Evolution of a Rapidly Learned Representation for Speech

Ramin Charles Nakisa and **Kim Plunkett**
Department of Experimental Psychology
Oxford University
South Parks Road
Oxford OX1 3UD, UK
{ramin.nakisa,kim.plunkett}@psy.ox.ac.uk

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.

1 Introduction

Precocious abilities in newborn infants are frequently taken as evidence for pre-specification of the representations that support those abilities. The prespecifications of these representations is innately determined, presumably in the genome of the individual. 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 a phoneme, such as /b/, are not discriminable, even though they differ *acoustically* by the same amount as /p/ and /b/ (although see (Kuhl, 1993)). 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, 1984a), 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. The ability to discriminate some non-native speech contrasts declines after the age of 10–12 months (Werker & Tees, 1984a).

Such rapid learning can be defined in terms of a taxonomy developed in the field of animal behaviour. Mayr (1974) suggested that programs of development form a continuum of flexibility in their response to environmental stimulation. He distinguished between "open" and "closed" programs of development. "Closed" programs of development rely on environmental input to a relatively small degree, producing highly stereotyped behaviour. Precedents of rapidly learned "closed" development abound and are also termed "innately guided" learning e.g. imprinting in geese and ducks and song acquisition in birds (Marler, 1991). "Open" programs, on the other hand, are responsive to a much broader range of stimulation and can produce a broader range of responses. The presence of one type of developmental program does not preclude the exis-

Ramin Charles Nakisa and Kim Plunkett (1997) **Evolution of a Rapidly Learned Representation for Speech.** In T.M. Ellison (ed.) *CoNLL97: Computational Natural Language Learning,* ACL pp 70-79.
© 1997 Association for Computational Linguistics

tence of the other, however. Just because a duckling has imprinted on its mother does not mean that it is unable to learn to recognise new objects later in life. Similarly, the rapid learning of speech sounds by infants does not preclude later tuning of the speech representation. In fact, we would argue that it aids such development by ensuring that later language-specific fine-tuning of the representation does not encounter local minima, which would be catastrophic for linguistic development.

To quote from a recent review (Jusczyk, 1992):

> Jusczyk and Bertoncini (1988) proposed that the development of speech perception be viewed as an innately guided learning process wherein the infant is primed in certain ways to seek out some type of signals as opposed to others. The innate prewiring underlying the infant's speech perception abilities allows for development to occur in one of several directions. The nature of the input helps to select the direction that development will take. Thus, learning the sound properties of the native language takes place rapidly because the system is innately structured to be sensitive to correlations of certain distributional properties and not others.

In order to make explicit what is meant by "innately guided learning" and "innate prewiring" we have developed a connectionist model of innately guided learning. The approach taken has been to encode an artificial neural network (ANN) in a genome which stores its architecture and learning rules. The genomic space of possible ANNs is searched for networks that are well suited to the task of *rapidly* learning to detect contrastive features of human speech sounds using unsupervised learning. Importantly, networks start life with a completely randomized set of connections and therefore have no representational knowledge about speech at the level of individual connections. The network must therefore use its architecture and learning rules in combination with auditory input to rapidly converge on a representation.

The model attempts to explain how innate constraints on a neural network could allow infants to be sensitive to a wide range of features so soon after birth, and to develop the same initial features whatever their target language. It also exhibits other features typically associated with human speech perception, namely categorical perception and patterns of phoneme confusability similar to that of humans. The model does not account directly for the much slower, roughly year-long process by which some featural distinctions are lost. It is possible that features are never lost and that units which represent information that is redundant in the target language are ignored by higher level processing, as suggested by Werker and Tees (Werker & Tees, 1984b).

