A Dynamic Neural Network for Syllable Recognition

Lin Zhong, Yuanyuan Shi and Runsheng Liu
Tsinghua University
Department of Electronic Engineering,
Beijing 100084, P.R.China
zhongl@hannah.ee.tsinghua.edu.cn

Abstract

A dynamic neural network architecture based on the Time-Delay Neural Network and the Convolutional Neural Network is originated. The dynamic network achieves much better performance than those of MLP and TDNN when dealing with syllable recognition. Such performance is also comparable to that of the more popular HMM method.

1. Introduction

Every Chinese Character is pronounced as an individual syllable. And one of the most important ways to achieve large vocabulary Mandarin speech recognition is to recognize the syllables [4]. Our work is aimed at achieving all syllables recognition by means of neural networks. TDNN[6] was originated to deal with the dynamic characteristics of speech and has proved to be some successful in speech recognition based on phonemes. However, it appears to be inept when we apply it to syllable recognition because of its very limited feature optimization and time integration abilities.

To improve the feature optimization ability, we substitute a CNN-like connection [3] for the standard TDNN-like connection from the input layer to the hidden layer. The CNN-like connection enables the network to further transform the speech feature into a feature space which permits easier classification by the upper layer. The CNN-like connection is adjusted through training. The network architecture is shown in figure 1. Each neuron on a lamination only takes *fwin* features from the $ft^{(0)}$ speech features of a speech frame as input. And a neuron altogether takes features from successive $wl^{(0)}$ frames of the $fr^{(0)}$ frames of an utterance. That is, a hidden neuron takes a rectangular field in the input frame-feature plane . Different neurons on the same lamination share the same parameters(weights and bias) but take the rectangular field as input at different position on the input frame-feature plane.

In section 2, details about the network are presented. Then section 3 offers the experiment results when the network is applied to syllable recognition. And our conclusion and future work are briefed in section 4.

2. The Dynamic Network

2.1 Dynamic Network Input

To further the dynamic characteristics of the network, it is permitted to take in the whole syllable no matter how long it is, instead of a fixed input length that the TDNN and CNN suffer. The network output after integration is normalized by the length of the utterance. Then, the new network is dynamic in the number of speech frames it takes in and the number of hidden layer neurons thus activated. Therefore, the new network obviates the notorious segmentation before classification and embraces the whole syllable to facilitate discrimination between confusing syllables. And it's worth noting that the dynamic network permits real time speech input and recognition.

2.2 Integration

TDNN's based on phonemes integrate the output layer by summing or squaring and summing. For the input speech length is constant, there is no need for normalization. On the contrary, a normalizing factor, $fr^{(2)}$, the number of frames in the output layer, is introduced in our network, i.e. $o_j = \frac{1}{fr^{(2)}} \sum_{i=1}^{fr^{(2)}} o_{ij}^{(2)}$ or $o_j = \frac{1}{fr^{(2)}} \sum_{i=1}^{fr^{(2)}} o_{ij}^{(2)^2}$, where o_j is the output after integration and $o_{ij}^{(2)}$ is the output of the output layer, *j* denotes the speech class, and *i* denotes the frame in the output layer. Note that $fr^{(2)}$ is different for different utterances. Moreover, we have found that a new way for integration is superior to them. That is $o_j = \left[\prod_{i=1}^{fr^{(2)}} o_{ij}^{(2)} \right]^{\frac{1}{fr^{(2)}}}$. Experiments have showed the new integrating strategy will ensure much quicker training and much better generalization. And it is also verified by theoretical analyses[8]. Hence we adopt the geometric mean as the integrating strategy .

2.3 Steepest Gradient Descent Training

The network is trained by an algorithm based on steepest gradient descent and error back propagation instead of the widely adopted back-propagating and averaging algorithm [6]. In the training process, the shared parameters are treated as one parameter and the updating quantity is calculated directly by GD and EBP. The following shows how to calculate the updating quantity for the connection from the input to the hidden layer.

