A STUDY ON THE RECOGNITION OF SPOKEN KOREAN LOCAL-NAMES USING SPATIO TEMPORAL

Do-sun Song *, Suk-dong Kim **, Haing-sei Lee ***

* Department of Computer Engineering, Joongkyoung College, Korea
** Department of Computer Science, Hoseo University, Korea
*** Department of Electronic Engineering, Ajou University, Korea

ABSTRACT

This paper is about an experiment of speaker-independent automatic Korean spoken words recognition using Multi-Layered Perceptron and Error Back-propagation algorithm. The words were not segmented into syllables or phonemes, and some feature components extracted from the words in equal gap were applied to the neural network.. This paper tried to find out the optimum conditions through various experiment which are comparison between total and pre-classified training.

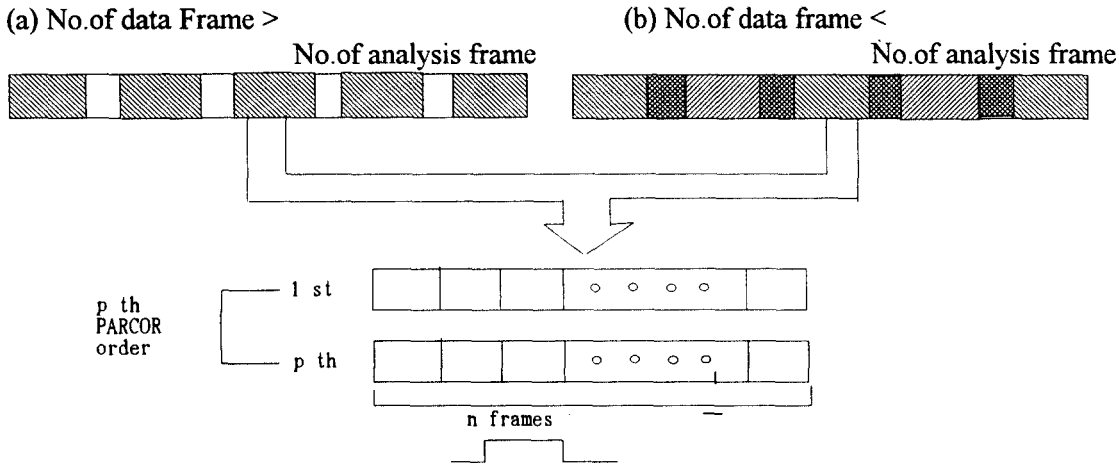
1. INTRODUCTION

The language is the most important communication means among the mankind. The language by information transport is classified into two kinds; the spoken language related to the transport means of a hearing sign information by the voice and the written language related to the transport means of a visual sign information by the character. The voice information is disappeared as soon as it was generated. Therefore the boundary of the propagation is very finite as far as an auxiliary device is not used. Recently the use of the voice language as the communication means between mankind and the machine is studied deeply into a subject thanks to the development of the information processing technology. The automatic speech recognition by the machine is very important because it takes too much time to put information into the computer[1]. If the computer is able to recognize the speech, the information processing is carried out more rapidly and comfortably.

The speech recognition has been studied with the synthesis technology and the transport of the speech using digital signal processing. The recognition of a number using formant frequency was carried out partially at Bell laboratory in 1952 first[2]. The speech recognition was studied in early 1970 as doing DARPA project[3]. The bottom-up approach has been chosen as the research method about the structure and the function of the hearing organism since the importance of the special knowledge was well known in middle era of 1970[4][5]. Nowadays the various methods such as ANN(Artificial Neural Network)[7-11], fuzzy theory and Viterbi algorithm of the HMM(Hidden Mark Modeling) [6]are used to recognize the speech. There are still much difficulty for developing the speech recognition algorithm. Thanks to the development of the machine and the biology, the structure and the function of the hearing organism are well known to us. But the function of the brain in order to analyze the signal sent from the hearing organism is mysterious to us. The problem to segment the boundary between syllables in a phoneme is not solved completely. Although the speech recognition is possible, we must have the knowledgement about phoneme to understand the speech contents as if the human recognize the speech The complete speech recognition using the machine is hindered by these problems. In the subsequent section, content and methodology of the speech recognition using the neural network are explained.

