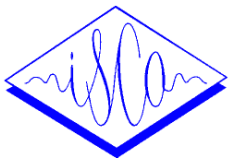

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PHONETIC VISUALIZATION FOR SPEECH TRAINING SYSTEM BY USING NEURAL NETWORK

Itaru Nagayama , Norio Akamatsu and Toshiki Yoshino

Faculty of Engineering, University of Tokushima
Tokushima-shi, Tokushima,770 Japan

Abstract

This paper describes an attempt to develop a new phonetic visualization method of speech training system for the speech impairment. A problem associated with the conventional speech training system is briefly reviewed. Reasons for attempting to develop the new method are given. The proposed method uses neural network approach to visualize phonetic patterns. The method and its display capabilities are described. Also the experimental result of visualizing five Japanese vowels /A/,/I/,/U/,/E/,/O/ by using the proposed method is described.

1. INTRODUCTION

Conventional speech training aids from a hand-held mirror to the electrical, mechanical and computerized systems have been developed[2][3]. These devices have been substantially valuable in teaching speech to profoundly hearing-impaired/deaf people. Especially, computer-aided speech training system with display terminal is widely used all over the world. The most useful application of the system in teaching speech to hearing-impaired/deaf people is to indicate the example of a speech in visual form. Waveforms and spectrum pattern are often used as the representation of sound.

However, the waveforms and spectrum pattern are complicated representation for human being, especially for children to recognize. The famous principle:*"For a speech training aid to be successful, it must be simple."* is essential even today[2]. It is not easy for us to recognize(or to distinguish) the complicated patterns of waveforms and spectrum which are displayed on the speech training system. Therefore, the man-machine interface of speech training system has a green hand.

Generally speaking, it is difficult for us to understand the multidimensional representation of patterns. Thus, it is important to visualize the multidimensional data in simple form. When multidimensional patterns are given, Nonlinear Mapping technique which has developed by Sammon is useful to represent the data in the lower-dimensional space, preserving the relationship among the patterns in the original space[1]. The Sammon mapping is carried out by using gradient descent optimization method to obtain the data distribution in the low dimensional space. But Sammon mapping has a weak point. If a new multidimensional pattern X is given, we must execute the optimization procedure again to know where the X is located in low dimensional space.

In this paper, we apply the artificial neural network for realizing the Sammon mapping. We propose a Neuro-Mapping, to build up a new interface for speech training system. We attempt to develop the prototype of speech training system which can easily visualize 5 Japanese vowels /A/,/I/,/U/,/E/,/O/. By using the artificial neural network, nonlinear mapping from multidimensional space to 2 dimensional(2-D) space can be done. The system displays the multidimensional pattern of vowel onto 2-D space corresponding to the original pattern. User can compare his/her vowel pattern distribution in the 2-D space with standard one during his/her training stage.

2. NONLINEAR MAPPING

2.1 Sammon mapping

In this section, Sammon mapping technique is reviewed according to [1]. Suppose that there are N vectors $X_i;(i=1,\cdots,N)$ in L dimensional space and corresponding to these there are N vectors $Y_i;(i=1,\cdots,N)$ in q(q=2 or 3) dimensional space. Let the distance between the vectors X_i and X_j in L dimensional space be defined by $d^*_{ij}=\mathrm{dist}[X_i,X_j]$ and the distance between the corresponding vectors Y_i and Y_j in q dimensional space be defined by $d_{ij}=\mathrm{dist}[Y_i,Y_j]$, in a sense of Euclidean metric. Next we select the initial comfiguration of q dimensional vectors $Y_i;(i=1,\cdots,N)$ in the space by randomly,

$$\left.\begin{aligned} Y_1&=(y_{11},y_{12},\cdots,y_{1q})\\ Y_2&=(y_{21},y_{22},\cdots,y_{2q})\\ &\cdots\cdots\\ Y_N&=(y_{N1},y_{N2},\cdots,y_{Nq}). \end{aligned}\right\} \quad (1)$$

Then we calculate all interpoint distances d_{ij} in the q dimensional space, which are used to define an error-evaluation function E,

$$E = \sum_{i>j}^{N} \{(d_{ij}-d_{ij}^{*})^2 / (d_{ij}^{*})^2\} \quad (2)$$

which represents how the present configuration of N points in the q dimensional space fits the N points in the L dimensional space. The next step is to adjust the variables y_{uv}, that lead to change the configuration of points in q dimensional space, in order to decrease and search for a minimum value of function E by using optimization procedure as like the steepest descent procedure.

