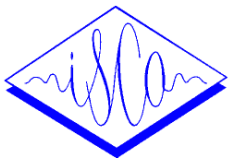


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

LIP-READING OF JAPANESE VOWELS USING NEURAL NETWORKS

Tomio Watanabe and Masaki Kohda

Department of Electrical and Information Engineering,Faculty of
Engineering,Yamagata University,Yonezawa,992 JAPAN

ABSTRACT

This paper proposes mappings of supervisory signals in layered neural networks for lip-reading the five Japanese vowels with the aim of enhancing recognition. The feature parameters of the width P_1 and height P_2 of the lip shape and the distance P_3 between the top of the upper lip and the bottom of the jaw are selected. Mappings from the input vector with the three feature parameters to the desired output vectors with various supervisory signals are discussed on the basis of the similarity between the pentagonal distribution of the $F_1 - F_2$ formant diagram and the $P_1 - P_2(P_1 - P_3)$ diagram. As a result of speaker-dependent lip reading experiments using twenty test sets of five vowels, the recognition rates of the mappings of supervisory signals based on the spatial relationship between vowels are several percent higher than the rates of the mappings disregarding the relationship. Finally, a mapping for generating the desired relationship between vowels in the hidden layer is proposed, and the effectiveness of the mapping is demonstrated.

I. INTRODUCTION

Lip-reading plays an important role in speech recognition, in particular when the acoustic signal is degraded[1]. Lip-reading is based on the recognition of vowels. The author has already proposed a method for the recognition of the five Japanese vowels (/a/,/e/,/i/,/o/,/u/) by machine lip-reading[2]. Compared to the feature parameters (height and width) of lip shape alone, adding the parameter of the distance between the upper lip and the bottom of the jaw revealed a significant difference in discrimination.

Neural networks automatically develop an internal structure that is appropriate for a task domain through a learning process. These are particularly interesting for cognitive tasks[3]. It has been reported that differences in input parameters to a neural network led to different performances in phoneme recognition[4]. It is suggested that recognition performance may depend on the mapping from the input vector to the supervisory signal vector in neural networks.

This paper focuses on supervisory signals in layered neural networks for lip-reading Japanese vowels, with the aim of enhancing recognition. Mappings from the input vector using three feature parameters (described in II) to the desired output vectors with various supervisory signals, are mainly discussed on the basis of the distribution of the feature parameters for the five vowels, taking note of the similarity between the acoustic and articulatory specification of the vowels in two-dimensional diagrams (the $F_1 - F_2$ formant diagram and the lip shape's height-width diagram).

II. LIP-READING MATERIALS

Video recordings were made, under good lighting conditions, of the faces of two male subjects (X and Y). Each subject uttered two hundred isolated vowels, forty sets of five vowels (repeating in the order /a/,/i/,/u/,/e/,/o/). The first twenty sets were considered the training sets, and the second were the testing sets. Each utterance was begun in a closed-mouth position to assist in detection. For measuring the vowels, three feature parameters of the width P_1 and height P_2 of the lip shape and the distance P_3 between the top of the upper lip and the bottom of the jaw were automatically extracted by an image processing system, in order to perform grayscale thresholding and contour coding as shown in Fig.1 [2].

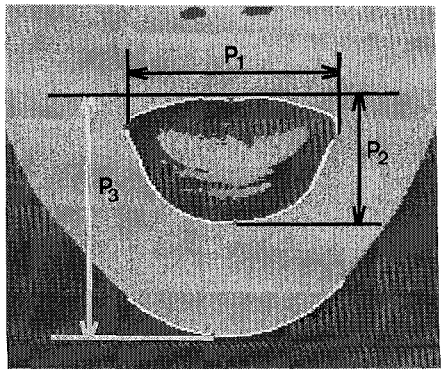


Fig.1 Feature parameters.

