ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 07, 2022

Speech Recognition Using Recurrent Neural Network

Shalini Gupta, Shubha Jain, Vijatashwa Awasthi, Amit Sabharwal, Mohit Kumar, Ajay Gaur, Waseed Mohamad Khan
Department of Computer Applications, Axis Institute of Higher Education, Kanpur, Uttar Pradesh, India

Abstract
— Speech recognition (SR) and understanding have been the subject of years of research. The vocalized method of human communication is called speech. A restricted collection of vowel and consonant speech sound units are phonetically combined to form each spoken word. SR is the process by which a computer or software recognizes phrases and words in spoken language and converts them into a computer-readable format. Modern technology has progressed to interpret meaning through the application of its own set of grammatical rules, even if it still uses continuous dictation. In this study, we suggest using the Recurrent Neural Network technique and the Mel Frequency Cepstral Coefficient (MFCC) for Speech Recognition (SR) to create an alphabetical list of words from the CORPUS database.
Index Terms: Principal Component Analysis, Feature Extraction, MFCC, Recurrent Neural Network, Speech Recognition

I. OVERVIEW

Speech is a complex audio signal that is affected by a number of factors, including the speaker's characteristics and the environment [1][3]. Before Automatic Speech Recognition (ASR), it is helpful to identify which audio segments include speech. Speech recognition is the process of converting a speech signal into a series of words [7]. Traditionally, the approach to continuous SR with a big vocabulary is to specify the word sequences and assume a simple probabilistic model of speech creation. Speech is the primary means of interpersonal communication. Speech signals are converted into words by a computer using the automated speech recognition (ASR) technology [6][8].

In this case, the fully recurrent neural network [14] is a Multilayer Perceptron (MLP) [9], and the network is fed back by the inputs shown in figure 1 as well as the previous set of hidden unit activations.

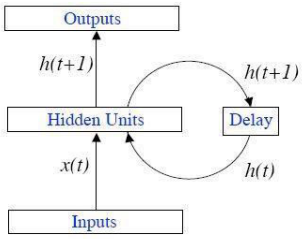

Figure 1: Fully Recurrent Neural Network

Keep in mind that the activations must be updated at each time step and the time t must be discretised. The temporal scale may be in line with how actual neurones function, or for artificial systems, any time step size suitable for the particular issue at hand may be employed. To store a delay unit must be inserted to hold activations until they are handled at the next time step [14].

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 07, 2022

I. RNN- BASED SPEECH RECOGNITION

Figure 2 illustrates the voice recognition system's flow.

We provide methods for enhancing speech recognition through the use of recurrent neural networks. To locate the voice recognition system, there are four distinct phases.

1) Feature extraction
2) Using artificial neural networks to extract features
3) The Language and Acoustic Models
4) Worldwide search methodology

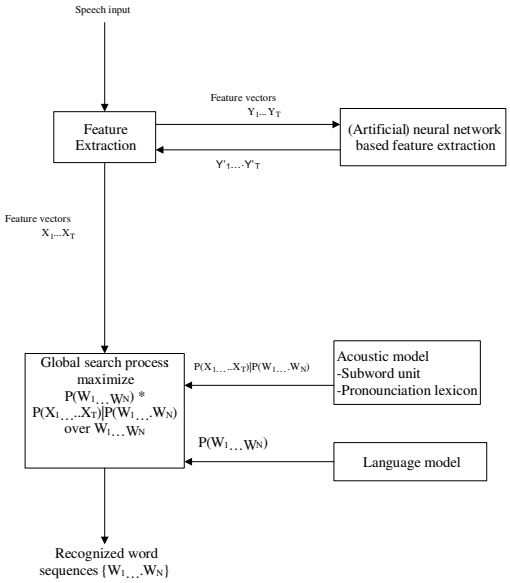

Figure 2: Speech recognition system

III. EXTRACTION OF FEATURES

A number of feature vectors are extracted. It explains how a sound signal is converted into a format that can be utilized in later steps. MFCC is a well-known characteristic that is used to characterize speech signals in addition to PCA. [10][11][12][13]. Technique of calculation Given that MFCC is based on short-term analysis,an vector is calculated for every frame. The MFCC feature extraction process involves several phases.

