Teaching a Minimally Structured Back-Propagation Network to Recognise Speech Sounds

T. K. Landauer
C. A. Kamm
S. Singhal
Bell Communications Research, Morristown, N.J. 07960

Abstract: An associative network was trained on a speech recognition task using continuous speech. The input speech was processed to produce a spectral representation incorporating some of the transformations introduced by the peripheral auditory system before the signal reaches the brain. Input nodes to the network represented a 150-millisecond time window through which the transformed speech passed in 2-millisecond steps. Output nodes represented elemental speech sounds (demisyllables) whose target values were specified based on a human listener's ability to identify the sounds in the same input segment. The work reported here focuses on the experience and training conditions needed to produce natural generalisations between training and test utterances.

The primary goal of this work is to explore the use of learning networks as an analytical tool for gaining insight into complex human pattern recognition processes. This goal motivates the choice of input representation, architecture, and training regimen. The strategy is to give a minimally structured network the experience and training it needs to perform a complex pattern recognition task of a kind that humans readily master, and then to use the way in which it learns, generalizes, and fails, and the internal weight organization which it adopts, as a means of studying ways in which the pattern classification in question can be accomplished. To the extent that the features of the input stimuli and the information processing that a successful network uses resemble those used by a human, these results may suggest new hypotheses about how humans accomplish the pattern recognition task.[1]

The input to our network model of speech recognition was processed to mimic several transforms that the peripheral auditory system imposes on the acoustic signal, but that are not specific to recognizing speech. Thus, the information in the input might be considered grossly analogous to information the brain receives from the inner ear. The network was required to learn to extract speech-relevant information from a continuous signal. Error feedback consisted only of information about which speech elements a human could detect in the signal, and no knowledge or theory about the process or mechanism of speech perception as such was embedded in the internal architecture of the learning network or the coding of its output.

The network readily learned to pick a small number of speech sounds, for example, the initial portions of the syllables do, re, mi, fa, so, la, and ti, out of continuous "sentences" consisting of only these syllables, and generalized successfully to other sentences composed of the same elements spoken in different orders. The success of generalization and the naturalness of the errors and partial recognitions evinced by the network appear to depend in interesting ways on the training set to which it was exposed and the discriminations required of its output. We will present both systematic data and impressions gained from experiments with these networks. First, however, we give necessary details on the input transformation, network configuration, learning rules, and training procedures.

1. Input Speech Transformation

The motivation for the signal processing of the input speech was to approximately simulate several of the modifications that the inner ear imposes on an acoustic signal. The processing consisted of five steps. First, the input speech was digitized and low-pass filtered to 5 kHz bandwidth. Spectral estimates were

1. While this is the primary goal, the research can also be viewed as an attempt to apply learning network methods to the problem of automatic speech recognition. In this respect, it asks whether the high-dimensional non-linear representation of such nets and the solution-optimization procedure of back-propagation will produce the same, or perhaps interestingly different, results from those obtained using more traditional pattern-matching or pattern-classification algorithms.

obtained using 128-point FFTs. To mimic the sensitivity of the human listener to rapid changes in high frequency components and to fine frequency distinctions at low frequencies, several FFTs were computed, using temporal windows of 4 to 20 ms, with a 2-ms frame shift. For each 2-ms frame, a composite amplitude spectrum was obtained by extracting low frequency components from the FFT with 20-ms window, the highest components from the FFT with 4-ms window, and intermediate components from FFTs computed with temporal windows between 4 and 20 ms. Second, the frequency scale was transformed to a Bark scale to reflect the frequency spacing along the basilar membrane of the inner ear (Schroeder, Atal & Hall, 1979). Third, the Bark-scaled spectra were convolved with an asymmetric filter simulating the spread of excitation along the basilar membrane (Schroeder & Hall, 1974), which results in a highly smoothed output spectrum. The combination of steps two and three serves to simulate the filtering known to occur in the peripheral auditory system. Fourth, to model the short-term adaptation of the peripheral auditory system, changes in the amplitude of each component were modified by applying a multiplicative function of the difference between successive frames, with exponential decay, producing an enhancement of spectrotemporal "edges". Fifth, the amplitude of each component was scaled relative to the overall minimum amplitude by a power function with exponent 0.6 to simulate the transform from acoustic pressure to relative loudness. Finally, 15 of the 128 transformed amplitudes, spaced at one Bark (approximately one critical band) intervals from 3 to 17 Barks (287 to 4884 Hz), were selected for input to the network.

