Evolution of a Rapidly Learned Representation for Speech

Ramin Charles Nakisa (RAMIN@PSY.OX.AC.UK)
Kim Plunkett (PLUNKETT@PSY.OX.AC.UK)
Department of Experimental Psychology
Oxford University
South Parks Road
Oxford OX1 3UD

Abstract

Newly born infants are able to finely discriminate almost all human speech contrasts and their phonemic category boundaries are initially identical, even for phonemes outside their target language. A connectionist model is described which accounts for this ability. The approach taken has been to develop a model of innately guided learning in which an artificial neural network (ANN) is stored in a "genome" which encodes its architecture and learning rules. The space of possible ANNs is searched with a genetic algorithm for networks that can learn to discriminate human speech sounds. These networks perform equally well having been trained on speech spectra from any human language so far tested (English, Cantonese, Swahili, Farsi, Czech, Hindi, Hungarian, Korean, Polish, Russian, Slovak, Spanish, Ukranian and Urdu). Training the feature detectors requires exposure to just one minute of speech in any of these languages. Categorisation of speech sounds based on the network representations showed the hallmarks of categorical perception, as found in human infants and adults.

Introduction

Precocious abilities in newborn infants are frequently taken as evidence of "hard-wired microcircuitry" that is innately specified. One such ability is that of newborn infants to be universal listeners, able to discriminate speech contrasts of all languages. This is all the more remarkable since the low-pass filtered speech sounds that foetuses hear *in utero* vary widely between different languages.

Eimas et al. (1971) showed that 1-4 month old infants displayed categorical perception of the syllables /ba/ and /pa/. That is to say, infants carve up the phonetic space into a set of categories with sharp boundaries. Variants of phoneme, such as /b/, are not discriminable, even though they differ *acoustically* by the same amount as /p/ and /b/. More recent research has shown that the categories are universal, so that English-learning infants can discriminate non-native contrasts in Czech (Trehub, 1973), Hindi (Werker, Gilbert, Humphrey, & Tees, 1981), Nthlakampx (Werker & Tees, 1984), Spanish (Aslin, Pisoni, Hennessy, & Perey, 1981) and Zulu (Best, McRoberts, & Sithole, 1988). This suggests that infants develop an initial representation of speech that is universal and largely insensitive to the particular language to which they are exposed.

Many connectionist models of language acquisition take a fully developed featural or phonemic representation of the speech signal as their input rather than spectra (Christiansen, Allen, & Seidenberg, In press; Elman, 1990). This side-steps the problem of acoustic variability and makes the task of acquisition considerably easier. Such models would be more convincing if it could be demonstrated that suitable features could be rapidly learned well before word comprehension begins.

Description of the Model

The model builds on interactive activation models, with three major modifications:

Learning Each network learns using many different, unsupervised learning rules. These use only local information, and so are biologically plausible.

Flexible Architecture Every network is split into a number of separate subnetworks. This allows exploration of different neuronal architectures, and it becomes possible to use different learning rules to connect subnetworks. Subnetworks differ in their time-constants, and therefore respond to information over a range of time-scales.

Genetic Selection Networks are evolved using a technique called genetic connectionism. Using a genetic algorithm allows great flexibility in the type of neural network that can be used. All the attributes of the neural network can be simultaneously optimised rather than just the connections. In this model the architecture, learning rules and time-constants are all optimised together.

Genome Design and Sexual Reproduction

The genome has been designed to have two chromosomes stored as arrays of numbers. One chromosome stores the attributes of each subnetwork, such as the number of units in the subnetwork, the subnetwork time constant and the indices of the other subnetworks to which the subnetwork projects. The other chromosome stores learning rules which are used to modify connections between individual units.

During sexual reproduction of two networks the two chromosomes from each parent are independently recombined. In recombination, a point within a chromosome array is randomly chosen, and all the information up to that point is copied from the paternal chromosome and the rest of the chromosome is copied from the maternal chromosome creating a hybrid chromosome with information from both parents. Clearly, the subnetwork and learning rule chromosomes must be the same length for sexual recombination to occur, so not all pairs of parents can reproduce. Parents must be sexually compatible i.e. must have the same number of subnetworks and learning rules.

