Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 7473–7476, August 1994
Neurobiology

A model of prenatal acquisition of speech parameters

BRADLEY S. SEEBACH*†, NATHAN INTRATOR‡§, PHIL LIEBERMAN¶, AND LEON N COOPER*§‖

Departments of *Neuroscience, ¶Cognitive and Linguistic Sciences, and ‖Physics, and §Institute for Brain and Neural Systems, Brown University, Providence, RI 02912; and ‡Sackler Faculty of Exact Sciences, Tel Aviv University, Ramat-Aviv 69978, Israel

Contributed by Leon N Cooper, April 18, 1994

ABSTRACT **An unsupervised neural network model inductively acquires the ability to distinguish categorically the stop consonants of English, in a manner consistent with prenatal and early postnatal auditory experience, and without reference to any specialized knowledge of linguistic structure or the properties of speech. This argues against the common assumption that linguistic knowledge, and speech perception in particular, cannot be learned and must therefore be innately specified.**

Chomsky's view that the "core" features of human linguistic ability are innate (1) is based in part on his assumption that linguistic knowledge, including the processes of speech perception, cannot be learned and thus must be preprogrammed. As this view is not universally accepted (2) and as there is some evidence for early alteration of phonetic perception by linguistic experience (3, 4), it appears useful to examine this assumption. In this paper we show that unsupervised neuron learning, as proposed to account for experimental data in visual cortex (5), can enable learned categorical perception of speech sounds with a reasonable approximation of the prenatal auditory environment. It thus follows that some aspects of early speech perception can be learned and therefore need not be innate.

Some of the strongest evidence for innate "linguistic" brain mechanisms comes from the study of speech. For example, the acoustic signals that differentiate stop consonants such as [b], [p], [d], [t], [g], and [k] are perceived categorically by adult (6) and infant (7) human listeners. The categorical behavioral responses of human adults and infants to these stop consonants has generally been interpreted as evidence for innate neural mechanisms tuned to the acoustic characteristics of speech (2, 7–10). In this view, linguistic development is not a learning process, but a process of selecting the discriminations useful to the maturing infant and forgetting those that are not useful (11). Supporting this belief is the discovery that infants, unlike adults, can discriminate phonetic units of languages they have never heard (12, 13). It is believed that such complex, cognitive behaviors of infants cannot arise from prenatal, experience-dependent modification of neurons.

However, such modifications have been shown to play a critical role in the development of neuronal selectivity. In visual cortex, for example, experience-dependent development proceeds rapidly from the onset of visual function through the so-called "critical period" and is strongly dependent on the visual environment in which the animal is raised. In the following study, we show that the prenatal auditory environment combined with a model of neuronal modification similar to that proposed for visual cortex can account for the acquisition of some basic speech contrasts as well as categorical perception of speech sounds.

The onset of hearing for humans begins as early as the 24th week of gestation (14, 15), raising the possibility that a lengthy "critical period" for auditory development may take place during the last several months of fetal life (16). Clearly, auditory experience in immature animals can alter frequency tuning (17) and spatial mapping (18) in auditory centers of the brain, and cognitive studies of human infants have shown that both prenatal (3) and postnatal (4) experiences may alter aspects of human speech perception prior to language acquisition. The fetus develops in an acoustically rich environment including the mother's voice. Low-frequency sounds dominate (19), whereas pure tones with higher frequencies (from external sources) are more attenuated. A certain amount of masking of low-frequency sounds is to be expected, though, due to the presence of low-frequency intrauterine noise, and tests of fetal hearing commonly use frequencies ranging from 500 Hz to 4 kHz. Low-frequency, broad-band noises are expected to be most efficient in producing responses in such tests (20). No adequate characterization of the transfer functions of the fetal middle ear exists (21).

The auditory periphery is characterized by broad bandpass tuning and poor phase-locking abilities during early mammalian development (22), though the "circuits" passing encoded information to auditory cortex appear to develop as functional units (for example, see ref. 23). Thus, encoded information reaching auditory cortical areas early on seems likely to consist of broad-band frequency information with consistent measures of intensity, but little or no phase information.

Fig. 1 shows an acoustic energy surface of the consonant–vowel (CV) syllable [ta] processed in a manner consistent with these constraints. A high degree of overlap in both time and frequency dimensions produces a "smooth" energy surface. Speech sounds typically display this type of energy surface, with peaks and valleys running in the general direction of the time axis. This stimulus is impoverished in comparison with those often used in speech studies but captures essential qualities of sounds that might be transmitted to the neonate's auditory centers: broad-band frequency information, little or no phase information, and fairly accurate representations of intensity. CV syllables processed in this manner were used as the speech data base for training and testing a neural network model.