2.2 Neural Network

Artificial neural network is a data-processing system which is suggested by a neuron's operation in the brain [5]. Figure 1 shows a multilayer neural network. The circles correspond to neurons which are variable taking a value ranging from −1 to 1. The data vector is input to an input layer and output from an output layer. Usually, the number of neurons in the input layer is equal to that of the number of elements of the data vector, while that of the output layer is equal to the number of categories. The number of neurons in the hidden layer is properly and carefully decided, because the number has a infuluence to the network operation. The neuron's operation can be expressed by Eqs.(3),

$$\left.\begin{aligned} O_j &= f(Y_j) \\ Y_j &= (\Sigma W_{ij} X_i) - \theta_j \end{aligned}\right\} \quad (3)$$

where X_i is one of the values of a neuron at the first or hidden layer, W_{ij} is an element of the weight matrix, expresses the weight value of interconnection between neurons i and j. θ_j is a threshold value of neuron j. The interconnection weights and thresholds are trained by applying the learning paradigm as like backpropagation algorithm[5][6]. f(X)=tanh(X) is a sigmoidal transfer function which expresses the neuron's operation.

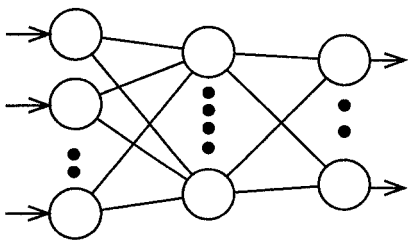

Fig.1 A three-layered neural network.

3. NEURO-MAPPING METHOD

Neuro-Mapping is a method to transform a multidimensional pattern to two or three dimensional pattern by using neural network. The schematic illustration of the Neuro-Mapping is shown in Fig.2. To construct the Neuro-Mapping for speech training system, spectrum patterns of 5 Japanese vowels are used as original feature. The spectrum pattern of a vowel is 16 dimensional(16-D) vector. At first step, 16-D vector patterns are mapped onto 2-D space throughout Sammon's procedure to obtain the location on the 2-D space. Then, the artificial neural network learns the nonlinear relationship between 16-D space patterns and 2-D space ones. The neural network trained using the backpropagation algorithm. After the training stage, the neural network can transform a new (unknown) 16-D pattern which is given to input layer, onto 2-D pattern which is easily compared with the others or the standards.

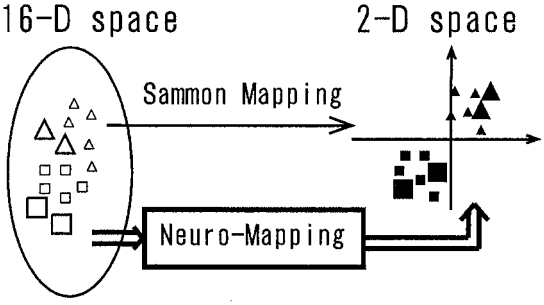

Fig.2 A schematic illustration of the Neuro-Mapping. Small marks are training patterns, large marks are new patterns.

4. EXPERIMENT AND RESULT

4.1 Sammon mapping

Five Japanese vowels, /A/,/I/,/U/,/E/,/O/ are essential for Japanese pronunciation. 20 samples of each vowel (total:100=20samples × 5vowels) are used as training data. For acoustic feature extraction in this experiment, a vowel data is sampled at 32KHz, Hamming windowed and 512-point FFT computed. Melscale coefficients are computed from the power spectrum and are normalized to fall between −1.0 and 1.0. Thus, a vowel is transformed to 16-D pattern. By using the Sammon's procedure, 100 patterns in 16-D space are mapped onto 2-D space. The result is shown in Fig.3. Those 100 vowel patterns are separated each other and 5 territories are made as in the figure. Denote that the 16-D pattern is expressed by Xi;(i=1,⋯,100), and corresponding 2-D pattern is expressed by Yi;(i=1,⋯, 100).

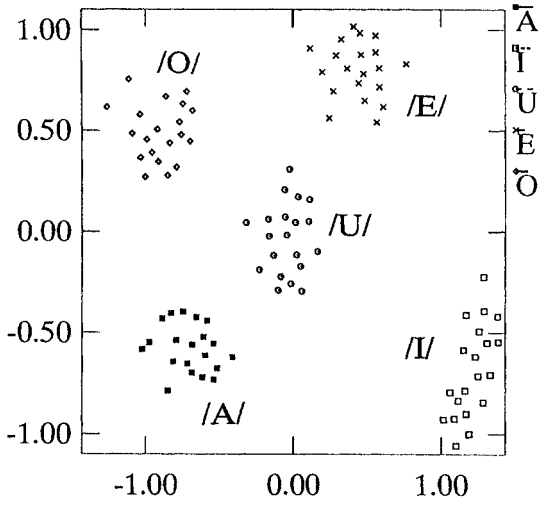

Fig.3 A result of Sammon mapping.

