ESTIMATION OF VOCAL TRACT SHAPE USING NEURAL NETWORKS

K.Shirai, T.Kobayashi and M.Yagyu*

Waseda University, Tokyo, JAPAN
* Hosei University, Tokyo, JAPAN

ABSTRACT

This paper discusses an application of neural networks (NN) to the problem of estimating a vocal tract shape from speech waves.

The experimental results show that the difference in estimated articulatory parameter values between the conventional method (MM) and NN is only 3 % of value range on average. For a few data, big differences arise between MM and NN, but this is due to mis-estimation in MM rather than NN. The percentage of mis-estimates in NN is less than 70 % of MM. By introducing recurrent nodes, the value is reduced to be 50 %. In this case, the spectral error is improved by 5 % .

1. INTRODUCTION

Coarticulation compensation and speaker adaptation, which are two major difficulties in speech recognition, might be considered most clearly and fundamentally from the view point of the speech production mechanism. On the basis of this idea, the model matching (MM) method was proposed to extract articulatory parameters, which represent a vocal tract shape, from speech waves [1] and the estimated parameters were used for speech recognition [2,3,4]. These parameters are effective for the above two problems but the conventional estimation method has some problems: one is calculation cost and the other is instability of estimation. Since this method is constructed on the basis of hill climbing methods, it requires many iterations to converge and sometimes finds only local minimum. For a speech recognition system, a faster and more stable estimation method is desired. To prove these problems, we are trying to apply neural networks. In the previous papers, we tried to utilize a simple four-layer feed-forward neural network [6,7]. In this paper, we apply a recurrent neural network to the problem.

2. ARTICULATORY MODEL

The total configuration of the mod' and the characteristics of the articulator parameters are shown in Fig.1 and Tab 1.

The first five parameters (X_{T1}, X_{T} X_J, X_L, X_G) determine the shape of th oral and pharyngeal cavity, and the nasa ization parameter X_N describes the cros sectional area of the velopharyngeal part

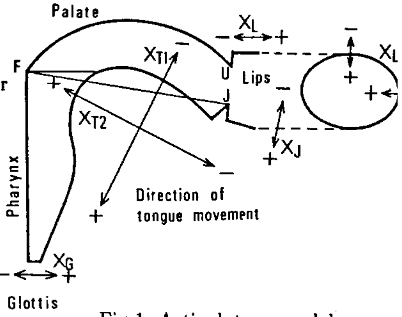

Fig.1 Articulatory model.

Table 1 Qualitative characteristics of the articulatory parameters.

parameter	organ	+	−
X_{T1}	tongue	back	front
X_{T2}	tongue	high	low
X_J	jaw	open	close
X_L	lip	rounded	unrounded
X_G	glottis	open	closed
X_N	velum	open	closed

3. MODEL MATCHING METHOD

The acoustic feature of speech waves (y) which are generated by the model (x) can be expressed by a nonlinear function of the articulatory parameters ($y = h(x)$). Therefore, the estimation problem is to solve a nonlinear function. The conventional technique for this problem is a nonlinear optimization method called the model matching method(MM). In this method, model parameters are iteratively changed to optimize a certain criterion function.

Let y_s be acoustic parameters measured from the speech wave after glottal and radiation characteristics are removed. Then, the estimate x of the articulatory parameters is obtained so as to minimize the following cost function.

$$J(x) = (y_s - h(x))^t P (y_s - h(x)) + x^t Q x + (x - x_0)^t R (x - x_0)$$

where P, Q, and R are the weight matrix, and x_0 is the estimate at the previous frame. This problem is solved by the following iterative form

$$x^{i+1} = x^i + \lambda \Big(\frac{\partial h(x^i)}{\partial x^i}^t P \frac{\partial h(x^i)}{\partial x^i} + Q + R \Big)^{-1} \Big(\frac{\partial h(x^i)}{\partial x^i}^t P(y_s - h(x^i)) - Qx^i - R(x^i - x_0) \Big)$$

4. ESTIMATION OF ARTICULATORY PARAMETERS USING NEURAL NETWORK

In our experiment, a four-layer feed-forward network and a four-layer recurrent network are adopted. Figure 2 shows the architecture of the recurrent network.