III. NEURAL NETWORK MODEL

We used a three layered neural network which has an input layer comprising three units corresponding to the feature parameters, a hidden (intermediate) layer of seven units, and an output layer of two, three, or five units, as shown in Fig.2. The particular number of units, seven, in the hidden layer was set on the basis of preliminary research aimed at an optimal result. Every unit has an output f_i which is a sigmoid function of its total input x_i with a bias θ_i.

$$f_i = \frac{1}{1 + exp\{-(x_i + \theta_i)\}} \qquad (1)$$

$$x_i = \sum_j w_{ij} \cdot f_j \qquad (2)$$

Learning of weight w_{ij} was performed by 5,000 sweeps using back-propagation[3].

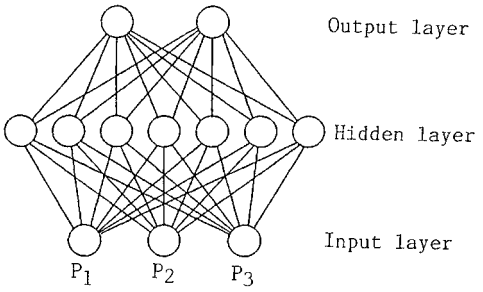


Fig.2 Neural network.

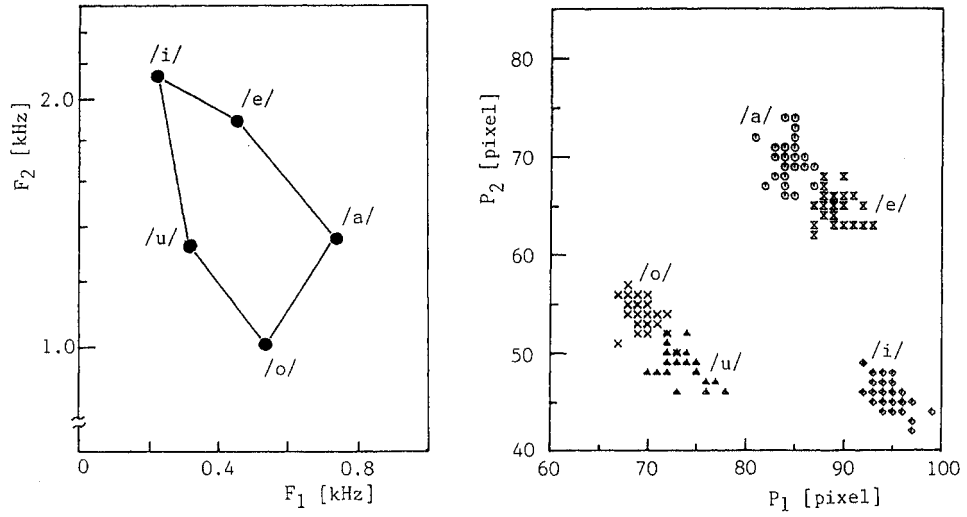
IV. RECOGNITION RESULTS AND DISCUSSION

4.1 *Effectiveness of Mappings Based on the Spatial Relationship Between the Position of Vowels on Two-dimensional Diagrams*

We will discuss the effectiveness of mappings based on the spatial relationship of the feature parameters of the five vowels, comparing two output units with five output units.

A typical acoustic specification of the vowels in a two-dimensional $F_1 - F_2$ formant diagram is shown in Fig.3(a) while an articulatory specification of the lip shape's width-height parameter $P_1 - P_2$ diagram is shown in Fig.3(b). These figures show the similarity between the pentagonal distribution of the five vowels. Thus, we examined the following three mappings of two output units based on the spatial relationship between the position of vowels: 1) $F_1 - F_2$ formant mapping <2-A> whose supervisory signals (F_1, F_2) are /a/=(0.90,0.55), /i/=(0.25,0.90), /u/=(0.35,0.55), /e/=(0.55,0.70), and /o/=(0.65,0.40) where F_1 and F_2 are normalized between 0 and 1 as shown in Table 1(a); 2) $P_1 - P_2$ mapping <2-B> with a normalized average vector $(\overline{P_1}, \overline{P_2})$ as shown in Table 1(b); 3) $P_1 - P_3$ mapping <2-C> with vector $(\overline{P_1}, \overline{P_3})$ as in Table 1(c) because the distribution of the $P_1 - P_3$ diagram also resembled that of the $F_1 - F_2$ formant diagram, in particular where discrimination between /o/ and /u/ increased [2].

