IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 09, Issue 8, August. 2019, ||Series -II || PP 01-08

A Stratified Approach for Spoken Word and Accent Recognition: Validation & Analysis

Mr. Ashok Shigli[1] Dr. Kunupalli Srinivasa Rao[2]
[1] *Research Scholar, Department of ECE, Rayalaseema University, Kurnool, A.P.,*
[2] *Research Supervisor, Dept. of ECE, Rayalaseema University, Kurnool, A.P.,*

Received 15 August 2019; Accepted 30 August 2019

Abstract: Speech translation is a process of both speech recognition and equivalent phonemic to word translation. Accent is a pattern which differentiates the pronunciation and acoustic features based on the specific language group. The recognition and classification of words with different accent is also challenging problem in speech recognition research. There are many issues which limit the performance of such systems since a word spoken by different persons can have different acoustic properties due to variation in physiological characteristics, emotional status, and cultural background. Speech recognition is a process of identifying phonemes from the speech segment which is affected by the accent of the speaker. Some of the English words spoken by South Indian people whose native language being other than English like Telugu, Kannada, Tamil, Malayalam and Marathi, etc., will have a typical accent pronounced under the influence of their mother-tongue are incorrectly recognized by most of the translating systems. In this paper, an illustrative attempt has been made to effectively increase the efficiency of the system adopting a robust speech and accent recognition methodology. The focus is on automatically identifying the dialect or accent of a speaker given a sample of their speech, and demonstrates how Syllable MFCC, HMM and FO-ANN algorithm, a stratified approach through MATLAB tool can be employed to improve Automatic Speech and Accent Recognition (ASAR).

Keywords: ASR, ANN, FA, FAR, HMM, MFCC, MLP-NN, syllable, VQ

I. INTRODUCTION

Every human has his/her own dialect and perception of his/her accent matters in day to day life right from infant to old age as rightly said by Mangner et al (1974) that “Language is a dialect with an army and navy”. With the advent of human-machine interaction, life along with communication has become easier for free to roam world population. Development of speaker recognition system began in early 1960’s with the exploration into voiceprint analysis. The ensuing years found that detection efficiency of speaker recognition systems gets severely affected by methods applied towards acoustic analysis and acoustic modelling which ensured the development of more robust and reliable method. Added to the problem is Accent Recognition. Speech speak about linguistic information and about the speaker himself. Techniques have been developed by using which the speakers nativity, age, gender etc., are technically determined. The accent analysis techniques incorporate Pre-processing, feature extraction, optimization and classification algorithms to overcome space, bandwidth, transmission rate limitations and retrieval. For classification any of the following techniques are incorporated viz., support vector machine, K-NN approach, Naïve Bayes classification, Decision trees, Genetic algorithms and Neural Networks etc. The performance criteria of speech processing tools are its robust classification accuracy which results in accurate human accent recognition. The approaches so far used and implemented still have some loop holes and needed to be worked minutely. Accent variation does not only stretch out in phonetic characteristics but, also in prosodic characteristics. The robustness of the algorithmic system software is judged by their adaptability to changing environs of age, emotions, speech storing aspect, dimension aspect, training of dataset, various metrics of classification etc. Speech signals largely suffer due to high individual speaker differences, emotion variations and noise disturbances and thus a robust invariant approach is needed for accurate classification. Classification narratives such as signal intermediate representation, minimal feature selection, maximum compression and optimized weights are some of the hierarchical bench marks set for this research work which ultimately help in for exact accent retrieval and classification.

