SRC TR 87-114

Neural Networks for Speech Processing and Recognition

by

S.A. Shamma

Neural Networks for Speech Processing and Recognition

Shihab A. Shamma
Department of Electrical Engineering and Systems Reserch Center
University of Maryland, College Park MD 20742
Mathematical Research Branch, NIH, Bethesda MD 20892

The auditory processing and recognition of speech and other environmental sounds can be conceptually viewed as consisting of three stages typical of all computational neural systems: (1) *The transformation* of the details and objectives of the computational task into a form suitable for processing with a neural network: In speech processing, this would entail the generation of new representations that encode the important signal parameters in a spatially distributed manner. (2) *The extraction of the features* that best and most efficiently describe the nature of the input signal: Non-adaptive neural networks are used to perform standard operations such as formant extraction, binaural cross-correlations, and phonemic segmentation. (3) *The learning and pattern recognition:* Here adaptive neural networks are employed to discover, classify, and store the unique or invariant features of the speech token. The resulting networks can subsequently be used in the recognition of these signals.

In this report, we shall outline a specific approach to the analysis and recognition of speech phonemes based on the fundamental principles of processing in the auditory nervous system. The system consists of particular implementations of the three conceptual stages mentioned above: The cochlear transformations of speech sounds into spatiotemporal patterns (stage 1); the subsequent feature extraction by the central neural networks (stage 2); and the use of various adaptive nets to identify the acoustic features of speech phonemes (stage 3). We shall illustrate the utility of this approach in identifying and organizing the invariant features of nine American English vowels.

I. The cochlear transformation of speech signals

This is the first stage of processing in the auditory nervous system. Its major consequence is the generation of speech signal representations that are naturally suited for the parallel, spatially distributed, structure of neural network processing. Our primary justification for adopting the *cochlear* representation of speech signals over more traditional representations such as spectrograms or LPC parameters is the belief that the apparent ability of human listeners to interpret speech effortlessly and reliably, is due in part to the unique representation of speech on the auditory nerve. From a neural network processing point-of-view, the same justification holds, i.e. that this representation of speech may be particularly suitable for neural network processing since it is precisely the type of signal that the brain sees at its input from the auditory nerve. As we shall briefly discuss below, there are immediate advantages that result from adopting this representation both in noise immunity and in tolerance to large sound-level changes.

A brief outline of the cochlear processing algorithm in one of 128 spatially distributed output channels is shown in Fig.1 (see Appendix A1 for more detail) [1,2]. It involves a complex multi-stage process that is modeled closely upon the mechanics of basilar membrane motion and the biophysics of hair cell transduction[3]. The one input - multiple output, highly overlapping,

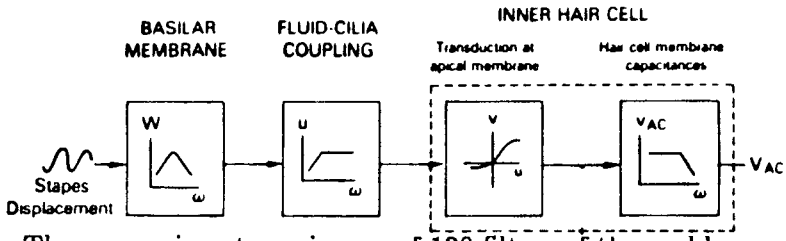

Figure 1: The processing stages in one of 128 filters of the cochlear model

cochlear filters convert the unidimensional signal into intricate spatiotemporal patterns of activity that encode the spectral parameters of the input in the interplay of spatial and temporal cues as described briefly later. There are two important features of cochlear processing that sets it apart from all other commonly used processing algorithms:

1. The cochlear filters have *highly asymmetrical* envelopes which peak and rapidly decay at specific frequencies depending on the spatial location of each filter along the cochlea. Because of the minimum-phase property of these filters, rapid accumulation of phase also occurs beyond the point of resonance.

2. The cochlear filters *preserve the fine temporal structure* of their output waveforms over much of the speech signal spectrum ($\leq$ 4KHz). For higher frequencies, only a measure of the output power is preserved, i.e. as in spectrogram displays.

