in a human learning study. It is concluded that this simple type of network based on correlational learning is able to effectively model early auditory processing.

1. Introduction

Speech is one of our most important forms of communication. Decoding the speech signal is a complex process requiring many neurophysiological stages, from simple signal decoding in the early auditory pathways, to higher cognitive processes that interpret the semantics and pragmatics of the intended message [Clark & Clark, 1977]. At the simplest level, speech decoding–as opposed to general auditory processing–consists of identifying specific phonemes of the listener's language. These phonemes must initially be learned, and then correctly identified in an uncertain environment. In this paper, we use a simplified model of early auditory processing to address issues of phonetic learning and of identification of phonemes in noisy inputs.

The model employs a Symmetric Diffusion Network (SDN), a class of multi-layer, parallel distributed processing model based on the principles of continuous, stochastic, adaptive, and interactive processing [Movellan & McClelland, 1993]. From a computational perspective, a SDN can be viewed as a continuous version of the Boltzmann machine; that is, time is intrinsic to the dynamics of the network. Furthermore, SDNs embody Bayesian principles in that they develop internal representations based on the statistics of the environment. One of the main advantages of SDNs is that they are able to learn probabilistic mappings (i.e., mapping from m→n, where m<<n) for a single input pattern, a task impossible for many other classes of neural network. One disadvantage of SDNs is that they have been difficult to train on larger data sets, and they have been mainly used with static inputs. Recent work [Medler & McClelland, 2001] showed that when biologically inspired constraints (i.e, activations within the range [0,1], positive between-layer projections, lateral inhibition) are applied to SDNs, their effective performance is increased substantially in terms of the number of patterns on which they can be trained,

the rate at which patterns are learned, and their ability to separate out independent sources in an unsupervised manner.

In addition to these biologically inspired architectural constraints, the network was trained using a biologically plausible process known as Contrastive Hebbian Learning (CHL) that embodies the maxim of "units that fire together, wire together"; that is, weights between coactive units are increased, whereas weights between uncorrelated units are decreased [Hebb, 1949]. To avoid some of the problems inherent in simple Hebbian learning, CHL uses positive and negative learning phases that act together to extract the signal from the environment [Peterson & Hartman, 1989; Movellan, 1990].

In these studies, we take advantage of the intrinsic temporal properties of SDNs to process time-varying input, which allows auditory input to be presented to the network as a continuous spectrotemporal stream. This approach stands in contrast to many previous computational models of speech perception that used abstract sets of phonetic features as input [e.g., Anderson, Silverstein, Ritz, & Jones, 1977; McClelland & Elman, 1986; but see Damper & Harnad, 2000]. The use of realistic spectrotemporal inputs allows the model to be presented with exactly the same stimuli (including signal and acoustic noise) given to the human participants. This approach is also more general in that it permits future modeling studies of tasks involving nonspeech auditory stimuli.

The model has been trained on two separate tasks: (i) a phonetic/nonphonetic discrimination task, and (ii) a signal detection task. In the first task, the model was trained to capture the processes of categorical and continuous perception of phonemic and nonphonemic stimuli [Liebenthal, Binder, Spitzer, Possing, & Medler, 2005]. The network was trained with dynamic spectrotemporal auditory forms representing the speech sounds /ba/ and /da/, and with the anchor points of an acoustically-matched non-phonetic continuum. The network was first trained on the phonetic stimuli. Training was stopped at specific times, and the model performance was assessed. The learning curves on the identification task showed a progression for continuous to categorical perception with overtraining. The network was then trained on the phonetic and non-phonetic stimuli with an interleaved ratio of 9:1 to simulate differential experience. Following training, the model showed categorical perception of the phonemes and continuous perception of the non-phonetic sounds. To test the model's prediction of the learning curves, participants were trained on the same non-phonetic stimuli, and tested on the identification task at specific training points. The participants produced qualitatively similar learning curves as the model, eventually showing categorical perception–as measured by the identification task–of the non-phonetic stimuli.

In the second task, the model was trained with the same dynamic spectrotemporal auditory forms representing the speech sounds /ba/ and /da/ as in the first simulations. After learning to identify these sounds, the model was tested on a discrimination task with varying amounts of masking noise added to the stimuli. Performance by the model and activity levels in the model units were compared to psychophysical data obtained previously from normal subjects tested on the same stimuli [Binder, Liebenthal, Possing, Medler, & Ward, 2004]. The network qualitatively reproduced the accuracy and reaction time data of the human participants.

2. General Methods

2.1 Network Dynamics and Learning

Symmetric Diffusion Networks are based on continuous time-varying activations and are governed by Equation (1)

$$\Delta a_i(t) = \Delta[net_i(t) - \hat{net}_i(t)] + \sigma \cdot \sqrt{\Delta t Z_i}(t) \qquad (1)$$

where, $net_i = h(\sum_{j=1}^{n} a_j w_{ij})$, $\hat{net}_i = 1/g_i \cdot f(a_i) = 1/g_i \cdot \log[(a_i - \min)/(\max - a_i)]$, gi is a gain function, h(u) = 1-exp(-u), and Zi(t) is the standard Gaussian variable with zero mean and unit variance. The last term in the equation provides the network with a stochastic and is essential for allowing the network to learn multiple outputs for a single input.