Dynamics

The dynamics of all units in the network are governed by the first order equation

$$\tau_n \frac{da_i^n}{dt} = \sum_{s,j} w_{ij}^{s \to n} a_j^s - a_i^n \qquad (1)$$

Where τ_n is the time constant for subnetwork n, a_j^s is the activity of the j^{th} unit in subnetwork s, a_i^n is the activity of the i^{th} unit in subnetwork n, $w_{ij}^{s \to n}$ is the synaptic strength between the j^{th} unit in subnetwork s and the i^{th} unit in subnetwork n. In other words, the rate of change in the activation of a unit is a weighted sum of the activity of the units which are connected to the unit i, minus a decay term. If there is no input to the unit its activity dies away exponentially with time constant τ_n. The activity of a unit will be steady when the activity of the unit is equal to its net input. Activities were constrained to lie in the range $0.0 \leq a \leq 1.0$. Network activity for all the units was updated in a synchronous fashion with a fixed time-step of 10 ms using a fourth order Runge-Kutta integration scheme adapted from Numerical Recipes (Press, Flannery, Teukolsky, & Vetterling, 1988).

Architecture

The architecture has to be stored in a "genome" to allow it to evolve with a genetic algorithm, and one very flexible method of encoding the architecture is to create a subnetwork connectivity matrix. If there are n subnetworks in the network, then the subnetwork connectivity matrix will be an n by n matrix. The column number indicates the subnetwork *from* which connections project, and the row number indicates the subnetworks *to* which connections project.

Complex architectures can be represented using a subnetwork connectivity matrix. The matrix allows diagonal elements to be non-zero, allowing a subnetwork to be fully connected to itself. In addition, the subnetwork connectivity matrix is used to determine which learning rule will be used for the connections between any pair of subnetworks. If an element is zero there are no connections between two subnetworks. A positive integer element indicates that subnetworks are fully connected and the value of the integer specifies which one of the many learning rules to use for that set of connections. A simple architecture is shown in Figure 1 alongside its corresponding subnetwork connectivity matrix.

Learning Rules

Learning rules are of the general form shown in equation 2. They are stored in the network genome in groups of seven coefficients k_0 to k_6 following the representation used by Chalmers (1990).

$$\Delta w_{ij} = l(k_0 + k_1 a_i + k_2 a_j + k_3 a_i a_j + k_4 w_{ij} + k_5 a_i w_{ij} + k_6 a_j w_{ij}) \qquad (2)$$

In Equation 2, Δw_{ij} is the change in synaptic strength between units j and i, l is the learning rate, a_i is the activity of unit i, a_j is the activity of unit j and w_{ij} is the current synaptic strength between units j and i. The learning rate l is used to scale weight changes to small values for each time step to avoid undesirably rapid weight changes. The coefficients in this equation determine which learning rule is used. For example, a Hebbian learning rule would be represented in this scheme with $k_3 > 0$ and $k_0 < 0$ and $k_1 = k_2 = k_4 = k_5 = k_6 = 0$. Connections between units using this learning rule would be strengthened if both units were simultaneously active. A network has several learning rules in its genome stored as a set of these coefficients. Weight values are clipped to avoid extremely large values developing over long training periods. The range used was $-1.0 \leq \Delta w_{ij} \leq +1.0$.

Training and Evaluation of Fitness

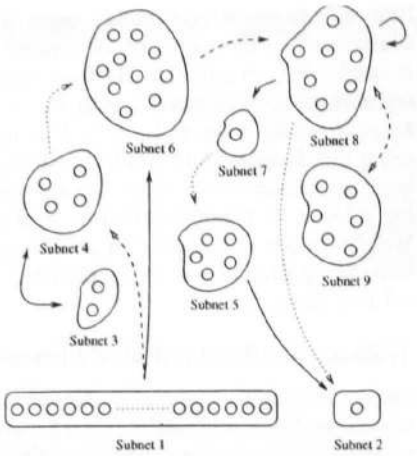
Networks were trained and evaluated using digitised speech files taken from the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) as described in Garofolo et al. (1990). All networks were constrained to have 64 input units because speech sounds were represented as power spectra with 64 values. This was an artificial constraint imposed by the format of the spectra. The power spectra were calculated with a modified version of the OGI speech tools program MAKEDFT with a window size of 10 ms and with successive windows adjacent to one another. For these simulations 8 output subnetworks were used to represent features because this is roughly the number claimed to be necessary for distinguishing all human speech sounds by some phoneticians (Jakobson & Waugh, 1979).