Fig.2a illustrates the consequences of these two properties in the responses to a simple two-tone stimulus. When plotted as spatiotemporal patterns, the responses due to each tone appear as travelling waves that start at the base of the cochlea (bottom of figure), peak and then decay with rapid phase shift at a specific location along the filter array. These effects combine to produce visual *edges or discontinuities* parallel to the temporal axis. The location of these edges along the *spatial* axis reflects the frequency of the underlying tone since the travelling waves due to different frequencies decay at characteristic locations in an ordered manner resulting in the frequency labeling of the spatial axis. This visual texture of the responses is completely preserved at much higher sound intensities where, because of their limited dynamic range, the cochlear channels become saturated (Fig.2b). Consequently, for any subsequent processing scheme utilizing these spatiotemporal features, e.g the edges (as we discuss below), no AGC mechanisms are needed [1].

For a complex signal like the speech vowel [I], the patterns (Fig.3) are surprisingly similar, i.e. the largest harmonics of the stimulus (those under the two formants) dominate the patterns, each producing a travelling wave which decays at the location characteristic of its frequency. Note also the overall periodic nature of the responses and the beating of the high-CF outputs at the fundamental frequency of the signal - typical features of all voiced stimuli. For the unvoiced fricative [s], the psuedo-random nature of the excitation is evident in the irregular pattern of the travelling waves. Note also the high frequency content and the the sharp spectral edge at $\approx$ 3KHz.

II. Feature extraction

Given the spatiotemporal representations illustrated above, neural networks can now be designed to extract the perceptually significant features of the speech signal. In the case of vowels, the overall shape of the spectrum [5] and the location (frequency) of its formants [6], constitute the primary features of the stimulus. They are expressed in the cochlear responses by

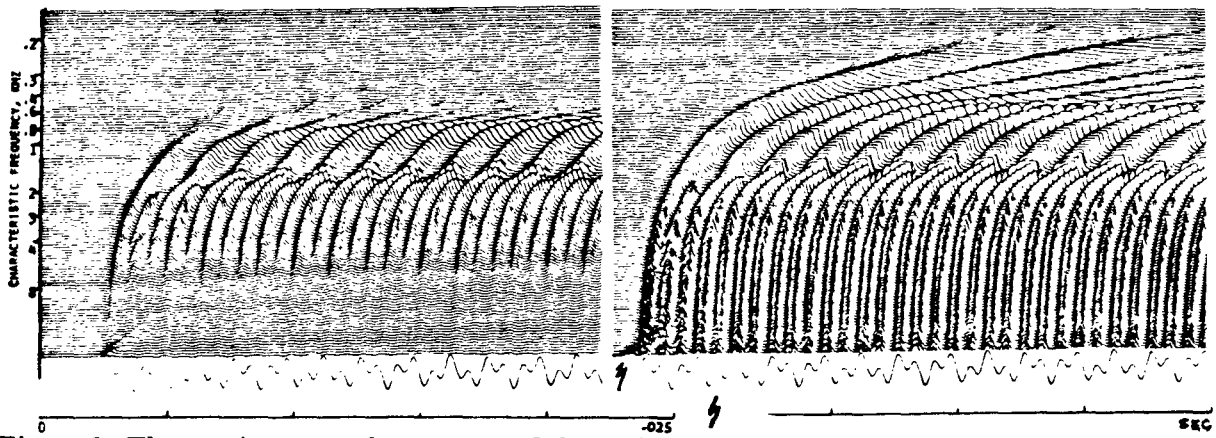

Figure 2: The spatiotemporal responses of the cochlear filters to a two-tone stimulus (600,1400 Hz). The ordinate is the spatial axis of the cochlea. (a) Tones at a moderate intensity. (b) Tones are at +40dB higher intensity than in (a). (Adapted from [4]).

