SPEECH PROCESSING USING ARTIFICIAL NEURAL NETWORKS

Roberto Togneri
Department of Electrical and Electronic Engineering

M.D. Alder
Department of Mathematics

Yianni Attikiouzel
Department of Electrical and Electronic Engineering
The University of Western Australia

ABSTRACT - A three layer perceptron network is used to classify the /i/ sound using isolated words from different speakers. A classification accuracy of 97% has been achieved. A map of phonemes is used to trace trajectories of utterances using the self-organising neural network. A crinkle factor is proposed which allows using the self-organising map to determine the inherent dimensionality of a set of points. By this technique speech data has been shown to possess an inherent dimensionality of at least four. A projection of the map and the speech data shows how the self-organising map fits the speech space.

INTRODUCTION

Neural networks have existed for a long time [1, 2] and have recently enjoyed a resurgence of interest, in particular their application to speech processing. If a set of utterances can be labelled (i.e. each frame is associated with a phoneme) then a supervised neural network like the multi-layer perceptron [3, 4] can be used. Without making any assumption concerning the labelling of the speech an unsupervised neural network, such as the Kohonen self-organising map [5, 6], can be used to classify a speech utterance.

A three layer multi-layer perceptron network is used in this paper to classify one phoneme sound using a training set of isolated words spoken by different speakers. Training is done using the back propagation algorithm [7].

In this paper we also use the self-organising map to segment a set of isolated digits. The dimensionality reducing properties of the Kohonen algorithm [5] are also examined. By measuring the extent that the map is crinkled the inherent dimensionality of the training set can be gauged. This is of interest in validating the beliefs of phoneticians that speech may be described with a small number of parameters. Finally, we attempt a simple projection of the set of speech and weight vectors and by performing limited rotations the way the self-organising map tries to span the speech space can be seen.

NEURAL NETWORK PARAMETERS AND TRAINING SET

The speech data was obtained by sampling at 10 kHz using 12-bit quantisation. Speech frames of 25.6 ms duration and spaced by 10 ms intervals were passed through a 256-point FFT. The dimensionality of the data was reduced by averaging the FFT coefficients into 16 overlapping mel-spaced intervals.

The multi-layer perceptron network had 16 input units corresponding to the 16 overlapping mel-spaced intervals. One output unit was used to classify the input frame presented to the network as the phoneme sound /i/ or not /i/. The number of hidden units were varied from 1 to 9 units. The speech data consisted of single syllable isolated words equally divided between words containing the /i/ phoneme sound and those containing other voiced and unvoiced sounds. These words were spoken by four male speakers producing a training set of around 6000 frames.

The network was trained by cycling through the training data and using the back propagation algorithm to adjust the weights. A record of the correct responses as the network was being trained was kept and used to determine the success of the training.

The self-organising map consists of a single layer of units arranged topologically as a grid. In this paper a 10x10 rectangular grid (i.e. 100 units) was used. Training of this network was done by presenting it with 16 input values from each consecutive speech frame from the utterance. The speech data training set consisted of isolated digits, isolated words and telephone numbers from a single speaker. This training set was different than that used for the multi-layer perceptron. After training the neural units are labelled according to which phonemic sound they respond to by presenting the map with speech frames from stationary phoneme samples. Neural units which consistently achieved a best match with a particular phoneme sample were labelled with that phoneme symbol.

RESULTS FOR THE MULTI-LAYER PERCEPTRON

The multi-layer perceptron was trained for 100 cycles of the training set data for each iteration. A response was correct if the output of the output unit was close to the desired response (0.9 for frames labelled /i/; 0.1 for frames labelled not /i/). As training proceeded the learning of the network stabilised and this gave the final training accuracy (ratio of correct responses / total responses). The results are summarised in Table 1.

