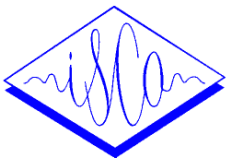

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A FUZZY PARTITION MODEL(FPM) NEURAL NETWORK ARCHITECTURE FOR SPEAKER-INDEPENDENT CONTINUOUS SPEECH RECOGNITION

Keiji Fukuzawa, Yoshinaga Kato and Masahide Sugiyama

ATR Interpreting Telephony Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 JAPAN

ABSTRACT

This paper proposes a Fuzzy Partition Model (FPM) neural network architecture for speaker-independent continuous speech recognition. Generally speaking, conventional TDNN(Time-Delay Neural Network) architecture in its training stage requires much computation time. Nevertheless, an FPM has a rapid training capability that is over two times faster than TDNN's training speed. FPM architecture is combined with an LR-parser and its recognition performance with 278 Japanese phrases is evaluated. The recognition rate of FPM-LR is higher than that of TDNN-LR. This paper also proposes a Multi-FPM-LR method. Using this method, the recognition rate is 77.5% for open speakers.

10.21437/ICSLP.1992-370

1 INTRODUCTION

Recently, neural network (Time Delay Neural Network: TDNN) based speaker-independent continuous speech recognition has been evaluated and its effectiveness with multiple speaker training has been reported[1]. Generally speaking, conventional TDNN architecture in its training stage requires much computation time. This is an especially serious problem for speaker independent continuous speech recognition because TDNNs requires a large amount of speech data from multiple speakers for speaker-independent training. An FPM(Fuzzy Partition Model) is a neural network with multiple input-output units[2]. The performance of FPMs for continuous speech recognition has been evaluated in speaker-dependent mode[3]. In this paper, FPM neural network architectures are applied to speaker-independent continuous speech recognition because FPM has a rapid training capability[4][5]. The speech recognition performance of an FPM is evaluated and compared with the performance of a TDNN. In the following section, FPM architecture and the FPM-LR speech recognizer are described. Next, the FPM-LR is evaluated.

2 SPEAKER INDEPENDENT SPEECH RECOGNITION USING FPMS

2.1 FPMs for continuous speech recognition

FPMs are neural networks with multiple input-output units. In each unit(n), output values($a_i^{(n)}$) are non-negative, and the sum of the output values in each unit is equal to 1.

$$a_i^{(n)} \geq 0 \quad (\forall_i), \tag{1}$$

$$\sum_{i=1}^{N} a_i^{(n)} = 1, \tag{2}$$

where $a_i^{(n)}$ is the ith output from the nth unit of a layer and N is the number of outputs in a unit, referred to as the "dimension" of an FPM unit (Fig.1). In this paper, a feed-forward type FPM architecture (Fig.2) is used and every unit in the same layer had the same dimensions. There can be various forms of the input-output function for (1) and (2). In this paper, inverse *logit transformation* [6] is applied, and is given by:

$$a_i^{(n)} = \frac{\exp(u_i^{(n)})}{1 + \sum_{k=1}^{N-1} \exp(u_k^{(n)})}, \quad (i = 1, \ldots, N-1)$$

$$a_N^{(n)} = \frac{1}{1 + \sum_{k=1}^{N-1} \exp(u_k^{(n)})}, \tag{3}$$

$$u_i^{(n)} = \sum_m \sum_j w_{ij}^{(nm)} a_j^{(m)}, \quad (i = 1, \ldots, N-1) \tag{4}$$

where $u_i^{(n)}$ is the ith input to the unit and $w_{ij}^{(nm)}$ is the weight connecting the jth output of the mth unit to the ith input of the nth unit. Using (3), when the FPM unit receives an $(N-1)$-dimensional vector, it takes out an N-dimensional vector. As seen in (1)-(4), an output value $a_i^{(n)}$ represents the proportion of an input pattern which belongs to the ith class when the number of output-layer units is one. The back-propagation algorithm[7] can be applied to FPMs as well as to multi-layer perceptrons[8]. As FPM-outputs have constrains such as a probabilistic distribution, the Kullback divergence, D, can be used as an error function, which is given by:

$$D((t^{(n)}) : (a^{(n)})) = \sum_n \sum_i t_i^{(n)} \log \frac{t_i^{(n)}}{a_i^{(n)}}, \tag{5}$$

