A Study on the Performance of TDNN-Based Speech Recognizer with Network Parameters

*Hojung Nam, **Y. Kwon, **Inchan Paek, **K. S. Lee, and *Sung-il Yang

Abstract

This paper proposes a isolated speech recognition method of Korean digits using a TDNN(Time Delay Neural Network) which is able to recognize time-varying speech properties. We also make an investigation of effect on network parameter of TDNN, ; hidden layers and time-delays. TDNNs in our experiments consist of 2 and 3 hidden layers and have several time-delays. From experiment result, TDNN structure which has 2 hidden-layers, gives a good result for speech recognition of Korean digits. Mis-recognition by time-delays can be improved by changing TDNN structures and mis-recognition separated from time-delays can be improved by changing input patterns.

Ⅰ. Introduction

Speech is a natural and easy method of communication among human beings. Even in the various new communication services to be provided in the future, speech will still play important roles. Recently, speech recognition technology has greatly progressed, in combination with AI, signal processing, statistical modeling and various other technologies. Speech recognition services are based on technology which has recently been realized or which will be realized in the near future. This services include voice dialer for cellular mobile radio and regular telephones, response systems for guidance and reservation, order receiving services using natural conversational speech, and voice-input word processors. Speaker recognition will also be used as a security control technique.

Speech is essentially a time-varying phenomenon. Exploring temporal variability in representation of speech is one of the outstanding problems in speech recognition. In the field of speech processing, the adaptive and learning ability of neural networks has been expected to provide excellent properties[1]. When a neural network is applied to speech recognition, none of updating template database is required. Conventional artificial neural networks are structured to deal with static patterns. Many neural networks cannot make accurate clusters and recognize correctly when the input data set holds time-varying characteristics[2][3]. Speech is inherently dynamic in time. Hence, some modifications to the simple structures are required. There is no known correct or proper way of handling speech dynamics within the framework. However several reasonable structures have been proposed and studied, and one solution to make use of time-varying characteristics is the time-delay neural network(TDNN). A TDNN is an MLP(Multi-Layer Perceptron) with fixed time delays. Each cell of a TDNN weights not only current input feature vector $f(t)$ but also N preceding vectors $f(t-n)$ [4][5].

This paper discusses TDNN suitable for speech recognition system and develops a network for a recognition of time-varying speech signals of Korean digits. The designed network can recognize characteristics of speech with time-varying properties. In experiments, the method to make use of time-varying characteristics is demonstrated by TDNN system. And we compare the performances of TDNN with those of VQ and MSVQ.

In Ch.2, we introduce Multi-Section Vector Quantization(MSVQ) which is a method to use time-varying properties. This is also time-normalizing input pattern in TDNN. Ch.3 and Ch.4 are devoted to explain TDNN. Ch.5 is concerned with LPC Cepstrum. Our experiments and Results are shown in Ch.6.

Ⅱ. Multi-Section Vector Quantization(MSVQ)

Conventional standard vector quantization approach that uses a single vector quantizer for the entire duration of utterance is not designed to preserve the sequential characteristics of speech class. It requires that codebook for a particular utterance speech be properly designed to minimize average distortion. Suppose there are M utter-

*Robust Speech Processing Laboratory,
**Acoustics Laboratory, Hanyang University.

Manuscript Received : July 7, 1997.

ance speech classes to be recognized. We collect M sets of training data. Each training set should contain a number of utterances of the same speech. M codebooks are designed by using minimum average distortion for the M utterance speech classes. Each codebook represents a characterization of each speech class. During the recognition operation, it doesn't require any explicit time alignment[6][7].

Lack of explicit characterization of sequential behavior can be remedied by treating each utterance speech as a concatenation of several utterance sub-classes, each of which is represented by a VQ codebook. We call this Multi-Section Vector Quantization. For an utterance speech, we divide the utterance into N_s section to decompose it into a concatenation of N_s.

Given a set of training utterances of known class, N_s sets of training data are formed and used to design N_s codebooks. These N_s codebooks have an implicit temporal order because they correspond to different portions of the utterances. Similar to the single codebook case, each set of Ns successive codebooks represents one class, and the average distortion incurred in encoding an unknown utterance with the corresponding successive vector quantizers is the discriminant score for the recognition decision[8][9].

Fig. 1 represents 5-section vector quantizer designed from 3 utterance speeches.