Table 1: Classification accuracy of the /i/ phoneme

Number of Hidden Units	Accuracy
1	92%
2	93%
3	96%
4	97%
5	97%
6	97%
7	97%
8	98%
9	98%

From the results the recognition accuracy improves as the number of hidden units increases. Examination of the results showed that all misclassifications were for frames incorrectly classed as /i/, mainly for frames before and after regions classed as /i/. This is not surprising since there was some doubt in the manual segmentation of these frames. There were also a few other regions which were not /i/ but were classed as /i/ by the network. These may be due to voiced phonemes which the network had difficulty separating from /i/.

RESULTS FOR THE SELF-ORGANISING NETWORK

The phonotopic map was presented with speech frames from the digit /three/. This was done for both data from the training set and new data from different speakers. As each frame is presented the winning neuron is highlighted. The winning neurons trace out a trajectory of the utterance (as shown by Figures 1 and 2). The underlying reason to expect a smooth trajectory is based on the fact that the human vocal tract varies slowly with time and this is reflected in slowly varying spectral frames. Since the phonotopic map orders the neural units by grouping similar sounds together it is expected that as frames are presented that adjacent neural units respond creating a smooth trajectory.

From Figure 1 it can be seen that for all speakers the units corresponding to the phonemes /r/ and /I/ are highlighted. The trajectory starts off in a region which corresponds to silence and unvoiced sounds (i.e. the /th/ sound) and ends in a similar region corresponding to the silence at the end of the word. The phonotopic map is unable to resolve such silence and unvoiced information, the effect is random movement of the trajectory which is clearly the case in Figure 1. By deliberately removing the silence the trajectories shown in Figure 2 demonstrate that the map can be used to perform limited segmentation of the speech data even when presented with new data.

DIMENSION OF THE SPEECH SPACE

The Kohonen algorithm [5, 8] adjusts the weights in such a way that the weight vectors try to span the input data space. For a two-dimensional self-organising map input data with an inherent dimensionality less than or greater than two will result in trained weight vectors that describe a highly crinkled or buckled map. We have proposed a measure of this crinkle [9] which is 0 for a flat grid and 1 for a highly buckled grid. By extending the self-organising map to higher dimensions the crinkle measure should be a minimum when the dimension of the neural grid is the same as the dimension of the input data.

This hypothesis was tested by presenting the map with input data from manifolds of known dimension, including planes, surfaces of spheres, and tori of various dimensions embedded in R^n. Results for both a two and three dimensional grid are shown in Table 2.

Table 2: Crinkle factor for known manifolds

Space	Dimension of space.	Dimension of enclosing R^n	32x32 grid	10^3 grid	8^4 grid	8^5 grid
Plane1	2	3	0.004	0.016	-	-
Plane2	2	12	0.004	0.017	-	-
Sphere	2	3	0.014	0.08	-	-
2-Torus	2	4	0.01	0.07	0.18	0.22
3-Torus	3	6	0.12	0.027	0.037	0.097
4-Torus	4	8	0.19	0.10	0.040	0.049

The planes, surfaces of the spheres and the 2-Torus are inherently two-dimensional and the crinkle factor is a minimum for a 32x32 grid. The 3-Torus is a three-dimensional manifold embedded in R^6 and the 10^3 grid exhibits the lowest crinkle factor compared to the 32x32, 8^4 and 8^5 grids. Finally the 4-Torus is a four-dimensional manifold in R^8 and, as expected, the crinkle factor is a minimum for the 8^4 grid. In fact for these test manifolds the crinkle factor calculated from self-organising maps of different dimensions exhibits a minimum at the expected dimension even where the manifold is non-linear. This is an encouraging result for trying to determine the inherent dimensionality of the speech space.

