REVIEW OF TDNN (TIME DELAY NEURAL NETWORK) ARCHITECTURES FOR SPEECH RECOGNITION

Masahide Sugiyama[†], *Hidehumi Sawai*[†] *and Alexander H. Waibel*[‡]

[†]ATR Interpreting Telephony Research Laboratories
Sanpeidani, Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

[‡] School of Computer Science, CMU, Pittsburgh, PA, 15213, U.S.A.

ABSTRACT

In this review, the TDNN architecture for speech recognition is described and its recognition performance for Japanese phonemes and phrases is explained. In comparative studies, it is shown that the TDNN yields superior phoneme recognition performance. The TDNN optimized for phoneme recognition, however, does not necessarily result in optimized word or phrase recognition performance, as overfitting to the specific phoneme data or recording conditions may occur. Care must therefore be taken to achieve robust integration, and several studies toward this goal will be reported.

1 Introduction

Recently Neural Network Modeling has been widely applied to various pattern recognition fields. Since one the of authors proposed a new architecture of the neural network model for speech recognition, TDNN (Time Delay Neural Network)[1], in 1987, it has been shown that neural network models have high performance for speech recognition. The great success of the TDNN[2] encouraged many speech researchers to concentrate on this approach. In this paper, the authors review the TDNN architectures and the recognition performance for Japanese phonemes and phrases, and describe the current topics for the neural network approach to speech recognition.

2 TDNN Architecture

The basic TDNN for $/b, d, g/$ phoneme recognition shown in Fig.1 is the four-layer feed-forward neural network model with the property: **Shift invariant connection**. In the input layer, the time series of mel-scaled 16-channel FFT spectrum patterns are fed into. One pattern contains 15 vectors (frames). Thus the input layer has 15 × 16 units. Three frames are connected to one frame in the first hidden layer. The first hidden layer has 8 × 13 units, and every 5 frames are connected to one frame in the second layer. The second hidden layer contains 3 × 9 units with the 9 frames connected to the output layer. The number of connection parameters in the input layer is 16 × 3 × 1 × 8, because of the **shift invariant connection**, that is, the shifted connection parameters in the following frames have the same values. The number of connection parameters between the first and second hidden layers, and between the second hidden layer and the output layer are 5 × 8 × 1 and 3 × 9, respectively. The total number of connection parameters is 6233. It is called **tied connection** and **shift invariant**. This structure has two advantages: reduction of the total number of independent connection parameters and invariance under shifts in time. Generally speaking, the increase in independent connection parameters requires the the amount of the training data. The first advantage helps to solve this problem. When the input pattern is well located for the recognizer, the recognition performance seems adequate. When the input pattern is not well located, it is not easy to recognize. **Shift invariance** helps to ensure successful recognition despite possible misalignments of the pattern in time. [1] This TDNN is trained using Back Propagation algorithm with word uttered speech. The number of training samples for each phoneme category is about 200.

Fig.2 shows a modular type TDNN system for recognizing 18 Japanese consonant, 5 vowels and silence. Each module is constructed and trained for the corresponding phoneme category. After separate training, the whole network is trained again. Recently, this large-scale network is directly trained without modular training using efficient neural network training software[8]. This network has a module corresponding to a consonant category identifier, and a vowel identifier and a silence/not silence identifier. Each module has the basic TDNN architecture shown in Fig.1. The total number of connection parameters is 87051.

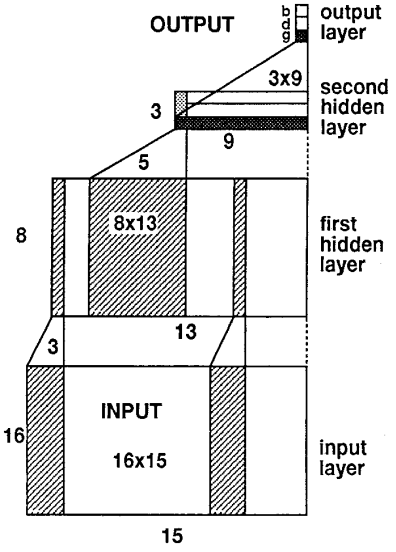

Figure 1: TDNN Architecture

[1] This property has now also been exploited in tasks other than speech recognition, such as optical character recognition[3], [4] and natural language processing[5].