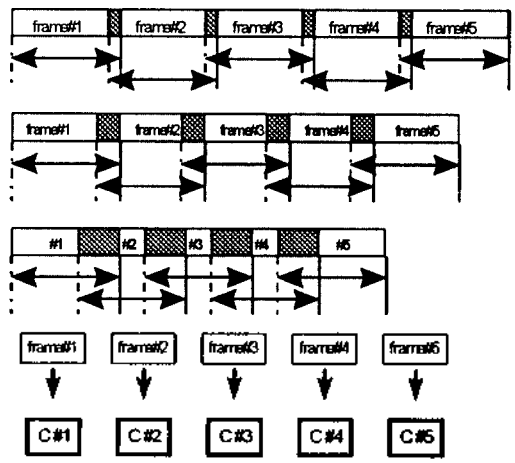

Figure 1. 5-MSVQ production

Ⅲ. TDNN(Time Delay Neural Network)

Neural Networks have been seriously considered for a wide range of problems: parallel computation, robustness or fault tolerance, nonlinearity. Conventional artificial neural networks are structured to deal with static patterns. However, since speech is inherently dynamic in time, it needs some modifications to the neural network[9][12]. There is no known correct or proper way to handle speech dynamics. However several reasonable structures have been proposed and studied. Perhaps the simplest neural network structure that incorporates speech pattern dynamics is the time delay neural network(TDNN)[4][5][12]. TDNN is a MLP with fixed time delays from lower layer to upper layer. We define input pattern as N speech sections. This structure extends input to each computational element to include N speech sections.

To completely specify a neural network, values for weighting coefficients and the offset threshold for each computation element must be determined, based on a labeled set of training data. By a labeled training set of data, we mean an association between set of input vectors $X_1, X_2, \cdots, X_{N-1}$ and desired vectors $d_1, d_2, \cdots, d_{N-1}$. Back propagation algorithm is used in TDNN. Using distortion between output result values and desired values, weighting coefficients are updated from output layer to input layer backwardly. The algorithm is as follows;

[STEP 1]

Initialize all weight coefficients and offsets

[STEP 2]

Input values $X_1, X_2, \cdots, X_{N-1}$ and desired output values $d_1, d_2, \cdots, d_{N-1}$.

[STEP 3]

Calculate output value at each node

$$net_j = \sum_{i=0}^{N-1} w_{ij} x_j , \tag{1}$$

$$y_j = \frac{1}{1 + e^{-(net_j - \theta_j)}} \tag{2}$$

where, i's are lower layer nodes,
j is an upper layer node,
and W_{ij} is a weight coefficient of the branch connecting node I and node j

[STEP 4]

Update weights

$$w_{ij}(t+1) = w_{ij}(t) + \eta_j \delta_j x_j + \alpha[w_{ij}(t) - w_{ij}(t-1)] \tag{3}$$

where η – training rate and α = momentum
for hidden layer,

$$\delta_j = y_j(1 - y_j)(d_j - y_j) \tag{4}$$

and for output layer

$$\delta_j = x_j(1 - x_j) \sum_k \delta_k w_{jk} \tag{5}$$

where k is an upper layer node than node j's

Ⅳ. TDNN Network Structures

It must specify four characteristics to implement an arbitrary neural network: number and type of inputs, network connectivity, choice of offset, and choice of non-linearity. Methods using Cepstral coefficients as a input pattern make an improvement on the recognition. But network connectivity is not so clear. Network connectivity involves the number of hidden layers and the number of nodes. Generally speaking, there is no good rule of thumb as to how large (or small) such hidden layers must be. In designing TDNN networks, the number of hidden layers and time-delay terms are very important network parameter. If we will add a hidden layer and increase or decrease time-delay, the structure is more complicated and has an important effect on other layers, time-delay, and recognition rate. This is due to the connectivity of the network. We will investigate influence of hidden-layer and time-delay on network. Changing these parameters, we observe recognition rate and mis-recognized digit.

Ⅴ. LPC Cepstrum

It can consider the basic model of speech production as a vocal tract filter $H(z)$ excited by a periodic excitation function $E(z)$ for voiced speech or white noise $E(z)$ for unvoiced speech. If, in the frequency domain, the product of the excitation and filter spectrum is transformed to the summation of these two spectra, the transformation from the frequency domain back to the time domain by Fourier transform results in the cepstrum. The Excitation $E(z)$ and vocal tract filter $H(z)$ are linearly separated by a complex logarithm operation. Then

$$logX(z) = logH(z) + logE(z)$$

The LPC cepstral coefficients C_n are defined as inverse transformation of the above log spectrum $logX(z)$. The cepstral coefficients, C_n, of the spectra obtained from LPC analysis can be computed recursively from the LPC coefficients, α_i.

$$c_n = -\alpha_n - \sum_{i=1}^{n-1} \frac{n-i}{n} \alpha_i c_{n-i}, \qquad n \geq 1$$

where $\alpha_i = 0$ when $i > p$ (p is the order of LPC analysis).