All the connections, both within and between subnetworks, were initially randomised to values between -1.0 and +1.0. Networks were then exposed to a fixed number of different, randomly selected training sentences (usually 30). On each time-step activity was propagated through the network of subnetworks to produce a response activity on the output units. All connections were then modified according to the learning rules specified in the genome. On the next time-step a new input pattern corresponding to the next time-slice of the speech signal was presented and the process of activity propagation and weight modification repeated. The process of integrating activities and weight updates was repeated until the network had worked its way through all the time-slices of each sentence.

In the testing phase activation was propagated through the network without weight changes. The weights were frozen at the values they attained at the end of the training phase. Testing sentences were always different from training sentences. When a time-slice corresponded with the mid-point of a phoneme, as defined in the TIMIT phonological transcription file, the output unit activities were stored alongside the correct identity of the phoneme. Network fitness was calculated using the stored output unit activities after the network had been exposed to all the testing sentences. The fitness function f was

$$f = \frac{\sum_i^N \sum_{j=i+1}^N dist(\vec{o_i}, \vec{o_j}) \cdot s}{N(N-1)} \qquad (3)$$

Where $s = +1$ if i and j are different phonemes and $s = -1$ if i and j are the identical phonemes, $\vec{o_i}$ and $\vec{o_j}$ were the output unit activities at the midpoint of all N phonemes and s was either +1 or -1 depending on whether phonemes i



$$C = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 1 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \end{pmatrix}$$

Figure 1: A network with 9 subnetworks. Subnetwork 1 and 2 are the input and output subnetworks, respectively. Arrows represent sets of connections and the type of learning rule employed by those sets of connections. There are three learning rules used; solid arrow (learning rule 1), dashed arrow (learning rule 2) and dotted arrow (learning rule 3). Some subnetworks are fully connected to themselves, such as subnetwork 8 (since $C_{88} = 1$), while others are information way-stations, such as subnetwork 5 ($C_{55} = 0$).

and j were different and $dist$ was euclidean distance. This fitness function favoured networks that represented occurrences of the same phoneme as similarly as possible and different phonemes as differently as possible. A perfect network would have all instances of a given phoneme type mapping onto the same point in the output unit space and different phonemes as far apart as possible. Note that constant output unit activities would result in a fitness of 0.0. An ideal learning rule would be able to find an appropriate set of weights whatever the initial starting point in weight space. Each network was trained and tested three times from completely different random initial weights on completely different sentences. This reduced random fitness variations caused by the varying difficulty of training/testing sentences and the choice of initial weights.

Evolution was carried out with a population of 50 networks. Genomes were initially generated with certain limits on the variables. All genomes had 16 input subnetworks and 8 output subnetworks with time constants randomly distributed in the range 100 ms to 400 ms. The input subnetworks had 4 units each and the output subnetworks had 1 unit each. Each network started with 10 different learning rules with integer coefficients randomly distributed in the range -2 to +2. Subnetwork connectivity matrices were generated with a probability of any element being non-zero of 0.3. If an element was non-zero, the learning rule used for the connections between the subnetworks was randomly selected from the 10 learning rules defined for the network. The networks were also constrained to be feed-forward.

Results

All results shown are from the best network evolved (fitness=0.45) after it had been trained on 30 English sentences corresponding to about 2 minutes of continuous speech. Figure 2 shows the response of this network to one of the TIMIT testing sentences. From the response of the feature units to speech sounds (see Figure 2) it was clear that some units were switched off by fricatives, and some units were switched on by voicing, so both excitation and inhibition play an important part in the functioning of the feature detectors. The feature unit responses did not seem to correlate directly with any other standard acoustic features (e.g. nasal, compact, grave, flat etc.). An analysis of the frequency response of the eight feature detectors (see Figure 3) showed that each unit had excitatory projections from several frequency bands. Generally, the frequency responses were mutually exclusive so that each unit responded to slightly different sounds, as one would expect.

