International Journal of Computer Applications Technology and Research
Volume 4– Issue 6, 475 - 480, 2015, ISSN:- 2319–8656

Performance Analysis of MLP, Modified ERNN and TDNN on Telugu Command Recognition

Smt. P. Prithvi
NIT Warangal
Warangal, India

B. Kishore Kumar
NIT Warangal
Warangal, India

Abstract: In this paper, Telugu speech recognition is implemented using MLP and dynamic neural networks in MATLAB. Ten Telugu commands are the words of interest for recognition. Speech samples are collected for ten Telugu words from 30 different speakers in a noise free environment. Front end processing and LPC feature extraction are applied to raw speech data. Data is divided into training and testing sets. This paper gives different topologies of Artificial Neural Networks are used to investigate the Automated Speech Recognition of Telugu speech. The neural network topologies considered are MLP, Modified ERNN and TDNN. The word models are created by giving training data set as inputs to these networks and trained using backpropagation algorithm. Each neural network is trained to identify and classify the words into the respective word models. The testing data set is used to analyze the performance of the network.

Keywords: Telugu command Recognition, MLP, Modified ERNN, TDNN and Backpropagation.

1. INTRODUCTION

The motive of speech recognition is to design a system that can be used to automate many tasks that previously required hands– on human interaction, such as recognizing simple spoken commands. A variety of methods and tools are available to implement speech recognition for small size vocabulary to voluminous dictionary applications. The simplest task in human machine communication is recognition of isolated words.

The concept of speech recognition started in early 1940s [3], practically the first speech recognition program has came into sight in 1952 at the bell labs, which was about to identify a digit in the clean environment [4], [5]. The work on speech recognition is extended to large vocabularies were used and the statistical analysis is introduced in the speech recognition with a wide range of networks for handling language structures were implemented [16]. The invention of hidden markov model (HMM) and the statistical models together allowed researchers to solve continuous speech recognition problem [3].

In 1990s, the major technologies used were statistical language understanding, acoustic learning and language models, and many methods were implemented for large vocabulary speech understanding systems.

In recent years, the speech recognitions developed for different languages. Syllable based speech recognition has been adopted for various languages in the literature. This had been developed for Japanese [12], Portuguese [13] and many others.

After so much of research, different methods in speech recognition have finally benefiting the users in variety of ways. The main goal of designing a system that acts like an intelligent human. So far, there has been less research on Telugu speech recognition compared to the other languages. So, I was motivated to find out the best recognition method for Telugu speech recognition.

To apply different neural network topologies i.e., MLP, TDNN and modified ERNN to implement the automatic speech recognition to detect the Telugu commands. To compare the accuracy of the different techniques used.

The thesis gives is organized in 5 parts. The part II gives the methodology in which different methods for speech recognition are explained. The part III framed by the result of used methods. The chapter IV framed by the conclusions and the future work. Finally, references are given in the chapter V.

2. METHODOLOGY

The standard architecture of speech recognition system is illustrated in Figure 1. The elements are as follows:

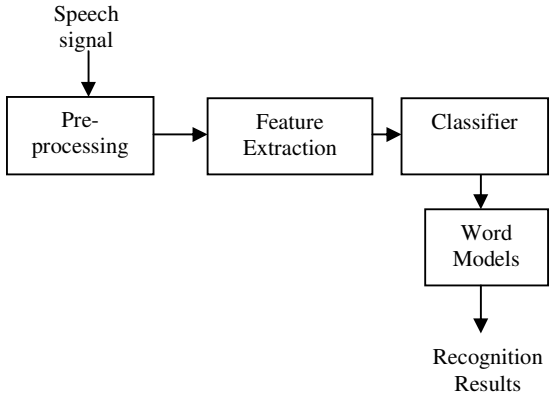

Figure. 1 Recognition system Architecture

Speech signal: acquired speech signal is sampled at a high frequency e.g., 16 KHz over a microphone or 8 KHz over a telephone. This gives the amplitude values of the speech signal over time.

Signal analysis: acquired speech initially transformed, to make the further processing steps easy. Signal analysis techniques which can extract the features from the speech signal.

Linear Predictive Coding (LPC) gives different coefficients of a linear equation that estimate the past history of the acquired speech signal.

Classifier: This block is to classify the different words. The classifiers considered are MLP and Modified ERNN.