The network input: $I_{ij}^{(0)}$ is the *j*th feature of the *i*th speech frame ,where *i*=1,2,…, $fr^{(0)}$ and *j*=1,2,…, $ft^{(0)}$.

The hidden layer input: $I_{ijt}^{(1)} = \sum_{k=1}^{wl^{(0)}} \sum_{l=1}^{fwin} w_{kjl}^{(1)} I_{i+k-1,t+l-1}^{(0)}$ and the hidden layer output: $o_{ijt}^{(1)} = sigmoid\ (I_{ijt}^{(1)})$,where *i*=1,2,…, $fr^{(1)}$;*j*=1,2,…, $hi^{(1)}$;and *t*=1,2,…, $ft^{(1)}$.

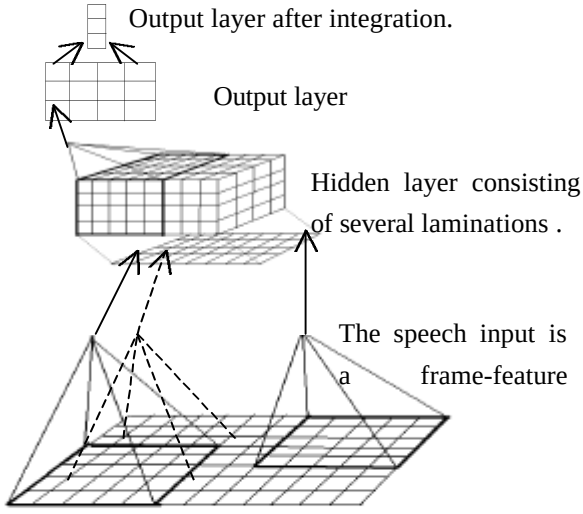

Figure 1.

The output layer input: $I_{ij}^{(2)} = \sum_{k=1}^{wl^{(1)}} \sum_{l=1}^{hi^{(1)}} \sum_{t=1}^{ft^{(1)}} w_{kjlt}^{(2)} o_{i+k-1,l,t}^{(1)}$

where *i*=1,2,…, $fr^{(2)}$;*j*=1,2,…, *c*. $fr^{(0)}$, $fr^{(1)}$ and $fr^{(2)}$ are the number of frames in the input, hidden and output layer, respectively. $ft^{(0)}$ and $ft^{(1)}$ are the number of features in the input and the hidden layer, respectively. $hi^{(1)}$ is the number of laminations in the hidden layer. And *c* is the number of speech classes. The updating quantity can be calculated as $-a \frac{\partial E}{\partial w_{kjl}^{(1)}}$ with a being the learning factor and *E* being the relative entropy instead of mean square error.

$$\frac{\partial E}{\partial w_{kjl}^{(1)}} = \sum_{i=1}^{fr^{(1)}} \sum_{t=1}^{ft^{(1)}} \frac{\partial E}{\partial I_{ijt}^{(1)}} \frac{\partial I_{ijt}^{(1)}}{\partial w_{kjl}^{(1)}} = -\sum_{i=1}^{fr^{(1)}} \sum_{t=1}^{ft^{(1)}} d_{ijt}^{(1)} I_{i+k-1,t+l-1}^{(0)}$$

where *k*=1,2,…, $wl^{(0)}$;*j*=1,2,…, $hi^{(1)}$;*l*=1,2,…,*fwin* ,with

$$d_{ijt}^{(1)} = -\frac{\partial E}{\partial I_{ijt}^{(1)}} = -\frac{\partial E}{\partial o_{ijt}^{(1)}} \frac{\partial o_{ijt}^{(1)}}{\partial I_{ijt}^{(1)}}$$