SDNs are trained with the Contrastive Hebbian Learning (CHL) algorithm [Peterson & Hartman, 1989; Movellan, 1990]. The CHL is based on two phases: (i) a plus phase in which the training patterns are clamped (i.e., clamped input and output units), and (ii) a minus phase in which the network is allowed to run free (i.e., when at least one set of units is unclamped). Furthermore, SDNs can be trained in either a supervisory mode, or an unsupervised mode. In supervised learning, an input and output pattern are clamped onto the relevant units, the network is allowed to settle for a pre-determined time, and co-occurrence statistics are collected (i.e., $\sum(a_i^+ a_j^+)/s$) where s is the number of collected samples. The output units are then unclamped, the network is allowed to settle again and co-occurrence statistics (i.e., $\sum(a_i^- a_j^-)/s$) are re-sampled. In unsupervised learning, the network is allowed to run completely free (i.e., no clamped patterns on either input or output units) during the minus phase [Medler & McClelland, 2001].

Once collected, the minus phase is subtracted from the plus phase. In this way, the base activities in the network (minus phase) are subtracted from the base activities plus the applied pattern to leave the activity of the pattern alone. Consequently, the change in weight between two units can be computed as

$$\Delta w_{ij} = \varepsilon\left(\sum(a_i^+ a_j^+)/s - \sum(a_i^- a_j^-)/s\right) \quad (2)$$

where ε is a small constant typically referred to as the learning rate parameter. Finally, the weight changes are mediated by a momentum term, α, which determines how much influence the current weight change has on the averaged weight change. That is, the weight at time t is computed as:

$$w_{ij}(t) = \alpha(\Delta w_{ij}(t)) + (1-\alpha)(\Delta w_{ij}(t-1)) \quad (3)$$

2.2 Network Architecture

The networks had 100 Input units (representing frequencies 1–40 Hz, 41–80 Hz,…, 3961–4000 Hz), 20 Intermediate units, and 10 Output units. In the first model, the four target patterns (/ba/, /da/, np1, np8) were assigned random binary patterns across the 10 units; in the second model, 2 banks of 5 units were used to classify the input pattern as either /ba/ or /da/. The network had lateral inhibition within each layer, positive connections between layers, and self-excitatory connections. Initial weights were randomized from a square distribution between [-1, 0] for inhibitory connections or [0, +1] for excitatory connections. Finally, unit biases were initialized between [-0.5, +0.5]. It should be recognized that this model does not try to capture all aspects of auditory processing, nor decision processes, but merely captures those putative processes in which we are specifically interested. Much more detailed models of auditory processing exist, but such models are often hard-wired [e.g., Husain, Tagamets, Fromm, Braun, & Horwitz, 2004], and are therefore limited in terms of addressing learning issues.

2.3 Stimuli

The same stimuli presented to the participants in the two studies [Binder et al., 2004; Liebenthal et al., 2005] were presented to the network. Audio files were sampled at 44100 Hz and processed through Praat 4.0.5 (http://www.praat.org/) to produce spectrograms, which are time-varying representations of power spectra. The stimuli were analyzed at 5 msec time steps using a 20 msec Gaussian filter, and frequency steps of 40 Hz ranging from 0-4000 Hz. The spectrograms were then presented to the network as time-varying input signals at 5 ms time steps broken into 100 equally spaced 40-Hz frequency bins. For the first study, 16 stimuli were created, the eight phonetic stimuli and the eight nonphonetic stimuli. In the second study, a total of 12 input patterns were created: two training patterns (veridical representations of /ba/ and /da/), and ten testing patterns (the five noise plus signal versions of /ba/ and /da/). Input activations were normalized for each individual input pattern; that is, activations ranged from 0.0 to 1.0 for each pattern. Sample training patterns for the phonetic and nonphonetic stimuli are shown in Figure 1. Formants are represented by the darker bands of activity.