Ⅵ. Experiments and Results

Speech data were sampled by 11 kHz, converted to 8 bits, pre-emphasized, covered with a 200-sample Hamming window and 10th order LPC coefficients were evaluated and converted to 10th order LPC cepstral coefficients. Speech data are 10 Korean digits: /young(0)/, /il(1)/, /i(2)/, /sam(3)/, /sa(4)/, /o(5)/, /yuk(6)/, /chil(7)/, /pal(8)/, /gu(9)/. To obtain 40 frames per each speech, we normalized utterance speech to 40 frames. 4 male and 1 female speakers uttered 20 times per each digit. 3 out of 20 speech data were used in training procedure, and the others were used in test procedure.

Our experiments were composed of 2 steps. At 1st step, we compared TDNN with MSVQ, and VQ method. This experiment had 2 network structures; 2 hidden layers for network 1, and 3 hidden layers for network 2. Time-delays were 6-frames. Each network was shown in Fig. 2 and Fig 3, respectively. At 2nd step, we constructed several TDNN structures which had 3 hidden layers and different time-delays. From this result we analyzed of time-delay effect on TDNN network. Table 1 shows these network. Total number of training iterations were 10,000 times, and the learning rate was fixed to 0.5. The initial node offsets and weights were randomly selected from −0.5 to 0.5. Total TDNN network structures are showed in Table 1.

Table 1. TDNN structures with 3 hidden layers, different time-delays

	Input-H1	H1-H2	H2-H3	H3-Output
Network 3	6	10	15	12
Network 4	3	6	14	20
Network 5	4	11	10	18
Network 6	8	10	15	10
Network 7	5	9	13	16
Network 8	7	10	17	9
Network 9	10	13	15	5
Network 10	8	13	16	6
Network 11	9	12	17	5
Network 12	10	13	15	5
Network 13	5	10	15	13

where H1 : Hidden Layer 1, H2 : Hidden Layer 2, H3 : Hidden Layer 3

1. Performance of TDNN

To evaluate performances of TDNN, we have run the same speech recognition experiments as VQ and 40-MSVQ with 10th order LPC coefficients. The size of codebook for vector quantizer was 16. The results of experiments are as follows: the recognition rates were 95.4% for VQ, 97.8% for 40-MSVQ, 98.7% for Network 1, and 98.1% for

Network 2. The best recognition rates were 97.6% for VQ, 98.8% for 40-MSVQ, 100% for Network1, and 99.4% for Network 2. The details of recognition rates can be found in Table 2. It is shown that the system using sequential characteristics of speech, that is, MSVQ and TDNN would give good results. The comparison of recognition rates between Network 1 and Network 2 would show that the performance of TDNN would not depend on the number of layers. Also, it should be noted that speech data recognized incorrectly were different for Network 1 and Network 2 (see Table 3). We can improve the recognition of TDNN using LPC cepstral coefficients instead of LPC. This results are found in Table 4.

2. Performance of TDNN structures with different time-delays

As we were varying time-delays with carefulness, the performance of TDNN is observed and shown in Table 4. From this results, when time-delays are 5, 6, and 7, performances of them are better than those of other time-delays. But Network 3 and 13 are similar networks. Only difference is time-delays between Input-H1 and H3-Output. Mis-recognitions is 3 for Network3 and 5 for Network 13. But they have no similarity of mis-recognized digits between two networks. It says that TDNN is very sensitive to time-delays.

When we observe the results of mis-recognized digits, we can find some similarity. All networks have the same mis-recognized digits; input pattern 8, 1, 4 or both. This patterns have no relation with time-delays. Time-delays influence only some mis-recognized. Changing TDNN structures, we can improve recognition rate, but the same mis-recognized digits, which have no relation with time-delays, can not be overcome only by changing TDNN network structures. This result are shown in Network 1 having a best result at Table 2(utterance B). When we use LPC Cepstrum, this problem was removed. At Table 2, B utterance has 2 mis-recognized digits; (6, 0) and (8, 1). Mis-recognition (8, 1) results from time-delays. When LPC cepstral coefficients are used, mis-recognized digits have 2; but both (6, 0). But changing from Network 1 to Network 2, this mis-recognized digits are removed; 100% recognition results(Table 3, B utterance). It says that we can improve recognition rate of the time-delay dependent digits by changing network structures. We can also improve time-delay independent digits by changing input pattern.