We used a neural network based on the work of Bienenstock, Cooper, and Munro (BCM) (5) to determine whether the neural circuitry essential to speech perception might develop inductively in cortical auditory centers. The BCM theory has been used to describe the outcome and kinetics of experience-dependent synaptic plasticity in kitten striate cortex (25, 26). The mechanism for learning used by BCM is one that may be active in immature auditory cortex, as it requires only experience with sensory information.

Consider a neuron with input vector d $(=d_1, \ldots, d_n)$, synaptic weight vector m $(=m_1, \ldots, m_n)$, both in R^n, and activity (in the linear region) $c = m \cdot d$. The essential properties of the BCM neuron are determined by a modification threshold Θ_m (which is a nonlinear function of the history of activity

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: CV, consonant–vowel; BCM, Bienenstock, Cooper, and Munro.

†Present address: Department of Neurobiology and Behavior, State University of New York, Stony Brook, NY 11794.

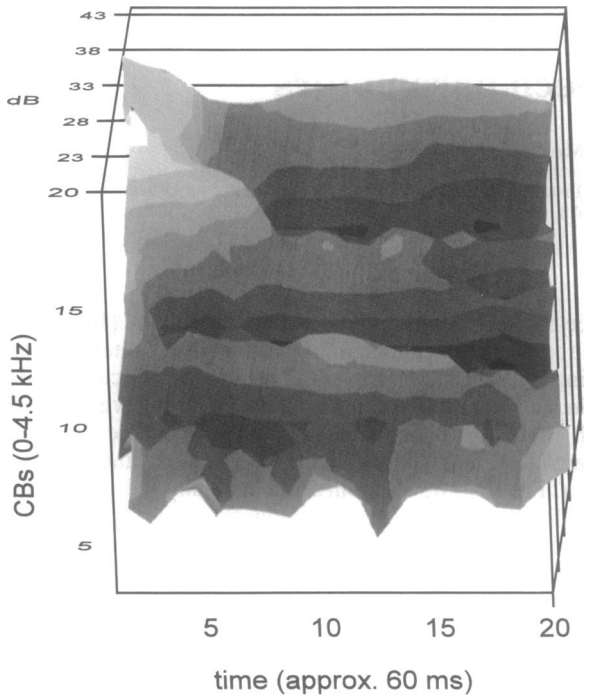

FIG. 1. The energy surface of approximately the first 55 ms of a pronunciation of the syllable [ta] by speaker BR. The ordinate is a perceptually defined scale of frequency known as critical bands (CBs) (24), and due to filter settings, represents frequencies from approximately 70 Hz to 4.5 kHz in a semilogarithmic manner. The abscissa is time, represented by 20 overlapping, 32-ms half-Hamming sampling windows. Each successive window was advanced 2 ms, so that the represented starting times for sampling windows ranged from 0 to 38 ms following the consonantal release. Higher decibel (dB) levels are represented by lighter-gray peaks.

of the neuron) and a ϕ function that determines the sign and amount of modification and depends on the current activity and the threshold Θ_m. The synaptic modification equations are given by

$$\frac{dm_i}{dt} = \mu\phi(c, \Theta_m)d_i, \qquad [1]$$

where in a simple form $\Theta_m = E[(m \cdot d)^2]$ and $\phi(c, \Theta_m) = c(c - \Theta_m)$.

In a lateral inhibition network of nonlinear neurons the activity of neuron k is given by $c_k = m_k \cdot d$, where m_k is the synaptic weight vector of neuron k. The inhibited activity and threshold of the kth neuron is given by $\bar{c}_k = \sigma(c_k - \eta \Sigma_{j \neq k} c_j)$ and $\hat{\Theta}_m^k = E[\bar{c}_k^2]$, for a monotone saturating function σ. The resulting stochastic modification equations for a synaptic vector m_k in such a network are given by

$$\dot{m}_k = \mu\left[\phi(\bar{c}_k, \tilde{\Theta}_m^k)\sigma'(\bar{c}_k) - \eta\sum_{j \neq k}\phi(\bar{c}_j, \hat{\Theta}_m^j)\sigma'(\bar{c}_j)\right]d. \qquad [2]$$

This network is actually a first-order approximation to a lateral inhibition network. Its properties were discussed by Intrator and Cooper (27).

