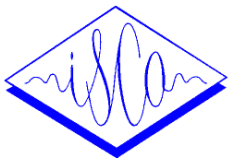

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Phoneme Recognition by Pairwise Discriminant TDNNs

Jun-ichi TAKAMI and Shigeki SAGAYAMA

ATR Interpreting Telephony Research Laboratories,
Sanpeidani, Inuidani, Seika-cho, Souraku-gun, Kyoto 619-02, Japan

Abstract

In this paper, a phoneme recognition method using Pairwise Discriminant Time-Delay Neural Networks (PD-TDNNs) is proposed. In conventional approaches to phoneme recognition based on neural networks, it was found that the difference between training data and testing data degrades recognition performance. To overcome this problem, we developed a phoneme recognition method using PD-TDNNs. Each PD-TDNN has the ability to discriminate between two phoneme categories. In this method, phoneme candidates are selected by judging multiple pair discrimination scores, each of which is obtained from the PD-TDNN. We tested this method on a phoneme recognition task for /b,d,g,m,n,N/. Testing on continuous speech using the PD-TDNNs which were trained with the phoneme data in isolated word utterances, we obtained a first candidate recognition rate of 81.6%, and 96.7% for the cumulative recognition rate up to third candidates.

1. Introduction

In recent years the investigation of phoneme recognition methods using Neural Networks[1] has received considerable attention. It has been shown that Time-Delay Neural Networks (TDNNs)[2][3] have high recognition performance for the data in isolated word utterances and the ability to tolerate the time lag caused by variations in the phoneme extraction positions. However, in conventional phoneme recognition based on TDNNs, it was found that the difference between training data and testing data degrades recognition performance. This is most likely because the networks are overly tuned with training data and very sharp discrimination boundaries are formed between phoneme categories. Under these conditions, output values calculated from such networks tend to concentrate near 1 or 0, thus they do not represent the likelihood of the phoneme categories. Consequently, the discrimination errors caused by such networks often result not only in the degradation of recognition accuracy for first candidates but also in the degradation of cumulative recognition accuracy for multiple candidates. This matter is well known as the so called over-learning problem.

To overcome this problem, we introduced a pair of discrimination techniques[4] into the process of phoneme recognition. The technique is to solve a complex recognition problem having multiple categories by judging multiple pair discriminant problems, each of which has only two categories. We then developed a Pairwise Discriminant Time-Delay Neural Network (PD-TDNN) to solve such pair discriminant problems with high accuracy and robustness. The method has the following two significant features:

- Discrimination between two phoneme categories is done by a PD-TDNN, which has the ability to do so with high accuracy.
- Phoneme candidates are selected by judging multiple discrimination scores which are calculated from the output values of the PD-TDNNs.

In this paper, we describe the phoneme recognition method using the PD-TDNNs, show the experiment results for 6 phonemes /b,d,g,m,n,N/ and discuss the significant features of this method.

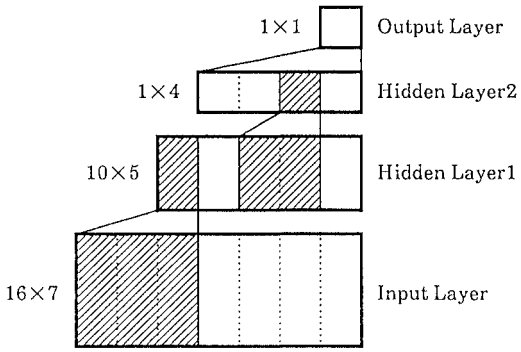

Fig.1. The architecture of a PD-TDNN.

2. Phoneme Recognition By PD-TDNNs

2-1. Architecture of a PD-TDNN

A PD-TDNN is a kind of TDNN consisting of four layers. The input layer consists of 112 units corresponding to input data (7 frames of 16 mel-scaled spectrum) and the output layer consists of only 1 unit that outputs a discrimination score lying between 1.0 and 0.0. Fig.1 shows the architecture of the PD-TDNN.

Although a conventional TDNN has 15 frames on its input layer, the PD-TDNN has 7 input frames because it was found by our preliminary observations that a network with 7 input frames achieves higher recognition performance for the data extracted from the center position of phonemes, than one having 15 input frames.

Now we designate a phoneme pair consisting of two different phonemes p_i and p_j as $(p_i:p_j)$, and a PD-TDNN which discriminates between the two phoneme categories p_i and p_j as Net(p_i/p_j). Net(p_i/p_j) is trained as follows: if input data belongs to category p_i, category p_j, or neither category, 1.0, 0.0, or 0.5 are given as the desired outputs, respectively. In this case, an output value of Net(p_i/p_j) represents a pair discrimination score for the category p_i, and the value calculated by subtracting the output value from 1.0 represents the pair discrimination score for the category p_j.