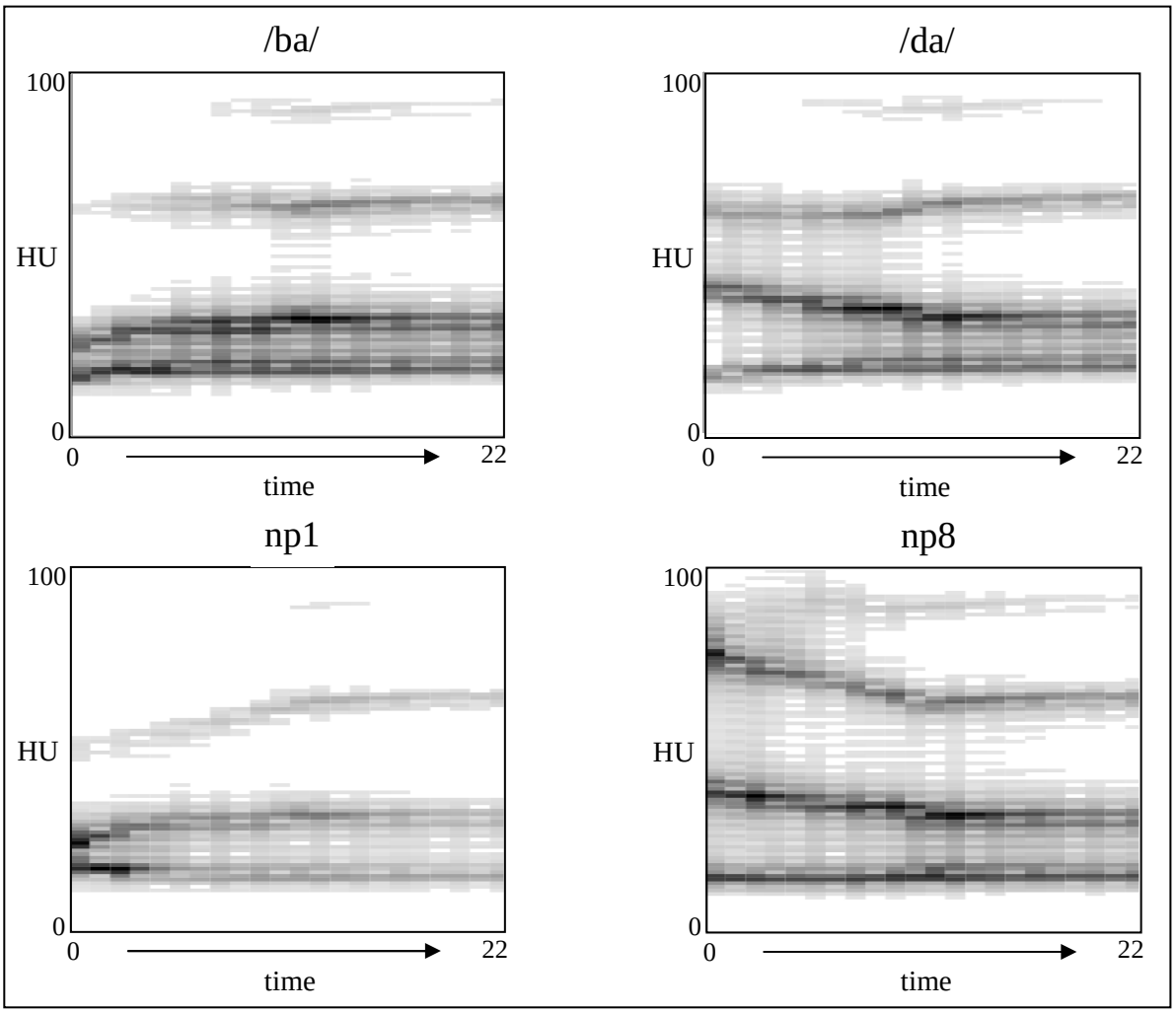

Figure 1. Sample inputs showing the time varying patterns presented to the networks. Amount of activation is indicated by the darkness of the square. The two phonetic stimuli are in the top row, and the two nonphonetic stimuli are in the bottow row.

3. Continuous and Categorical Perception [Liebenthal et al., 2005]

One phenomena closely associated with speech perception is categorical perception. Categorical perception refers to a poor ability to discriminate between members of the same category, yet a good ability to discriminate between members of different categories [MacMillan, Kaplan, & Creelman, 1977; Pisoni, 1971]. Categorical perception may be a necessary component of speech recognition as multiple acoustical instances (either by different speakers or by slight mechanical variations in the vocal tract or tongue placement of the same speaker) must be mapped to the same phoneme for proper identification. Consequently, as applied to speech, categorical perception implies good discrimination between different phonemes produced by the same speaker and poor discrimination between two tokens of the same phoneme spoken by the same speaker. Continuous perception, on the other hand, can be described as an approximate adherence to Weber's law, according to which the difference limen is a constant fraction of stimulus magnitude.

One way of assessing categorical perception is to train a participant on two end points of a continuous auditory spectrum (say /ba/ and /da/) and then to test the participant on identification and

discrimination of stimuli that successively move from one end of the spectrum to the other (e.g., ba+0, ba+1, ba+2, ba+3, da+3, da+2, da+1, da+0; where ba+0 and da+0 are the anchor points of the spectrum). In ideal categorical perception, participants should produce a step function for identification accuracy in that all ba+ stimuli should be categorized as /ba/ and all da+ stimuli should be categorized as /da/ stimuli. For continuous perception, however, accuracy should follow a linear function from one end of the continuum to the other.

Liebenthal et al. [2005] tested participants' auditory perception while undergoing functional magnetic resonance imaging on both a synthetic phonemic continuum and a spectrally matched nonphonemic continuum. The phonemic continuum was based on the phonemes /ba/ and /da/ which form a continuous auditory spectrum where the second formant transforms from a rise for /ba/ to a fall for /da/. Eight equally spaced stimuli from /ba/ to /da/ (e.g., ba+1, ba+2, ba+3, ba+4, da+4, da+3, da+2, da+1; where ba+1 and da+1 are the anchor points of the spectrum) were created by systematically modifying the F2 formant. The nonphonemic continuum was created by spectrally inverting the first formant (F1) of the phonemic sounds, increasing the slope of the third formant (F3) transition, and adding a dip to F1 of the second anchor. These latter two manipulations increased the discriminability of the nonphonemic continuum to be comparable to the phonemic continuum. The nonphonemic continuum was similarly interpolated by varying the center frequencies of the first five formants of the nonphonemic anchor points during the transition segment in equal steps. The basic behavioral results from Liebenthal et al. [2005] showed the classic identification and discrimination curves (See Figure 2).

Previous models of auditory categorical perception have been based on associative memories, and showed that categorical perception would occur when models were trained on two end points of a continuum and then tested on equally spaced points between the proto-types [Anderson et al., 1977]. The TRACE [McClelland & Elman, 1986] model used both bottom-up and top-down influences to model categorical perception at the phonetic level, but did not cover continuous perception. Still, other researchers [Guenther & Gjaja, 1996; Bauer, Der, & Herrmann, 1996] have used self-organizing maps trained with differential inputs (reflecting differences in auditory experience) to produce categorical and continuous perception of sounds. Finally, in a comprehensive review of models of categorical perception [Damper & Harnad, 2000], it has been argued that categorical perception is an emergent property of learning systems in general, and that computational models provide a much needed way of studying categorical perception.