Word Models: Word models are created by training the network. The training is done setting target to each word.
In testing phase, the word models are used to recognize the word.

2.1 Speech Signal

The speech signal is acquired in the quit environment i.e., in laboratory. Recording is done with skull microphone.

Sampling rate: 8000 samples/sec
Duration: 2secs
Total no.of samples: 16000 samples.

2.2 Pre - processing

Pre-processing is the first part of the recognizer that the speech signals have to go through. Only the useful speech information in the pre-processing part, will give the good recognition results.
Pre-emphasis is one part in the pre-processing [7]. It will compensate the lip radiation and immanent attenuation of high frequencies in the sampling process. Components at high frequencies are emphasized and components at low frequency are attenuated.
The pre – emphasis will do:

1. The information in the high frequencies of speech is enhanced.
2. The effect of energy decrease in higher frequencies is opposed to enable desired analysis on the complete speech signal spectrum.

The Z-transform representation of pre-emphasis filter is shown below.
$H(z) = 1 - az^{-1}$
Where a is filter coefficient and it is ranges from 0.9 to 1. Typical value of a is 0.95.

2.3 Feature Extraction

Feature extraction, also known as front end processing is performed in both testing and training phase. Feature extraction gives the some sets of numerical vectors called feature vectors which represent the speech signal in numerical manner [7].
LPC furnish an excellent speech signal model which is shown in Figure 2. The partially stationary voiced regions of speech in which the LPC which is all pole model gives a good estimation to the vocal tract spectral envelope. For the non stationary regions of speech, the LPC model is less sensitive than for stationary voiced regions. The implementation of LPC will produce much better separation of source and vocal tract.
The principle behind the LPC model is that a given speech sample at time n, s (n) can be estimated as a linear combination of the past p speech samples. Such that

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots\ldots + a_p s(n-p)$$

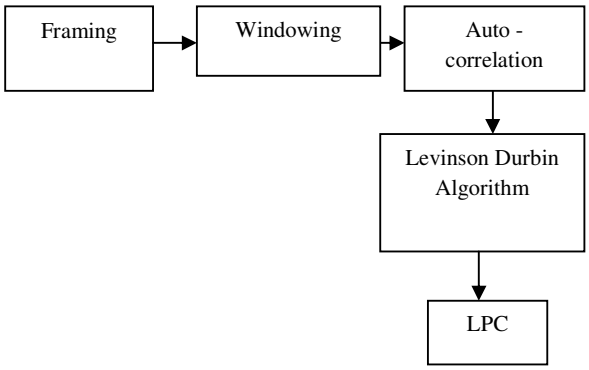

Figure. 2 Feature Extraction

The pre-emphasized speech samples are divided into 30-ms frame having 240 samples for each frame i.e., Number of samples in a Frame = Fs * Frame Length.
In addition, adjacent frames are separated by 80 samples (1/3 of the original frame), with 160 overlapping samples. The number of samples for separation and overlapping depends on frame length. The frame length is the choice according to the sampling rate. The larger the sampling rate, larger the frame length.
The windowing technique for each frame is product of the impulse response and a window function to implement a corresponding filter, which tapers the ideal impulse response.

Where $\alpha = 0.54$ and $\beta = 1-\alpha = 0.46$.
The Auto-correlation LPC block determines the autocorrelation of the input i.e., windowed speech signal. These auto-correlations in turn useful in finding the linear predictor coefficients for the time series in each frame of input signal by reducing the error in the least square sense. Levinson–Durbin algorithm is a procedure in linear algebra to recursively calculates the solution to normal equation. It is computationally efficient to calculate the prediction coefficients.

2.4 Classifier

The classifier is basically to divide words into certain groups. The classifiers used are MLP, TDNN and Modified ERNN.

2.4.1 Multilayer Perceptron (MLP)

An MLP is a fully artificial neural network model that maps group of input data onto a group of appropriate outputs [18]. An MLP consists of more than two layers of nodes in a directed graph, with each layer fully connected to the next one. MLP diagram is shown below in Figure 3.