2. THE METHOD OF EXPERIMENTATION

In this section we introduce a method of experimentation of the back-propagation algorithm for layered feed forward networks. The feature of voices is calculated from the finite number of samples in a frame after a file is divided into 10 or 40 frames at the same intervals. The amount of sample to calculate the feature from the divided frame is ranged from 30% to 120% of original data. It is considerable amount considering with the high correlation of the neighbor speech sample although the amount of the extracted data is 30% of voice data. The more analysis frame can be obtained as the divided frames are overlapped. Consistency between the overlapped frame and the other frame can be controlled automatically according to the length of the voice file. As doing the above processing, the pronunciation duration time of the voice compensates for the effect of the length change at each file. Including the change of the voice length is very important for developing voice recognition technology. PARCOR coefficients are used for feature of the speech recognition. Fig.1 shows the extracting processing of the feature calculating the PARCOR coefficient using autocorrelation method based on autoregressive method.

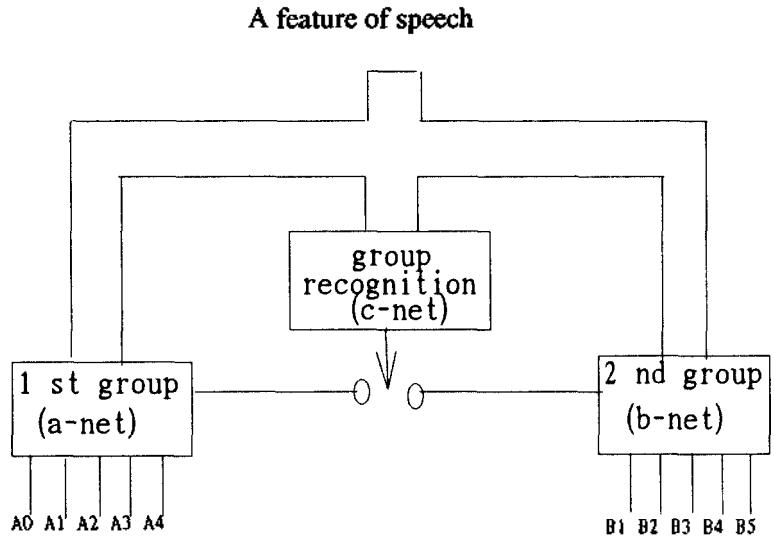


[Fig 1] process of extracting feature components

An experimentation data is obtained every time a 20-aged-man pronounce a word. The number of the learning data is 350 which means seven experimentation data multiplied by 50 signed as cityname and the number of the recognizable data is 150. The learning rate and the momentum rate of the neural network are given as 0.1 and 0.6 respectively. The learning is to be completed when the range of the error E between the output of the neural network and the desired value is within 0.2, otherwise to be stopped although the number of the iteration is up to 1500 times.

-Total Learning and Pre-classified Learning

Fifty sorts of the words obtained as pronouncing 50 citynames are chosen at this experimentation. The learning method is classified into two kinds: total learning related that almost all of data are put into one neural network at a time and pre-classified learning related that they are put into several neural networks after separating them once. It is rather than adapt pre-classified learning considering with the learning time and the finite number of input neuron. The third neural network is needed because a data doesn't know that which route of the networks is chosen. The third neural network has one output neuron layer. If the output value of the layer is 0, the first group is selected. Otherwise the second group is selected. Fig.2 shows the pre-classified learning method using three neural networks.



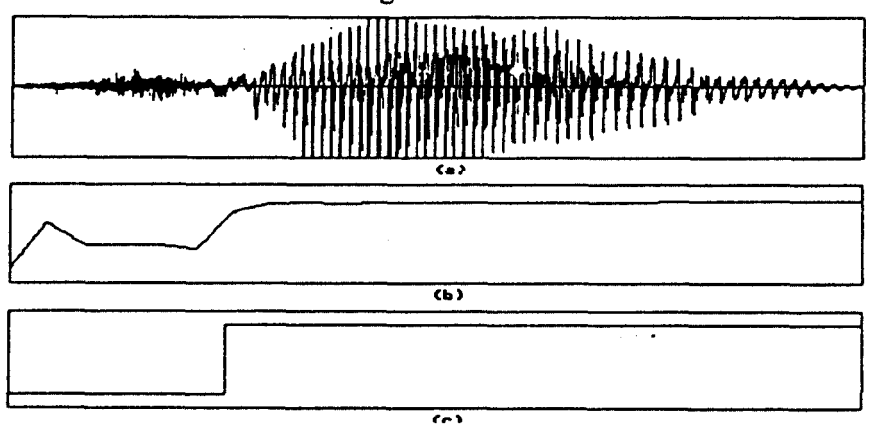
[Fig 2] pre-classifying method

As the number of the connecting line between neurons, which means the weighing factor, is diminished, the third neural network is suitable for making decision whether the input of 0(A-net) or 1(B-net). Fifty sorts of data has to be divided into the two groups to take pre-classified learning. The separation reference is decided whether the unvoiced sounds which is divided into two or three parts are included in the voice element of the word. We know that how much the unvoiced sounds is included to the word as calculating the autocorrelation because the autocorrelation of the unvoiced sounds is less than one of the voiced sounds generally. Equation shows the calculation of the autorrelation.