One of the most salient aspects of previous models and human studies of categorical and continuous perception is that training on the novel stimuli (continuous stimuli) often is stopped after the model/participant reaches a specific criterion such as 90% correct. In this model and subsequent experiment, we use overtraining on the novel stimuli under the assumption that regular speech sounds (i.e., sounds that produce categorical perception) are essentially overtrained from everyday exposure. Furthermore, we focus our training on the anchor points of the stimulus continuum, as opposed to the full continuum.

3.1 Method

The stimuli and network architecture are described in the general methods section. For training purposes, we set $t = 0.1$, $g = 1.0$, $\sigma = 0.05$, $\varepsilon = 0.001$, and $\alpha = 0.1$. Two different simulations were actually performed for this stimulus set. In the first simulation, only the anchor points of the phonetic continuum were presented to the network for training. The network was initially given unsupervised training on the speech phonemes for 20 epochs, followed by supervised training for 1000 epochs (overtraining). The network performance was evaluated on the entire dataset at five different time points during training; during this testing phase, the network did not learn (i.e., weights were not updated).

In the second simulation, we took advantage of the probabilistic properties of the SDN, and trained on interleaved speech and non-speech stimuli. We presented the stimuli with a 9:1 ratio such that the speech stimuli would be overtrained by the time the nonspeech stimuli reached a criterion of greater than 90% correct on anchor points. Following training, the network was tested for identification of all 16

stimuli (8 phonetic, and 8 nonphonetic stimuli). In both simulations, we trained 10 separate, randomized networks and averaged their responses.

To evaluate network performance, responses were recorded using the equivalent of a two-alternative forced choice paradigm. A differentiation score [McClelland & Chappell, 1998] was computed for the two alternatives

$$D = \prod_{i=1}^{n} \left|1 - \mathrm{P}(T_i) - \mathrm{P}(A_i)\right| \tag{4}$$

where P(Ti) is the probability of the target pattern, and P(Ai) is the actual probability of the generated pattern. In this case, the target pattern was one of the anchor points. A decision was made when the absolute value of the log ratio of the two scores exceeded a threshold.

$$\left|\log(D_{/ba/}) - \log(D_{/da/})\right| \geq CRIT \tag{5}$$

Note that this decision process is based solely on the activity levels in the network and an external criterion that is set at a constant level for each individual network.

3.2 Results and Discussion

3.2.1 *Simulation 1: Learning Phonetic Categories*

In simulation 1, the network was simply trained on the phonetic pair to see whether or not an SDN would learn categorical perception. Figure 2a shows the performance of the network as a function of the number of training trials. Before supervised training (0 epochs), the network is essentially at chance as it has not learned to map the patterns onto the responses yet. By 250 epochs of supervised training, the network exhibits continuous perception of the stimuli. It should be noted that by this time, the network was identifying the anchor points with approximately 95% accuracy. If we had stopped training at this point because the performance had reached a set criterion, then we would have concluded that the network was unable to show categorical perception of stimuli. With further training, however the network begins to show categorical perception, as defined by a nonlinearity in the identification curve. After 1000 epochs of training, the network shows the classical categorical perception identification function (see Figure 3). It should be emphasized that the network was only ever trained on the anchor points (ba+0 and da+0), and never on any of the intermediate stimuli (ba+1, ba+2, ba+3, da+3, da+2, da+1).

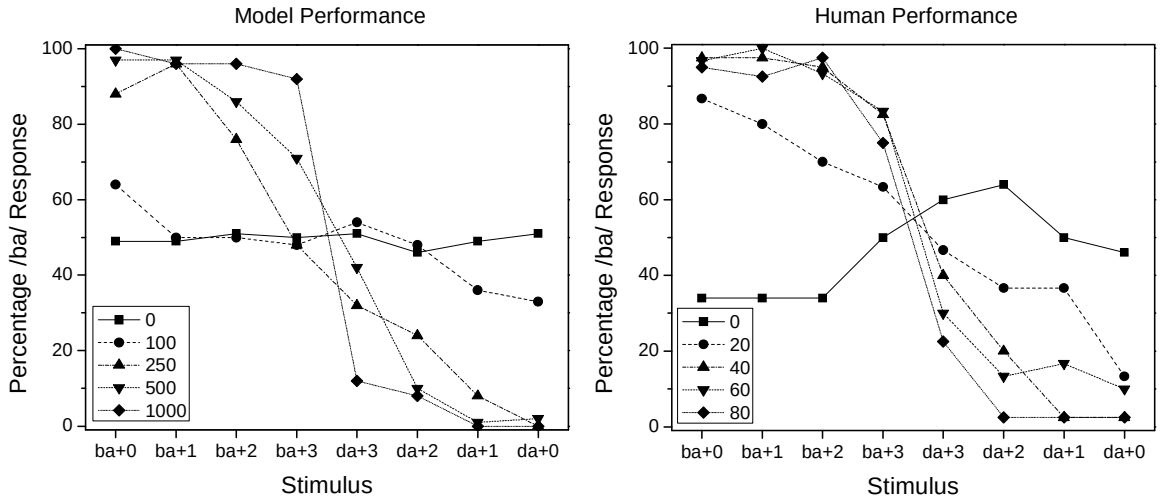

Figure 2. (a) Performance of the model trained on the phonetic stimuli. Note the progression from continuous to categorical perception with increased training. (b) Performance of humans on the nonphonetic stimuli showing progression from continuous to categorical perception with overtraining.