II. PROBLEM AND OBJECTIVE SOLUTION

Around 6,500 languages are spoken in the world but, its diversity makes the world beautiful place to live in. As per *Ethnologue*‘s 2019 edition, published by SIL International of United States, the world is diversified in 90 most spoken languages in which at least 10 million are first language speakers and as far as India is concerned and according to Swedish encyclopedia, National encyklopedin, there are 365 million native speakers in English, 310 million native speakers in Hindi, 76 million speakers in Telugu, 73 million native speakers in Marathi, 70 million in Tamil, 38 million in Kannada and 38 million in Malayalam. The most challenging aspect of Automatic Speech Recognition (ASR) system is to handle the acoustic differences and accent differences among speakers which are geographically based. In era of globalization, every human is allowed to roam freely and there comes the need for accent recognition since after gender accent is seen as a major factor which reduces accuracy of speaker dependent recognition system. Natural language of any region has hundreds of accents which create difference in pronunciation and intonation of speech as a result of which the system accuracy comes down heavily. The major challenge is to understand speech by non-native speakers and vice-versa. Accent is seen as basic source of within world and outside world speaker variability. Accented speech results in phonemes that are not typical of a language which makes speech recognition difficult. Other problem is variation caused due to differences in individual speaker characteristics, emotion variation, age factor variations, noise and physical noise cord disturbances. The solution of the problem lies with designing accent recognition of various language speakers by taking care of inter and intra speaker variability problem. A four-pronged objective strategy is laid down in order to arrive at the solution:

- Pre-processing using DC component removal, Pre-emphasis and framing.
- A novel feature extraction technique using Syllable Mel-Frequency Cepstrum Coefficient (SMFCC) with energy, zero crossing and formant frequency.
- Coding of these features by Hidden Marko Model (HMM) for spoken word recognition.
- A Firefly optimized Artificial neural network classifier (FO-ANN) to evaluate accented word recognition system performance.

III. DESIGN APPROACH

The below figures outline the basic flow, Neural Network with its functions and simulated MLP-NN.