Mapping <5-A> with five output units simply mapped correct one of the five vowels, that is, the supervisory signals were set to 0.85 for the desired output and 0.15 for the other four outputs. These particular values were selected instead of 1 and 0 so as to obtain good recognition with reference to a previous report



(a) F_1-F_2 formant diagram (b) P_1-P_2 parameter diagram

Fig.3 Similarity between the pentagonal distribution of the F_1-F_2 formant diagram and the P_1-P_2 diagram.

[5]. This mapping is irrespective of the spatial relation between vowels.

Table 1 Mapping of two output units based on the spatial relationship between the position of vowels on two-dimensional diagrams.

(a) Mapping <2-A>

	F_1	F_2
a	0.90	0.55
i	0.25	0.90
u	0.35	0.55
e	0.55	0.70
o	0.65	0.40

(b) Mapping <2-B>

	Subject X		Subject Y	
	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_1$	$\overline{P}_2$
a	0.57	0.85	0.67	0.85
i	0.85	0.15	0.85	0.15
u	0.21	0.15	0.29	0.21
e	0.64	0.51	0.75	0.50
o	0.15	0.48	0.15	0.24

(c) Mapping <2-C>

	Subject X		Subject Y	
	$\overline{P}_1$	$\overline{P}_3$	$\overline{P}_1$	$\overline{P}_3$
a	0.57	0.70	0.67	0.85
i	0.85	0.21	0.85	0.17
u	0.21	0.15	0.29	0.15
e	0.64	0.46	0.75	0.56
o	0.15	0.85	0.15	0.46

Table 2 Recognition results.

Number of output units	Mapping	Recognition rate (%)			
		Subject X		Subject Y	
		Training	Testing	Training	Testing
2	<2-A>	93	90	95	77
	<2-B>	93	91	95	76
	<2-C>	97	87	96	76
5	<5-A>	96	82	97	73

For discrimination of the output vector with two output units, the output vector was identified by taking the vowel corresponding to the nearest supervisory signal vector to the output vector in Euclidean distance. Table 2 shows the recognition results. In the case of both subjects(X and Y), the testing recognition rates of mappings <2-A>, <2-B> and <2-C> are several percent higher than that of mapping <5-A>. There are no differences in recognition rates between mappings <2-A>, <2-B> and <2-C> compared with the rate of mapping <5-A>. The results demonstrate the effectiveness of mappings based on the spatial relationship between vowels in achieving a high degree of recognition.

4.2 *Appropriate Dimension of the Supervisory Vector*

Here we discuss whether the appropriate dimension of the supervisory vector is two or three, in terms of the combination of the feature parameters based on the spatial relationship.

Mapping <3> with three output units is a $P_1 - P_2 - P_3$ mapping with a normalized average vector $(\overline{P}_1, \overline{P}_2, \overline{P}_3)$ based on Tables 1(b) and (c).

Table 3 shows the results when comparing mappings <2-B> and <2-C> with mapping <3>. In both subjects, the testing recognition rate of mapping <3> is a little lower than that of mappings <2-B> and <2-C>, though it is higher than the rate of mapping <5-A> in Table 2. This indicates that appropriate dimension of the supervisory vector is two.

Table 3 Recognition results.