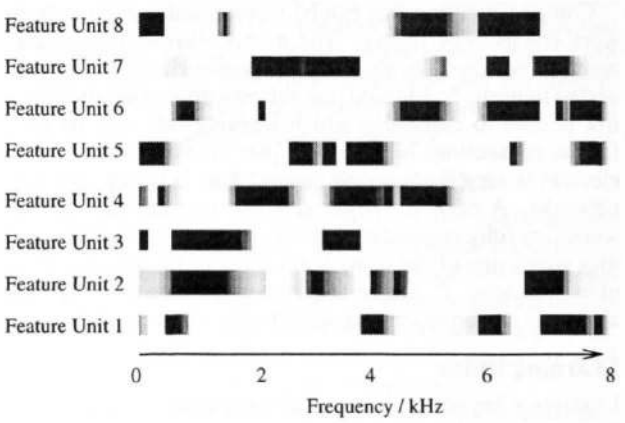


Figure 3: Complex frequency response of all eight feature units to pure tones. Feature units 2 and 3 receive strong excitatory inputs from low frequencies (below 4 kHz) and are therefore activated by voicing.

In order to determine the cross-linguistic performance of the "innate" features evolved on English speech, sound files of the news in several languages were obtained from the Voice of America FTP site (`ftp.voa.gov`). Since phonological transcription files were not available for these files they could not be used to test the network, because the times of the phoneme mid-points were unknown. All the VOA broadcast

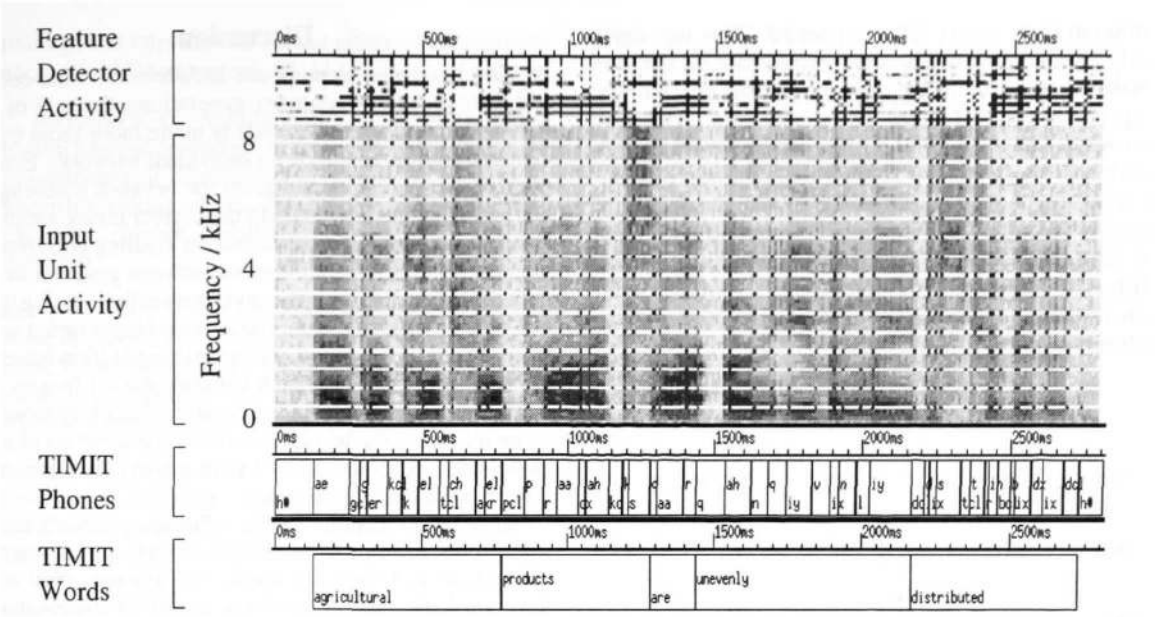


Figure 2: Network response to the sentence "Agricultural products are unevenly distributed" (TIMIT speech file test/dr3/fkms0/sx140). Input units are fed with sound spectra and activate the feature units. Activity is shown as a greyscale (maximum activity is portrayed as black) with time on the horizontal axis. Phone and word start and end times as listed in TIMIT are shown in the bottom two panels. This is the same network as shown in Figure 3.

languages[1] were used as training files, and the network was tested on 30 American English sentences found in the TIMIT speech files. The time-course of development for four languages are shown in Figure 4. Maximum fitness was reached after training on any language for roughly 20 sentences (each lasting about 3 seconds).

All of the human languages tested seemed to be equally effective for training the network to represent English speech sounds. To see whether *any* sounds could be used for training, the network was trained on white noise. This resulted in slower learning and a lower fitness. The fitness for a network trained on white noise never reached that of the same network trained on human speech. An even worse impediment to learning was to train on low-pass filtered human speech.