2 Overview of the Model

The goal of the model is to create a neural network that takes speech spectra as input and develops the same representation of speech whatever the language it is exposed to. Furthermore we avoid hard-wiring the connections in the network. Rather, the network employs a set of unsupervised learning rules that converge on the same representation whatever the initial set of connection strengths between neurons in the network. It is important that the learning is unsupervised as the developing infant has no teaching signal as to the contrasts present in speech. In essence this model of early speech perception embodies Waddington's (1975) principle of epigenesis, or what Elman et al. (1996) have more recently described as architectural/computational innateness.

The approach we have taken is to encode the properties of neural networks in a genome and to evolve, by a process called a genetic algorithm, a population of neural networks that respond in the appropriate way to speech spectra. Initially, a population of 50 genomes are randomly generated. Each of these networks is presented with speech spectra and we quantify how well its neuronal engram of speech encodes the incoming signal. This number is called the "fitness" of a network. For the task of representing speech sounds we want a network that is responsive to the salient aspects of the speech signal, in particular those necessary for identification of speech segments. A network that is good at representing speech will encode tokens of the same acoustic segment as similarly as possible and different segments as differently as possible.

The initial population performs very poorly on the task, but some networks perform better than others. Two parents are randomly selected from the population with a probability that increases with increasing fitness. The parental genomes are spliced together to form one child network that is then tested to find its fitness. The child network then replaces the network that has the lowest fitness in the population. Each gene also has a small chance of mutating to a new value after sexual reproduction, so that new genes are constantly entering the collective gene pool, otherwise the evolutionary process would simply be a re-shuffling of genes present in the initial population. The process of parental selection, sexual reproduction, mutation of the offspring and evaluation of the offspring is repeated for several thousand generations. Genes that are useful for the task at hand, as specified by the fitness function, increase in frequency in the population, while genes that are not useful decline in frequency. Within a few hundred generations the networks in the population develop representations that have a high fitness value, as illustrated in Figure 1.

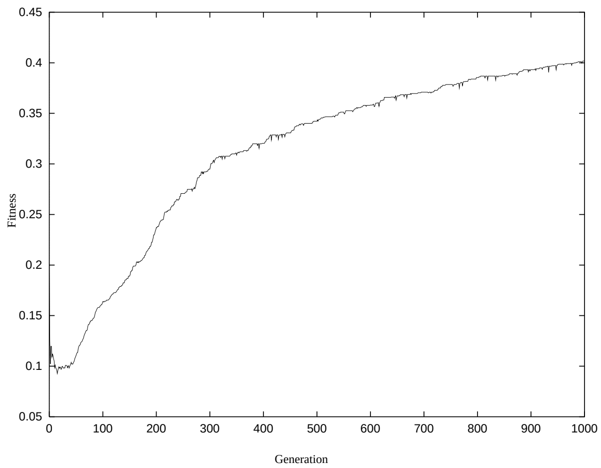


Figure 1: Increase in the mean fitness of the population with increasing number of generations, where a generation is defined as the production of one new network. Initially networks perform very poorly, but selection improves the population rapidly.

Clearly, the encoding scheme used to store the properties of neural networks critically affects how well the networks may perform on any given task. The encoding scheme we have chosen is very flexible, storing information about the architecture of a network and its learning properties. Architecture defines what neurons may be connected to other neurons, and this presupposes some way of grouping neurons such that these gross patterns of connectivity can be defined. For the purposes of defining network architecture, therefore, the network is subdivided into subnetworks. The genome specifies how many subnetworks there are, how many neurons are in each subnetwork what subnetworks are connected to one another, and given that two subnetworks are connected, what learning rule is used in connections between neurons in those subnetworks.

3 Description of the Model

The model builds on previous connectionist models, particularly the broad class of models known as interactive activation with competition (IAC) models (see Grossberg (1978) for review). An IAC network consists of a collection of processing units divided into several competitive pools. Within pools there are inhibitory connections and between pools there are excitatory connections. Connections are *interactive* because one pool interacts with other pools and in turn is affected by those pools. Because of these interactions the activity of units in IAC networks develop over time, sometimes settling into steady patterns of activation. Inhibitory connections within a pool mean that one unit at a time dominates the others in a winner-take-all fashion. The TRACE model of speech perception is possibly the most successful and best known example of such models (McClelland & Elman, 1986).