$$= -\frac{\partial o_{ijt}^{(1)}}{\partial I_{ijt}^{(1)}} \sum_{m=\max(1,i-wl^{(1)}+1)}^{\min(i,fr^{(2)})} \sum_{n=1}^{c} \frac{\partial E}{\partial I_{mn}^{(2)}} \frac{\partial I_{mn}^{(2)}}{\partial o_{ijt}^{(1)}}$$

$$= \left(1 - o_{ijt}^{(1)}\right) o_{ijt}^{(1)} \sum_{m=\max(1,i-wl^{(1)}+1)}^{\min(i,fr^{(2)})} \sum_{n=1}^{c} \left[d_{mn}^{(2)} w_{i-m,n,j,t}^{(2)} \right]$$

where i=1,2,…, $fr^{(1)}$;j=1,2,…, $hi^{(1)}$;t=1,2,…, $ft^{(1)}$.

2.4 Training Strategies

To accelerate training, researches are conducted to find out several strategies which successful save time in training. These strategies include gradually increasing training samples, gradually reducing the target error at which training is terminated, and arranging the training samples in a sequence that accelerates training, etc. [8].

3.Experimental Results

Mandarin digit speech recognition [1], a very popular task, is adopted to test the dynamic network when applied to syllable recognition. The Mandarin digits consist of nearly all kinds of syllables and there is remarkable confusion, as showed in table 1. Mandarin digit speech recognition is then considered exemplar for syllable recognition and is the first step toward all syllables recognition.

Table 1

DIGIT	0	1	2	3	4
Pronunciation	[ling]	[yao]	[er]	[san]	[si]

DIGIT	5	6	7	8	9
Pronunciation	[wu]	[liu]	[qi]	[ba]	[jiu]

Table 2 summarizes the results of our experiments on speaker dependent Mandarin speech digit recognition:

Table 2

NEURAL NETWORKS (SPEECH FEATURE USED)	ACCURACY (%) WITHOUT REJECTION	NUMBER OF FREE PARAMETERS
MLP(12MFCC)	95.7	8,302
TDNN(12MFCC)	92.9	1,342
	88.7	4,006*
DNN(12LPCC)	93.6	3,492
DNN(12MFCC)	96.0	3,492
DNN SYSTEM (12MFCC AND 8LPCC)	97.7	5,674

*The number of free parameters is increased by adding neurons in the hidden layer.

The TDNN that we tested is different from the dynamic network only in the connection from input layer to hidden layer. The structural details of both have been optimized through experiments. The dynamic network system consists of two subnets which are dynamic networks with slightly different structural details, especially, fewer laminations. The subnets are trained separately using the same training speech samples but different speech features. When recognition, the outputs of two subnets are combined by multiplying. The rationale to use both LPCC's and MFCC's is that LPCC's are based on the human articulatory model while MFCC's on the auditory model and these two kinds of speech features are mutual complementary to some extent. The dynamic network combines LPCC's and MFCC's with ease. All the networks are trained with 3 tokens for each of the 10 monosyllabic digits by the speaker and tested by other tokens by the same speaker. 5 speakers are tested and the accuracy is averaged. In all the experiments conducted for this paper, silences of utterances are removed by an automatic end point detecting algorithm. The not so outstanding performance is due to a lot of end point detecting errors. The performance also demonstrates the dynamic network's robustness to such errors.

Figure 2 and table 3 present the results of experiments conducted on the speaker independent Mandarin digit speech recognition. Increasing the number of laminations of the hidden layer boosts the network performance, which is obvious from Figure 2:

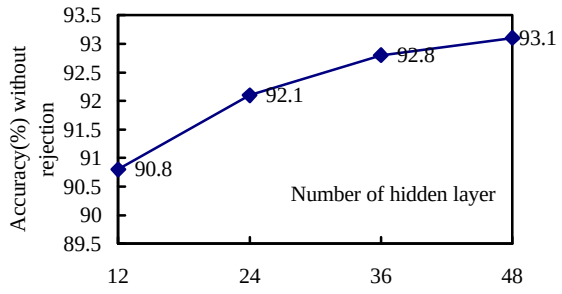

Figure 2.