3.2.2 *Simulation 2: Learning Continuous and Categorical Perception*

In simulation 2, the network was trained with both the phonetic and non-phonetic stimuli. The network was trained with the phonetic stimuli on 90% of the trials, and the non-phonetic on 10% of the stimuli, until the non-phonetic stimuli reached criterion of 90% correct on the anchor points. Again, the network was only trained on the anchor points of the continuum, and not on the intermediate stimuli.

Figure 3 shows the performance of the model on the phonetic (P) and the nonphonetic (NP) stimuli. For comparison, the human performance data from Liebenthal et al. [2005] are shown as well. The first thing to note is the fairly good correspondence between the performance of the model and the performance of the human participants. There are two discrepancies in the performance of the model and the participants. The first is on the da+3 stimuli for the phonetic stimuli, where participants were around 50% in calling it a /ba/ stimulus, whereas the model was more likely to identify it as a /da/ stimulus. The second discrepancy is on the fact that participants had a harder time identifying the nonphonetic stimuli on the /da/ end of the scale, as shown by the fact that participants were only at approximately 20% for identifying the da+0 nonphonetic stimuli. More importantly, however, this simulation shows that a network can exhibit both categorical and continuous perception of stimuli. This result is due mainly to overtraining of a stimulus to produce categorical perception, and training to criterion to produce continuous perception.

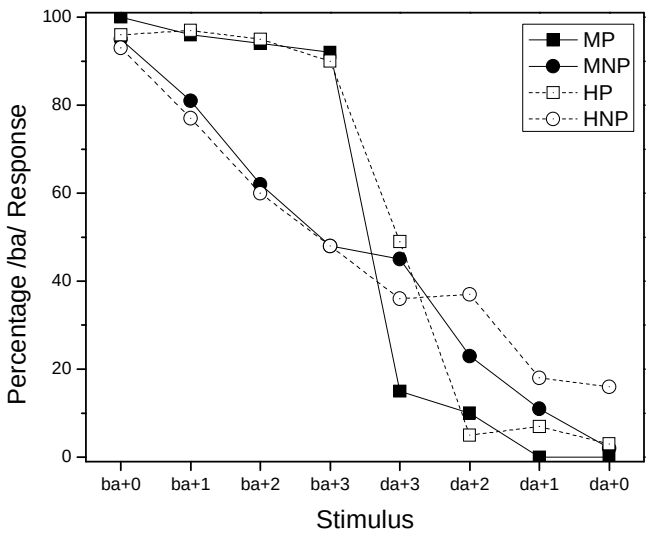

Figure 3. Performance of the human participants (H) and the model (M) on the phonetic (P) and nonphonetic (NP) identification task.

3.3 From Continuous to Categorical Perception

One of the predictions that can be taken from the model in Simulations 1 and 2 is that overtraining on the anchor points of a stimulus should be able to produce categorical perception. This is in contrast to previous research that found that categorical perception could not be learned simply by training on the anchor points [e.g., Guenther, Husain, Cohen, & Shinn-Cunningham, 1999]. This previous research, however, only trained participants to criterion; they did not overtrain their participants.

In this pilot experiment, participants were presented with the same nonphonetic stimuli as Liebenthal et al. [2005]. Initially, participants were given minimal exposure to the stimuli, and then tested on the identification task. Participants were then trained on the anchor points and tested after set learning periods.

3.3.1 *Methods*

Participants were six normal, healthy volunteers (3 females) with no reported hearing difficulties. Informed consent was obtained from all participants. The stimuli were the same nonphonemic anchor points and testing stimuli as in Liebenthal et al. [2005], and used in the modeling simulations. All auditory stimuli were presented over KOSS UR20 headphones (Milwaukee, WI). Presentation of stimulus and data collection was controlled by E-Prime software (v 1.1; Psychology Software Tools, Inc.) run on an IBM A22m laptop computer.

The experiment had three components: (i) initial exposure to the anchor points, (ii) identification testing of the full continuum, and (iii) training on the anchor points. During the initial exposure, participants were given three samples of each of the two anchor points labeled as "sound 1" or "sound 2"; participants passively listened to these samples. Following this initial exposure, participants were tested on the full continuum of sounds. Each sound was randomly presented a total of 10 times (80 trials total). Participants were asked to identify the sound as either "sound 1" or "sound 2" by using either the left or the right mouse button. No response feedback was given during testing. Response selection and reaction times were recorded.

After initial testing, participants were then given 20 random samples of the two anchor points (10 of sound 1, 10 of sound 2) and were asked to identify the sounds as sound 1 or sound 2. Accuracy feedback was provided during training. Participants were then tested again on the full continuum. This training/testing procedure was for a total of four times. The proportion of "sound 1" responses was then calculated for each testing period and plotted.

3.3.2 Results and Discussion

The results for the five testing period are shown in Figure 2b. As can be seen, after limited exposure to the anchor points, participants are basically at chance for identifying the individual stimuli. With training, however, participants begin to show continuous perception of the stimuli (time points 20 and 40). It should be noted that participants in the Liebenthal et al. study were given a total of 40 training trials. Further training on the stimuli, however, show a progression towards more categorical perception, as defined by a step-like function between stimuli "ba+3" and "da+3". The entire testing and training session approximately 20 minutes, indicating that categorical perception of novel stimuli can be learned rapidly.