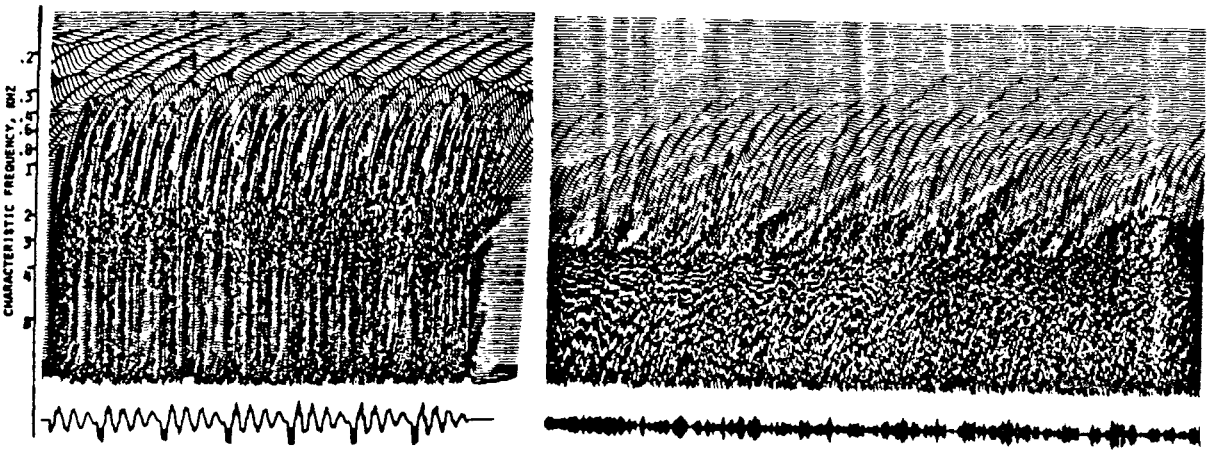

Figure 3: The spatiotemporal responses of the vowel [I] and the fricative [s].

the edges in the spatiotemporal patterns, and can be extracted by networks that perform edge detection along the spatial axis of the patterns. Such a single-layer network is shown in Fig.4. Each element in the network is modeled by a first order nonlinear differential equation of the following form:

$$\frac{du_i(t)}{dt} + u_i(t) = \sum_j m_{ij} x_j(t) + \sum_j w_{ij} y_j(t) \tag{1}$$

where $y_i(t) = g(u_i(t))$ is the i^{th} neuron output, $g(.)$ is a nonlinear threshold function, $x_i(t)$ is the neuron's external input, $u_i(t)$ is the neuron's *membrane* potential, and the time constant is assumed unity. The network has two types of interconnectivities, feedforward m_{ij} and feedback w_{ij}.

By careful tailoring of the connectivities, the network function can be altered to perform a wide range of operations [7–9][1]. One specific and particularly useful form of these matrices is

[1]See Appendix A2 for the values and the rational behind choosing the network connectivities used in the simulations shown in this report.

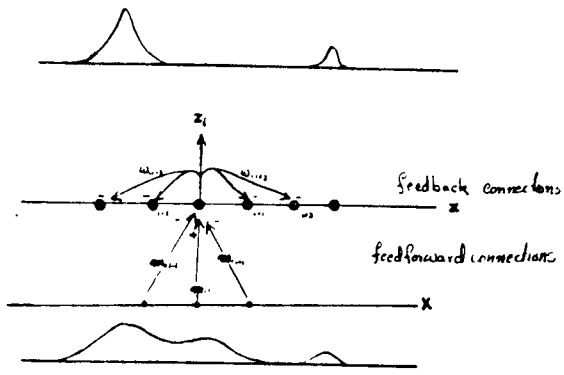

Figure 4: Neural network for spectral feature extraction and phonetic segmentation

obtained if we set all off-diagonal elements to be negative (inhibitory) and the diagonal elements to be zero or positive (self excitatory). This form of connectivities is also known in the biological literature as the *lateral inhibitory network (LIN)*, where it is found in the retina and in the somatosensory systems.

The feedforward connections: Spatial edge detection

Consider first the feedforward operation of this LIN net (setting all $w_{ij} = 0$) The network will combine the elements of its input vector according to the m_{ij} weights, essentially subtracting from each element a weighted sum of its neighbors. Therefore, the network will roughly perform a *spatial differentiation* on its input patterns $\mathbf{x}$, operating as an edge detector. This is precisely the way this net can extract the formants and spectral edges of the cochlear speech patterns. In Figs.2-3 (section I) we saw that the frequencies of tones and vowel formants are marked in the cochlear patterns by the locations of *edges* due to the two unique features of the cochlear filters. Fig.5a illustrates the network feedforward detection of these edges in a series of vowels. The output peaks correspond approximately to the formants of the different vowels. The same operation was performed on various other speech phonemes and words, e.g. unvoiced fricatives [10] and stop consonants [1,2].