1) Prior to emphasis
2) Blocking frames
3) Using Hamming Windowing
4) Quick Fourier analysis
5) Mel-scale
6) Transform the discrete cosine
7) Logarithmic energy
8) Delta cestrum

II. LANGUAGE MODEL AND ACOUSTIC MODEL

Language Model: It gives an a priori probability of the word sequence and implicitly covers the language's syntax and semantics. The expression for the language model probability P () under all model assumptions is as follows:

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 07, 2022

$$P(W_1^N) = \prod_{n=1}^{N} p(W_n | W_1^{n-1}) \quad (1)$$

Acoustic Model: The acoustic model is a statistical model that yields the probability P for a series of acoustic characteristics given a word sequence. Syllables, phonemes, or phonemes with context are examples of sub-word models that are used in large vocabulary continuous voice recognition systems.

in place of modelling entire words. A pronunciation lexicon gives the relationship between a series of subword units and entire words mapped out. A hidden Markov model has an observable observation sequence and an unobservable state sequence. Consequently, the probability Pis was expanded by a few (hidden) random variables that represented the model's states [2][4][5]:

$$P(x_1^T | W_1^N) = \sum_{S_1^T : W_1^N} (X_1^T, S_1^T | W_1^N) \quad (2)$$

Worldwide search procedure: The search module of the voice recognizer is its most important part. As seen in Figure, the search encompasses all knowledge sources. determining the word combination for a given feature that maximizes the posteriori probability P vector sequence is the aim of the search module.

$P(w_1 \ldots w_N) * P(x_1..x_T | w_1 \ldots w_N)$ over $w_1 \ldots w_N$ (3)

Then, finally recognized word sequences.

III. IMPLEMENTATION

Grab the audio from the specified CORPUS database. As seen in figure 3, we use an English alphabetic alphabet from A to Z, etc.

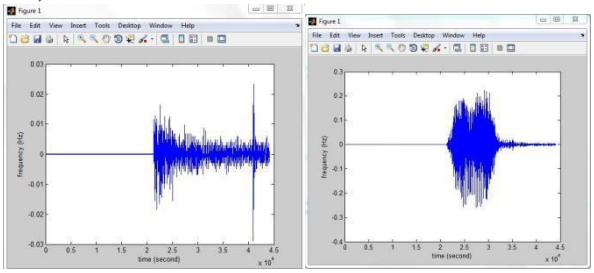

Figure 3: Sound MFCC feature extraction

IV. RESULT ANALYSIS

The regression graph for a single sound for the training, validation, test, and all sets is displayed in Figure 4 below.

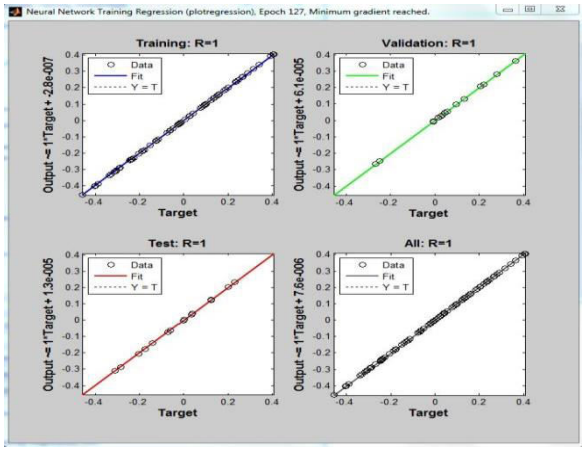

Figure 4: regression analysis

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 07, 2022

A specific data input speech's performance curve is shown in for example, "A," which has the best validation performance at 0.089797 at epoch 55.Figure 5.

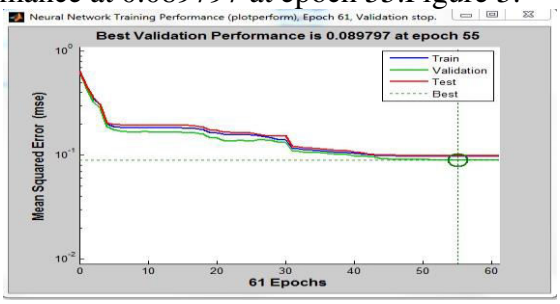

Figure 5: Performance analysis

The following Figure 6 shows the errors analysis.