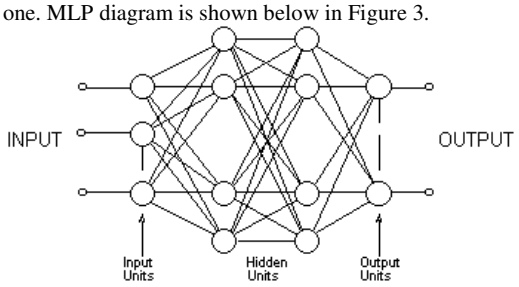

Figure. 3 MLP Architecture

The input to the MLP is 13 LPC coefficients. So, the input layer has 13 nodes. 10 telugu commands are classified using

the MLP. Hence, the output layer consists of 10 nodes. The hidden layers are variable with 15 nodes in each layer.

Back propagation algorithm:

The backpropagation algorithm shown in Figure 4 trains a given MLP for a given set of input patterns with known targets in turn gives the specific classification. When the specific input sample is given to the network, the network determines its output to the specific input pattern. The response is then compared with the known and target output and the mean square error value are calculated. The connection weights of the MLP are adjusted based on the mean square error value.

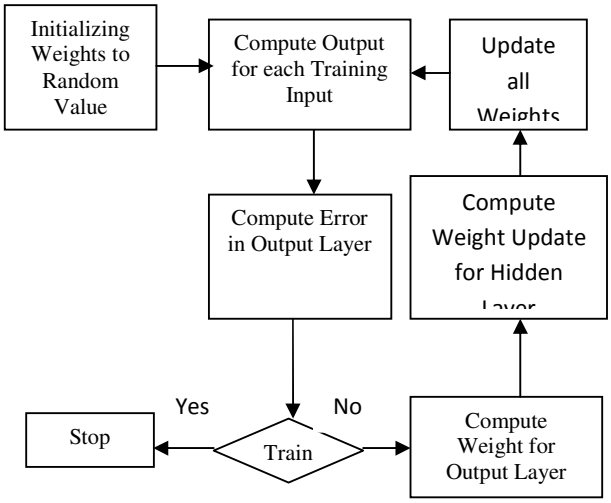

Figure. 4 Backpropagation Algorithm

Figure. 4 Backpropagation Algorithm

For each neuron i in every layer j = 1,2, the output of neuron from input to output layer is calculated using following equation.

Where

Error value is calculated for every neuron i in each layer in backward order j say L, L-1,…,2,1, from output to input layer, followed by weight adjustments. For the output layer, the error value is:

And for hidden layers,

The adjustment of weights is done for each connection from neuron k in layer i-1 to each neuron i in each layer i:

Where β represents the weight adjustment factor. It is used normalize the weights between 0 and 1.

2.4.2 Time delay Neural Network (TDNN)

TDNN architecture [20] was originally implemented for analyze the speech sequence pattern in time series with local time shifts.

TDNN used has the delay of 1 sample period between each input sample. The delayed input can be applied by using the tapped delay line. The tapped delay line architecture is shown below in Figure 5.

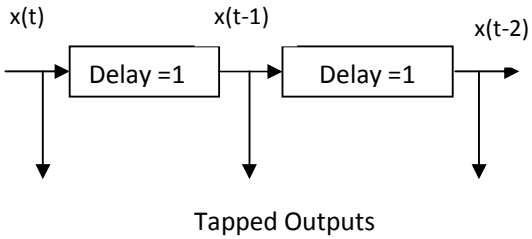

Figure. 5 Tapped Delay Line

TDNN is a multilayer Perceptron with delays in the input shown in Figure 6. Delay in the input is created by using the above delay line. These delayed inputs are given to the MLP. The training phase is same as the MLP.

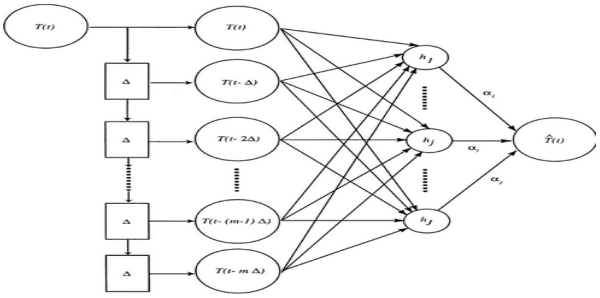

Figure. 6 TDNN Architecture

Delay Used: 1 sample period.