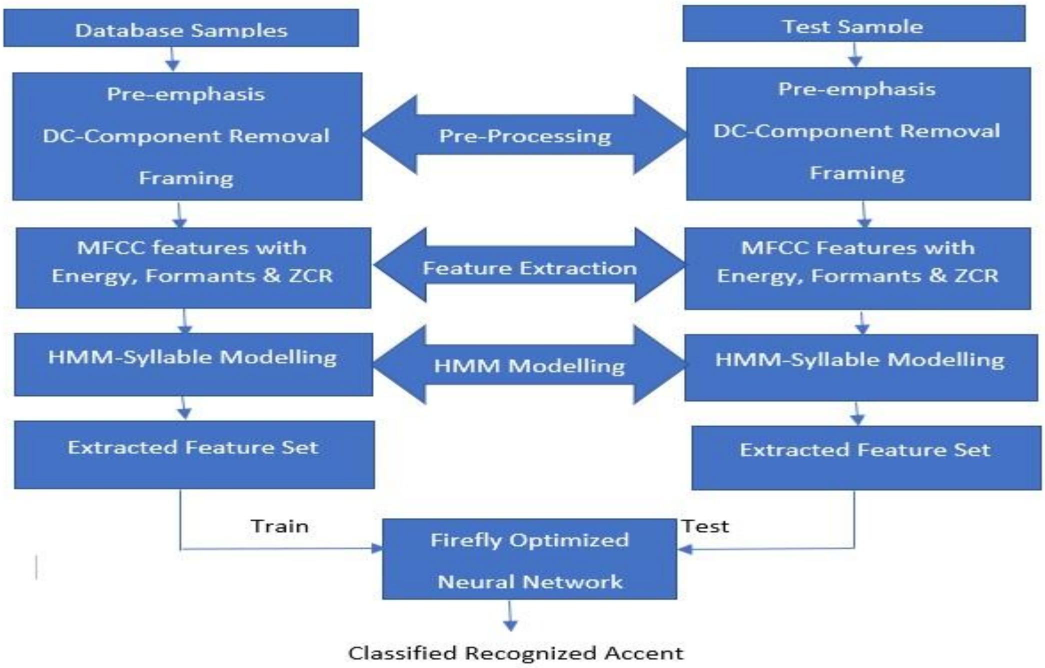


Figure 1: Basic Flow Diagram

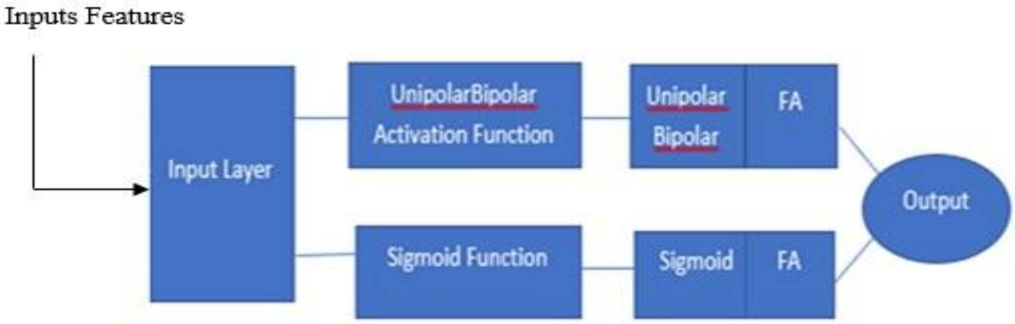


Figure 2: Neural Network with its functions

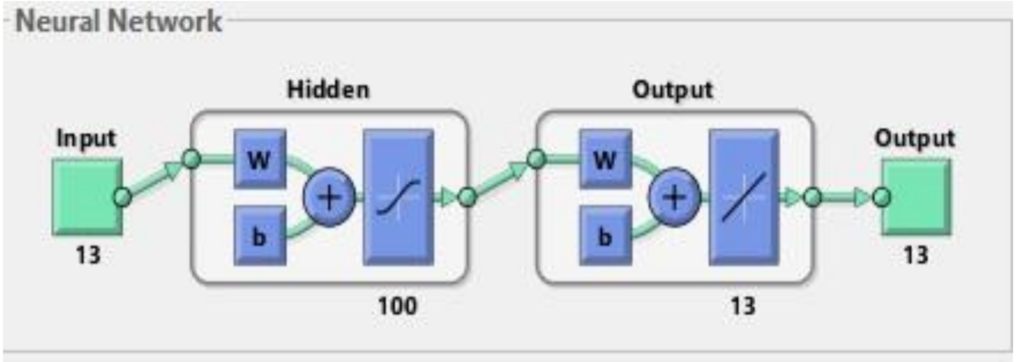


Figure 3: Typical Simulated NN Architecture

The methodology initially involves, pre-processing of the speech signals using pre-emphasis, windowing and syllable segmentation algorithm. MFCC features are then extracted along with the energy, formants and zero crossing rate. The features so gathered are subjected to optimization using Firefly algorithm (FO). These optimized feature patterns are subjected to classification and recognition using artificial Neural Network (ANN). In ANN, the required weights are updated using firefly optimization. The performance of the proposed FO-ANN is evaluated with the performance measures such as Execution time, Precision, Recall, F-measure, Recognition rate and Accuracy.

IV. SYSTEM VALIDATION AND ANALYSIS

For this paper, the suggested approach is considered for validation for 740 words with 50 samples of different languages namely Telugu, Kannada, Tamil, Malayalam and Marathi. 30 Speakers with different age group. Different features like frequency, energy, zero-crossing rate and FFT magnitude variance are only used to evaluate the system performance. The Syllable MFCC features are extracted and are then applied to Firefly Optimized Neural Network for classification employing 25 number of fireflies, absorption coefficient of 0.1, learning rate of 1.2 (up) and 0.5 (down). A typical simulated NN is shown in figure 3 above with one hidden layer and various parameters such as number of epochs 100, number of inputs 13, number of target outputs 15, number of hidden neurons 100 with Rosenbrock error backpropagation algorithm which is given by:

f (x, y) = (1-x) ^2 + 100(y-x^2) ^2

where, x and y are two functions and it has a global minimum at (x, y) = (1, 1) where f (x, y) =0. Validations and performance are subjected to Execution time, Recognition rate, Recall, F-Measure, Precision and Accuracy but, for this paper only execution time and accuracy as considered for system performance validations. The Accuracy parameter is mathematically expressed as:

$$A = \frac{(TP+TN)}{(\,TP+FP+FN+TN)} \qquad (1)$$

where TP and TN denote true positive and true negatives whereas FP and FN denote false positive and false negatives. Accuracy describes percentage of correct recognition. To validate the system performance critical parameters such as background noise, pronunciation under stress, rate of utterances is only considered as they dominate the criteria in assessing system performances.