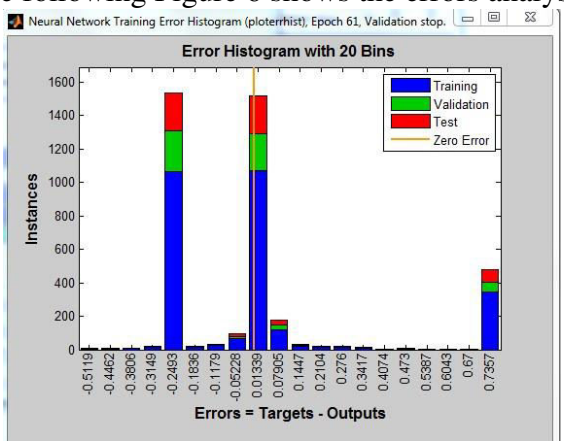

Figure 6: Error analysis

V. CONCLUSION

Recurrent neural networks and the MFCC approach, which we have presented, are utilised to identify speech. Both frequency and unknown noise are reduced with the use of MFCC. We discovered that RNN enhances speech.

In a subsequent analysis, we will use a new sound database that can yield better results and boost performance by lowering the word error rate. Additionally, research is being done to enhance the RNN system that has been suggested by combining system analysis with various modules.

REFERENCES

[1] Thad Hughes and KeirMierle, "Recurrent Neural Networks for Voice Activity Detection," Google inc., USA. IEEE, 2013, pp. 7378-7382.

[2] G.Heigold, V. Vanhoucke, A. Senior, P.Nguyen, M. Ranzato, M. Devin and J. Dean, "Multilingual Acoustic Models Using Distributed Deep Neural Networks," Google Inc., USA. IEEE, 2013, pp. 86198623.

[3] Vincent Vanhoucke, Matthieu Devin, Georg Heigold, "Multiframe Deep Neural Network or Acoustic Modeling," Google, Inc., USA. IEEE, 2013, pp. 7582-7585.

[4] Yik-Cheung Tam, Yun Lei, Jing Zheng and Wen Wang, "ASR Error Detection Using Recurrent Neural Network Language Model and Complementary ASR," 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014, pp. 2331-233 .

[5] Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan "Honza" **Ceremony**, **Sanjeev Khudanpur**, "Extensions of Recurrent Neural Network Languag Model," 2011 IEEE ICASSP, 2011, pp. 5528-5531.

[6] Ms. Vrinda, Mr. **Chandra Shekhar**, "Speech Recognition System for English Language, " International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1, January 2013, pp. 919-922, ISSN 2278-1021.

[7] Sanjivani S. Bhabad, Gajanan K. Kharate, An Overview of Technical Progress in Speech Recognition, International Journal of Advanced Research in Computer Science and Software Engineering Vol. 3, Issue 3, March 2013, pp 488- 497, ISSN 2277 128X.

[8] IlyaSutskever, James Martens, Geoffrey Hinton,"Generating Text with Recurrent Neural Network, " Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011, pp 1017-1024.

[9] Lawrence R. Rabiner, "Applications of Speech Recognition in the area of Telecommunications," AT&T Labs IEEE 1997, pp. 501-510.

[10] Abhishek Thakur, Rajesh Kumar, Naveen Kumar, "Automatic Speech Recognition System for Hindi Utterances with Regional Indian **Accent**: A Review," **IJET** Vol. 4, Issue Spl – 3, April – June 2013, pp. 38-43, ISSN 2230-7109.

[11] Andrew L. Maas, Quoc V. le, Tyler M. O'Neil, **Oriol Vinyals**, Patrick Nguyen, Andrew Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," Google, Inc., USA 2013, pp. 12901293.

[12] M. Sundermeyer, I. Oparin, J.-L. Gauvain, B. Freiberg, R. Schluter, H. Ney, "Comparison of Feedforward and Recurrent Neural Network Language Models," 2013 IEEE ICASSP, 2013, pp.8430-8434.

[13] Vidushi Sharma, **Sachin Rai**, **Anurag Dev**, "A Comprehensive Study of Artificial Neural Networks," IJARCSSE Vol. 2, Issue 10, October 2012, pp. 278284, ISSN 2277 128X.

[14] Simon Haykin, "Neural networks and learning machines," Pearson publications, third edition, 2013, pp.798, ISBN 971-81-203-4000-8.