The networks in the Figure 2 are trained by one token for each of the 10 digits by every one of the 24 different males, and are tested using utterances by males out of the training set. The feature used is 12MFCC. It should be noted that increasing hidden laminations directly increase the network free parameters.

Table 3 shows that increasing training samples, i.e., using utterances by more people, will improve the network performance too, which accords with the intuition. It also demonstrates the dynamic network is superior to MLP[9] and TDNN when applied to syllable recognition.

Table 3.

NEURAL NETWORKS (SPEECH FEATURE USED)	ACCURACY WITHOUT REJECTION(%) Number of Training tokens(number of training speakers×10)			NUMBER OF FREE PARAMETRS
	24×10	30×10	40×10	
MLP(12MFCC)	87.7	88.2		6,920
TDNN (12MFCC)	75.4	76.2		2,098
DNN(12MFCC)	92.1	92.8	93.0	6,994
DNN SYSTEM (12MFCC AND 8LPCC)	93.8	94.2	94.6	5,684

The best performance, 94.6%, is achieved by the dynamic network system which combines LPCC and MFCC. Such performance is significant for the networks are trained by utterances by only 40 people and each subnet has only 12 hidden laminations. The same speech database is used to test an HMM system. The HMM system, after fine tunings, achieves the accuracy of 95.4% without rejection when trained by utterances by 70 people, much more than those used in training the networks.

4. Conclusion and Future Work

The experiments on Mandarin digit speech recognition have proved the dynamic network's superiority to MLP and TDNN in dealing with syllable recognition. It is also comparable to the more popular HMM method when applied to syllable recognition. Moreover, the dynamic neural network recognizer requires much less storage than recognizers based HMM do, and its simple algorithm architecture is very suitable for hardware fulfillment. We expect further experiments to be conducted on more syllables. The final goal is to achieve all syllables recognition and to integrate the networks into a continuous mandarin speech recognition system. We do not think it is desirable or practical to achieve all about 408 mandarin syllables recognition just using one network. A hierarchical system of networks is necessary as in [5][7].

Reference:

[1]Gu, L. and Liu, Runsheng(1997), "Mandarin Digit Speech Recognition: State of the Art, Difficult Points Analysis and Methods Comparison", *Journal of Circuits and Systems*, In Chinese, vol.2, no.4, pp.32-39.

[2]Lang, K.L. et al (1990), "A Time-delay Neural Network Architecture for Isolated Word recognition", *Neural Networks*, vol.3, pp.23-43.

[3]le Cun, Y. et al (1989), "Handwritten Digit Recognition: Applications of Neural Network Chips and Automatic Learning,", *IEEE Communication Magazine*, November, pp.41-46.

[4]Lee, Y. And Lee, L.-S.(1993), "Continuous Hidden Markov Models Integrating Transitional and Instantaneous Features for Mandarin Syllable Recognition", *Computer Speech &Language,* vol.7, pp.247-263.

[5] Poo, G.S. (1997) , " Large Vocabulary Mandarin Final Recognition Based on Two-level Time-Delay Neural Networks(TLTDNN)," *Speech Communication*, vol.22,pp.17-24.

[6]Waibel, A. et al(1989), "Phoneme Recognition Using Time-delay Neural Networks", *IEEE Trans. ASSP*, vol37,pp.1888-1897.

[7]Waibel, A. et al(1989), "Modularity and scaling in large phonemic neural networks", *IEEE Trans. ASSP*,vol37,pp.328-339.

[8]Zhong, Lin(1998), “On Artificial Neural Networks for Small Vocabulary Chinese Speech Recognition”, *Bachelor’s Thesis*, Tsinghua University .

[9]Zhong, Lin and Liu, Runsheng(1999), “Multilayer Perceptron for Isolated Mandarin Digit Speech Recognition”, accepted by *Journal of Circuits and Systems*, in Chinese.