CASE 1: Evaluation and validation with and without background noise.

1A: Input Word: “Nearest” with no background noise in Telugu Accent.

```
Input Word:
nearest
```

Figure 4: Accented word display

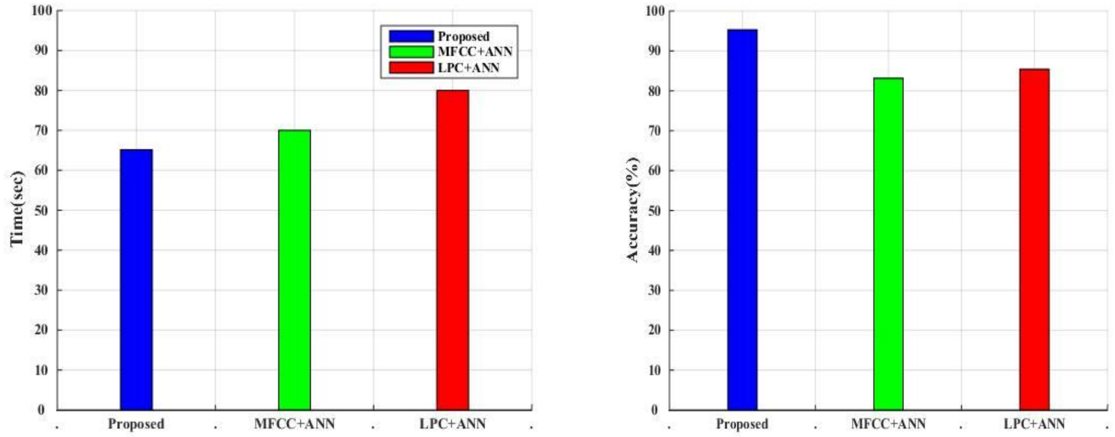


Figure 5: Execution Time and Accuracy Simulated Plots

As can be seen from the simulated plots the execution time without noise for the accented word is 66 sec and accuracy rate is 96 %. The accuracy can be validated as per the simulated data for accented word “Nearest” which shows: TP= 97, TN= 74, FN= 2 and FP= 6. By substituting the above simulated values in the equation 1 for accuracy, the value comes out to be 95.5 %. (≈96 %).
1B: Input Word: “Nearest” with background noise in Kannada Accent.

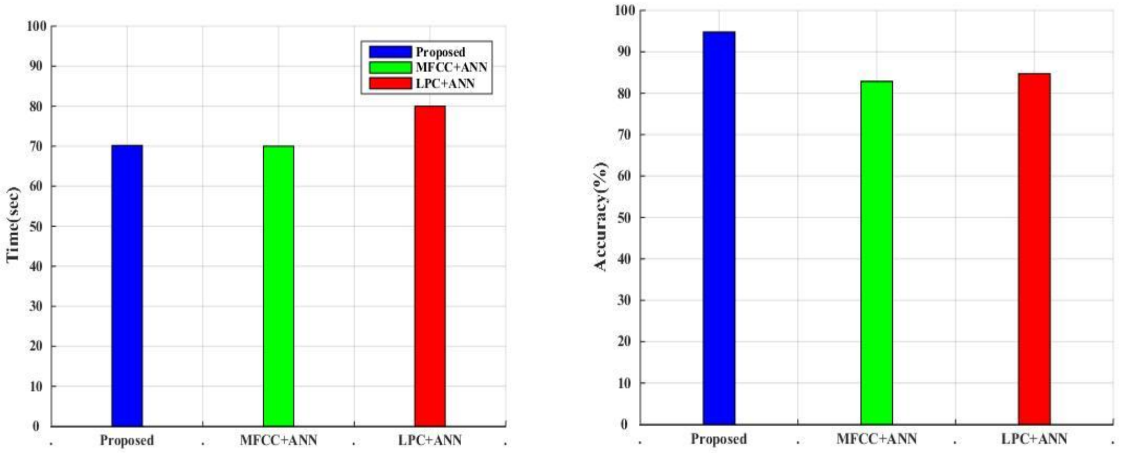
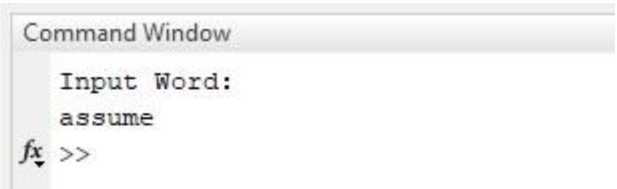