Although similar to IAC networks, the models described here have 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" i.e. respond to information over different time-scales.

Genetic Selection Networks are evolved using a technique called genetic connectionism (Chalmers, 1990). 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.

3.1 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.

3.2 Dynamics

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

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

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).

3.3 Architecture

Architecture defines the gross pattern of connectivity between groups of units. 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 2 alongside its corresponding subnetwork connectivity matrix.

3.4 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).

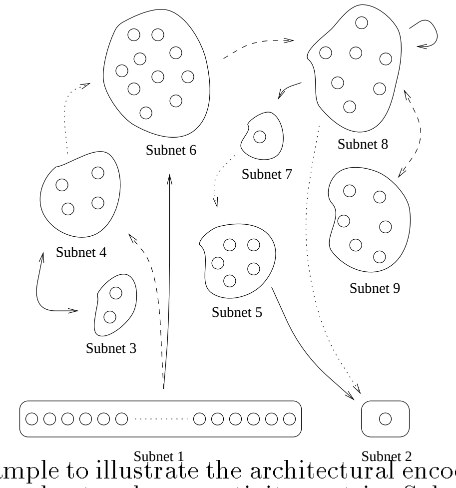
$$\Delta w_{ij} = l(k_0 + k_1 a_i + k_2 a_j + k_3 a_i a_j + k_4 w_{ij} + k_5 a_i w_{ij} + k_6 a_j w_{ij}) \tag{2}$$

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 $-1.0 \leq w_{ij} \leq +1.0$.

3.5 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 the OGI speech tools program MAKEDFT[1] (modified to produce the correct output format) 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

[1] Available from http://www.cse.ogi.edu.



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

Figure 2: Example to illustrate the architectural encoding scheme showing a network of 9 subnetworks and its corresponding subnetwork connectivity matrix. 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 ($C_{55} = 0$).

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 initialised with random weights in the range -1.0 to +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

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

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 mid-point of all N phonemes 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, as shown in Figure 3.

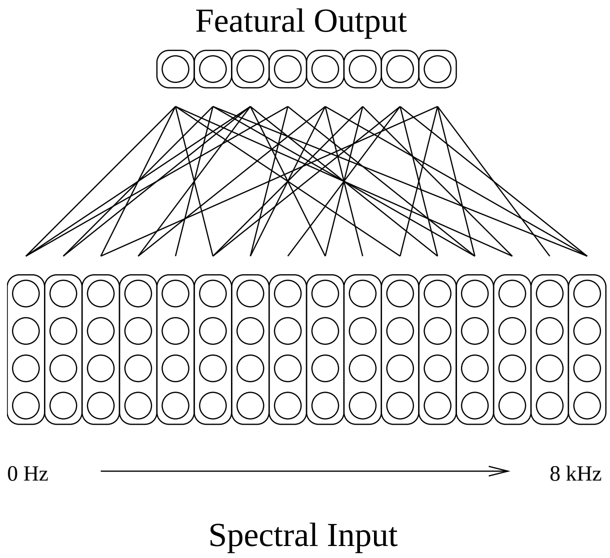


Figure 3: Architectural constraints on the evolutionary process. The networks were all feed-forward, with no "hidden" units and a fixed number of input units (64) and output units (8). Input units were grouped into subnets of 4 units each and each input unit carried information from one of the 64 frequency values in the speech spectra ranging from 0 to 8 kHz.

4 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 4 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 4) 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 5) 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.