We showed in these simulations how a simple computational model can learn to discriminate between spectrograms of the synthetic speech sounds /ba/ and /da/. Initial perception of these phonemes was continuous as measured by the identification curve. With overtraining, however, perception of these phonemes became categorical. We then trained the model on the phonetic stimuli and spectrally similar non-speech sounds, until criterion was reached for the non-speech sounds. It is important to note that the speech sounds were being overtrained during this period as they were being presented nine times as often as the non-speech stimuli. Following training, the model showed categorical perception of the speech sounds, and continuous perception of the non-speech sounds.

A pilot study with human participants was then conducted to test the predictions of the model in terms of the learning curves showing continuous to categorical perception. Participants were repeatedly trained on the anchor points of the nonspeech continuum and then tested on the entire continuum. Results show that initially, participants only showed continuous perception of these nonspeech sounds. With repeated exposure, however, these nonspeech sounds came to be perceived categorically. These results are somewhat at odds with previous studies that have shown training on the anchor points is insufficient to produce categorical perception [Guenther et al., 1999]. These previous studies, however, only trained participants to a specific criterion. These current results suggest that categorical perception may be due in part to (over)exposure to auditory stimuli.

4. Auditory Object Identification Study: Binder et al. [2004]

Binder et al. [2004] conducted a fMRI study in which participants had to discriminate between two synthetic speech syllables, /ba/ and /da/, which differed only in their second and higher formant transitions (see Figure 1a & 1b). Participants were given a target syllable–either /ba/ or /da/–to monitor. Trials consisted of presentation of one of the syllables, followed by a 200-ms interstimulus interval, then presentation of the other syllable. Participants indicated whether the first or second syllable presented was the target syllable. To manipulate task difficulty, the test stimuli were presented at a constant 65 dB, and a simultaneous white noise mask was presented at varying levels that produced signal/noise ratios (SNR) of

+24 dB, +14 dB, +4 dB, -1 dB, or -6 dB. A simple auditory detection task (respond to a sinewave tone) was also presented to the participants to provide a baseline measure for comparisons to the SNR conditions. Participants completed 100 trials per scanning run (20 at each noise level), and a total of six runs. The behavioral results from this study are recapped in Figure 4a. As the level of the noise mask increased (i.e., a decrease in the SNR), accuracy decreased in accordance with a standard signal detection curve, whereas RT followed an inverted-U-shaped function.

We designed a computational model of auditory identification and tested the model on a task analogous to the one used by Binder et al. (2004). The model was first trained to criterion on the veridical representations of /ba/ and /da/, and then tested on the noise-masked stimuli. Model performance was behaviorally evaluated in terms of accuracy and RT.

4.1 Training and Testing Procedure

The stimuli and network architecture are described in the general methods section. For training purposes, we set $t = 0.1$, $g = 1.0$, $\sigma = 0.05$, $\varepsilon = 0.001$, and $\alpha = 0.1$. Initially, the network was presented with 25 epochs of unsupervised training. The network was presented with a training pattern clamped and no output pattern during the plus phase, and no clamped patterns during the minus phase. Within each epoch, each pattern was sampled once (1 restart), and the network activations were updated over 10 time steps (with co-occurrence statistics collected over the last five time steps). Weights were updated after each pattern presentation. The network was then presented with supervised training for 100 epochs (input and output patterns clamped during the plus phase, only input pattern clamped during the minus phase). Again, each pattern was sampled once during the epoch, and allowed to settle for 10 time steps with co-occurrence statistics collected over the last five time steps. Weights were modified after each pattern presentation. Following training, the network discriminated the veridical training patterns with 100% accuracy. The same network parameters given above were used for the test stimuli, although weights were not updated during testing. Ten networks, with different initial starting weights, were trained on the problem. The 10 test patterns (2 syllables, each mixed with 5 levels of noise) were then presented to each network 50 times for testing, and responses were recorded using the differentiation score elaborated in Equations 4 and 5 to simulate the two-alternative forced choice paradigm.

Sample decision curves for the /ba/ and /da/ stimuli at the various noise levels are shown in Figure 5. One notable difference between the model and participants in Binder et al. [2004] is that the model does not have any learning processes or criterion adjustments during the testing phase, whereas the

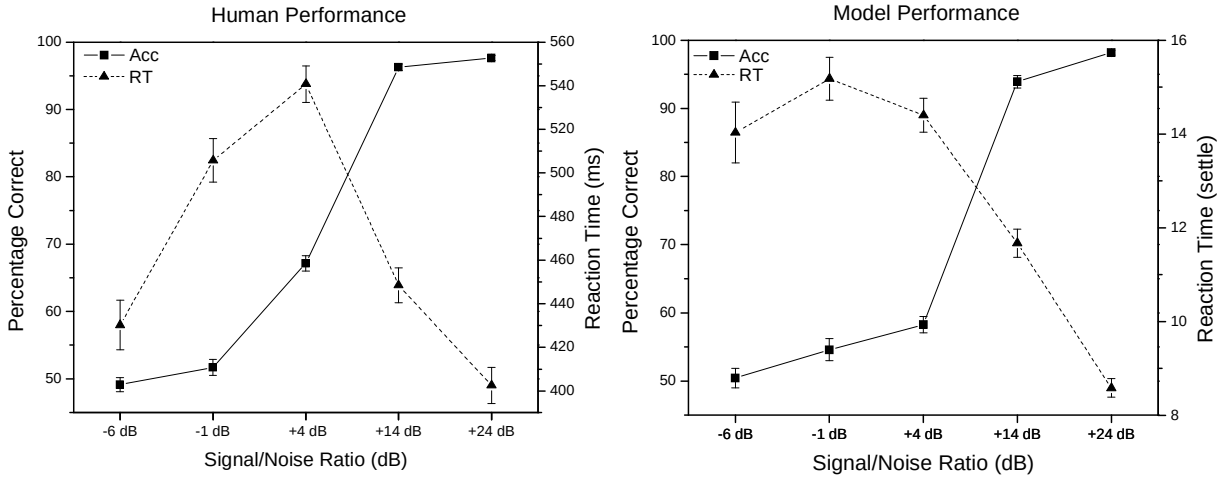

Figure 4. (a) Accruacy and RT performance from the participants in Binder et al. [2004]. (b) Performance of the model on an analogous task.

participants were free to adjust their decision criteria and task strategy. This discrepancy between the actual experiment and the simulation is addressed in the Results and Discussion section.