CH 3006-4/91/0000 – 0582 $1.00 © IEEE

Fig.3 shows the TDNN output firing pattern for a Japanese word, (*/ikioi/*). The speech analysis specification is shown in Table 1. Speech is uttered by one male speaker. 15 frames with a length of 150 ms is an input pattern to the TDNN. Input speech is sampled at 12kHz and analyzed by 256-point FFT. Considering human hearing characteristics, the FFT spectrum is divided into 16 channels (about 140 Hz - 5400Hz). The upper block in Fig.3 is the time series of mel-scaled 16-channel FFT spectrum patterns. The filled black rectangle indicates the higher intensity part and the gray rectangle indicates the lower part. The size of the rectangle corresponds to the intensity. The firing position for each input pattern is aligned to the center of 15 frames for the TDNN. The lower block is the firing pattern for 23 Japanese phonemes $/b, d, g, p, t, k, m, n, N, s, sh, h, z, ch, ts, r, w, y, a, i, u, e, o/$ and silence $/Q/$. The size of the rectangle indicates the strength of firing. The medium part shows the phoneme labels and segmentation boundaries given by expert labelers. This figure shows that TDNN firing pattern corresponds to each phoneme. The first firing is silence and the following firing is vowel $/i/$. Next is the closure (silence part) of unvoiced stop $/k/$ and consonant $/k/$, and so on. It seems to be easier to read the TDNN firing pattern than to read the FFT spectrogram series. It can be seen that most of the phonemes can be detected easily.

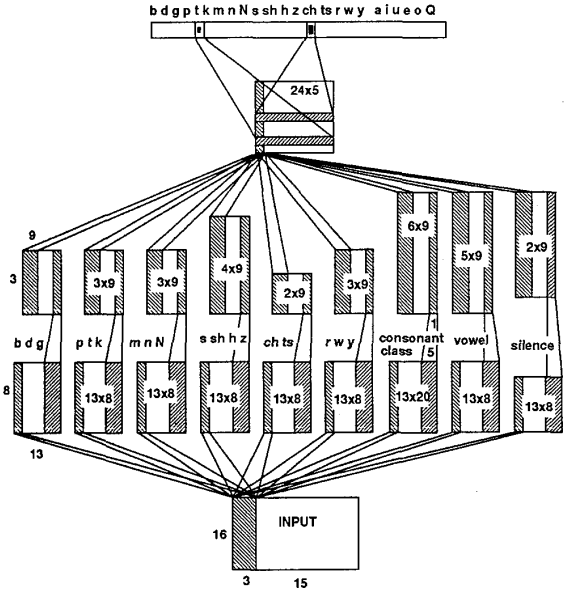

Figure 2: The Modular TDNN 18 Consonant, 5 Vowel and Silence Recognition Architecture

Table 1: Speech Analysis Specification in a TDNN Recognition System

speaker	1 male
sampling frequency	12kHz
windowing	256 point Hamming window
frame rate	10ms
feature parameter	16 channel FFT mel-scaled spectrum 15 frames
power normalization	normalized to lie between -1.0 and +1.0 with the average at 0.0

Considering the TDNN firing pattern, it seems that TDNN is applicable to word and continuous speech recognition. Fig.4 is the architecture of the TDNN-LR continuous speech recognition system. The linguistic information, which is written in a context-free grammar, is converted into an LR parsing table, and the LR parser predicts the next phoneme candidates using the previous phoneme recognition results. Accordingly by using the LR parser, linguistic hypotheses are generated and the output firing pattern produced by the TDNN is verified. Then the hypothesis with the highest output score is recognized. To calculate the total score for input speech, the dynamic time warping technique is applied and the summation of the log activations along the best path is computed.