Note that, when a common sigmoid function is used in the output unit, it is difficult to decrease the averaged error between outputs from a network and the desired outputs. This is because the sigmoid function has a large differential coefficient at the point corresponding to the output value of 0.5, thus the error for the data desired to output the value of 0.5 does not decrease easily. Hence, we introduced a new non-linear function into the output unit to overcome this problem. It is as follows:

$$f(x)=\begin{cases}\dfrac{g(x+\alpha)}{2g(\alpha)} & (x<0)\,,\\[2ex] 1-\dfrac{g(-x+\alpha)}{2g(\alpha)} & (x\geqq 0)\,,\end{cases} \qquad (1)$$

where g is: $g(x)=\dfrac{1}{1+e^{-x}}$

Fig.2 shows the form of this non-linear function.

Where α is a value to control the differential coefficient at the point corresponding to the output value of 0.5. When α is 0.0, this function coincides with the sigmoid function. The value of 3.0 was used as α during our observations.

2-2. Phoneme Recognition Method

Since ${}_NC_2$ phoneme pairs are existent in N phoneme categories, ${}_NC_2$ PD-TDNNs are formed by training. Although a PD-TDNN for discrimination between p_i and p_j will have two variations, Net(p_i/p_j) and Net(p_j/p_i), according to which category corresponds to the desired output of 1.0, it is sufficient if only one of these networks is formed because of their symmetrical relationship.

Respective Pair discrimination scores of p_i and p_j are calculated as follows:

$$S(p_i|p_i:p_j) = \mathrm{Out}(p_i/p_j) \ , \qquad S(p_j|p_i:p_j) = 1\text{-}\mathrm{Out}(p_i/p_j) \ . \tag{2}$$

Where $S(p_i|p_i:p_j)$ represents the pair discrimination score for category p_i under the opposing viewpoint between p_i and p_j, $\mathrm{Out}(p_i/p_j)$ represents an output value of $\mathrm{Net}(p_i/p_j)$.

All pair discrimination scores obtained from ${}_NC_2$ PD-TDNNs are arranged in the table shown in Fig.3.

The values on horizontal lines of this table represent the pair discrimination scores, which are obtained under various opposing viewpoints, for an individual phoneme category. A phoneme recognition score for category p_i, $T(p_i)$, is obtained by summing up these scores as follows:

$$T(p_i) = \sum_{j \neq i} S(p_i|p_i:p_j) \ . \tag{3}$$

Multiple phoneme candidates are selected in descending order using the $T(p_i)$.

3. Recognition Experiments

Phoneme Recognition experiments were performed using 6 phonemes, /b,d,g,m,n,N/. For discrimination between these 6 phonemes, 15 (${}_6C_2$) PD-TDNNs are necessary.

In addition, experiments using a single TDNN were also performed to compare the performance. Fig.4 shows the architecture of the single TDNN. The single TDNN has 7 frames on its input layer, as same as the PD-TDNN, to compare these performance under the same condition.

3-1. Utterances and Input Data

For testing, we used the phoneme data in the Japanese Large Vocabulary Database uttered by a native male speaker[5]. All utterances in this database are digitized at 12kHz. Input data for the Neural Network are calculated by the following process: a speech utterance is Hamming windowed and a 256-point FFT is computed every 5ms, 16 mel-scaled spectrum coefficients are computed from the power spectrum, adjacent coefficients in time are averaged resulting in an overall 10ms frame rate and the input data coefficients (7 frames of the mel-scaled spectrum coefficients) are normalized to lie between -1.0 and +1.0 with the average at 0.0.

3-2. Network Training

Each PD-TDNN was trained with 500 training data per phoneme. Each training data was extracted from the center position of the phoneme in even-numbered isolated words of the 5,240 common Japanese words. During the training, the fast back-propagation learning method[6] is used. The training was stopped when the averaged error between outputs from the network and the desired outputs became 0.01. At this moment, the first candidate recognition rate was 98.75% for the training data.

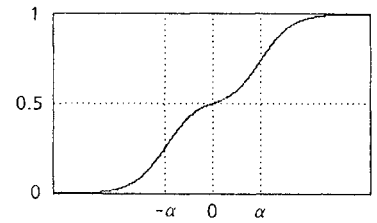

Fig.2. Form of a non-linear function in output unit.

	p_1	·····	p_j	···	p_N	*Total*
p_1	---------	·····	$S(p_1\|p_1:p_j)$	···	$S(p_1\|p_1:p_N)$	$T(p_1)$
p_i	$S(p_i\|p_1:p_i)$	·····	$S(p_i\|p_i:p_j)$	···	$S(p_i\|p_i:p_N)$	$T(p_i)$
p_N	$S(p_N\|p_1:p_N)$	·····	$S(p_N\|p_j:p_N)$	···	---------	$T(p_N)$

Fig.3. Table for arrangement of all pair discrimination sores.