Speech data was tested using both 12 and 16 overlapping mel-spaced intervals as well as 12 and 16 LPC co-efficients. This was done to ensure that the results were not dependent on the size of the initial speech space (12 or 16) nor on the method used to represent the speech signal (FFT or LPC analysis). The speech samples were also chosen to be as general as possible. Connected speech from six male and six female speakers was taken giving 16000 frames. These frames represent the points in the speech space, either in R^{12} or R^{16}. Note that there is no need to restrict the data to single speakers or isolated words as in the previous cases, since no

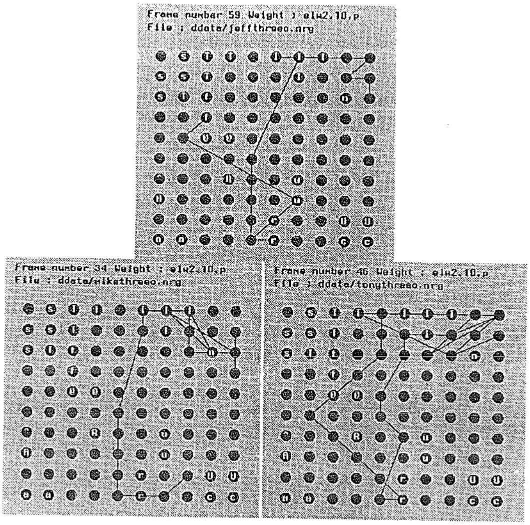
Frame number 59 Weight : elw2.10.p
File : ddata/jeffthreeo.nrg
Frame number 34 Weight : elw2.10.p
File : ddata/mikethreeo.nrg
Frame number 46 Weight : elw2.10.p
File : ddata/tonythreeo.nrg

Figure 1: Trajectory for the utterance /three/ on a 10x10 phonotopic map. Speaker jeff was used for the training set, speakers tony and mike represent new data.

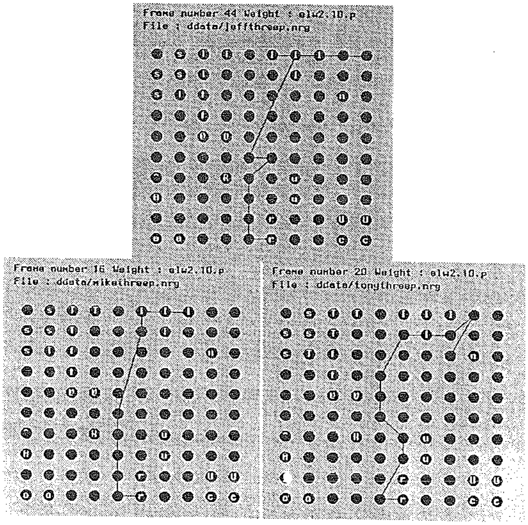
Frame number 44 Weight : elw2.10.p
File : ddata/jeffthreep.nrg
Frame number 16 Weight : elw2.10.p
File : ddata/mikethreep.nrg
Frame number 20 Weight : elw2.10.p
File : ddata/tonythreep.nrg

Figure 2: Trajectory for the utterance /three/ on a 10x10 phonotopic map with silence regions removed.

segmentation of the speech is necessary. The results are shown in Table 3. The results were compared with a set of points in R^{12} chosen at random (noise).

From Table 3 the crinkle factor using points from the speech space is a minimum for the 8^4 grid and is higher for the 8^5 grid. The set of points generated randomly, however, have the lowest crinkle factor for the 8^5 grid and this is expected to continue decreasing for higher dimensioned grids.

Table 3: Crinkle factor for speech data

Space	32x32 grid	10^3 grid	8^4 grid	8^5 grid
R^{12}(noise)	0.27	0.21	0.18	0.17
R^{12}mel speech	0.09	0.05	0.04	0.06
R^{16}mel speech	0.10	0.06	0.05	0.07
R^{12}LPC speech	0.11	0.09	0.06	0.08
R^{16}LPC speech	0.12	0.10	0.08	0.10

The result of using the self-organising map as a means of estimating the inherent dimensionality of a set of points has been proven to behave as expected for manifolds of known dimensionality. The results with speech [10] are very interesting and seem to indicated that the inherent dimensionality of speech is at least four.

PROJECTION OF THE SELF-ORGANISING MAP

With two-dimensional input data it is easy to plot the input vector data set together with the weight vectors from the neural grid. However, of real interest is the speech space which is of dimension 16 as described previously. It would be very interesting to examine a projection of this space onto a two dimensional screen together with the weight vectors from a trained two dimensional grid.