A neuronal delay-line mechanism similar to that proposed by Jeffress (28) is assumed to provide the network with a sequence of acoustic events beginning with the syllable onset. Many species, including humans (29), use neuronal delay lines for sound localization.

The phonetic features distinguishing English stop consonants are place of articulation and voicing (6–8, 10). Place of articulation can be viewed as trinary (labial, alveolar, or velar) and voicing as binary (voiced or unvoiced). The problem then becomes to determine whether neurons can learn to detect these phonetic features in an appropriate environment without explicit preprogramming. The training paradigm we used is aimed at finding subphonemic features distinguishing between the different places of articulation independently of the voicing information. This is conceptually different from learning to recognize consonants as if they are unrelated, indivisible units, and requires a different training paradigm. For example, if one wishes to use a "supervised" model to recognize the stop consonant [k], one might train the network on a data base containing examples of all stop consonants in a variety of contextual situations (for example, ref. 30). In this manner, the net might be "taught" to accept all [k] sounds and reject all non-[k] sounds in many different environments. However, if one's purpose is to develop a neuron's selectivity for a subphonemic feature that distinguishes the consonant [k] from the consonants [p] and [t] (place of articulation), then the best training set is one that contains only that feature distinction.

Such a reduction of available phonetic distinctions in the training data base may typify a real process of development. It may be useful to think of the developing, peripheral sensory system as passing to fetal cortex a very simplified training set in the earliest stages—poorly focused frequency information, no phase locking, etc., in our example—and gradually increasing the complexity of available information, allowing for a progressive, cumulative development and refinement of neuronal selectivities.

A BCM network containing five "cells" was trained on a set of 74 pronunciations of [pa], [ka], and [ta] (unvoiced-stop syllables plus [a]). These syllables were pronounced by a single speaker and were not normalized for loudness or speaking rate. Averages of the three syllable types are shown as gray-scale images in Fig. 2*A*. The network was trained by random presentations of the 74 tokens until changes in neuronal selectivity became minimal. In the experiments reported here, 5 features were extracted from the 440-dimensional original space. The synaptic weights developed by each of the five cells have been reconstructed graphically (Fig. 2*B*) with the same axes as the training syllables (Fig. 2*A*), so that comparisons between the auditory inputs and the selectivities that developed might be made.

After training, cell 4 responded strongly to CV syllables containing labial stops and had excitatory synaptic weights corresponding to a large, low-frequency burst area and the region of the second formant frequency (F2) for [a]. These are the distinctive features of [pa] in the training set (Fig. 2*A*). Excitatory synaptic weights did not develop corresponding to strong but nondistinctive features of [pa], such as high energy in mid-range frequencies in the earliest time frames. Cells 1 and 5 responded most strongly to alveolar stop CV syllables. Both captured the short, high-frequency burst of [ta] and had negligible or inhibitory synaptic weights in high-frequency regions following the burst. An alveolar stop was identified when both of these cells responded strongly. Cells 2 and 3 together produced a strong response to velar-stop CV syllables. Both cells developed excitatory synaptic weights in extensive high-frequency regions, corresponding to the long duration of high-frequency burst energy for velar stops. The synaptic development of these cells differed in some respects: cell 2 had excitatory weights corresponding to a mid-frequency burst (often associated with [ka]), whereas cell 3 had excitatory weights in high-frequency areas that faded into the region of F3.

The BCM network effectively reduced the dimensionality of the original problem space from 440 dimensions to the 5 dimensions corresponding to the cells' selective responses. These cell selectivities correspond to subphonemic features. In order to interpret the results, a statistical classifier was trained to "phonemically" classify the output of the net's five cells as they responded to a testing data base. Although the