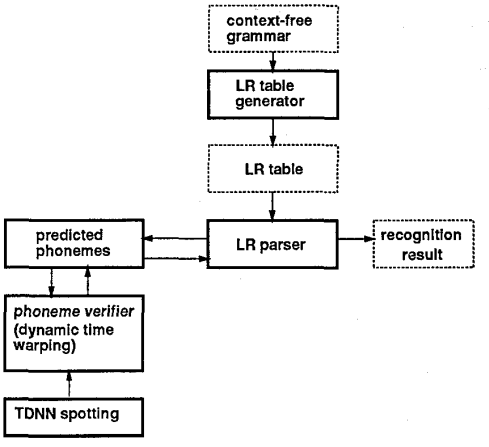

Figure 4: TDNN-LR Speech Recognition System

3 Speech Recognition Performance

Based on the previous TDNN phoneme firing performance, Japanese phoneme recognition and phrase recognition experiments have been carried out. Table 2 shows recognition performance of 18 Japanese consonant using several speech recognition algorithms[6]. HMM (Hidden Markov Model) is the most popular recognition technique for speech and most speech recognition systems have been built based on this technique. In HMM technique, speech is modeled as probabilistic events trained from sufficient speech data. This table shows that the TDNN provides higher recognition performance than the traditional discrete HMM.

Table 2: Japanese Phoneme Recognition Rates (18 consonants)

methods	rank	word	short phrase	continuous
		5.7 mora/s	7.1 mora/s	9.6 mora/s
Discrete HMM	$= 1$	93.1	81.4	71.6
	≤ 3	99.5	95.9	92.4
Continuous HMM	$= 1$	98.1	79.7	66.6
	≤ 3	99.8	94.6	86.9
TDNN	$= 1$	96.2	76.2	56.6
	≤ 3	99.6	91.5	78.5
LVQ	$= 1$	97.9	81.7	61.6
	≤ 3	99.9	96.8	87.7
LVQ+HMM	$= 1$	97.2	80.6	69.6
	≤ 3	99.6	94.6	89.3
Fuzzy LVQ+HMM	$= 1$	94.4	80.8	74.0
	≤ 3	99.7	96.8	93.3

Considering this encouraging phoneme recognition performance, phrase recognition experiments have been performed[7]. Japanese continuous