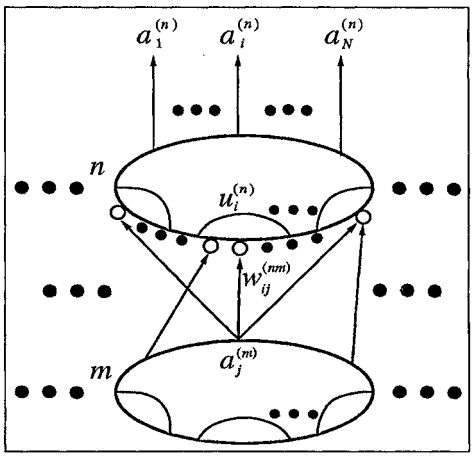

Figure 1: Fuzzy Partition Units with N dimensions

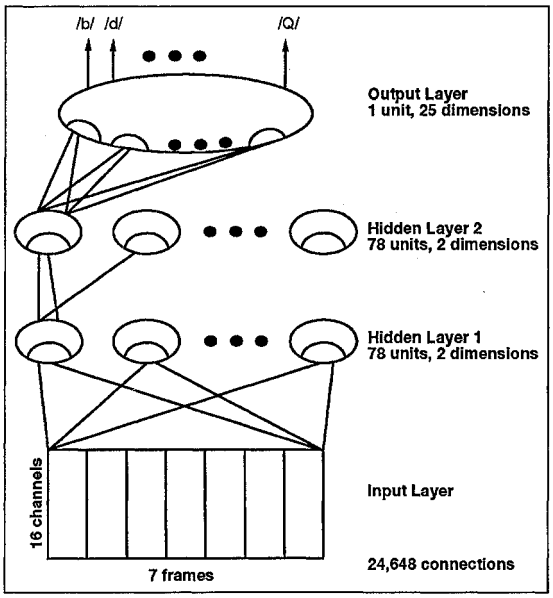

Figure 2: Four-layer Fuzzy Partition Model applied to Speech Recognition

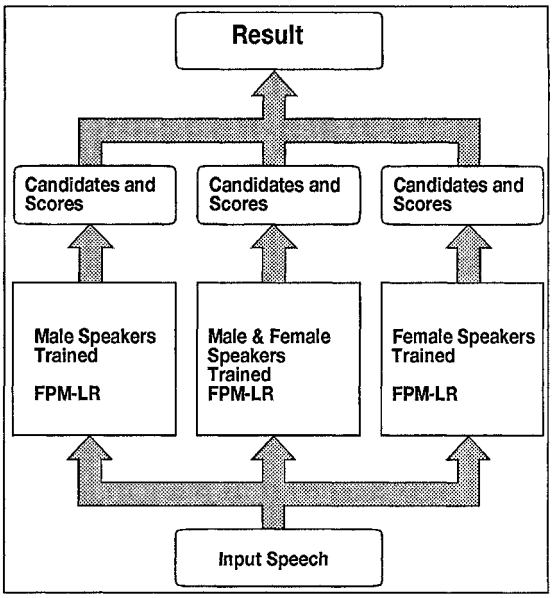

Figure 3: Architecture of Multi-FPM-LR

where $t_i^{(n)}$ is the desired output value and D represent the distance between the desired and actual distributions. It has been shown that D is a more important factor for improving learning speed than the conventional mean squared error[4][5]. The search space for output value is restricted because FPM has the constraint that the sum of its output values is constant. Consequently, FPM's training convergence is superior to that of conventional neural networks[5]. Therefore, FPMs can train phoneme classification quickly even with a large amount of multiple-speaker training samples.

TDNN-LR is proposed as a continuous speech recognizer using a neural network[9]. In this work, FPM architecture is combined with an LR-parser[3]. Phoneme duration values(the length of reference-phonemes[9]) are set as the training speaker's averages: these are extracted with 2,620 training-words uttered by the speakers.