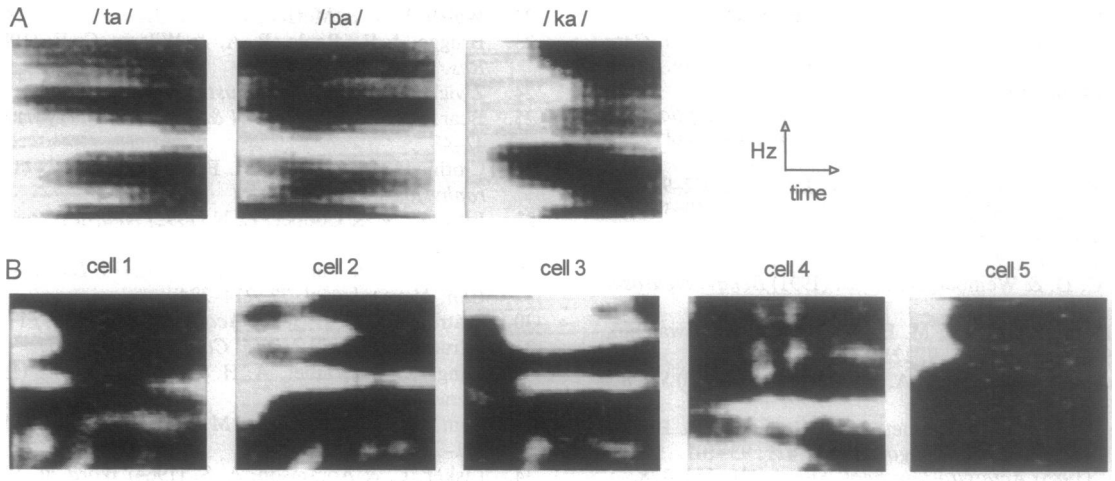

FIG. 2. (*A*) Average energy contour for each of the three training syllable types for speaker BR, shown as gray-scale images. The lighter areas of these images represent the presence of significant energy, with the ordinates of each image representing increasing frequency on the critical-band scale, and the abscissa of each image representing increasing time, which goes from 0 to 38 ms as marked by the start of each sampling window. (*B*) Gray-scale images of synaptic weights for the five cells of a BCM network following training.

BCM network was trained only with the unvoiced tokens of a single speaker, the classifier was trained on (5-dimensional) voiced and unvoiced data from the other speakers as well. The unsupervised feature extraction/classification method was discussed by Intrator (31). Results of the classification of training-set stimuli and testing stimuli from all three speakers are presented in Table 1.

Novel unvoiced-stop syllables from two different speakers, one male and one female, were correctly classified at a 98% rate. In contrast, neural network speech recognizers working on phoneme identification tasks have been highly successful in speaker-dependent tasks on entrained syllable types (30), and utterance recognition systems have been successful with multiple, novel speakers over very limited vocabularies (32, 33). Novel voiced-stop syllables were also successfully classified 96% of the time, despite the absence of voiced-stop syllables in the training data base. The generality of solutions is atypical for speech recognition systems and indicates that the BCM net is discovering features that yield categorical place-of-articulation distinctions.

Such an ability to internalize distinctive features of environmental sounds could explain an infant's ability to discriminate features of languages they have never heard. Phonemic features are defined perceptually. For example, Hindi has types of stop consonants that are distinguished by the presence or absence of aspiration (34). An adult speaker of Hindi can reliably produce and perceive this distinction. Adult speakers of English produce aspirated stop consonants under certain circumstances, though they no longer perceive a difference between an aspirated stop and an unaspirated stop. Because our supposed prenatal auditory neural network learns distinctive features of its sound environment, the child of an English-speaking woman could learn to make phonetic distinctions which his or her mother produces but cannot perceive.

Table 1. Classification results

	% correct (*n*)		
Speaker	Unvoiced stops	Voiced stops	Total by speaker
BR	99 (74)*	95 (73)	97 (147)
LN	98 (45)	91 (44)	94 (89)
JS	99 (75)	100 (75)	99 (150)
Total	98 (194)	96 (192)	97 (386)

*Unvoiced stops from speaker BR were the training set for this sample run.

The network and its training paradigm present a different approach to speaker-independent speech recognition. In this approach the speaker variability problem is addressed by training a network that concentrates on the distinguishing features of a single speaker, as opposed to training a network that concentrates on both the distinguishing and common features, on multi-speaker data.

Although we cannot yet be sure that the features discovered are invariant, the high degree of generalization across states of voicing, loudness, and speakers of both sexes gives reason to believe that neuronal selectivities such as those that develop in this model might provide a basis for perceptual abilities seen in early infancy. It thus appears that the assumption that a complex, cognitive behavior such as categorical perception of speech sounds cannot have its roots in prenatal experience is incorrect.

The questions become empirical. What is the appropriate prenatal auditory experience? Is this sufficient to produce the phonetic perceptual abilities underlying infant phonetic discriminations? Finally, what is really happening in this phase of early auditory development?