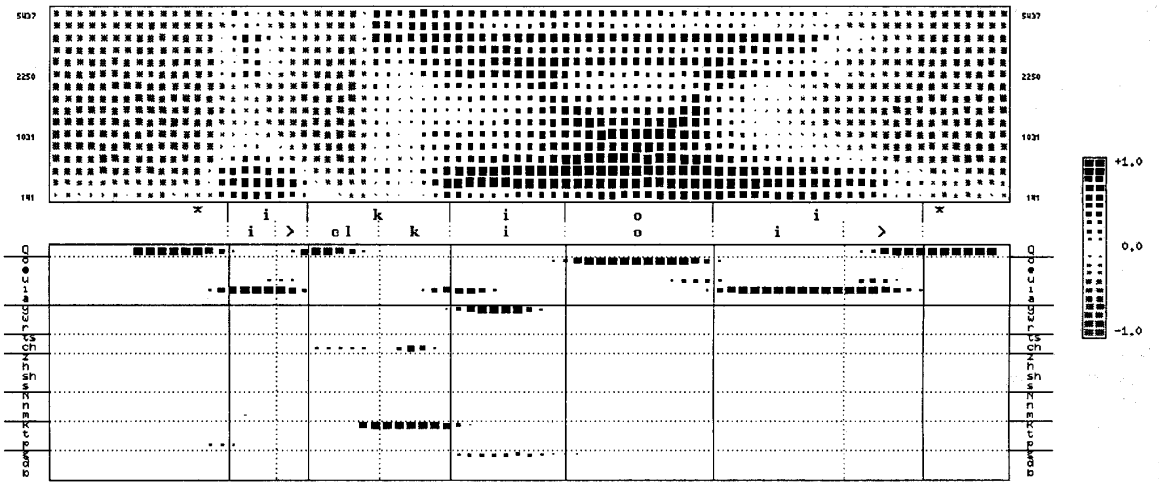

Figure 3: The TDNN Output Firing Pattern for the Word */ikioi/*

speech can be uttered phrase by phrase, and this utterance style is the most promising for current speech recognition technology. Table 3 shows the recognition performance for 279 Japanese phrases. This is the **Conference Registration Task** whose phoneme perplexity is 5.9. After training was performed on phonemes from isolated word data, testing on continuous speech data resulted in lower performance, than for an HMM. This is presumably due to the tighter fit of the TDNN to the specific acoustic phonetic features found in the training data. While this results in better classification performance within speaking modality, it leads to suboptimal results across speaking rates/modalities. Several techniques aimed at improving generalization across modalities, were explored and will be described below. Comparing this result with the HMM-based approach, TDNN does not provide adequate performance. In particular, the performance within the top five candidates is considerably lower than the HMM performance. One reason is that the TDNN which over-trained to word utterance speech data has less robustness than the conventional HMM technique. In the next section, several approaches solving this problem are described.

Table 3: Rates for Japanese Phrase Recognition by TDNN-LR (279 phrases)

rank	phrase recognition rate (%)		
	TDNN	HMM Single VQ	HMM Separate VQ
1	55.0	72.0	83.2
≤ 2	70.1	85.3	93.9
≤ 3	76.6	91.8	96.4
≤ 4	81.3	94.3	97.5
≤ 5	82.7	95.3	98.6

Single VQ: Power + WLR
Separate VQ: Power + WLR + DCEP
beam width in LR search$\leq$ 18(local); $\leq$ 250(global)

The TDNN-LR recognition system was applied to large-vocabulary recognition[9]. The performance for a 5240-word recognition task is 92.6%, and the rates within the top two and top three are 97.6% and 99.1%, respectively. This performance is quite high, and shows that the TDNN architecture has the potential to handle difficult speech recognition tasks.

4 Extended Recent Studies of the TDNN

To increase the robustness of the TDNN for continuous speech recognition, the following improvements are studied. The first study shows how to provide smoothing effects in the training and/or recognition stage. The 2nd is a study of the optimal TDNN architecture. The 3rd study improves the TDNN recognition performance by using phrase uttered speech data. Table 4 shows the effectiveness of the incremental training. The 4th study provides robustness of the TDNN using multiple pairwise decisions (PD-TDNN architecture), which provides pairwise discriminant ability. Table 5 shows that PD-TDNN provides more robust for short phrase and continuous uttered speech than the conventional TDNN. The next two studies (5 and 6) propose new architectures to capture more time information. All studies are speaker-dependent speech recognition tasks, that is, the input speaker is limited to the speaker who was used in training. The last two studies evaluate and propose speaker independent-type speech recognition. The 7th study show that the TDNN based speech recognition (Modular SID) has possibility to give high performance for a speaker-independent recognition task. Table 6 shows that proposed TDNN has high performance for a */b, d, g/* recognition task. The 8th study proposes segment-based (time-frequency) speaker adaptation using a NN. Input speech is filtered by a speaker-mapping NN and recognized using the TDNN designed for a standard speaker. The 9th study proposes adaptive time-delay in TDNN. The 10th study proposes multi-state TDNN for continuous speech recognition. The 11th studies propose the combination architecture with HMM and NN.