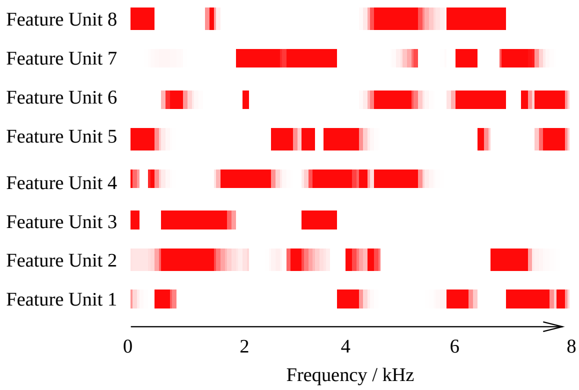


Figure 5: Complex frequency response of all eight feature units to pure tones. Feature units 2 and 3 receive strong excitatory inputs from low frequencies (below 4 kHz) and are therefore activated by voicing.

4.1 Cross-Linguistic Performance

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 languages[2] were used as training files, and the network was tested on 30 American English sentences found in the TIMIT speech files. The time-courses of development for four languages are shown in Figure 6. 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.

4.2 Categorical Perception

Categorical perception of some phonemes is a robust phenomenon observed in both infants and

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

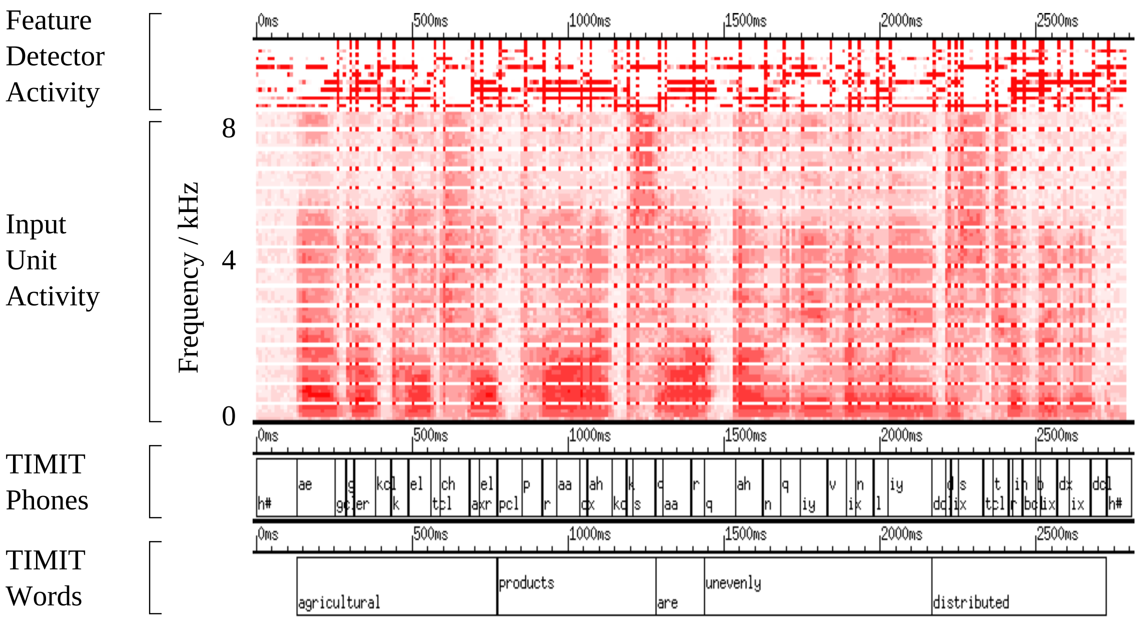


Figure 4: Network response to the sentence "Agricultural products are unevenly distributed" (TIMIT speech file test/dr3/fkms0/sx140). Input units are fed with sound spectra and activate the feature units. Activity is shown as a greyscale (maximum activity is portrayed as black) with time on the horizontal 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 5.