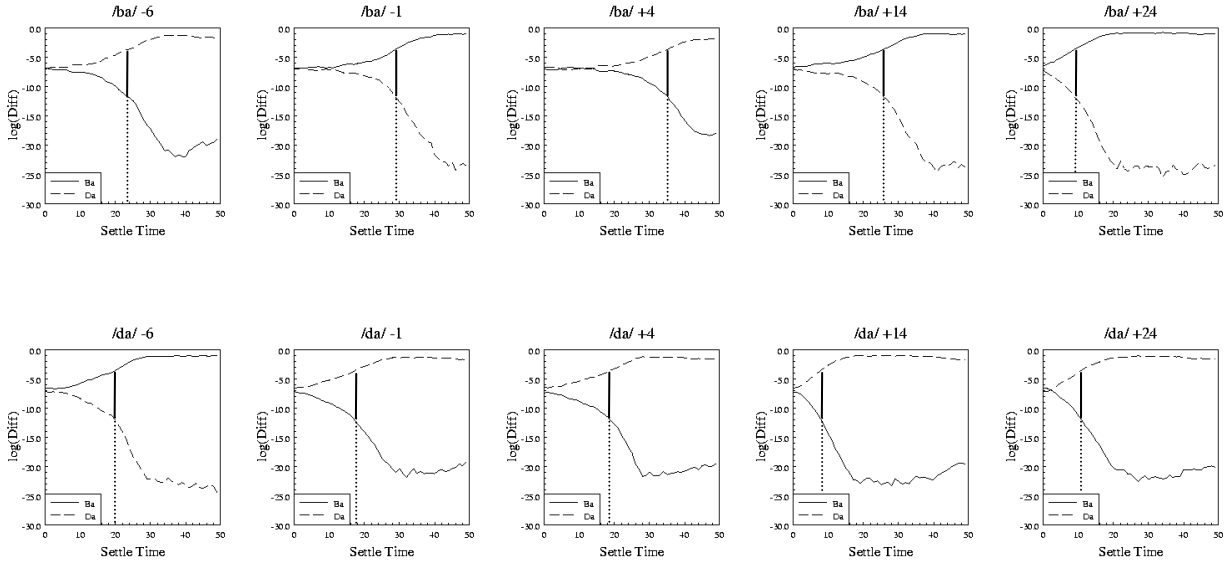

Figure 5. Sample differentiation scores and decision criterion for the /ba/ and /da/ stimuli with varying noise.

4.2 Results and Discussion

Ten networks, with different initial starting weights, were trained to discriminate the /ba/ and /da/ stimuli. Each network was given unsupervised training for 25 epochs, and then 100 epochs of supervised training. Following training, all networks discriminated the veridical training patterns with 100% accuracy. The network performance results from simulations using the noise-masked inputs are shown in Figure 4b. The network produced the same qualitative behavioral performance–in terms of accuracy and RT–as the human participants shown in Figure 4a. Network accuracy followed a standard signal detection curve, with 100% accuracy at the low noise level and chance performance at the higher noise levels. Reaction time, as measured by the number of network time steps before response criterion was reached, exhibited an inverted-U-shape function, but with one notable difference compared to the human participants. Specifically, the peak of the RT function produced by the network was at the -1 dB SNR condition rather than at -4 dB.

The small discrepancy in the simulated RT response over the output units could be due to the design of the original fMRI study. Recall that in the fMRI study, participants completed 600 trials across six scanning runs (100 trials/run). During this extensive testing, it is likely that subjects learned strategies for optimizing speed-accuracy trade-offs. For example, as the

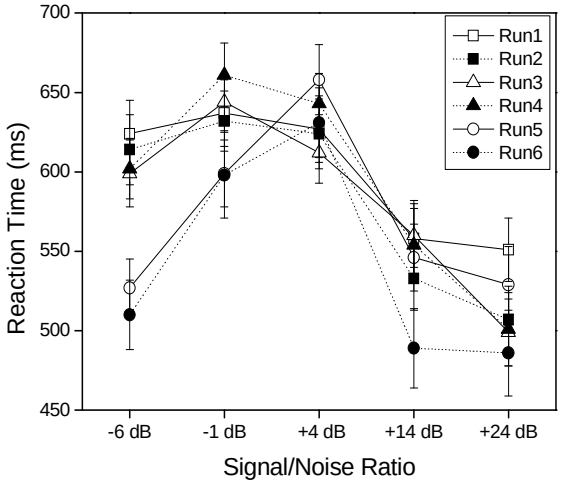

Figure 6. Reaction time performance of human participants as a function of experimental run. Note that the first 4 runs match the model data quite well.