Figure 6: Execution Time and Accuracy Simulated Plots

As can be seen from the plots the execution time for the accented word with noise is 69 sec and accuracy rate is almost same at 96 %. The accuracy can be validated as per the simulated data for accented word “Nearest” which shows: TP= 97, TN= 74, FN= 2 and FP= 6. By substituting the above simulated values in the equation 1 for accuracy, the value comes out to be 96.47 %. (≈96 %). Other than accuracy the other parameters of performance are: Precision 96 %, Recall Rate 97.95%, F-Measure 96.96%.

CASE 2: Evaluation and validation with Stressed accent.
2A: Input Word: “Assume” in Malayalam Accent.



```
Input Word:
assume
>>
```

Figure 7: Accented word display

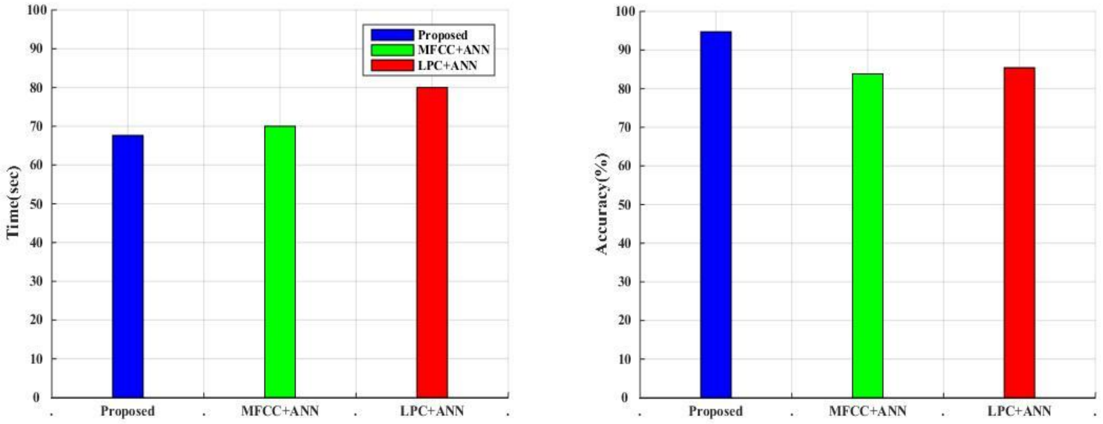


Figure 8: Execution Time and Accuracy Simulated Plots

The execution time with stressed accented word is 67 sec and accuracy rate is 95 %. The accuracy can be validated as per the simulated data for accented word "Assume", TP= 96, TN= 65, FN= 2 and FP= 6 and by substituting the above simulated values in the equation 1 for accuracy, the value comes out to be 95.26 % (≈95 %). Other than accuracy the other parameters of performance are: Precision 94.11 %, Recall Rate 97.95%, F-Measure 96%.

2B: Input Word: "Assume" in Marathi Accent.