The feedback connections: Enhancement and temporal edge detection

The recurrent connectivities add a new range of functional possibilities to the network, because of the rich dynamic behavior that they facilitate in the network equations. Preliminary theoretical investigations have already shown that such networks are capable of simulating such operations as temporary storage and oscillations [7]. In Eq.1, the feedforward outputs ($\sum_j m_{ij} x_j$) can be thought of as the input pattern to a purely recurrent net. Again, if lateral inhibitory connections are used, the resulting net computes a much more enhanced version of its input patterns, essentially selecting and preserving the maximum peak in each *local* region of its input. Fig.5b illustrates the extraction and enhancement of the spectral edges of the series of vowels. The recurrent network can also highlight the spectral changes in time that often occur at the boundaries of different phonemes. This is accomplished by introducing appropriate temporal delays in the path of the *recurrent* inhibition. Thus, when the spectral pattern at the network input abruptly changes *in any frequency band (i.e. location along the net)*, no inhibition is fedback initially, thus allowing for large outputs to mark these instants.

Relationship to articulatory features

An important aspect of the results of Fig.5 is the systematic change along the vowel sequence

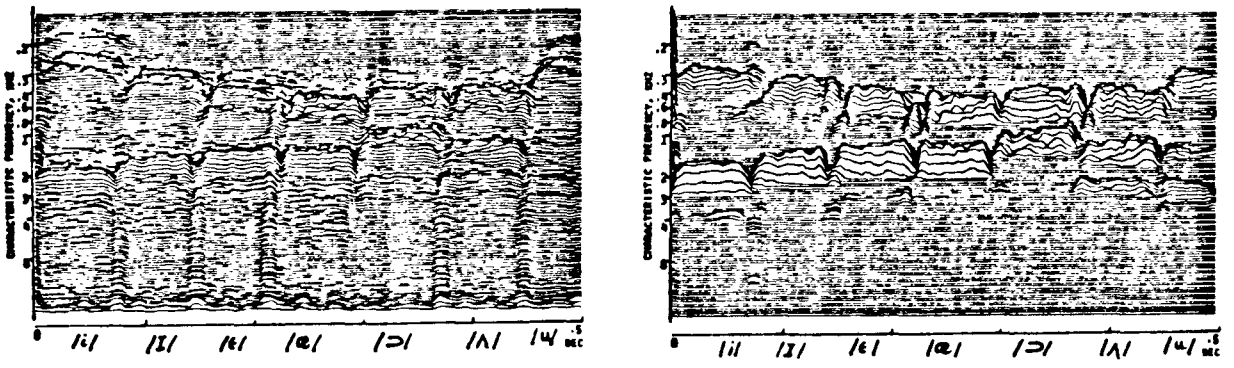

Figure 5: (a) Nonrecurrent processing of a vowel series by a feedforward network; (b) Full processing of the same series

of the relative amplitudes of the extracted peaks [10]. This reflects a systematic change in articulation. Thus for the close vowels at either end of the sequence([i] and [u]), the high-CF peak is relatively large when the tongue constriction is fronted (as in [i]) and vice versa in the back vowel [u]. The open vowels [æ] and [ɔ] occupy an intermediate position in that the two peaks are comparable. This *spectral-shape* change apparently reflects the progressive shift of the so-called *Front Cavity Resonance* [11] which, as we argue below, may be used as a reliable indicator of the vowel identity.

The vowel variability problem

We have also examined the resulting spectra of nine American English vowels from 10 male and female speakers. It has been postulated earlier that the long observed variability in vowel spectra across speakers and sexes may be significantly reduced if cochlear-like processing is employed [5]. In the LIN outputs, however, the variability persists especially across the sexes [10]. Thus, outputs due to the same vowel may exhibit formant shifts and cancellations while still perceived (on playback) as the target vowel. Nevertheless, despite the variability in the patterns of a given vowel, the earlier observation regarding the relative levels of the output peaks largely holds for all outputs. This has important consequences in the recognition of these patterns as we discuss in the following section.