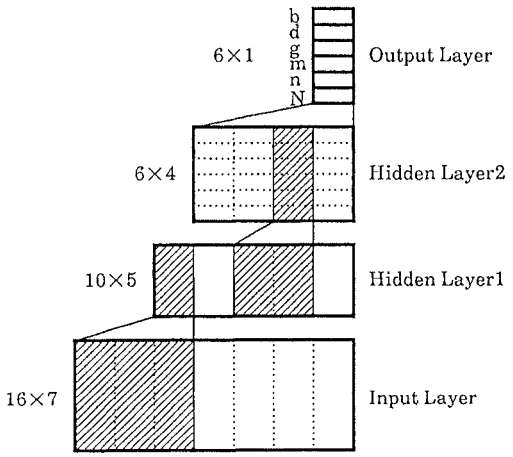

Fig.4. The architecture of a single TDNN.

The training of the Single TDNN was also done with the same training data using the following conventional training method: the desired output of 1.0 was given for the output unit corresponding to the correct phoneme category and that of 0.0 was given for the other output units. The training was stopped when the averaged error became 0.005 because the first candidate recognition rate was almost the same (98.85%) as that obtained using the PD-TDNNs for training data at that moment.

3-3. Testing Data

We used testing data extracted from the following utterances[5]:

- Odd-numbered isolated words of the 5,240 common Japanese words. ("word" : 5.68 mora/sec)
- Conversational sentences from a task called "The International Conferences Secretarial Service" uttered in each compound phrase. ("phrase" : 7.72 mora/sec)
- Conversational sentences from a task called "The International Conferences Secretarial Service" uttered in each short phrase. ("short phrase" : 7.14 mora/sec)
- Conversational sentences from a task called "The International Conferences Secretarial Service" uttered continuously. ("continuous" : 9.56 mora/sec)

Table1 Recognition rates for various utterances. (unit : %)

testing data	1st	2nd	3rd
"word"	97.25 95.19	99.21 99.15	99.79 99.79
"phrase"	88.81 84.84	96.85 92.68	99.19 96.54
"short phrase"	86.24 84.51	95.41 91.23	98.06 95.31
"continuous"	81.64 77.79	93.31 87.12	96.65 92.19

upper : PD-TDNNs lower : single TDNN

Table2 Recognition rates for shift tokens. (unit : %)

time lag	1st	2nd	3rd
-20ms	72.16 62.37	90.86 77.64	95.98 85.68
-10ms	94.77 87.42	98.31 95.40	99.42 98.36
0ms	97.25 95.19	99.21 99.15	99.79 99.79
10ms	94.19 88.42	98.68 96.35	99.58 98.36
20ms	70.03 60.68	83.51 76.53	95.14 93.60

upper : PD-TDNNs lower : single TDNN

4. Results

4-1. Recognition Accuracy for Various Utterances

Table1 shows the experiment results. From these, it was found that the method using PD-TDNNs has high accuracy and robustness for various speaking manners compared with one using the Single TDNN.

4-2. Confirmation of Shift Tolerance

It has been found that a conventional TDNN has high shift tolerance under 15 input frames[7]. However 7 input frames were used in the architecture of the PD-TDNN. We then tested the shift tolerance under 7 input frames using both the PD-TDNNs and the Single TDNN. The recognition rates were calculated using the testing data extracted, every 10ms from -20ms to +20ms, from isolated word utterances (odd-numbered isolated word utterances).

Table2 shows the experiment results. From these, it was found that the method using PD-TDNNs has enough shift tolerance for a time lag between ±10ms. Moreover, it also has higher performance for a time lag exceeding ±10ms.

5. Discussion

5-1. Control of Discrimination Boundaries

The significant feature of the PD-TDNN is that it is trained not only with values of 1.0 and 0.0 but also with the value of 0.5 as the desired outputs. In conventional TDNN training, the network is trained not with the value of 0.5 but only with the values of 1.0 and 0.0. Hence, very sharp discrimination boundaries are formed in the network to classify each category into several separated spaces, which correspond to the desired outputs of 1.0 or 0.0. Therefore, the decrease of the averaged error tends to be accomplished by greater use of the saturated areas on the sigmoid function in each output unit. Consequently, such network outputs a values near 1.0 or 0.0 for almost all input data.