We wish to thank the members of the Institute for Brain and Neural Systems for many helpful discussions. Speech preprocessing was done at the Cognitive and Linguistic Science Department of Brown University. Research was supported by the National Science Foundation, the Office of Naval Research, and the Army Research Office.

1. Chomsky, N. (1986) *Knowledge of Language: Its Nature, Origin and Use* (Prager, New York).
2. Lieberman, P. (1991) *Uniquely Human: The Evolution of Speech, Thought, and Selfless Behavior* (Harvard Univ. Press, Cambridge, MA).
3. DeCasper, A. J. & Spence, M. J. (1986) *Infant Behav. Dev.* **9**, 133–150.
4. Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. & Lindblom, B. (1992) *Science* **255**, 606–608.
5. Bienenstock, E. L., Cooper, L. N & Munro, P. W. (1982) *J. Neurosci.* **2**, 32–48.
6. Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. (1967) *Psychol. Rev.* **74**, 431–461.
7. Eimas, P. D., Sisqueland, E. R., Jusczyk, P. & Vigorito, J. (1971) *Science* **171**, 304–306.
8. Blumstein, S. E. & Stevens, K. N. (1979) *J. Acoust. Soc. Am.* **66**, 1001–1017.
9. Lieberman, P. (1984) *The Biology and Evolution of Language* (Harvard Univ. Press, Cambridge, MA).
10. Chomsky, N. & Halle, M. H. (1968) *The Sound Pattern of English* (Harper & Row, New York).

11. Piattelli-Palmarini, M. (1989) *Cognition* **31**, 1–44.
12. Eimas, P. D., Miller, J. L. & Jusczyk, P. W. (1987) in *Categorical Perception*, ed. Harnad, S. (Cambridge Univ. Press, New York), pp. 161–195.
13. Kuhl, P. K. (1987) in *Handbook of Infant Perception*, ed. Salapatek, P. & Cohen, L. (Academic, New York), Vol. 2, pp. 275–381.
14. Murphy, K. P. & Smyth, C. N. (1962) *Lancet* **5**, 972–973.
15. Johansson, B., Wedenberg, E. & Westin, B. (1963) *Acta Oto-Laryngol.* **57**, 188–192.
16. Eggermont, J. J. (1986) *Acta Oto-Laryngol. Suppl.* **429**, 5–9.
17. Condon, C. D. & Weinberger, N. M. (1991) *Behav. Neurosci.* **105**, 416–430.
18. Moore, D. R., Hutchings, M. E., King, A. J. & Kowalchuk, N. E. (1989) *J. Neurosci.* **9**, 1213–1222.
19. Armitage, S. E., Baldwin, B. A. & Vince, M. A. (1980) *Science* **208**, 1173–1174.
20. Graniere-Deferre, C., Lecanuet, J. P., Cohen, H. & Busnel, M. C. (1985) *Acta Oto-Laryngol. Suppl.* **421**, 93–101.
21. Rubel, E. (1985) *Acta Oto-Laryngol. Suppl.* **421**, 114–128.
22. Walsh, E. J. & McGee, J. (1987) *Hearing Res.* **28**, 97–116.
23. Brugge, J. F., Reale, R. A. & Wilson, G. F. (1988) *Hearing Res.* **34**, 127–140.
24. Zwicker, E. (1961) *J. Acoust. Soc. Am.* **33**, 248.
25. Bear, M. F., Cooper, L. N & Ebner, F. F. (1987) *Science* **237**, 42–48.
26. Clothiaux, E. E., Bear, M. F. & Cooper, L. N (1991) *J. Neurophysiol.* **66**, 1785–1804.
27. Intrator, N. & Cooper, L. N (1992) *Neural Networks* **5**, 3–17.
28. Jeffress, L. A. (1948) *J. Comp. Phys. Psychol.* **41**, 35–39.
29. Jones, S. J. & Van der Poel, J. C. (1990) *Electroencephalogr. Clin. Neurophysiol.* **77**, 214–224.
30. Watrous, R. L. (1990) *J. Acoust. Soc. Am.* **87**, 1753–1772.
31. Intrator, N. (1992) *Neural Comput.* **4**, 98–107.
32. Lang, K. J., Waibel, A. H. & Hinton, G. E. (1990) *Neural Networks* **3**, 23–43.
33. Tom, M. D. & Tenorio, M. F. (1991) *Neural Networks* **4**, 711–722.
34. Lisker, L. & Abramson, A. S. (1964) *Word* **20**, 384–442.