Clearly this transformation does not perfectly represent the signal sent from ear to brain - for one thing it carries much less information - but it does have several of the important characteristics of that signal. Thus, the information the network learns to extract from this input to recognise speech sounds should have a fair chance of being information the human brain could also extract.

2. Network Architecture and Training Procedure

There were 1,125 input nodes representing the 15 transformed amplitudes for each of 75 successive 2-millisecond frames of a 150-millisecond window of speech. On each training cycle a real value between 0 and 1, proportional to the amplitude of each spectral component, was applied to each input node. Various numbers of hidden nodes, usually 20, were fully interconnected with the input nodes and with a varying number of output nodes depending on the number of speech-sound elements that were to be detected. Training proceeded by stepping the 150-millisecond speech window across an utterance in 2-millisecond steps, at each one calculating the results of forward activation to the output nodes, calculating an error signal, and updating weights by the standard back-propagation procedure (Rumelhart, Hinton and Williams, 1986).

To specify targets, judgments of whether each of the speech sounds to be trained was or was not present in the input window were made by listening to 150-millisecond segments of the signal and by visual inspection of the speech waveform and a speech spectrogram of the signal. Judgments were made on a nine-point confidence scale, from "possibly" to "definitely" present. The error signal at each output node was adjusted in proportion to this confidence value. A separate judgment was made as to whether the system should be trained on a particular sound element for that window, or allowed to produce an output without error being propagated back from the corresponding output unit (this strategy is related to the "don't care" procedure of Jordan, 1986).

3. Results, Observations, Modifications, Comparisons and Lessons.

In early trials with a small number of artificially synthesized speech-like stimuli, it quickly became apparent that the system would learn a training set readily, but did not generalise well to the same nominal "speech" sounds in other contexts, and that the errors it made were not always sensibly related to the confusions between sounds that a human listener would make. Two features of the procedure seemed at fault. First, we thought that the system was not being exposed to enough different speech sounds in enough different contexts to be able to extract the important features and structure of the stimuli. Second, the error calculation procedure was forcing the system to discriminate strongly between sounds that would be perceived as similar by a human, such as "fa" and "va". Therefore, for our main investigation, we used natural speech input with a greater variety of speech sounds in many different contexts, although still a very tiny subset of English. A professional announcer spoke 14-syllable "sentences" composed of two

tokens of each of seven syllables (do, re, mi, fa, so, la, ti) strung together in a Latin square design such that over the sentences studied each syllable followed and preceded every other syllable, including itself, equally often. The speaker did not sing the "sentences", but spoke them with wide variations in intonation, duration and phrasing. (Examples of the recordings will be played.) The total set contained 39 demisyllables (7 initial and 32 final demisyllables) of the roughly 800 to 1,000 demisyllables needed to transcribe all of English speech (Fujimura, Macchi & Lovins, 1977). Two representative sentences are given in Table 1. We typically trained the system to recognize the subset of initial demisyllables.

Table 1. Sample Input "Sentences"

1. do do re so mi re fa la so mi la ti ti fa.
2. mi mi fa ti so fa la do ti so do re re la.

In addition to using this richer training set in the main experiments, we also altered the rule for error calculation to encourage more generalization. In particular, the error for output nodes corresponding to demisyllables not present in the window was multiplied by a constant between 0 and 1, so that a high value on a node corresponding to an absent element was not as strongly corrected as a low value on a node corresponding to a present element. For a given residual criterion training error, this causes the system to move towards a solution in which it adopts high output values for all positive patterns but allows itself moderate output values for other output nodes that tend to be excited by the same inputs.