2.2 Multi-FPM-LR method

It is naturally assumed that when an input speaker's voice is spoken into a suitable FPM, the FPM outputs the right phoneme firing pattern and the FPM-LR's recognition-result score (showing the result's reliability) is high. On the other hand, when an input speaker's voice is spoken into an unsuitable FPM, the FPM outputs an incorrect phoneme firing pattern and the recognition-result score is low. If an input speaker's voice is spoken into a multiple FPM-LR and these recognition candidates are arranged in order of their scores (Multi-FPM-LR method), these arranged recognition results will score higher than that of single FPM-LR.

In a preliminary experiment, the method that used one LR-parser for the averaged firing pattern of multiple FPM outputs, showed lower performance than the Multi-FPM-LR method that used separate LR-parsers for each FPM's firing pattern. Therefore, the Multi-FPM-LR method is used. Fig.3 shows the architecture of a Multi-FPM-LR. Since preliminary experiment showed that using three FPMs (a male speakers trained FPM-LR, a female speakers trained FPM-LR, and a male & female speakers trained FPM-LR) gives better performance than using two FPMs (a male speakers trained FPM-LR, and a female speakers trained FPM-LR), the male & female trained FPM-LR is added to the Multi-FPM-LR.

2.3 Acoustic parameter

Table 1 shows speech analysis conditions. 16-channel mel-scaled coefficients are computed form the log FFT power spectrum for every 10ms period. 7frames(70ms) of mel-scaled spectral coefficients are normalized to lie between 0.0 and 1.0, with the average at 0.5. The input acoustic parameter for a FPM consists of the normalized 7frames coefficients.

Table 1: Speech Analysis Conditions

Sampling Frequency	12 kHz
Window Function	Hamming
Window Length	21.3ms
Analysis Interval	10ms
Analysis Method	256 point FFT

For comparison, 7-frame's power or Δspectrum are added to the input parameters. The 7-frame's power values(p_t) are calculated by:

$$p_t = \sum_{k=1}^{M} s_{tk}, \quad (t = -L, \ldots, L)\ (M = 16, L = 3), \tag{6}$$

The Δspectrum(Δs_k) is calculated by:

$$\Delta s_k = \frac{\sum_{t=-L}^{L} t w_t s_{tk}}{\sum_{t=-L}^{L} t^2 w_t}, \quad (k = 1, \ldots, M), \tag{7}$$

$$w_t = 1 - \frac{|t|}{(L+1)}. \tag{8}$$

where s_{tk} is the the log FFT power spectrum(non-normalized) that is the kth spectrum of the tth frame and w_t is a weight parameter in the tth frame. The power and Δspectrum parameters are expected to be robust parameters for different speakers.

3 PHRASE RECOGNITION EXPERIMENTS

3.1 Comparison of FPM-LR and TDNN-LR

The training-times and recognition performances of FPM and TDNN are compared. FPM and TDNN are trained with phoneme segment data (25 phonemes × 250 samples × 8 speakers = 50,000 samples) : it is extracted according to hand labels from 2,620 Japanese words uttered by 8 male speakers. The number of free-weight-parameters in TDNN is 24,825, slightly higher than the 24,648 number in FPM(Fig.2). 278 Japanese phrases from a task called "The International Conference Secretary Service" are used. The complexity of the task is shown in Table 2. Evaluations of phrase recognition are performed with 8 male closed speaker and 2 male open speaker. The training times are shown in Table 3 and the phrase recognition performances are shown in Table 4.

Table 3 shows that the FPM can be trained in less than half the time needed for TDNN. For training, TDNN uses a fast back-propagation learning method[10] that is a thousand times faster than the normal back-propagation procedure. On the other hand, FPM uses a normal back-propagation procedure. If FPM uses a faster procedure, the difference between TDNN and FPM training times would be greater. Table 4 shows that recognition performance of FPM-LR is slightly better than that of TDNN-LR. These results suggest the superiority of FPM-LR in speaker-independent continuous speech recognition.