On the other hand, we will discuss training with the desired output of 0.5 for training data distributed over broad areas, as in this method. In this case, if the network has very sharp discrimination boundaries, it is probably difficult to concentrate output values near 0.5, for input data which have a desired output of 0.5. Hence, the effect of suppression against the formation of sharp discrimination boundaries is raised by training with many data each of which has the desired output of 0.5.

In order to prove such an effect, we observed which areas on the non-linear function standing in each output unit are used to represent the output values for the data /b/ and /d/, using both the output unit Net(b/d) (one of 15 PD-TDNNs) and the output unit corresponding to /b/ of the Single TDNN. Both of these output units are similar with regard to the following points: if the input data belongs to category /b/, or category /d/, 1.0, or 0.0 are the respective desired outputs. The difference between these units is which value, 0.5 or 0.0, is given as the desired output for the input data that belong to neither category.

Figs.5(1) and 5(2) show the experiment results. (a) shows a histogram calculated every 0.1 steps for the training data, and (b) shows that for the "continuous" data. Here, the upward and downward histograms correspond to /b/ and /d/, respectively.

As shown in this figure, when PD-TDNN was used, the histogram of the training data was concentrated in narrow areas. Thus it was found that the saturated areas are not used very much. Moreover, a similar tendency was also shown for the "continuous" data, so that the fatal errors did not occurred very often.

On the other hand, when the Single TDNN was used, the histogram of the training data tended to expand over broad areas. Thus, it was found that the outputs are represented by considerable use of the saturated areas. In addition, the discrimination errors occurred significantly more often than when

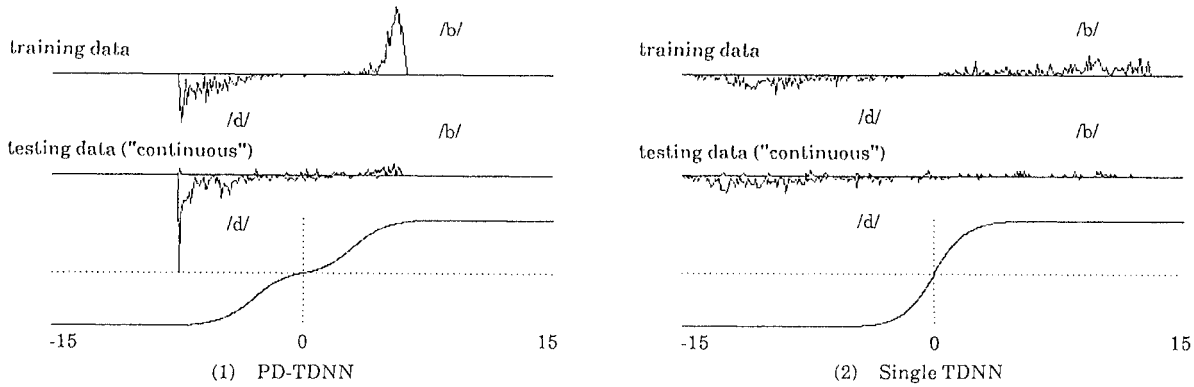

Fig.5. Representation for both /b/ and /d/ on the respective output units.