Number of layers used: 1 input layer, 2 and 4 hidden layers and 1 output layer

Number of nodes used: 13 input nodes, 15 nodes for every hidden layer and 10 output layer..

Activation function: hidden layers use sigmoid function and output layer use linear function as activation function.

Learning algorithm used: Backpropagation algorithm

2.4.3 Modified Elman Recurrent Neural Network (MERNN)

The Elman Recurrent Neural Network (ERNN) is the recurrent neural network [23] which had only two layers, and used a sigmoid activation function for the hidden layer and a linear activation function for the output layer.

The Modified ERNN is the generalized model of Elman network to have an arbitrary number of layers and to have arbitrary activation functions in each layer. The Modified ERNN uses same gradient-based backpropagation algorithm used in the multilayer Perceptron. The following Figure 7 illustrates a two-layer Modified ERNN.

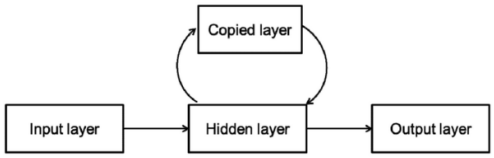

Figure. 7 MERNN Block Diagram

The backpropagation algorithm adds the delay line in the measuring the error during training of the input.

Delay Used: 1 sample period.

Number of layers used: 1 input layer, 2 hidden layers and 1 output layer

Number of nodes used: 13 neurons in input layer, 15 nodes for each hidden layer and 10 neurons in output layer.

Activation function: Sigmoid function for hidden layers and output layer.

Learning algorithm used: Backpropagation algorithm through time [19]

Computation of output:

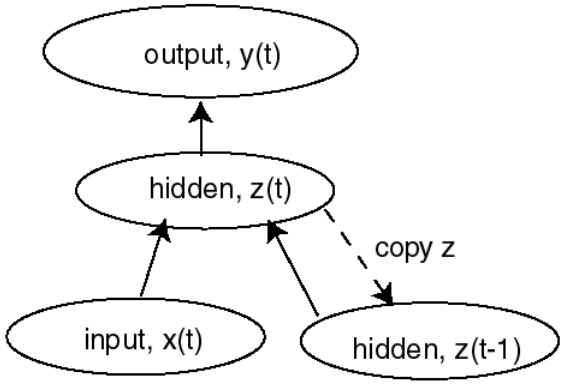

Figure. 8 Computational Architecture of MERNN

3. RESULTS

The recognition with TDNN with 4 – layers and 300 samples of data is shown Figureure. It recognizes the word 'aagu' 8 times out of 10 samples.

```
aagu    aagu    aagu    aagu    aagu    aagu    aagu    aagu    emiti   akkada  akkada
>> outputs

outputs =

  Columns 1 through 8

    0.6331    0.7513    0.7601    0.2782    0.7872    0.8470    0.6131    0.6283
    0.3941    0.0871    0.0776    0.0078    0.0810    0.0739    0.0472    0.0497
    0.1393    0.1617    0.2120    0.2543    0.1383    0.2293    0.2130    0.2259
    0.0687    0.0683    0.0628   -0.0325   -0.0085    0.0079    0.0058    0.0002
    0.1642    0.0783    0.0317    0.0173    0.1067    0.0026    0.0191    0.0175
   -0.1986   -0.1905   -0.1966    0.0211   -0.2037   -0.2295   -0.1199   -0.1317
   -0.0798   -0.0957   -0.0400    0.1940   -0.1193   -0.0379    0.0248    0.0349
    0.0197   -0.0265   -0.0366    0.0739   -0.0186   -0.0631   -0.0137   -0.0106
   -0.0325    0.0655   -0.0095   -0.0019    0.2007    0.0664    0.0886    0.0558
   -0.0847   -0.0561   -0.0133    0.1988   -0.0739   -0.0170    0.0458    0.0533

  Columns 9 through 16

    0.2462    0.0687   -0.0059    0.0020    0.0001    0.0751    0.0347    0.0264
    0.0651    0.8477    1.1000    0.6671    0.8456    0.7381    1.4161    1.1846
    0.0355    0.1282    0.1133   -0.0098    0.0037    0.0823    0.1557    0.0571
   -0.0209    0.0720   -0.0817    0.2690    0.1241    0.1598   -0.3640   -0.1558
    0.3821    0.1574    0.2314    0.2694    0.3262    0.1831    0.2964    0.4089
   -0.0166    0.0055   -0.0080    0.0664    0.0238    0.0225   -0.0969   -0.0728
   -0.0586    0.1118    0.1536   -0.0673   -0.0221    0.0375    0.2464    0.0864
    0.2454    0.0633    0.1095    0.0985    0.1403    0.0647    0.1587    0.2097
```