Table 2: Complexity of 278-Phrase Recognition Task

Number of Words	1,035
Number of Rules in CFG	1,672
Number of States in LR	4867
Phoneme perplexity	5.9

Table 3: Comparison of FPM and TDNN training-times

Architecture	Time* [hour]	Rate (/FPM)
FPM	59.8	1.0
TDNN	138.7	2.32

*HP Apollo 9000/730 (76 MIPS)

Table 4: Comparison of phrase recognition performances of FPM-LR and TDNN-LR for male speakers

Architecture	Recognition rate (%)	
	closed	open
FPM	74.8 (89.3)	71.0 (87.9)
TDNN	72.7 (91.2)	68.3 (89.6)

( ): rank $\leq$ 5

3.2 FPM-LR phrase recognition performance in several training modes

The phrase recognition performances of FPM-LR are evaluated in four modes: speaker-dependent mode, multiple male-speaker training mode (male-only), multiple female-speaker training mode (female-only) and multiple male-and-female training mode(male-and-female). These evaluations are performed as indicated in Table 5. FPMs are trained with 50,000 phoneme segment samples in speaker-dependent, male-only and female-only modes. In the male-and-female mode, the FPM is trained with 100,000 phoneme segment samples and a network of 49,938 weighting parameters, which is about two times larger than for the other modes. Training samples are extracted by using hand labels from 2,620 Japanese words.

Table 6 shows phrase recognition performances. In closed speaker, the performances of the male-only and female-only modes are close to speaker-dependent performance. The recognition performance of the male-and-female mode is lower than that of the male-only and female-only modes. It is supposed that the male and female patterns are so different that their training is difficult for a single FPM.

Table 5: Evaluation conditions

Mode	Training	Testing	
		closed	open
male-only	8 males	8 males	2 males
female-only	8 females	8 females	2 females
male-and-female	8 males & 8 females	8 males & 8 females	2 males & 2 females
speaker-dependent	2 males, 2 females		

Table 6: Phrase recognition performances in several training modes

Mode	Recognition rate (%)	
	closed	open
male only	74.8 (89.3)	71.0 (87.9)
female only	76.9 (91.1)	68.5 (88.8)
male-and-female	72.0 (88.5)	64.5 (83.3)
speaker-dependent	76.3 (91.6)	-

( ): rank $\leq$ 5

3.3 Comparison of single FPM-LR and Multi-FPM-LR

Male-only trained FPM-LR, female-only trained FPM-LR and male-and-female trained FPM-LR are used for the Multi-FPM-LR (Fig.3). These FPMs are trained as indicated in Table 5. The male-and-female trained FPM-LR is applied to a single FPM-LR. Table 7 shows a comparison of phrase recognition performance between single FPM-LR and Multi-FPM-LR. The recognition rate of the Multi-FPM-LR is 70.0%, 5.5% better than the rate of the male-and-female trained single FPM-LR. This shows the effectiveness of Multi-FPM-LR method for speaker-independent continuous speech recognition.

Table 7: Comparison of phrase recognition performance between single FPM-LR and Multi-FPM-LR

Method	Recognition rate (%)
	open (2 males, 2 females)
Single FPM-LR	64.5 (83.3)
Multi-FPM-LR	70.0 (88.5)

( ): rank $\leq$ 5

3.4 Using various speech data for training and adding acoustic parameters

To improve recognition performance for open speakers, various speech data are used for training. Also, 7-frame's power parameters and Δspectrum parameters are added to the input parameters. Table 8 shows a comparison of 2,620 word-data trained FPM-LR and 216 phoneme-balanced words & 597 phrases data trained FPM-LR performances. The phrase data for training is different from testing phrase data. The numbers of training samples are 50,000 for both FPM. The performances are evaluated in the male-only mode. When using balanced-word data and phrase data for training, the recognition rate is 74.8%, 3.8% higher than the rate when using only words data for training. This shows that using various data for training improves phrase recognition performance.

Table 9 shows a comparison of phrase recognition performance with different parameters. 216 phoneme-balanced words and 597 phrases are used for the training data. These performances are evaluated in the male-only mode. By adding 7-frame's power the recognition rate is 75.9%, and adding Δspectrum achieves a recognition rate of 76.4%. These recognition rates are slightly better than the rate without using these parameters. This shows that using these parameters improves phrase recognition performance for open speakers.