Results of this training procedure were quite encouraging. The seven plots in Figure 1 show the output values for the seven units corresponding to the initial demisyllables as a function of time for a test sentence after training on just one other sentence. The syllables at the top of the figure, and the lines at the top of each panel indicate the positions in the utterance where each target syllable occurred. The figure shows that, in all instances, the output unit corresponding to the target syllable had the highest output of the seven units. There are several instances where a second output unit also had relatively high activation (e.g., the "so" unit shows activation of about 0.8 when the "do" syllable is in the input window). Such evidence of similarity (or confusability) typically occurred for output units sharing the same vowel sound (e.g., "do" and "so", "fa" and "la").

We are also studying the effect of variation in the training set by testing on one set after having trained on one or more others. Preliminary data suggest that more varied training produces somewhat but not dramatically better generalization. A more detailed characterization of these effects awaits the completion of additional experiments.

Another technique that we are exploring in an attempt to improve the generalization of the network involves training the system to simultaneously recognize the set of speech elements and perform an encoder or auto-association function. In this network, the output layer consists of output nodes of the set of speech elements and 1,125 additional output nodes, fully interconnected to the hidden layer. The error on each of the latter nodes is calculated as the difference between its output and the activation applied to a corresponding input node. The rationale for this combination of directed training and auto-association is that this architecture should supply a much greater degree of constraint on the representations adopted by the hidden nodes, and that these additional constraints may result in a solution that demonstrates better generalization for identifying speech sounds. The system is required to maintain its ability to represent faithfully the raw spectral information at the same time it is learning to recognise a particular speech element, and should thus be less likely to generate an idiosyncratic representation which distorts its overall representation of sound. Preliminary results using this network suggest that the performance and the apparent naturalness of the generalization are improved. Figure 2 shows time functions displaying the activation of each inital demisyllable output node for a training run of this combined direct-training/auto-associative network. Notable in Figure 2 is that demisyllables with common vowel portions appear to be slightly better differentiated than the same demisyllable pairs in Figure 1, and that the silent