Number of output units	Mapping	Recognition rate (%)			
		Subject X		Subject Y	
		Training	Testing	Training	Testing
2	<2-B>	93	91	95	76
	<2-C>	97	87	96	76
3	< 3 >	94	85	97	74

4.3 *Generating the Desired Relationship in the Hidden Layer*

Now we know that a mapping of a two-dimensional supervisory vector based on the spatial relationship between vowels is preferred. With the aim of realizing mapping <2-A> as an example of a two-dimensional supervisory vector in the second hidden layer, we propose the following mapping using combined neural networks where the first network with two output units is combined with a second network with five output units, as shown in Fig.4. We trained the first network using mapping <2-A>, and

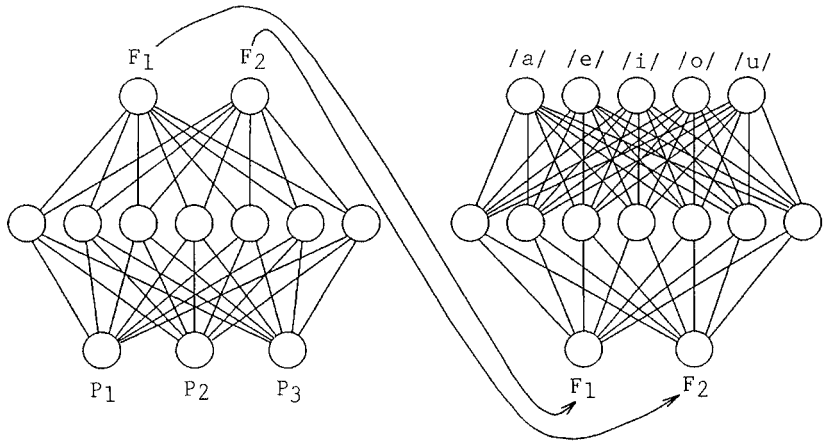


Fig.4 Combined neural networks.

trained the second network using the mapping of (F_1, F_2) signals as input to the supervisory signals of <5-A>.

To compare this combined mapping, we examined mapping <5-B> where the network has two hidden layers: the first layer of seven units and the second layer of two units, and an output layer of five units using mapping <5-A>.

Table 4 shows the results. For both subjects, the testing recognition rate of the proposed combined mapping is enhanced recognition, and the rate of mapping <5-B> is as low as that of mapping <5-A>. This demonstrates the effectiveness of generating the desired relationship between vowels in the hidden layer.

Table 4 Recognition results.

Number of output units	Mapping	Recognition rate (%)			
		Subject X		Subject Y	
		Training	Testing	Training	Testing
2	<2-A>	93	90	95	77
5	<5-A>	96	82	97	73
	<5-B>	96	83	95	73
	<2-A> +<5-A>	86	94	95	80

V. CONCLUSION

We discussed supervisory signals in layered neural networks for lip-reading Japanese vowels, taking note of the similarity between the pentagonal distribution in the $F_1 \doteq F_2$ formant diagram and the lip shape's height-width diagram. It was concluded that the mapping of supervisory signals based on the spatial relationship between vowels enhances recognition. Finally, we proposed a mapping for generating the desired relationship between vowels in the hidden layer, and demonstrated the effectiveness of this mapping.

ACKNOWLEDGMENT

The authors are grateful to the Telecommunications Advancement Foundation for its financial support.

REFERENCES

[1] E.D.Petajan, "Automatic Lipreading to Enhance Speech Recognition," IEEE Commun. Technol. GLOBECON, pp.265-272,1984.

[2] T.Watanabe, "Vowels Recognition by Machine lip Reading," Trans.Jpn.Soc.Mech.Eng., Vol.53, pp.2613-2616, Dec. 1987.

[3] D.E.Rumelhart et al., "Learning Representations by Back-Propagating Errors," Nature, Vol.323, pp533-536, Oct. 1986.

[4] S.Nakamura and K.Shikano, "Comparison of Input Parameters for Time Delay Neural Network(TDNN) and Implementation of VQ-based Speaker Adaptation," IEICE Technical Report Vol.89,pp.43-50,June 1989.

[5] M.Kohda,"A Study on Supervisory Signals in Vowel Recognition Using Neural Networks," Proc. of the Acoust.Soc.Japan Conf.'89, pp.27-28, Oct. 1989.