Weight coefficients are determined as follows: Firstly, vowels are selected from training dataset. Then, using MM, articulatory parameters are estimated for all frames in these data (including glides). 12th order LPC cepstral coefficients are also calculated for the same data and cepstrum - articulatory parameter pairs are prepared. Finally, applying backpropagation to these data pairs, the weights of the network are determined.

All vowel frames of phone balanced 216 tokens in the ATR word database(speaker ID: MAU) except data whose articulatory parameters are different from average estimate of corresponding vowel by more than 20 in mahalanobis distance measure are used for the training.

To estimate articulatory parameters from speech waves, cepstral coefficients are calculated by LPC analysis, and then, these data are input to the neural network.

5. EXPERIMENT

The evaluation test is performed using the vowel data in 5200 tokens in the ATR word database.

Table 2 shows the difference in estimated articulatory parameter values between MM and NN/RNN.

The difference is only 0.1 on average (3 % of value range). It can be seen that the neural networks work well to estimate articulatory parameters.

Figure 3 shows the estimated articulatory parameters and spectra. The solid line denotes the articulatory parameters obtained by MM and the dashed line denotes that by NN. Data is /niou/.

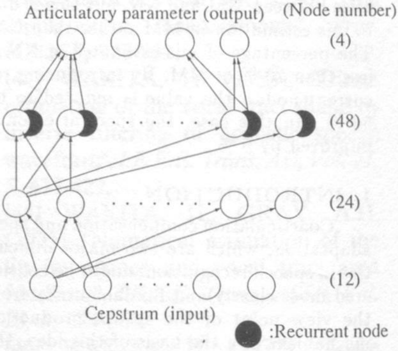

Fig.2 Structure of the neural network for the articulatory parameter estimation.

Table 2 Average and standard deviation of estimation error.

	Traning data		Test data	
	average	std.dev.	average	std.dev
NN	0.08	0.12	0.11	0.13
RNN	0.08	0.12	0.13	0.14

(value range : [-1.5,1.5])

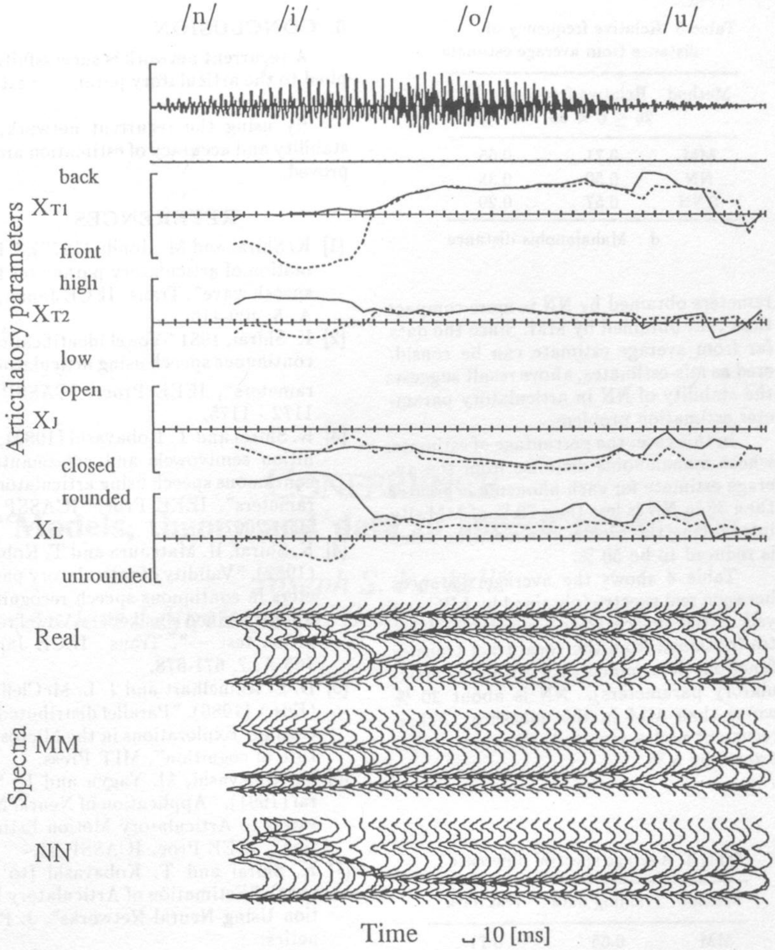

Fig.3 Movements of articulatory parameters obtained by the model matching method ( ) and the neural network ( ).