Table 8: Comparison of word-data trained FPM-LR and balanced-word & phrase data trained FPM-LR for male speakers

Training samples	Recognition rate (%)
	open (2 males)
word-data	71.0 (87.9)
balanced-word & phrase data	74.8 (89.9)

( ): rank ≤ 5

Table 9: Comparison of input acoustic parameters for male speakers (using balanced-words and phrase data for training)

Parameters	Recognition rate (%)
	open (2 males)
mel-scaled 16-channel × 7-frames	74.8 (89.9)
mel-scaled 16-channel × 7-frames + 7-frame's power (p_t)	75.9 (90.8)
mel-scaled 16-channel × 7-frames + Δspectrum (Δs_k)	76.4 (90.3)

( ): rank ≤ 5

Since using training data from various sources, such as phoneme-balanced words and phrases improved the recognition performance, it was expected that adding 2,620 words to the training data would increase recognition performance. In the next evaluation, 2,620 words, 216 phoneme-balanced words and 597 phrases are used for training data. Table 10 shows the results. The phrase recognition performances are evaluated with 2 male and 2 female open speakers in the male-only, female-only and male-and-female modes and with the Multi-FPM-LR. The input parameters are 7-frame mel-scaled 16-channel FFT outputs. The FPMs are trained with 75,000 phoneme segment samples in the male-only and female-only modes. In the male-and-female mode, the FPM is trained with 100,000 phoneme segment samples. By using word, balanced-word and phrase for training data, FPM-LR recognition performance is improved. The recognition rate with the Multi-FPM-LR is 77.5% for open speakers. It is also expected that adding the 7-frame's power and the Δspectrum parameters will further increase recognition performance.

Table 10: Phrase recognition performance using various speech data for training (words, balanced-words and phrase data)

Method		Recognition rate (%)
		open
Single FPM-LR	male-only	80.1 (93.5)
	female-only	77.8 (93.8)
	male-and-female	71.0 (90.0)
Multi-FPM-LR		77.5 (93.3)

( ): rank ≤ 5

4 CONCLUSIONS

This paper proposed FPMs for speaker-independent continuous speech recognition and reported the evaluation of FPM-LR phrase recognition performance. It has been shown that FPM is more effective than TDNN in both training time and recognition performance. The effectiveness of the Multi-FPM-LR is shown for speaker-independent speech recognition. Using 278 phrases, recognition performance of 77.5% is achieved for open-speakers with the Multi-FPM-LR.

ACKNOWLEDGMENTS

The authors would like to thank Dr. A. Kurematsu and Mr. S. Sagayama for their support of this research. Thanks is also owed to all of the members of the Speech Processing Department for their discussions and encouragement.

References

[1] K. Fukuzawa, Y. Komori and M. Sugiyama, "A Comparison between Multi-Speaker Trained TDNN and Speaker Adaptation Neural Network in a TDNN-LR Continuous Speech Recognition System," *Proc. of ASJ*, (Mar. 1992)(in Japanese).

[2] Y. Tan and T. Ejima, "A Network With Multipartitioning Units," *Proc. of IJCNN*, (Jun. 1989).

[3] Y. Kato and M. Sugiyama, "Fuzzy Partition Models and Their Effects in Continuous Speech Recognition," *to appear in Proc. of IEEE Workshop NNSP*, (Aug. 1992).

[4] Y. Tan, Y. Kato and T. Ejima, "Error Functions to Improve Learning Speed in PDP Models ," *Trans. IEICE, Vol. J73-D-II, 12*, pp.2022-2028, (Dec. 1990)(in Japanese).

[5] Y. Kato, Y. Tan and T.Ejima, "A Comparative Study with Feedforward PDP Models for Alphanumeric Character Recognition," *Trans. IEICE, Vol. J73-D-II, 8*, pp.1249-1254, (Aug. 1990)(in Japanese).

[6] J.T.Kent and K.V.Mardia, "Spatial Classification Using Fuzzy Membership Models," *IEEE Trans. Pattern Anal. Mach. Intell., Vol. PAMI-10*, pp.659-671, (Sep. 1988).

[7] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning Internal Representations by Error Backpropagations," *Parallel Distributed Processing Vol. 1, MIT Press*, (1986).

[8] M.Minsky and S.Papert, "Perceptrons," *MIT Press*, (1969).

[9] M. Miyatake, et al., "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks," *Proc. ICASSP90*, S8.10, pp.449-452 (Apr. 1990).

[10] P. Haffner, et al., "Fast Back-Propagation Learning Methods for Phonemic Neural Networks," *Proc. Eurospeech89, Vol. 2*, pp.553-556, (Sep. 1989).