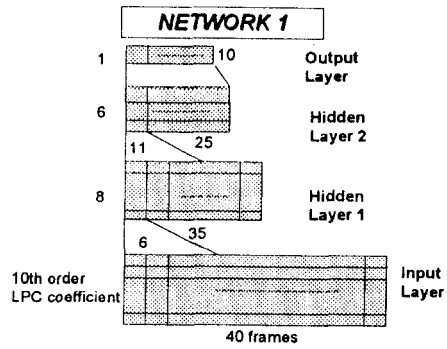

Figure 2. TDNN Network 1

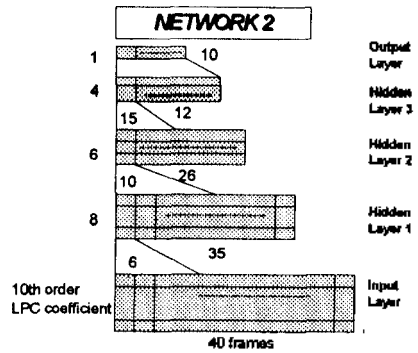

Figure 3. TDNN Network 2

Table 1. Speech Recognition Rate of Korean Digits

	VQ	MSVQ	Network 1	Network 2
A	95.3	96.5	97.6	98.8
B	92.9	98.8	99.4	98.2
C	96.5	97.1	98.2	97.6
D	94.7	98.2	98.2	96.5
E	97.6	98.8	100	99.4
Average	95.4	97.8	98.7	98.1

Table 2. Incorrectly recognized speech data by TDNN

	Network1			Network2		
	Input Digit	Output Digit	the number of mis-recognized digit	Input Digit	Output Digit	the number of mis-recognized digit
A	5	9	2	8	1	1
	9	5	2	8	4	1
B	6	0	1	7	8	1
	8	1	1	8	1	1
				8	4	1
C	6	1	2	0	3	1
	6	5	1	5	6	1
				6	1	1
				6	0	1
D	6	0	2	3	0	1
	8	1	1	4	1	1
				4	3	3
				5	0	1
E	X	X	X	0	4	1

Table 3. Improvement Recognition Rate using Cepstral Coefficients

	Network 1		Network 2	
	LPC	LPC Cepstral	LPC	LPC Cepstral
A	97.6	98.2	98.8	98.8
B	98.8	98.8	98.2	100
C	98.2	99.4	97.6	100
D	98.2	98.8	96.5	98.8
E	100	100	99.4	100
Averag	98.6	99	98.1	99.5

Table 4. Performance of TDNN with different time-delays(B utterance)

	mis-recognition	input pattern	mis-recognized digit
Network3	3	7	8
		8	1
		8	4
Network4	6	1	8
		3	-
		5	4
		6	...
		8	1
		8	4
Network5	5	3	4
		6	
		8	1
		8	4
Network6	3	6	-
		8	1
Network7	3	6	...
		8	1
		8	4
Network8	4	3	
		6	
		8	1
		8	4
Network9	4	6	
		6	9
		8	1
Network10	6	3	...
		6	
		6	9
		8	1
Network11	6	-	6
		1	7
		3	...
		6	-
		8	1
Network12	6	-	6
		1	7
		7	8
		8	1
Network13	5	1	8
		5	4
		6	9
		8	1

Ⅶ. Conclusion

In this paper, we proposed Korean digits recognition system. With the help of TDNN structure, we could use sequential characteristics of speech. The recognition rate to TDNN is 98.6% in speaker-dependent, and can be reached to 99.5% with LPC cepstral coefficients. Time-delays gave influence some recognition rate. Mis-recognition dependent on time-delays was improved by changing network structures. Mis-recognition independent on time-delays was improved by changing input patterns.