Figure 9 Recognition result of word 'aagu'

The table-1 and Table-2 shows the recognition rates of speech recognition system implemented using MLP with 300 samples of data in which 200 samples for training and 100 samples for testing. Recognition results are tabulated for different number of layers i.e., 4 – layer and 2 – layer network.

Table 1, Recognition rate Using MLP with 300 samples

Word	No. of Hidden Layers	
	4 Layer	2 Layer
Aagu	100	70
Akkada	100	100
Avunu	100	100
Ela	100	80
Emiti	100	100
Kadu	100	100
Muyu	100	100
Teruchu	100	60
Tesuko	100	100
Vellu	100	100
Total	**100**	**91**

Table 2, Recognition rate Using MLP with 150 samples

Word	No. of Hidden Layers	
	4 Layer	2 Layer
Aagu	100	80
Akkada	100	100
Avunu	100	100
Ela	100	100
Emiti	100	100
Kadu	100	60
Muyu	100	60
Teruchu	100	100
Tesuko	100	100
Vellu	100	80
Total	**100**	**88**

The recognition results of the modified ERNN are shown below. Table-3 and 4 shows the recognition results with 4 – layer and 2 – layer modified ERNN with 300 and 150 samples

Table 3, Recognition rate Using Modified ERNN with 300 samples

Word	No. of Hidden Layers	
	4 Layer	2 Layer
Aagu	100	90
Akkada	100	100
Avunu	90	90
Ela	100	100
Emiti	100	80
Kadu	90	90
Muyu	80	80
Teruchu	100	100
Tesuko	100	100
Vellu	100	100
Total	**96**	**93**

Table 4, Recognition rate Using Modified ERNN with 150 samples

Word	No. of Hidden Layers	
	4 Layer	2 Layer
Aagu	100	80
Akkada	100	80

Avunu	100	100
Ela	100	100
Emiti	80	60
Kadu	80	80
Muyu	60	60
Teruchu	80	100
Tesuko	100	100
Vellu	100	80
Total	**90**	**84**

The recognition results of the TDNN are shown below. Table-V and VI shows the recognition results with 4 – layer and 2 – layer TDNN with 300 samples and 150 samples respectively.

Table 5, Recognition rate Using TDNN with 300 samples

Word	No. of Hidden Layers	
	4 Layer	2 Layer
Aagu	90	80
Akkada	90	80
Avunu	100	90
Ela	100	100
Emiti	100	100
Kadu	90	70
Muyu	100	100
Teruchu	100	80
Tesuko	100	100
Vellu	90	80
Total	**96**	**88**

Table 6, Recognition rate Using TDNN with 150 samples

Word	No. of Layers	
	4 Layer	2 Layer
Aagu	80	80
Akkada	100	100
Avunu	80	60
Ela	80	80
Emiti	80	100
Kadu	60	80
Muyu	80	80
Teruchu	100	80
Tesuko	80	60
Vellu	100	80
Total	**84**	**80**

4. CONCLUSION AND FUTURE SCOPE

Speech recognition system for Telugu command recognition is implemented. This implementation is done by using three neural network architectures i.e., MLP, TDNN and modified ERNN. Recognition accuracy is calculated for different number of hidden layers (n= 4 and 2) with varied number of samples. It is observed that as the number of hidden layers increases, the accuracy of the recognition system is increased for all three networks i.e., modified ERNN, MLP and TDNN for small vocabulary.

Also, the overall recognition rate is good for MLP (100%) with 4 – hidden layers compared to Modified ERNN (96%) and TDNN (96%). The recognition rates with 2 – layer MLP gave 91% accuracy for 300 samples and 88% for 150 samples, where as 2 – layer TDNN gave 80 % for 150 samples and 84% for 300 samples. 2 – layer Modified ERNN gave 93% for 300 samples and 84% for 150 samples. Overall, 4 – layer MLP gave best (100%) for the small vocabulary of size 300 samples.