1. Smoothing Techniques for Robustness[10], [11], [12].
2. Optimization of Network Architecture[13].
3. Incremental Training for Phrase Recognition[14].
4. Pair Discriminant NN (PD-TDNN)[15].
5. Time-State NN(TSNN)[16].
6. Time-Frequency Shift-Tolerant TDNN[17].
7. TDNN Applied to Speaker-Independent Recognition[18], [19], [20].
8. Speaker Adaptation NN[21], [22].
9. Adaptive Time-Delay NN[23]

10. Multi-State TDNN [24]

11. Combination with HMM and NN [25], [26], [27]

5 Conclusions

In this review, the TDNN architecture for speech recognition is described and its performance for Japanese phonemes and phrases is explained. It is shown that the TDNN yields superior phoneme recognition performance. The TDNN optimized for phoneme recognition, however, does not necessarily result in optimized word or phrase recognition performance, as overfitting to the specific phoneme data or recording conditions may occur. Several extended studies to solve this problem have been cited. In the near future a more robust TDNN will be introduced.

Acknowledgments

We would like to thank Dr.Kurematsu, President of ATR Interpreting Telephony Research Labs., for his encouragement, and Mr. S.Sagayama, Head of the Speech Processing Dep., for his useful discussions, and Dr. K.Shikano, Head of the Speech Processing Dep., NTT Human Interface Lab., for his useful suggestions.

References

[1] A.Waibel, Phoneme Recognition Using Time-Delay Neural Networks, Report of Speech Committee, SP87-100, pp.19-24 (Dec. 1987).

[2] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, K.J.Lang, Phoneme Recognition Using Time-Delay Neural Networks, Trans., ASSP-37, No.3, pp.328-339 (Mar. 1989).

[3] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard and L.D. Jackel, Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation , Vol.1, no.4 (1990).

[4] I. Guyon, P. Albrecht, Y. LeCun, J. Denker and W. Hubbard, Design of a Neural Network Character Recognizer, Pattern Recognition (1990) (submitted for publication. Report available from AT&T Bell Laboratories, Holmdel, NJ).

[5] A. Jain and A.H. Waibel, Robust Connectionist Parsing of Spoken Language, ICASSP90, 40.S11.8 (Apr. 1990)

[6] Y.Minami, T.Hanazawa, H.Iwamida, E.McDermott, K.Shikano, S.Katagiri, M.Nakagawa, On the Robustness of HMM and ANN Speech Recognition Algorithms, Proc. of ICSLP90, 31.3, pp.1345-1348 (Nov. 1990).

[7] Y.Minami, M.Miyatake, H.Sawai, K.Shikano, Continuous Speech Recognition Using TDNN Phoneme Spotting and Generalized LR Parser, Proc. ASJ, 3-1-11, pp.97-98 (Oct. 1989) (in Japanese).

[8] P.Haffner, A.Waibel, K.Shikano, Fast Back-Propagation Learning Methods for Neural Networks in Speech, Rec. Fall Meet. Acoust. Soc. Jpn. 2-P-1, pp.203-204 (Oct. 1988).

[9] H.Sawai, The TDNN-LR Large-Vocabulary and Continuous Speech Recognition System, Proc. of ICSLP90, 31-4 (Nov. 1990).

[10] T.Kawabata, Generalization Effects of k-Neighbor Interpolation Training, Proc. of ASJ spring meeting, 2-P-21, pp.161-162 (Mar. 1990) (in Japanese).

[11] Y.Minami, S.Tamura, H.Sawai, K.Shikano, Output Smoothing for TDNN Using Information on Input Vectors Neighborhood in Input Layer, Hidden Layer, Proc. ASJ, 1-3-18, pp.35-36 (Mar. 1990) (in Japanese).

[12] Y.Komori, A.Waibel, S.Sagayama, A New Fuzzy Training Method for Phoneme Identification Neural Networks, Proc. of ASJ, 1-5-15 pp.33-34 (Mar. 1991) (in Japanese).

[13] Y.Minami, M.Nakagawa, H.Sawai, Comparison of Performance in Various TDNN Architectures, Proc. ASJ spring meeting, 1-5-5, pp.13-14 (Mar. 1991) (in Japanese).

[14] H.Sawai, TDNN-LR Continuous Speech Recognition System Using Adaptive Incremental TDNN Training, Proc. of ICASSP91, 8.S2.4 (May 1991).