Categorical perception of some phonemes is a robust phenomenon observed in both infants and adults. We tested the network on a speech continuum ranging between two phonemes and calculated the change in the representation of the speech tokens along this continuum. Note that this model simply creates a representation of speech on which identification judgements are based. It does not identify phonemes itself. All that the model can provide is distances between its internal representations of different sounds. Categorical perception can be exhibited by this network if the internal representation exhibits non-linear shifts with gradual changes in the input i.e. a small change in the input spectrum can cause a large change in the activity of the output units.

Using a pair of real /ʃ/ and /s/ spectra from a male speaker, a series of eleven spectra were created which formed a lin-

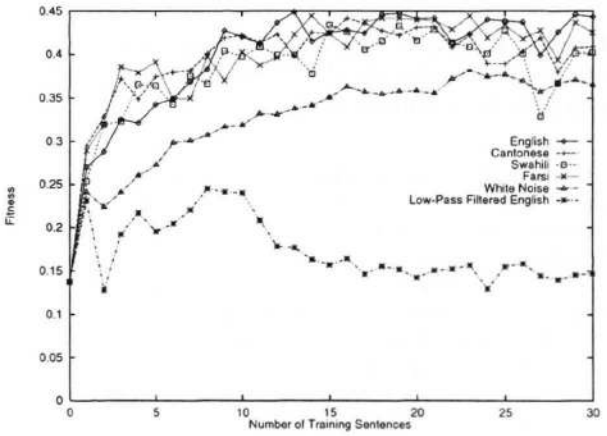


Figure 4: Network performance increases to its final value after presentation of just 20 sentences regardless of the language used to train the network. The six curves show the learning curves for a network tested on 30 sentences of English having been trained on English, Cantonese, Swahili, Farsi, white noise and low-pass filtered English.

[1]English, Cantonese, Swahili, Farsi, Czech, Hindi, Hungarian, Korean, Polish, Russian, Slovak, Spanish, Ukranian and Urdu.

ear continuum from a pure /ʃ/ to a pure /s/. This was done by linearly interpolating between the two spectra, so the second spectrum in the continuum was a linear sum of 0.9 times the /ʃ/ spectrum plus 0.1 times the /s/ spectrum. The next spectrum was a linear sum of 0.8 times the /ʃ/ spectrum plus 0.2 times the /s/ spectrum, and so on for all nine intermediate spectra up to the pure /s/. Each of the eleven spectra in the continuum were individually fed into the input of a network that had been trained on 30 sentences of continuous speech in English. The output feature responses were stored for each spectrum in the continuum. The distances of these feature vectors from the pure /ʃ/ and pure /s/ are shown in Figure 5.

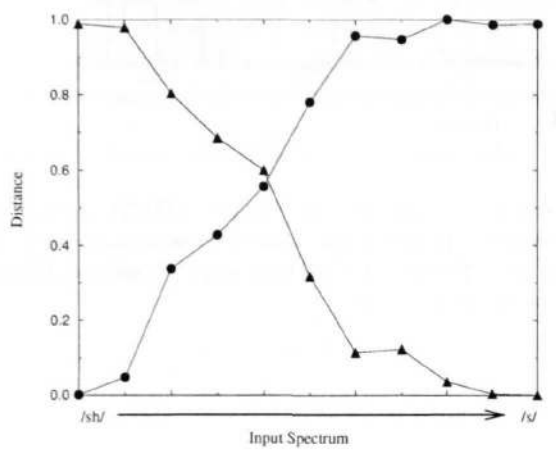


Figure 5: Response of the network to input on a /ʃ/ - /s/ continuum. Circles show the distance from a pure /ʃ/ and triangles show the distance from a pure /s/.

Clearly, the distance of the pure /ʃ/ from itself is zero, but moving along the continuum, the distance from the pure /ʃ/ increases steadily until it reaches a maximum for the pure /s/ (distances were scaled such that the maximum distance was 1). It is clear from Figure 5 that the representation is non-linear. That is, linear variations in the input spectrum do not result in linear changes in the activity of the feature units. Compared to the spectral representation of the /ʃ/ /s/ continuum, the network re-represents the distances in the following ways:

- There is a discontinuity in the distances which occurs closer to the /ʃ/ than the /s/.
- The distance from the representation of a pure /s/ remains small for spectra that are a third of the way toward the pure /ʃ/.