III. Learning and recognition

The reliable classification of the vowel patterns by human listeners suggests the existence of unique or invariant features that characterize each vowel, and that a certain measure of robustness must be associated with these features. In order to isolate these cues, we have employed supervised algorithms to train a variety of neural networks to classify correctly the vowel patterns, and then examined the distributed representation of these patterns in the resulting network connectivities or outputs. For instance, in a preliminary investigation of American English vowels, we trained two neural networks to recognize 9 vowel spectra generated using the processing stages discussed earlier. Each pattern was the average of 10 male and female speakers[2]. In one case, the *backward propagation algorithm* [9] for feedforward networks was

[2]This is a case of supervised learning since the pattern labels are used explicitly to generate the average of

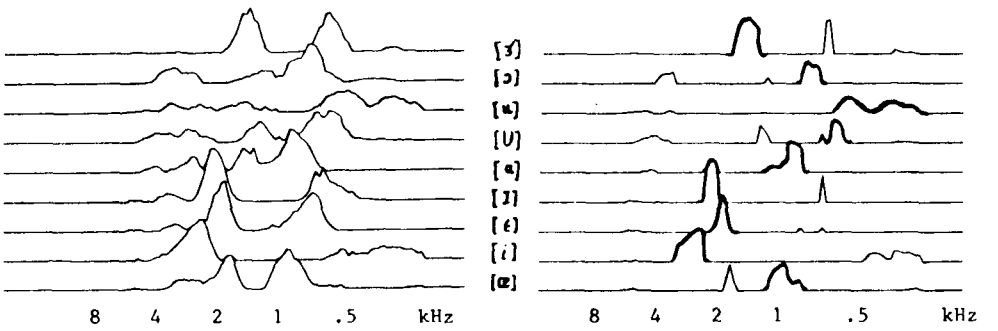

Figure 6: (a) The average patterns of 9 American English vowels. (b) The recurrent network representation of the vowel patterns. Note the reverse frequency representation of the spatial axis

used to associate each vowel with a different unit vector (i.e. a two-layer network - 128 input and 9 output units - with no hidden units [9]). In the other case, two new algorithms were developed to train a single-layer recurrent network to orthogonalize its outputs to the different vowels (a form of competitive learning as described in Appendix A3.). At the end of the learning phase, we compared the vowel patterns with their network representations which, in the case in the recurrent net, are the final orthogonalized outputs, and in the case of the feedforward network, are the vowel patterns as seen by their appropriate output units, i.e. the input patterns weighted by the network connectivities (see Appendix A3). The final representations in both networks were quite similar, showing an emphasis on selected features of the vowel patterns. Specifically, the results indicated an interesting subdivision of the vowel space (Fig.6) - namely that in front vowels (e.g. [i],[I],[ϵ]), the high-CF peak is heavily overrepresented, while in more back vowels (e.g. [u],[U],[a]) the first formant is more weighted [10].

The emphesised formants in the final vowel patterns have an important articulatory correlate in the resonance of the front cavity which appears to determine the energy distribution in the vowel spectra, and hence their overall shape [10,11]. This is already evident in the relative heights of the peaks in the input patterns as mentioned earlier. Thus, by forcing a spatial orthogonalization and categorization, the learning network has accentuated and made apparent the larger peaks of the input patterns (e.g. [i],[I],[a]). Furthermore, it has de-emphasized the peaks that overlap heavily with the peaks of other patterns (e.g. in [_]), thus preserving the unique features of each pattern[3]. We are currently examining the ability of these networks to generalize such a representation to vowels in a variety of contexts.

These results suggest that for the recognition of these vowels, the most reliable feature is the location (or frequency) of the front cavity resonance. This measure combines efficiently both frequency (formant) and spectral shape cues. It can be derived using the above described networks, or simply in hindsight, as approximately the frequency of the high-CF formant of [i],[I],[ϵ],[_] and the low-CF formant of [æ],[a],[$\supset$],[U],[u]. It is interesting to note that recent extensive examination of vowel spectra have shown that these formants exhibit the least variability

each vowel pattern

[3]Note that in /burt/, /bit/, and /book/ the low-CF formant peaks heavily overlap. Since this is the largest peak in the /book/ vowel and since the other two patterns possess unique high-CF formant peaks, the network *assigns* this peak to /book/ and largely de-emphasizes it in the others.

across speakers [12].

In summary, we have illustrated the use of various neural networks in conjunction with a model of cochlear processing. The non-adaptive networks take advantage of the peculiar properties of the cochlear filters to extract robust estimates of the spectral parameters of speech signals. In the case of American English vowels, adaptive networks are used both to organize these features and to recognize them.

Acknowledgment: This work is supported partially through an NSF initiation award (ECE-85-05581), by NSF Grant (CDR-85-00108), and by the Naval Research Laboratory.