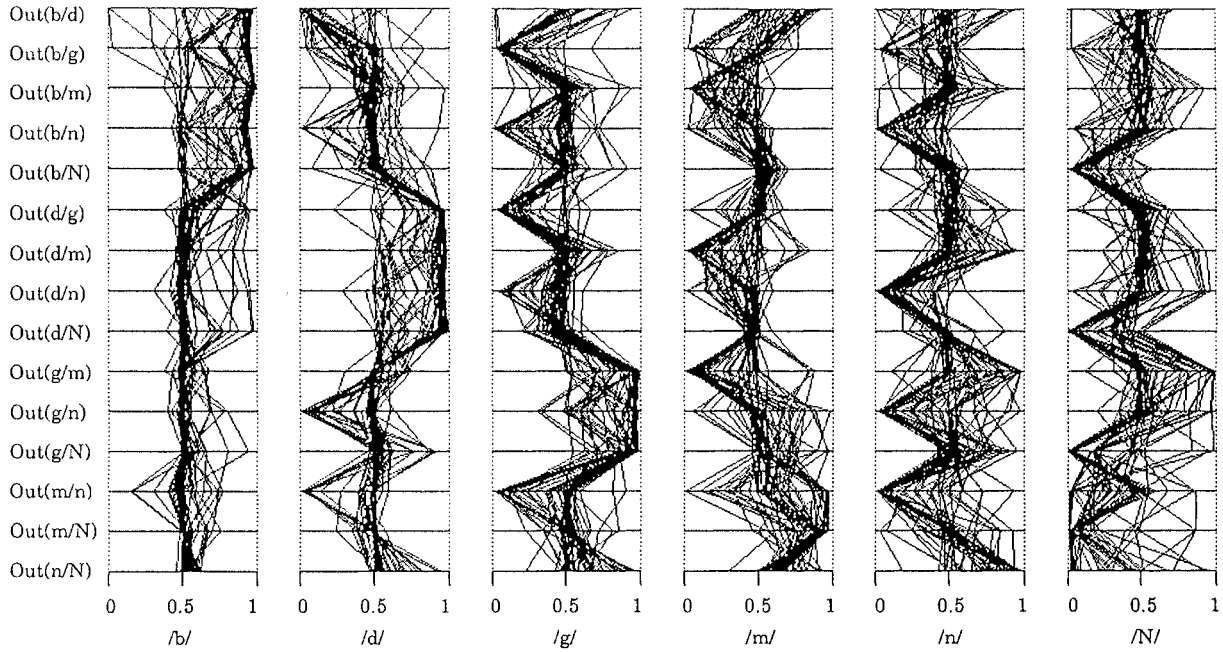

Fig.6. Output values from every PD-TDNNs for "continuous" testing data.

using the PD-TDNN for the "continuous" data.

Consequently, it was found that our training method is effective for suppressing the formation of sharp discrimination boundaries.

5-2. Contribution of Multiple Network Outputs

The PD-TDNN is trained while one side of the phoneme pair influences the other side. Hence, it is most likely that various discrimination boundaries are formed in the networks. Each of the networks has the same phoneme category own one side of corresponding phoneme pair because of the influence of the phoneme category own the other side. This shows that each network discriminates between phoneme categories. Therefore, it is most likely that the final recognition rate increases by judging all pair discrimination scores because all output values are smoothed even if wrong outputs are obtained from some networks.

To confirm this effect, we observed which outputs are obtained from every PD-TDNNs for the "continuous" data. Fig.6 shows the output values from each network.

As shown in this figure, although the output values have some dispersion and some of these include a large error, on the average, they are essentially correct. Consequently, it was found that the effect of smoothing is enhanced by total of all network outputs.

6. Conclusions

In this paper, we proposed a phoneme recognition method using Pairwise Discriminant Time-Delay Neural Networks (PD-TDNNs). This method can simultaneously accomplish the following:

- Formation of a network with a less sharp discrimination boundary.
- Formation of multiple networks each of which has various discrimination boundaries.

Moreover, it was found that this method shows higher accuracy and robustness than a single TDNN for various speaking manners.

Issues that need further research are as follows:

(1) Analysis of the factors responsible for high performance.

(2) Investigation of ways to reduce calculation time.

Acknowledgments

We would like to thank Dr.A.Kurematsu, President, ATR Interpreting Telephony Research Laboratories, for his continuous support of this work. We also acknowledge all the members of the Speech Processing Department for their discussion and encouragement.

References

[1] D.E.Rumelhart and J.L.McClelland, "Parallel Distributed Processing ; Explorations in the Micro Structure of Cognition," MIT Press (1986).

[2] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano and K.Lang, "Phoneme Recognition Using Time-Delay Neural Networks," Tech. Rep. TR-I-0006, ATR Interpreting Telephony Research Laboratories (1987-10).

[3] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano and K.Lang, "Phoneme Recognition Using Time-Delay Neural Networks," IEEE Trans Acoust., Speech, Signal Processing, vol.37, pp. 328-339 (1989-3).

[4] A.Amano, T.Aritsuka, N.Hataoka and A.Ichikawa, "On The Use of Neural Networks and Fuzzy Logic in Speech Recognition," IEEE International Joint Conference on Neural Networks, vol.1, pp. 301-305 (1989-6)

[5] P.Haffner, A.Waibel, K.Shikano, "Fast Back-Propagation Learning Methods for Neural Networks in Speech," ASJ fall-meeting, 2-P-1 (1988-10).

[6] K.Takeda, Y.Sagisaka and S.Katagiri, "Acoustic-Phonetic Labels in a Japanese Speech Database," European Conference on Speech Technology, pp.13-16 (1987-8).

[7] H.Sawai, A.Waibel, M.Miyatake and K.Shikano, "Spotting Japanese CV-Syllables and Phonemes Using Time-Delay Neural Networks," ICASSP'89, S1.7, pp.25-28 (1989-5).