4.2 Neuro−Mapping

Neural network consists of an input−layer(16 neurons), hidden−layer(12 neurons) and output layer(2 neurons) is trained by using backpropagation algorithm to respond Y_i when X_i is given to the input−layer. Training iterations are updated by 20000 epochs. After the training, the neural network is examined by using test pattern. The test patterns, which are 16 dimensional patterns, are newly sampled and formed by above described procedure as same as the training pattern. Figure 4 illustrates the result of Neuro−Mapping for test patterns by the trained neural network. Obviously, test patterns are mapped into their own territories each other. Notice that the result is scaled to same size of Fig.3 by multipling 1.5.

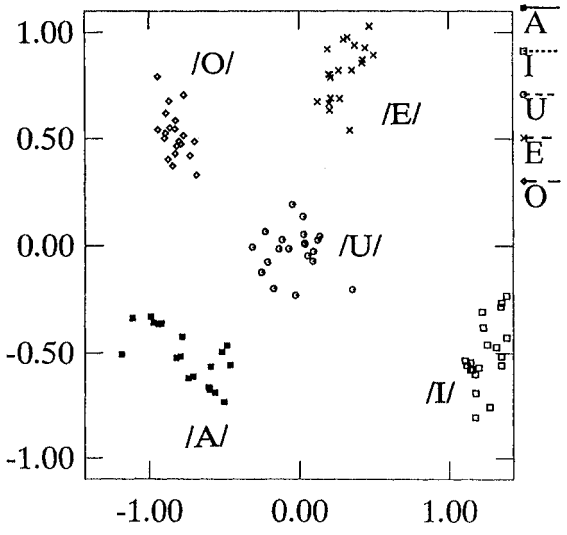

Fig.4 A result of Neuro−Mapping for test patterns.

5. DISCUSSION

5.1 Similarity of patterns

By using the Neuro−Mapping method, we can see and compare the mapped vowel patten with the others and with the standard(which is defined as the center of territory.). However, the similarity of patterns must be preserved even throughout the Neuro−Mapping procedure. That is, the distance from the center to the vowel region in the original 16−D space must be in proportional to the distance from the center to the vowel in the mapped 2−D space. The distance is in a sense of Euclidean. Figure 5 indicates the relation of a radius r of spherical surface in the 16−D space and the one which is mapped into the 2−D space. Mapped examples by the Neuro−Mapping method are shown in Fig.6. Obviously, the distance between the center of each region and their patterns in 16−D space is in proportional to the distance in 2−D space. However, if there is an overlapping area between two categories in the 16−D space, the overlapping area is mapped into the 2−D space as it is.

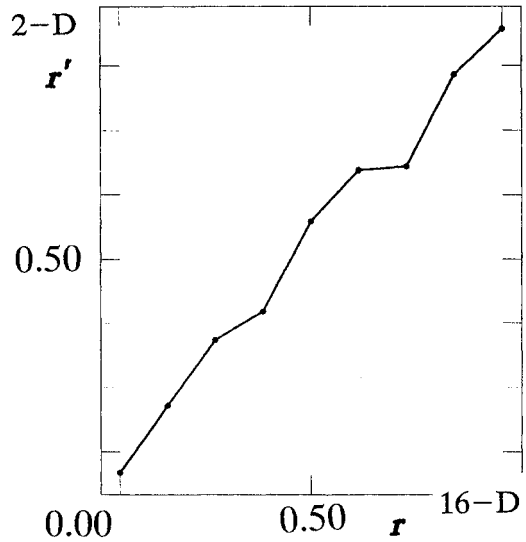

Fig.5 Proportional relation between 16−D and 2 −D space by the Neuro−Mapping.

5.2 Distortion of mapping space

The transformation from 16−D space to 2−D space is a nonlinear mapping. Therefore, a straight line which is mapped into 2−D space from the 16−D space may be distorted. If the distortion of the space is intensitively, the proper comparison between unknown patterns and standard patterns cannot be done. The distortion intensitivity by Neuro−Mapping is investigated. A straight line from a point $\boldsymbol{B}$ to another point $\boldsymbol{C}$ in 16−D space can be expressed by

$$\boldsymbol{l} = (1-t)\boldsymbol{B}+t\boldsymbol{C},\ (0 \leqq t \leqq 1), \qquad (4)$$