Appendix

A1. The cochlear model:

The cochlear spatiotemporal patterns are computed using digital algorithms based on a detailed biophysical model of the cochlea [3,13]. At each of 128 location along the basilar membrane, the transfer function is computed and used in an FFT-based overlap-and-add method to generate the membrane's response to the stimulus. This output is then highpass filtered (y_n=x_n-.8x_{n-1}; modelling both outer ear and fluid-cilia coupling stages) and compressed by a sigmoidal function of the form: $y = M{\cdot}(1/(1+be^{-ax}))$, where a, b, and M are parameters of the nonlinearity, and y, x are the output and input respectively. Finally, a lowpass filter smooths the output (time constant = .1 msec). The parameters of the compressive nonlinearity should be such that approximately 30 dB of linear gain is available between threshold and saturation (defined as .1 - .9 of maximum output level, M) and that the output is saturated at moderate sound levels (approximately 60 dB SPL).

A2. The lateral inhibitory networks:

The cochlear outputs are processed by a single-layer Lateral Inhibitory Network (LIN), which effectively operates as a nonrecurrsive stage followed by a recurrent net.

LIN.I - The first stage (which can be thought of as dendritic) is implemented by a homogeneous inhibitory profile which effectively operates as a highpass, linear phase, FIR filter (symmetric coefficients m_{ij} : .02, .08, -.3, .4, -.3, .08, .02). Each of the 128 resulting outputs is then half-wave rectified, and time-window averaged (window width=10 msec, computed at 4 msec intervals) to generate the final traces shown in Fig.3a.

LIN.II - The second recurrent stage sharpens further the LIN.I outputs. The outputs are computed at the steady state of the network, given by the set of equations:

$y_i = g(v_i - \sum_j w_{ij}\, y_j)$, for i = 0 .. 127,

where g(u) = max(u,0), v_i is the input to the network at the i^{th} location, and y_i is the corresponding i^{th} output which satisfies the mapping above. At each time instant, the input is applied and a set of outputs is computed (the y vector) by iterating the mapping randomly and asynchronously from zero initial conditions. The inhibitory profile of connectivities w_{ij} is chosen such that within any local region (approximately equivalent to the spatial width of the inhibitory profile) only the largest output peak survives (i.e. a local *winner-take-all* strategy). This is an extension of the network analyzed (for global connectivity) by [7]. The symmetric w_{ij} profile of inhibitory connection used is given by the set: 0 (midpoint), .02, .05, .1, .15, .2, .25, .25, .2, .15, .1, .0 and its reflection.

A3. The learning algorithms 4

The results of the vowel classifications discussed in section III were duplicated in three independent networks and learning algorithms. In all, competition emerges as a common theme, primarily due to the desire to separate (categorize) the vowel patterns into nonoverlapping patterns.

1. *Feedforward network and back-propagation algorithm [9]:* A two-layer network with 128 input and 9 output units is trained to associate each of the averaged vowel patterns with a different unit vector (i.e. the ON state of each output unit acts as label for one of the vowels). Hidden layers can also be used although interpreting the results becomes more difficult with the proliferation of network connectivities. At the end of the training phase, each of the averaged vowel patterns is applied, and its representation at the input of its output unit is examined. This representation is weighted by the network connectivities and presumably preserves the features of the input pattern that the output unit utilizes to achieve its ON state (recognition)[4]. The resulting patterns of the nine averaged vowels look quite similar to those shown in Fig.6 (derived from the recurrsive net below). We have also trained the feedforward network on the individual non-averaged patterns, i.e. by associating the multiple realizations of a given vowel from different speakers with one output unit. The average of the weighted patterns again replicate closely the earlier results.

2. *A gradient-descent algorithm for single-layer recurrsive networks:* The averaged vowel patterns are applied to a single-layer lateral inhibitory network which, through a competitive learning algorithm, associates each of the patterns with unique stable state of the network. A gradient-descent algorithm is derived by minimizing (with respect to the connectivities) the following function of the network outputs:

$$E = g(< y^s, y^k >) \cdot (1 - g(< y^s, y^k >)) \tag{2}$$

where y^k represents the steady state output of the network when a pattern x^k is applied. Here the constraint employs the usual similarity measure between the two patterns y^k and y^s. The resulting algorithm forces the network outputs to coalesce into orthogonal classes depending on the parameters (a,b) of the nonlinearity g(x) (a sigmoid of the form $\frac{1}{1+be^{ax}}$). Initially, the network connectivities are set to zero and the outputs are identical to the input patterns. The connectivities are then updated following the application of the entire set of patterns. This is repeated until E(.) reaches a minimum. The final outputs are orthogonalized versions of the input patterns which preserve the unique elements of each vowel. As shown in Fig.6, the surviving peaks of a given pattern are either the largest features of the corresponding input pattern (e.g. as in [i] and [I]), or they overlap least with the peaks of other patterns (as in [_]). The resulting network can also be used for the recognition of an unknown vowel pattern by simply identifying the output towards which the network iterates[5].

---

[4]It should be emphasized that this approach should in general be complemented by a full examination of the weighted representations of the other vowels at the same output unit, especially where exceedingly overlapping patterns are involved.

[5]Unused neurons do not establish connectivities in this algorithm, and therefore need to be suppressed in the recognition phase. This can be effectively accomplished by activating all neurons during the training phase, e.g. through using positively biased versions of the vowel patterns, which result in global inhibitory inputs.

3. *Heuristic algorithm for recurrent nets:* This algorithm is more firmly rooted in physiological models, and is closely related to the *lateral inhibitory* topology described above. The basic idea is to train the decision network as follows: Each time a pattern is presented, the connectivity is modified in the direction of setting up an inhibitory field surrounding the output (in analogy with the inhibitory fields set-up around each neuron in the LIN). Thus, patterns with overlapping fields begin to compete, while closely spaced patterns merge. By varying the range and shape of the interactions, it is possible to control the degree of feature overlaps in the final set of classes. The iterative algorithm is given by:

$$\delta w_{ij} = -\beta w_{ij} + \alpha G(y_i^s - y_j^s) \cdot \overline{y_j^k} \tag{3}$$

where $\overline{y_j^k}$ is the estimate of the expectation of the output at the j^{th} neuron; β is the so called inertial (or forgetting) term, $\beta \ll \alpha$. This algorithm was applied to the averaged vowel patterns (Fig.6) and the final output representations of the vowels were very similar to those of the feedforward and feedback networks discussed above.

References 4

[1] S. A. Shamma, "The auditory processing of speech," *Proceedings of the Symposium on Speech Recognition*, Montreal (1986).

[2] S. A. Shamma, "Encoding the acoustic spectrum in the spatio-temporal responses of the auditory-nerve," in *Auditory Frequency Selectivity*, B. C. J. Moore & R. Patterson, eds., Plenum Press, Cambridge, 1986, 289–298.

[3] S. A. Shamma, R. Chadwick, J. Wilbur, J. Rinzel & K. Moorish, "A biophysical model of cochlear processing: intensity dependence of pure tone responses," *J. Acoust. Soc. Am.* (1986).

[4] S. A. Shamma & K. Morrish, "Synchrony suppression in complex stimulus responses of a biophysical model of the cochlea," *J. Acoust. Soc. Am.* (1987 (in press)).

[5] A. Bladon, "Using auditory models for speaker normalization in speech recognition," *Proceedings of the Symposium on Speech Recognition*, Montreal (1986).

[6] C. G. Fant, "Acoustic description and classification of phonetic units," in *Speech Sounds and Features*, MIT, Cambridge, MA, 1973.

[7] I. Morishita & A. Yajima, "Analysis and simulation of networks of mutually inhibiting neurons," *Kybern.* 11 (1972), 154–165.

[8] D. Tank & J. Hopfield, "Neural computation of decisions in optimization problems," *Biol. Cybern.* 52 (1985), 141–152.

[9] D. E. Rummelhart & J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* #1, Bradford Books, 1986.

[10] S. A. Shamma, "The acoustic features of speech phonemes in a model of auditory processing: Vowels and unvoiced fricatives," *J. Phonetics* (1987 (in press)).

[11] G. M. Kuhn, "On the front cavity resonance and its possible role in speech perception," *J. Acoust. Soc. Am.* 58(2) (1975), 428–433.

[12] Ann K. Syrdal & H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. Acoust. Soc. Am.* 79(4) (Apr. 1986), 1086–1100.

[13] M. H. Holmes & J. D. Cole, "Cochlear mechanics: analysis for a pure tone," *J. Acoust. Soc. Am.* 76 (1984), 767–778.