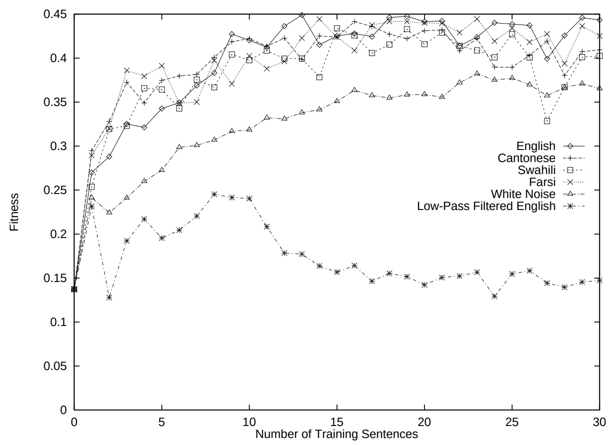


Figure 6: Network performance increases to its final value after presentation of just 20 sentences regardless of the language used to train the network. The six curves show the learning curves for a network tested on 30 sentences of English having been trained on English, Cantonese, Swahili, Farsi, white noise and low-pass filtered English.

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 linear 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 7.

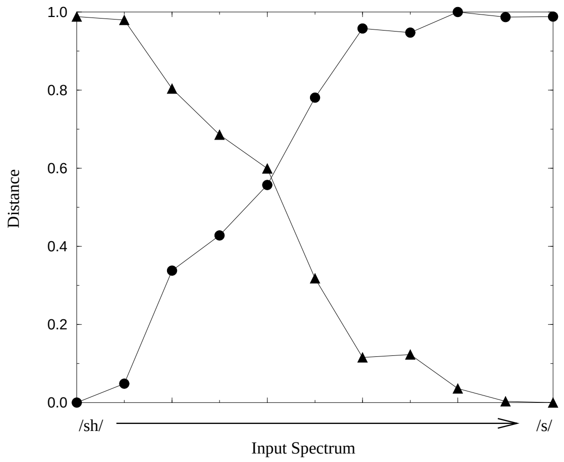


Figure 7: Response of the network to input on a /ʃ/ - /s/ continuum. Circles show the distance from a pure /ʃ/ and triangles show the distance from a pure /s/.

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). Figure 7 shows 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.

4.3 Similarity Structure of the Representation

A consequence of any representation is its effect on similarity judgements. Miller and Nicely (1955) used this fact in an elegant experiment designed to infer the manner in which humans identify sixteen English consonants. They asked subjects to identify CV pairs where the consonant was one of the sixteen being tested and the vowel was /ɑː/, as in *father*. By adding noise to the stimuli at a constant loudness and varying the loudness of the speech they could control the signal to noise ratio of the stimuli and measure the number and type of errors produced. Subjects produced a consistent pattern of errors in which certain pairs of consonants were more confusable than others. For example, the following pairs were highly confusable: m–n, f–θ, v–ð, p–t–k, d–g, s–ʃ, z–ʒ. When clustered according to confusability the consonants formed three groups: voiceless, voiced and nasal consonants. Confusability was greatest within each group and smallest between groups.

Since our model did not classify phonemes it was not possible to create a phoneme confusability matrix using the same method as Miller and Nicely. However, it *was* possible to create a clustering diagram showing the similarity structure of the representations for each phoneme. If given noisy input, phonemes whose representations are closest together in the output space will be more easily confused than phonemes that lie far apart. Since a cluster analysis of many thousands of phoneme tokens would not be clear, a centroid for each phoneme type was used as the input to the cluster analysis. Centroids were calculated by storing the input and output representations of phonemes in 1000 TIMIT sentences. Cluster analyses for the spectral input representation and the featural output representation are shown in Figure 8.[3]

From Figure 8 it is clear that the featural output representation broadly preserves the similarity structure of the spectral input representation despite the eight-fold compression in the number of units. In both the input and output representations the phonemes can be divided into three classes: fricatives/affricates, vowels/semi-vowels, and other consonants. Some phonemes are shifted between these broad categories in the output representation, e.g. t, θ and f are moved into the fricative/affricate category. The reason for this shift is that t occurs with

[3] It should be noted that for stops, TIMIT transcribes closures separately from releases, so /p/ would be transcribed /pcl p/. The results shown here are for the releases, hence their similarity to fricatives and affricates.

a high token frequency, so by pulling it apart from other frequently occurring, spectrally similar consonants, the fitness is increased.