Projection of any higher dimensioned data is simply done by plotting two of the vector components. For the space R^n we can also rotate around n-2 axes (e.g. in 3D we rotate around one axis, in 4D we rotate around two axes, etc.). The number of possible rotations is C_2^n.

The set of speech data and weight vectors in 12 dimensions were taken and the x1 and x6 (i.e. the first and sixth component) co-ordinates were plotted on the screen. By patiently rotating the picture different views of the neural grid structure were obtained. In most cases the shape of the enclosing speech space changed as the picture was rotated. In all these instances the flat nature of the neural grid was evident and, importantly, the grid itself followed the shape of the speech space. This was a very reassuring result and is a very convincing demonstration that the self-organising map is indeed mapping itself to the input training set.

The set of speech data points from an utterance was also examined. Consecutive speech frames were analysed and the resulting 12 dimensional vector was plotted. Successive points were connected to highlight the trajectory of the utterance in the speech space. This was tested for the isolated digit /three/. The speech data points from the initial frames were closer to the origin than those from the last frames. This is expected since the fricative /th/ has a lower energy than the voiced /I/ sound. The density of the points on the trajectory were indicative of sustained sounds. This was especially evident for the /I/ sound at the end of the utterance.

CONCLUSIONS

The multi-layer perceptron provides a way of classifying speech sounds. Separate neural networks can be trained to identify different phonemes. This results in a highly parallel architecture which can be exploited by the correct hardware (i.e. transputers). Further testing of this network is needed, especially using speech data which is not part of the training data, and using more training data. The main bottleneck is expected to be obtaining sufficient labelled speech data for training.

The phonotopic map provides a technique for segmenting speech sounds into their basic units. Our map manages to segment the limited training set successfully. Further testing is needed with more training data and different size/dimension maps.

Close examination of the Kohonen process using higher dimension maps has shown that the crinkle or buckle in the neural grid becomes a minimum when the dimension of the neural grid is the same as the dimension of the input training data. For speech the minimum crinkle factor for a 8^4 grid indicates that the minimum number of features for representing speech is at least four. No indication of what these features are is given at present and their identification is needed.

Plotting the speech data points together with the self-organising map weight vectors has shown that the Kohonen process does indeed try to approximate the speech space. Views of the projection from different angles confirm the flat nature of the grid and its enclosure by the speech data points.

References

[1] Widrow G. and Hoff M.E. (1960) "Adaptive switching circuits", Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4, 96-104.

[2] Nilsson N.J. (1965) "Learning Machines: Foundations of Trainable Pattern-Classifying Systems", McGraw-Hill Systems Science Series.

[3] McCulloch N., Ainsworth W.A. and Linggard R. (1988) "Multi-layer perceptrons applied to speech technology", Br Telecom Technol J, Vol. 6, No. 2, 131-139.

[4] Waibel A., Hanazawa T., Hinton G., Shikano K. and Juang K.J. (1989) "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. Acoustics, Vol. 37, No. 3, 328-339.

[5] Kohonen, T. (1988) "Self-Organization and Associative Memory", 2nd Edition, Springer-Verlag, Series in Information Sciences, Vol. 8, Berlin-Heidelberg-New York.

[6] Kohonen T., Makisara K and Saramaki T. (1988) "Phonotopic maps - Insightful representation of phonological features for speech recognition", ISCAS'88 Tutorial Lectures.

[7] Rumelhart D.E. and McClelland J.L. (1986) "Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations", Cambridge, MA: Bradford Books/MIT Press.

[8] Alder M., Togneri R., Lai E. and Attikiouzel J. (1990) "Kohonen's algorithm for the numerical parametrisation of manifolds", Pattern Recognition Letters 11, 313-319.

[9] Alder M., Togneri R. and Attikiouzel J. (1990) "Dimension of the Speech Space", IEE Comm, Speech and Vision. Accepted for publication.

[10] Togneri R., Alder M. and Attikiouzel J. (1990) "Dimension and Structure of the Speech Space", IEE Comm, Speech and Vision. Submitted for publication.