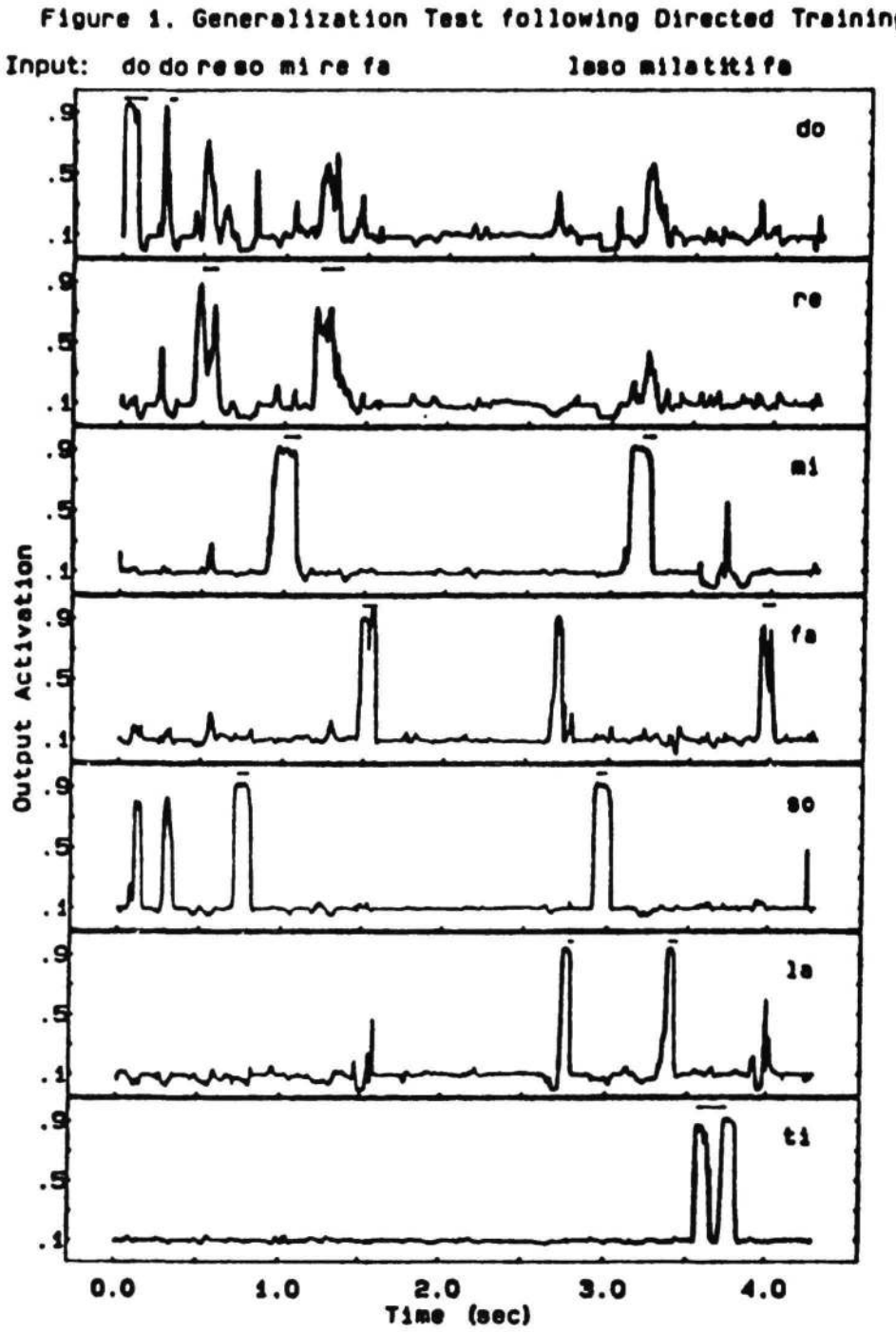

Fig 1. The network whose recognition performance is shown here had been trained on just one other "sentence" consisting of the same syllables in a different order. Shown in each horizontal frame is the activation level of an output node trained to respond to the presence of a particular demisyllable, as a function of time as the continuous test sentence was stepped through the 150 msec input window. Labels across the top of the figure indicate the demisyllables judged present by a human observer; the small horizontal lines indicate intervals in which the highest confidence of presence was assigned.