References

1. Cabral Jr., E. F. and Tattersall, G. D., "Trace-Segmentation of Isolated Utterances for Speech Recognition," *Proceedings of the ICASSP*, Michigan, Detroit, pp. 364-368, May, 1995.
2. E. Levin, "Hidden control neural architecture modeling of nonlinear time varying systems and its applications," *IEEE Trans. Neural Networks*, vol. 4, no. 1, pp. 109-116, 1993.
3. D. P. Morgann and C. L. Scofield, *Neural Networks and Speech Processing*, 1991, Kluwer Academic Publishers.
4. A. Waibel, et al., "Phoneme Recognition: Neural Networks vs. Hidden Markov Models," *Proceedings IEEE International Conference on Acoustic, Speech and Signal and Speech Processing*, New York, 1988, pp. 107-110.
5. K. J. Lang and G. E. Hinton, "The Development of the Time-Delay Neural Network Architecture for Speech Recognition," *Tech. Rep. CMU-CS-88-152*, Carnegie-Mellon University, Pittsburgh, PA, 1988.
6. J. E. Shore and K. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Information Theory*, IT-29(4): 473-491, July, 1983.
7. L. R. Rabiner, C. K. Pan, and F. K. Soong, "On the performance of isolated word speech recognizers using vector quantization and temporal energy contours," *AT&T Tech. J.*, 63(7): 1245-1260, 1984.
8. D. K. Burton, J. E. Shore and J. T. Buk, "Isolated-word speech recognition using multisection vector quantization codebooks," *IEEE Trans. On ASSP*, vol. 33, NO. 4, Aug., 1985.
9. L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, 1993, Prentice Hall.
10. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition using Time Delay Neural Network," *IEEE Trans. Vol. ASSP-37*, Aug., 1989.
11. D. J. Burr, "Experiment on Neural Net Recognition of Spoken and Written Text," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-36, pp. 1162-1168, July, 1988.
12. R. P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Mag*, Vol. 4, pp. 4-22, 1987.
13. L. R. Rabiner and R. W. Schafer, *Digital Processing of*

Speech Signals, Prentice Hall, 1978.

14. H. Nam, Y. Kwon, I. Paek, K. S. Lee, S. Yang, "A Study on Speech Recognition of Korean Digits using TDNN," *Proc. Conference on Acoustics, Korean Institute of Acoustics*, Vol. 15, pp 87-90, Nov, 1996, the Acoustical Society of Korea.

▲Hojung Nam

He was born in Seoul, Korea, in 1972. He received the B.S. and M.S. degrees in control and instrumentation engineering from Hanyang University, Ansan, Korea in 1995 and 1997, respectively. He is a graduate student for Ph.D currently in the course of Ph.D. degree at Hanyang University. His current research interests include digital speech and image processing, wavelets, neural network, and speech recognition.

▲Sung-il Yang

He was born in Goesan, Chungbuk, Korea in 1956. He received in B.S. degree in Electronics Engineering with the greatest honors from Hanyang University, in Seoul, Korea, and his M.S. and Ph.D. degrees in Electrical & Computer Engineering from the University of Texas at Austin, Austin, Texas, in 1986 and 1989, respectively.

Since 1990, he has been with Dept. of Control and Instrumentation Engineering, Hanyang University. He is now an associate professor and his current research interests include speech recognition, digital signal processing, microwave drying, and responsible technology.

He is also a member of IEEE, Korea Institute of Telematics and Electronics, and the Acoustical Society of Korea.

▲Inchan Paek

He was born in Daegu, Korea, in 1972. He received the B.S. degree in physics from Kyoungwon University, Korea in 1994 and the M.S. degree in Physics from Hanyang University, Korea in 1996, respectively. He is currently in the course of Ph. D. degree in Physics at Hanyang University. His current research interests include neural network, hidden Markov model, and speech recognition.

He is also a member of the Korean Physical Society, and the Acoustical Society of Korea.

▲Y. Kwon

He was born in Seoul, Korea in 1961. He received in B.S. degree in Mathematics from Hanyang University, in Seoul, Korea, and his M. S. and Ph.D. degrees in Physics from the University of Rochester at Rochester, NewYork, in 1986 and 1987, respectively.

Since 1995, he has been with Dep. of Physics, Hanyang University, Ansan, Korea. He is now an associate professor and his current research interests include mathematical physics, geometry, artificial intelligence.

He is also a member of the Korean Physical Society.

▲K. S. Lee

He was born in Yong-in, Kyounggi-do, Seoul, Korea in 1945. He received in B.S., M.S., and Ph.D. degrees in Physics from Hanyang University, in Seoul, Korea, in 1969, 1977, and 1985, respectively.

Since 1981, he has been with Dep. of Physics, Hanyang University, Ansan, Korea. He is now a professor and his current research interests include noise/vibration, speech analysis, and speech recognition.

He is also a member of the Korean Physical Society, and the Acoustical Society of Korea.