A classifier system using this representation would therefore shift the boundary between the two phonemes toward /ʃ/ and be relatively insensitive to spectral variations that occurred away from this boundary. These are the hallmarks of categorical perception.

Discussion

By developing an appropriate architecture, time-constants and learning rules over many generations, the task of learning to represent speech sounds is made more rapid over the course of development of an individual network. Evolution does all the hard work and gives the network a developmental "leg-up". However, having the correct innate architecture and learning rules is not sufficient for creating good representations. Weights are not inherited between generations so the network is dependent on the environment for learning the correct representation. If deprived of sound input or fed acoustically filtered speech input, the model cannot form meaningful representations because each network starts life with a random set of weights. But given the sort of auditory input heard by an infant the model rapidly creates the same set of universal features, whether or not it is in a noisy environment and whatever the language it hears.

We envisage that this method of creating a quick and dirty initial representation of sounds by innately guided learning is not specific to humans. Clearly, humans and other animals have not been selected for their ability to discriminate the phonemes of English. But we would expect results similar to those presented here if the selection criterion were the ability to discriminate a wide range of spectrally dissimilar sounds in the environment from only limited exposure to their patterns of regularity e.g. discrimination of the maternal call from other conspecific calls, and the sound of predators from everyday environmental noises. It is therefore unsurprising that animals have been found, after suitable training, to discriminate some phonemes in similar ways as do humans (Kuhl & Miller, 1975).

The advantages of innately guided learning over other self-organising networks are that it is much faster and is *less* dependent on the "correct" environmental statistics. It also offers an account of how infants from different linguistic environments can come up with the same featural representation so soon after birth. In this sense innately guided learning as implemented in this model shows how genes and the environment could interact to ensure rapid development of a featural representation of speech on which further linguistic development depends. In terms of the taxonomy of "ways to be innate" offered by Elman et al. (1996), this model is lacking in any form of representational innateness — there is no hard-wiring of the microcircuitry. On the other hand, the model exemplifies what Elman et al. call "architectural/computational innateness" — innate processing biases in the network make it ideally suited to extracting structural information from speech input when the opportunity presents itself. Speech offers the network a nutritious environment in which to grow.

Acknowledgements

Ramin Nakisa was supported by a Training Fellowship from the Medical Research Council. Further support was provided by research project grants from the EPSRC and ESRC to Kim Plunkett.

References

Aslin, R., Pisoni, D., Hennessy, B., & Perey, A. (1981). Discrimination of voice-onset time by human infants: New

findings and implications for the effect of early experience. *Child Development, 52*, 1135–1145.

Best, C., McRoberts, G., & Sithole, N. (1988). The phonological basis of perceptual loss for non-native contrasts: Maintenance of discrimination among Zulu clicks by English-speaking adults. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 345–360.

Chalmers, D. (1990). The evolution of learning: An experiment in genetic connectionism. In D. Touretzky, J. Elman, T. Sejnowski, & G. Hinton (Eds.), *Connectionist models: Proceedings of the 1990 summer school* (pp. 81–90). Morgan Kaufmann Publishers, Inc.

Christiansen, M., Allen, J., & Seidenberg, M. (In press). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes.*

Eimas, P., Siqueland, E., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*, 303–306.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179–212.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, Massachusetts: The MIT Press.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., & Dahlgren, N. (1990). *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM* (Tech. Rep. No. NISTIR 4930). National Institute of Standards and Technology, USA.

Jakobson, R., & Waugh, L. (1979). *The sound shape of language.* Bloomington: Indiana University Press.

Kuhl, P., & Miller, J. (1975). Speech perception by the chinchilla: Voiced–voiceless distinction in alveolar plosive consonants. *Science, 190*, 69–72.

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1988). *Numerical recipes in C: The art of scientific computing.* Cambridge, England: Cambridge University Press.

Trehub, S. (1973). Infants' sensitivity to vowel and tonal contrasts. *Developmental Psychology, 9*, 91–96.

Werker, J., Gilbert, J., Humphrey, K., & Tees, R. (1981). Developmental aspects of cross-language speech perception. *Child Development, 52*, 349–353.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganisation during the first year of life. *Infant Behaviour and Development, 7*, 49–63.