In the part of /ou/, articulatory parameters estimated with NN is almost equal to that estimated with MM. As for /i/, the big difference can be seen between MM and NN. The contour of the spectra (formant structure etc.) obtained by NN is more similar to the real one than that by MM. Parameters $(X_{T1}, X_{T2}, X_J, X_L)$ should be (front, high, close, unrounded) in /i/ sound. The estimated articulatory parameters using NN satisfy this term but those by MM do not. Estimated articulatory parameters using NN can be regarded as more appropriate for /i/.

As we can see in this example, big differences arise between MM and NN for a few data. However, this result does not mean the mis-estimation in NN.

Table 3 shows the distribution of the estimated articulatory parameters. The values in the table denote the relative frequencies of the distances from the average estimate of each vowel.

The distribution of the articulatory pa-

Table 3 Relative frequency of distance from average estimate.

Method	Relative frequency [%] $25 \leq d < 36$	 $36 \leq d$
MM	0.71	0.55
NN	0.50	0.38
RNN	0.57	0.29

d : Mahalanobis distance

rameters obtained by NN is more compact than that obtained by MM. Since the data far from average estimate can be considered as mis-estimates, above result suggests the stability of NN in articulatory parameter estimation problem.

In this case, the percentage of estimates whose mahalanobis distance from the average estimate for each phoneme is greater than 36 in NN is less than 70 % of MM. By introducing the recurrent nodes, the value is reduced to be 50 %.

Table 4 shows the average distances between real spectra (obtained by LPC analysis of the speech wave) and model spectra (obtained from the vocal tract transfer functions determined with estimated articulatory parameters). NN is about 30 % worse than MM in this measure. When recurrent nodes are used, the value is improved by 5 %.

Table 4 Average spectral difference.

Method	Training data	Test data
MM	0.65	0.71
NN	0.87	0.94
RNN	0.79	0.89

Thus, it is proved that the NN method is not suitable for the strict articulatory parameter estimation but has little risk of making serious errors. Moreover, this accuracy and stability are improved when recurrent nodes are introduced. This little risk is due to the characteristics of neural networks as associative memories. The strong constraints among articulatory parameters are embedded in the network structure on the training process. So the unnatural combination of parameters can be automatically excluded.

6. CONCLUSION

A recurrent network is successfully applied to the articulatory parameter estimation problem.

By using the recurrent network, the stability and accuracy of estimation are improved.

REFERENCES

[1] K. Shirai and M. Honda (1978), " Estimation of articulatory parameter from speech wave", Trans. IECE Japan, J61-A, 5, 409-416.

[2] K. Shirai, 1981 "Vowel identification in continuous speech using articulatory parameters", IEEE Proc. ICASSP 81, 1172 - 1175.

[3] K. Shirai and T. Kobayashi (1982), "Recognition semivowels and consonants in continuous speech using articulatory parameters", IEEE Proc. ICASSP 82, 2004-2007.

[4] K. Shirai, H. Matsuura and T. Kobayashi (1982), "Validity of articulatory parameters in continuous speech recognition for unspecified speakers — Vowel recognition test —", Trans. IECE Japan, J65-A, 7, 671-678.

[5] D. E. Rumelhart and J. L. McClelland (Eds.) (1986), "Parallel distributed processing: Explorations in the Microstructure of cognition", MIT Press.

[6] T. Kobayashi, M. Yagyu and K. Shirai (1991), "Application of Neural Networks to Articulatory Motion Estimation", IEEE Proc. ICASSP 91.

[7] K. Shirai and T. Kobayashi (to appear), "Estimation of Articulatory Mortion Using Neural Networks", J. Phonetics.