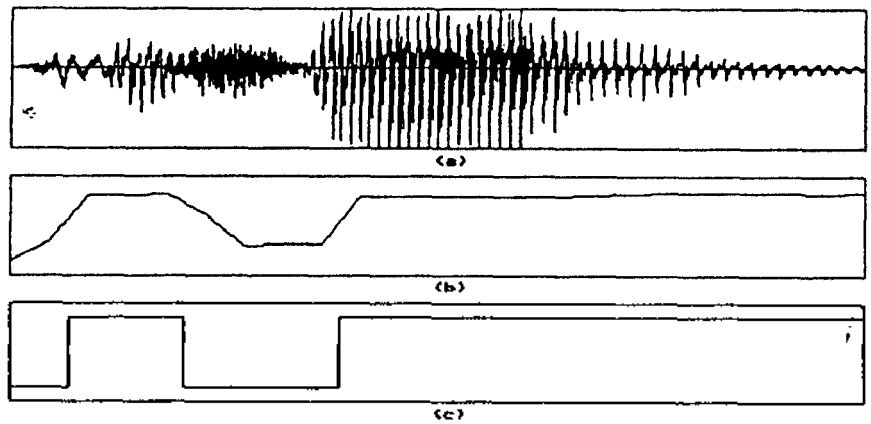
$$y(n,k) = x(m)\, w(m)\, x(m+k)\, w(m+k), \qquad k=0,1,\ ,M-1$$
$$r(n) = \max[\, y(n,k)/\, y(n,0)], \qquad n=1,2,\ ,N$$

Where w is Hamming window, M is the number of the samples, 320, consisting of one frame, and N is the number of frame. Fig. 3 shows the speech waveform of the word "Seoul", autocorrelation, and the thresholding result of Seoul.

A phoneme "S" which consists of the front of the voice has less correlation and phonemes "eouls" consisting of the voice sound has higher correlation. u(w-0.82) is the function changing (b) of the fig.3 into (c) of fig.3. w of u(w-0.82) is the value of autocorrelation. If the value of autocorrelation is over 0.82, 1 is assigned. Otherwise, 0 is assigned. The first group consists of 22 citywords and the second group consists of 28 shown in table 2 and table 3. Fig.4 shows that the first group has one of "1" interval and the second group has more than one of "1" interval as the result of thresholding autocorrelation.



[Fig 3] For /SEOUL/ (a)speech waveform (b)autocorrelation (c)after thresholding



[Fig 4] For/busan/(a)speech waveform (b)autocorrelation (c)after thresholding

As the above method, the target words are classified into the two groups. Comparing the recognition ratio obtained as fifty words are put into the C-net, A-net and B-net separately using pre-classified learning with one they are put into one network at a time using total learning is focused on at first experimentation. Comparison of the total learning and the pre-classified learning is done as follows: the first case related to extracting the tenth PARCOR coefficient from a frame about twenty frames and the second case related to extracting the tenth PARCOR coefficient from a frame about fifteen frames. As there are fifty words to be separated in case of the pre-classified learning and total learning, six neurons are enough to separate fifty words by the binary coding method because they can separate sixty-four number of output. All of the output neurons except the fourth experimentation adapts binary coding mechanism. The pre-classified learning is used except the first experimentation.