The work can be extended by implementing the hybrid models along with the artificial neural networks. Dynamic models use an external representation of time. It is also possible to design neural models in which time is internally managed by the network. The accuracy may be improved by using unsupervised learning methods and the reinforcement learning method.

5. REFERENCES

[1] R.Cardin,Y.Normandin and E.Millien,Inter-word coarticulation modeling and MMIE training for improved connected digit recognition,ICASSP,p243-246,1994.

[2] Overview of the Speech Recognition Technology Jianliang Meng,Junwei Zhang,Haoquan Zhao.

[3] Applications Mohamed Atri, Fatma Sayadi, Wajdi Elhamzi, Rached Tourki, “Efficient Hardware/Software Implementation of LPC Algorithm” in Speech Coding, Journal of Signal and Information Processing, 2012, 3, 122-129

[4] "Fundamentals of Speech Recognition". L. Rabiner & B. Juang. 1993.

[5] D. Jurafsky, J. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition", 2000.

[6] Ehsan Khadangi, Alireza Bagheri, “Comparing MLP, SVM and KNN for predicting trust between users in Facebook”, 2013 IEEE.

[7]. Rashmi. M, Urmila. S, M. Thakare “Speech Enhancement Using Pitch Detection Approach For Noisy Environment”, IJSET,Vol.3,No.2, Feb-2011.

[8] Hermansky, H. and M. Pavel, "Psychophysics of speech engineering systems", Invited paper, 13th International Congress on Phonetic Sciences, Stockholm, Sweden, pp. 42-49, 1995.

[9] Malayath, N., H. Hermansky, and A. Kain, "Towards decomposing the sources of variability in speech" , Proc. Eurospeech 97, Rhodos, Greece, 1997.

[10] Hermansky.H., "Modulation spectrum in speech processing", in Signal Analysis and Prediction, Boston 1998.

[11] Burcu Can, Harun Artuner, “A Syllable-Based Turkish Speech Recognition System by Using Time Delay Neural Networks (TDNNs)” IEEE-2013

[12] Y. A. Jun Ogata, “Syllable-based acoustic modelling for japanese spontaneous speech recognition,” in Proceedings of Interspeech, International Symposium on Computer Architecture (ISCA), 2003.

[13] J. a. P. N. Hugo Meinedo, “The use of syllable segmentation information in continuous speech recognition hybrid systems applied to the portuguese language,” in Proceedings of Interspeech, International Symposium on Computer Architecture (ISCA), 2000, pp. 927–930.

[14] Indonesian Speech Recognition System Using Discriminant Feature Extraction – Neural Predictive Coding (DFE-NPC) and Probabilistic Neural Network Untari N. Wisesty, Thee Houw Liong, Adiwijaya 2012,IEEE.

[15] B.H. Juang & Lawrence R. Rabiner , “Automatic Speech Recognition – A Brief History of the Technology Development”, 2004

[16] Ney and A. Paeseler, “Phoneme-based continuous speech recognition results for different language models in the 1000-word spicos system,” Speech Communication, vol. 7, no. 4,pp. 367–374, December 1988.

[17] Stephen Cook “”Speech Recognition HOWTO” Revisionv2.0 April 19, 2002
[18] Saurabh Karsoliya, “Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture”, International Journal of Engineering Trends and Technology-Volume3, Issue6, 2012.
[19] T. Mikolov, M. Karafiat, L. Burget, J. Cernockly, and S. Khudanpur. “Recurrent neural network based language model”. In INTERSPEECH, pages1045-1048, 2010.
[20] Burcu Can, Harun Artuner, “A Syllable-Based Turkish Speech Recognition System by Using Time Delay Neural Networks (TDNNs)”, IEEE – 2013.
[21] Hamdi A. Awad, “A Novel Version of ELMAN Neural Networks for Modeling and controllingMedical Systems” Advances in Neural Networks-Theory and Applications, 2007.
[22] Malay Speech Recognition in Normal and Noise Condition C. Y. Fook, M. Hariharan, Sazali Yaacob, Adom AH , 2012 IEEE.
[23] Diamantino Caseiro, Andrej Ljolje, “Multiple Parallel Hidden Layers And Other Improvements To Recurrent Neural Network Language Modeling”, 2013 IEEE.