experiment progressed, it is likely that subjects learned that the very low SNR conditions could not be performed accurately and so acquired a strategy based on "fast guessing" [Binder et al., 2004]. The model, on the other hand, was presented with a single trial per testing block and possessed no mechanisms for adjusting speed of responses based on likelihood of accuracy. Follow-up analyses of the original human data showed an interesting evolution in the response pattern across runs. Figure 6 shows that RTs to the highest noise conditions (-6 dB and -1 dB SNR) became faster over the last two runs (1159 msec) when compared to the first four runs (1227 msec) ($F = 12.75$, $p < .01$). Furthermore, the peak of the RT response occurred at -1 dB SNR for the first four runs and shifted to the +4 dB SNR condition only on the last two runs. This suggests that participants shifted strategies over the course of the experiment, effectively trading a negligible chance at discrimination accuracy for much faster RTs at the highest noise levels. The current model does not have this type of strategy shift capability, and therefore is only a model of the first run of trials; regardless, it models the initial human RTs quite well.

5. General Discussion

In this paper, we showed how a simple symmetric diffusion network trained with contrastive Hebbian learning is able to successfully model early auditory processing. Specifically, the model was trained and tested on two different auditory tasks. In the first task, the model was trained on phonetic and nonphonetic speech stimuli presented to participants in Liebenthal et al. [2005]. It was shown that when trained to criterion, the model showed simple continuous perception of the stimuli; however, when the model was overtrained on the anchor points of the stimulus continuum, then the model showed categorical perception of the stimuli as assessed by identification curves. Both continuous perception of nonspeech stimuli and categorical perception of speech stimuli was then shown in a single model; this was accomplished by interleaving the training of the stimuli such that the speech stimuli were presented nine times as often as the nonspeech stimuli. It should be noted at this point that the model had no preconceived notion of what a speech sound was and what was a nonspeech sound (as both stimuli were spectrally similar). The results of the network were due simply to the training paradigm presented. Indeed, follow-up simulations in which the speech and nonspeech stimuli were trained in a 1:9 ratio showed the opposite effects; that is, the model showed categorical perception with the nonspeech sounds and continuous perception with the speech sounds. Importantly, the model was able to predict that human participants should also be able to show categorical perception of nonspeech sounds when overtrained on the anchor points of the nonspeech continuum.

In the second auditory task, the model was trained on the speech sounds /ba/ and /da/ and then tested on noised-versions of these stimuli in a task analogous to Binder et al. [2004]. Again, the model was able to qualitatively capture the accuracy data of the participants. An interesting discrepancy in the reaction time data between the model and the participant data led to a reanalysis of the original human data. This reanalysis showed that the model was able to qualitatively capture the reaction time data of the participants before the participants were able to implement a strategy shift.

In conclusion, this simple model of early auditory processing [cf., Husain et al., 2004] that used realistic time varying inputs was able to correctly model existing data and predict human learning patterns of novel stimuli. Further models that are more biologically realistic in terms of auditory processing stages are currently in development, as well as a method of mapping the network activities onto a model of the blood oxygenation level dependent signal to match the imaging data of the two studies presented here.

Acknowledgements

The author thanks Jeffrey R. Binder, and Einat Liebenthal for helpful commentary and original behavioral and imaging data. This research was supported by National Institute of Neurological Diseases and Stroke grant R01 NS33576, National Institute for Deafness and Communication Disorders grant 2 R01 NS/DC33576-08, National Institutes of Health General Clinical Research Center grant M01 RR00058, and Medical College of Wisconsin RAC grant 403-14.

Bibliography

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413-451.

Bauer, H.-U., Der, R., & Herrmann, M. (1996). Controlling the magnification factor of self-organizing feature maps. *Neural Computation, 8,* 757-771.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience, 7,* 295-301.

Clark, H. H. & Clark, E. V. (1977). *Psychology and Language*. New York: Harcourt Brace Jovanovich.

Damper, R. I., & Harnad, S. R. (2000). Neural network models of categorical perception. *Perception & Psychophysics, 62,* 843-867.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America, 100,* 1111-1121.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *The Journal of the Acoustical Society of America, 106,* 2900-2912.

Hebb, D. O. (1949). *The Organization of Behaviour*. New York: John Wiley & Sons.

Husain, F. T., Tagamets, M.-A., Fromm, S. J., Braun, A. R., & Horwitz, B. (2004). Relating neuronal dynamics for auditory object processing to neuroimaging activity: a computational modeling and an fMRI study. *NeuroImage, 21,* 1701-1720.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, in press*.

MacMillan, N. A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review, 84,* 452-471.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724-760.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86.

Medler, D. A. & McClelland, J. L. (2001). Improving the performance of Symmetric Diffusion Networks via biologically inspired constraints. In K. Marko & P. Werbos (Eds.), *IJCNN'01: Proceedings of the INNS-IEEE International Joint Conference on Neural Networks* (pp. 400-405). Washington, DC: IEEE Press.

Movellan, J. R. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretsky, J. L. Elman, T. J. Sejnowski, & G. E. Hinton (Eds.), *Connectionist Models: Proceedings of the 1990 Summer School (San Mateo, CA, 1990)* (pp. 10-17). Morgan Kauffmann.

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with Symmetric Diffusion Networks. *Cognitive Science, 17,* 463-496.

Peterson, C., & Hartman, E. (1989). Explorations of the mean field theory learning algorithm. *Neural Networks, 2,* 475-494.

Pisoni, D. B. (Ed.). (1971). On the nature of categorical perception of speech sounds .New Haven, CT, Haskins Laboratories.