[15] J.Takami, S.Sagayama, A Pairwise Discriminant Approach to Robust Phoneme Recognition by Time-Delay Neural Networks, Proc. of ICASSP91, 8.S2.13 (May 1991).

[16] Y.Komori, Time-State Neural Networks (TSNN) for Phoneme Identification Considering Temporal Structure of Phonemic features, Proc. of ICASSP91, 8.S2.22 (May 1991).

[17] H.Sawai, Frequency-Time-Shift-Invariant Time-Delay Neural Networks for Robust Continuous Speech recognition, Proc. of ICASSP91, 8.S2.1 (May 1991).

[18] J.Hampshire, A.H.Waibel, The Meta-Pi Network: Connectionist Rapid Adaptation for High-Performance Multi-Speaker Phoneme Recognition, Proc. of ICASSP90, 12.S3.9, pp.164-168 (Apr. 1990).

[19] S.Nakamura, H.Sawai, A Preliminary Study on Neural Network Architectures for Speaker Independent Phoneme Recognition, Technical Report of Speech Committee, SP90-61, pp.33-40 (Dec. 1990) (in Japanese).

[20] H.Sawai, S.Nakamura, K.Fukuzawa, M.Sugiyama, On Connectionist Approaches to Speaker-independent Recognition, Proc. ASJ spring meeting, 1-5-17, pp.37-38 (Mar. 1991).

[21] K.Fukuzawa, H.Sawai, M.Sugiyama, Speaker Adaptation Using Identity Mapping by Neural Networks, Proc. ASJ, 1-8-16, pp.31-32 (Sep. 1990) (in Japanese).

[22] M.Sugiyama, Fukuzawa, Sawai, Sagayama, Unsupervised Training Methods for Set Mappings Using Neural Networks, Proc. ASJ, 2-P-10, pp.149-150 (Sep. 1990) (in Japanese).

[23] U.Bodenhausen, The Tempo Algorithm: Learning in a Neural Network with Adaptive Time-Delays, Proc. of the IJCNN, pp.597-600 (Jan.1990).

[24] P. Haffner and A.H. Waibel, Integrating Time Alignment in a Neural Network for Continuous Alphabet Recognition, Proc. of ICASSP, 8.S2.17 (May 1991).

[25] N.Morgan and H.Bourland, Continuous Speech Recognition using Multilayer Perceptrons with Hidden Markov Models, Proc. ICASSP90, S8.1, pp.413-416 (1990-04).

[26] Les T. Niles and H.F. Silverman, Combining Hidden Markov Model and Neural Network Classifiers, Proc. ICASSP90, S8.2, pp.417-420 (1990-04).

[27] M.A. Franzini, K.F. Lee and A.H. Waibel, Connectionist Viterbi Training: A New Hybrid Method for Continuous Speech Recognition, Proc. of ICASSP90, 26.S8.4 (Apr. 1990).

Table 4: Incremental Training for Phrase Recognition

rank	phrase recognition rate (%)		
	before training	after training	
		100/category	200/category
1	55.0	64.4	65.1
≤ 2	70.1	79.5	78.4
≤ 3	76.6	81.7	87.1
≤ 4	81.3	86.0	88.1
≤ 5	82.7	88.8	88.8

Table 5: 18 Phoneme Recognition Rates Using PD-TDNN

speaking style	recognition rate (%)		
	PD-TDNN		TDNN
	1	≤ 3	1
word	93.9	99.8	96.5
phrase	86.1	97.1	75.3
continuous	76.8	92.9	56.7

Table 6: Speaker Independent /b, d, g/ Recognition Using Several TDNN Architectures, (tested by 4 speakers)

architecture	recognition rate (%)			
	speakers for training			
	6 speakers		12 speakers	
	closed	open	closed	open
Monolithic	93.2	85.1	-	-
Modular	97.7	92.1	97.1	95.5
SID	96.4	80.0	85.8	85.8
Meta-pi	96.9	82.0	95.4	85.9
Modular SID	97.2	89.7	97.3	95.6