[Table 1] words of group 1

SEOUL	BA-LAN	EI-WANG	WON-DANG	YUNMU DAI
DAE-GU	SOO-WON	HA-NAM	BO-EUN	DONG-KWANG
GO-YANG	SUNG-NAM	DO-GYE	YOUNG-WOL	
KWANG-MYANG	AN-YANG	SU-LAK	CHUL-WON	
MI-GUM	YONG-IN	YANG-GU	TAI-BAK	

[Table 2] words of group 2

BU-SAN	HWA-CHUN	YWANG-JU	PA-JU	CHOON-CHUN	JANG-HOWON
KWANG-JU	MUN-SAN	YOU-JU	HWA-SUNG	PWANG-CHANG	JUMUN-JIN
DAI-JUN	SI-HWEUNG	YOUN-CHUN	WON-JU	HONG-CHUN	JANG-SEUNGPO
KWANG-HWA	AN-SAN	O-SAN	IN-JEI	NAM-YANGJU	
KWA-CHUN	AN-SUNG	IL-SAN	JUNG-SUN	EIJ-JUNGBU	

3. EXPERIMENTATION RESULT

The number of iteration, learning time and recognition rate carrying out total learning and pre-classified learning are given in table 3.

[Table 3] experimental result of and pre-classifying learning method and Total learning..

			it	ti	R.r(%)
N.f 20 P.O 10 N.N.h 20	T.L		1500	19:19	78.7
	PC	C-net	69	0:32	98.7
		A-net	101	0:21	92.4
		B-net	1500	6:47	88.1
		sum	1669	7:30	89.1
N.f 15 P.O 10 N.N.h 20	T.L		1500	16:28	68.0
	PC	C-net	55	0:20	94.0
		A-net	74	0:12	78.8
		B-net	103	0:22	82.0
		sum	231	0:54	75.6

Where N.f is No. of frame, P.O is PARCOR order, N.N.h is No. of nuron in hidden layer, T.L is Total learning, it is iteration ti is Time R.r is recognition rate. The recognition rate of A-net and B-net has not any more meaning independently without considering with C-net. The recognition rate is calculated using pre-classified learning as follows:

$$\frac{\text{R.r of C-net} * (\text{R.r of A-net} + \text{R.r of B-net})}{2}$$

The recognition rate of each network is dependent on how much 150 files, consisting of voice data of fifty sorts multiplied by 3, can be separated. The recognition rate is 90% if 15 data of 150 files is not be able to be separated. The pre-classified learning is more excellent for speech recognition than the total learning as shown at table 3.

4. CONCLUSION

This paper presents the speech recognition of two or three syllables of a word a twenty-aged man pronounced using the artificial neural network. The recognition rate using pre-classified learning has higher than one using total learning has. One-to-one mapping has higher recognition rate than binary coding method as the type of the neuron output. But the learning time of binary coding method is more efficient than one of one-to-one mapping because the number of output neuron is increased proportionally to the input pattern when many vocabularies are used for learning. As the feature consisting of 15 frames multiplied by tenth order PARCOR coefficient is applied to the neural network that has 150 input layer and 20 hidden layer whose output type is one-to-one mapping, the highest recognition rate, 89.6% is obtained among 14 sorts of simulations.

REFERENCE

[1] S.Furui *Digital Speech Processing, Synthesis, and Recognition* 1992 Marcel Dekker Inc.
[2] K.H.Davis, R.Biddulph, and S.Balashek, *Automatic Recognition of Spoken Digits J. Acoust. Soc. Am.*,

Vol.24, 1952.
[3] R.D.Peacoke and D.H.Graf *An Introduction to Speech and Speaker Recognition* *Computer* Vol.23 No.8 August 1990
[4] L.R.Rabiner and R.W.Schafer *Digital Processing of Speech Signals* 1978 Prentice-Hall Inc.
[5] H.Sakoe and S.Chiba *Dynamic Programing Optimazition for Spoken Word Recognition*" *IEEE Trans. Acoust., Speech, Signal Processing*. Vol. ASSP-26, pp. 43-49, Feb. 1978.
[6] Lee, Kai-Fu and H.W. Hon, *Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM Proceedings IEEE ICASSP*. pp.123-126, 1988.
[7] D.E. Rumelhart, H.L. McClelland, *Parallel Distributed Processing.*, The MIT Press Cambridge, Massachusetts London, England, 1986.
[8] T. Kohonen, *The 'neural' Phonetic Typewriter IEEE Computer Special Issue on Neural Networks and Neural Computing*, pp. 11-22, March 1988.
[9] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang , *Phoneme Recognition using Time-Delay Neural Network IEEE Trans. Vol. ASSP-37*, No. 8, Aug. 1989.
[10] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang : *Phoneme Recognition: Neural Networks vs. Hidden Markov Models Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY, April 1988
[11] R.P.Lippmann *Review of Neural Networks for Speech Recognition* *Readings in Speech Recognition* 1990 Morgan Kaufman, Inc