Both spectral and featural representations showed a high confusability for m–n, f–θ, d–g, s–ʃ, as found in the Miller and Nicely experiments. There were discrepancies, however: the stops p–t–k were not particularly similar in either the input or output representations due to an artifact of the representations being snapshots at the mid-points of the stop release. In human categorisation experiments, phonemes are judged on the basis of both the closure and the release, which would greatly increase the similarity of the stops relative to other phonemes. In the input representation, v–ð are fairly close together, but are pulled apart in the output representation. Both these phonemes had low token frequencies, so this difference may not be a result of random variation. In Figure 8 ʒ is not shown because it occurred very infrequently, but the centroids of z–ʒ were very close together, as found by Miller and Nicely.

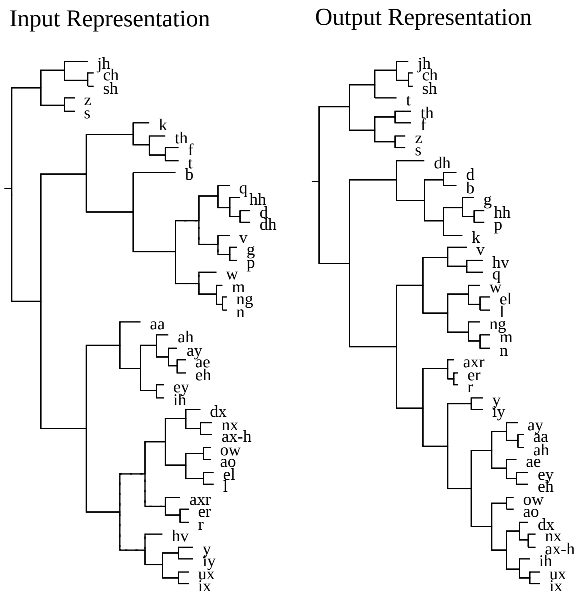


Figure 8: Similarity structure of the spectral and featural representations. Labels are TIMIT ASCII phonemic codes: dx–ɾ, q–ʔ, jh–ʤ, ch–ʧ, zh–ʒ, th–θ, dh–ð, em–m̩, en–n̩, eng–ŋ̍, nx–ɾ̃, hh–h, hv–ɦ, el–l̩, iy–iː, ih–ɪ, eh–ɛ, ey–ej, aa–ɑ, ay–aj, ah–ʌ, ao–ɔ, oy–ɔj, uh–ʊ, uw–uː, ux–u, er–ɹ̩, ax–ə, ix–ɨ, axr–ɚ, ax-h–ə̥.

5 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.

6 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). Examination of perceptual reorganization for nonnative speech contrasts - Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human journal = Perception and Performance*, *14*(3), 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.

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

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.

Grossberg, S. (1978). In L. Leeuwenburg & H. Buffart (Eds.), *Formal theories of visual perception.* New York: Wiley.

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

Jusczyk, P. (1992). In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 17–64). Timonium, Maryland 21094, USA: York Press, Inc.

Jusczyk, P., & Bertoncini, J. (1988). Viewing the development of speech perception as an innately guided learning process. *Language and Speech*, *31*, 217–238.

Kuhl, P. (1993). Developmental speech-perception - implications for models of language impairment. *Annals of the New York Academy of Sciences*, *682*(2), 248–263.

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

Marler, P. (1991). Song-learning behaviour: The interface with neuroethology. *Trends in Neurosciences*, *14*(5), 199–206.

Mayr, E. (1974). Behaviour programs and evolutionary strategies. *American Scientist*, *62*, 650–659.

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

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, *27*, 338–352.

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.

Waddington, C. (1975). *The evolution of an evolutionist.* Ithaca, NY: Cornell University Press.

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

Werker, J., & Tees, R. (1984b). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, *75*(6), 1866–1878.