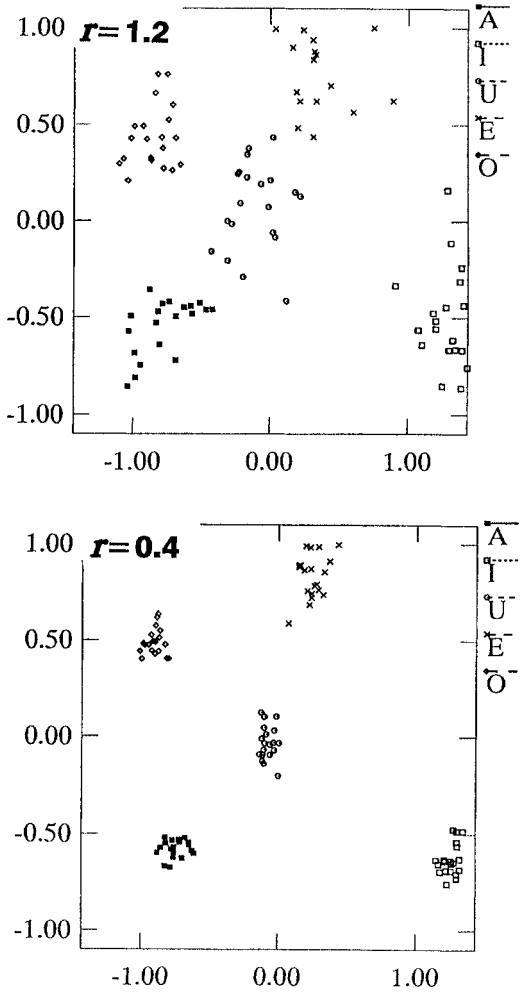

Fig.6 Mapped examples by the Neuro−Mapping.

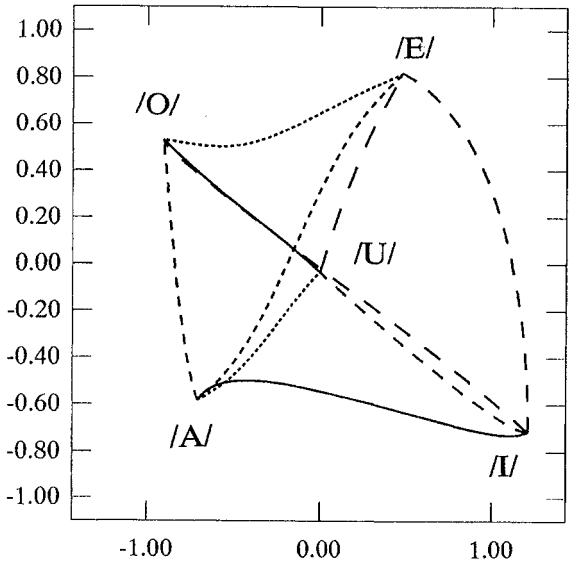

Fig.7 Distortion of space by Neuro−Mapping. The neural network has 4 hidden neurons.

where ***C*** is the center of the each category and ***B*** is the corresponding test pattern. How is the straight line ***l*** mapped onto 2−D space? Straight lines connecting each pair of 5 categories in the 16−D space are mapped onto 2−D space throughout the Neuro−Mapping. As shown in Fig.7, the neural network which has 4 hidden neurons appears faint distortion. According to our repearted experiments, setting the number of hidden neurons as 4,8,12, and 16, the distortion intensitivity is proportional to the size of the network. Thus, the simple structure is relevant to use.

5.3 A new speech training system

A schematic block diagram of a new speech training system which we attempt to fabricate is shown in Fig.8. The system has two main components, spectrum processing unit and Neuro−Mapping unit. In the spectrum processing unit, rejection signal can be sent to VDT. If a very noisy vowel is collected by microphone, the noisy pattern is deleted before sending to Neuro−Mapping unit by looking up the reference table.

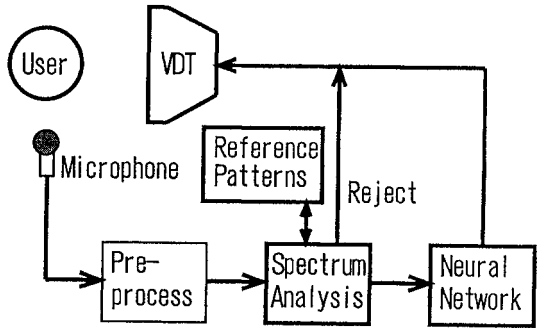

Fig.8 A schematic block diagram of the new speech training system.

REFERENCES

[1] Sammon J.W.:"A Nonlinear Mapping for Data Structure Analysis",IEEE Trans. on Computer,C−18, 5,pp.401−409(1969).

[2] Levitt H.:"Technology and speech training:An affair to remember",The Volta Review,91(5),pp.1−6,1989.

[3] Watanabe A.,Ueda Y.,Shigenaga A.:"Color Display System for Connected Speech to be Used for the Hearing Impaired",IEEE Trans. on ASSP−33,No.1, 164−173,1985.

[4] Raymond S.N. and Kenneth N.S.:"Teaching Speech to the Deaf:Can A Computer Help?",IEEE Trans. on AU−21,No.5,pp445−455,1973.

[5] Rumelhart D.E.,McClelland J.L. and the PDP Research Group:"Parallel Distributed Processing," vol.1,MIT Press,1986.

[6] Lippmann R.P.:"Pattern classification using neural networks",IEEE.comm,pp.47−64,Nov.,1989.