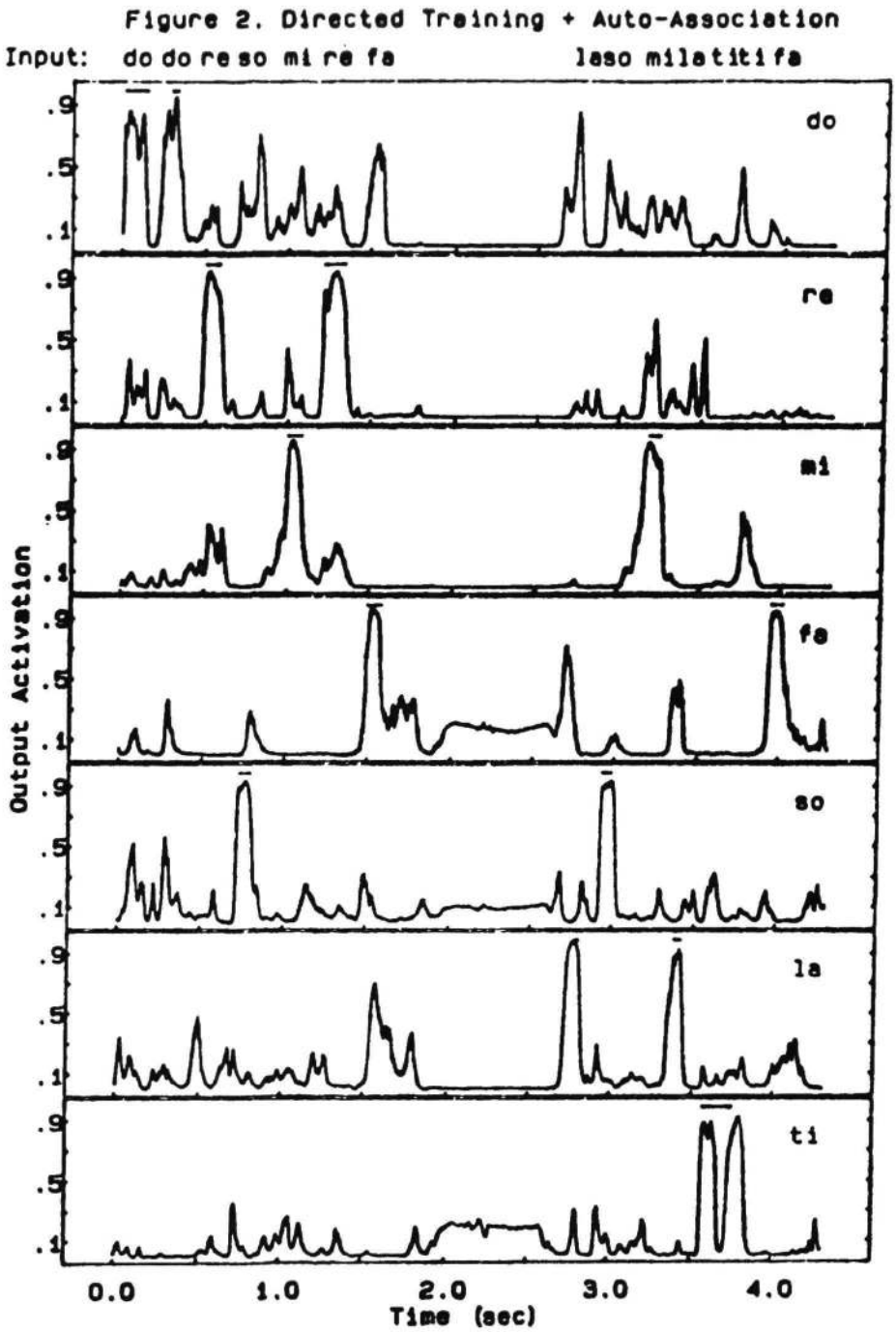

Fig 2. Performance after a moderate degree of training for a network that was simultaneously learning to recognize seven demisyllables and to reconstruct the activation levels of its input nodes. Only the demisyllable recognition node activities are shown. The training (and test) sentence is the same as the test sentence of fig 1.

period (at approximately 1.7 - 2.6 seconds) contains only low levels of activation, the highest of which are from output units corresponding to demisyllables with voiceless initial consonants containing frication noise or noise bursts (/f/, /t/, /s/).

This work clearly represents only the first few steps toward the goal of being able to use the network to analyse the important constituents of speech sounds and their processing. It does appear, however, that, with sufficient experience and proper training techniques, such a complex yet minimally-structured system can learn to use approximately natural input to recognise speech sounds in a way that leads to fairly natural generalization, as evidenced by the recognition performance for non-training tokens of the syllables and the representation of perceptual "similarities" in the output activation patterns. This work also encourages the development of methods for studying what the network is extracting from the transformed speech signal. Experiments along this line may well resemble ones that might be performed using human observers, for example making various systematic modifications of the speech signal and observing the result, or trying various systematically arranged generalization tests. Among the advantages of using networks as opposed to using human listeners would be that such experiments could be run very rapidly and that the internal states of the network, including the continuous activation functions on all of the different output nodes, as well as the auto-association function, are available for analysis.

4. References

Fujimura, O., Macchi, M. J. and Lovins, J. B. Demisyllables and affixes for speech synthesis. Paper presented the 9th International Congress on Acoustics, Madrid, Spain, July 4-9, 1977.

Jordan, M. I. Serial Order: A Parallel Distributed Processing Approach. ICS Report No. 8604, Institute for Cognitive Studies, UCSD, La Jolla, California, 1986.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing, Vol. 1,* 318-362, 1986.

Schroeder, M. R., Atal, B. and Hall, J. L. Optimizing digital speech coders by exploiting masking properties of the human ear. *J. Acoust. Soc. Am., 66,* 1647-1652, 1979.

Schroeder, M. R. and Hall, J. L. Model for mechanical to neural transduction in the auditory receptor. *J. Acoust. Soc. Am., 55,* 1055-1060, 1974.