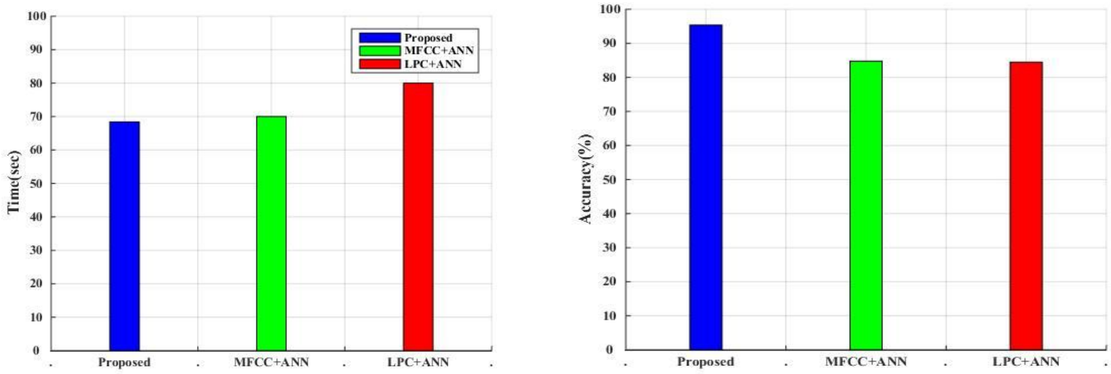
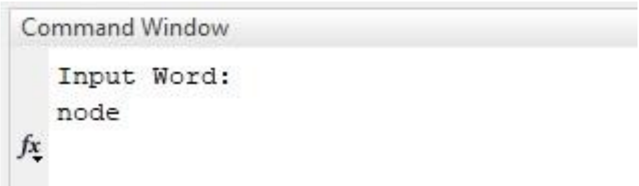


Figure 9: Execution Time and Accuracy Simulated Plots

The plots demonstrate the execution time with stressed accented word in Marathi accent is 68 sec and accuracy rate is 96 %. The accuracy can be validated as per the simulated data for accented word "Assume, TP= 98, TN= 72, FN= 4 and FP= 4. By substituting the above simulated values in the equation 1 for accuracy, the value comes out to be 95.50 %. (≈96 %). Other than accuracy the other parameters of performance are: Precision 96.07 %, Recall Rate 96.07%, F-Measure 96.07%.

CASE 3: Evaluation and validation with Rate of Speech.

3A: Input Word: "Node" in Telugu Accent at 30/min**.**



```
Input Word:
node
```

Figure 10: Accented word display

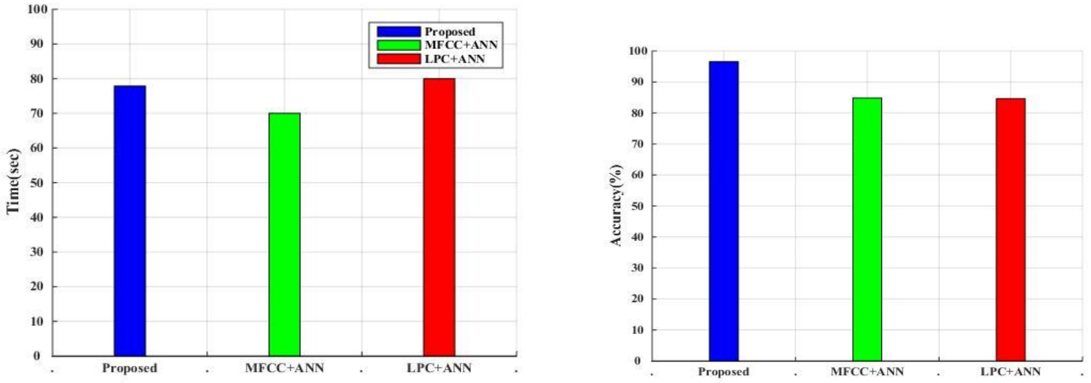


Figure 11: Execution Time and Accuracy Simulated Plots

The execution time with 30/min rate of speech in Telugu accent. is 78 sec and accuracy rate is 97 %. The accuracy can be validated as per the simulated data, TP= 98, TN= 70, FN= 2 and FP= 4. for the accented word "Node". By substituting the above simulated values in the equation 1 for accuracy, the value comes out to be 96.55 %. (≈97 %). Other than accuracy the other parameters of performance are: Precision 96.07 %, Recall Rate 98%, F-Measure 97.02%.

3B: Input Word: "Node" in Tamil Accent with 90/min rate of speech.

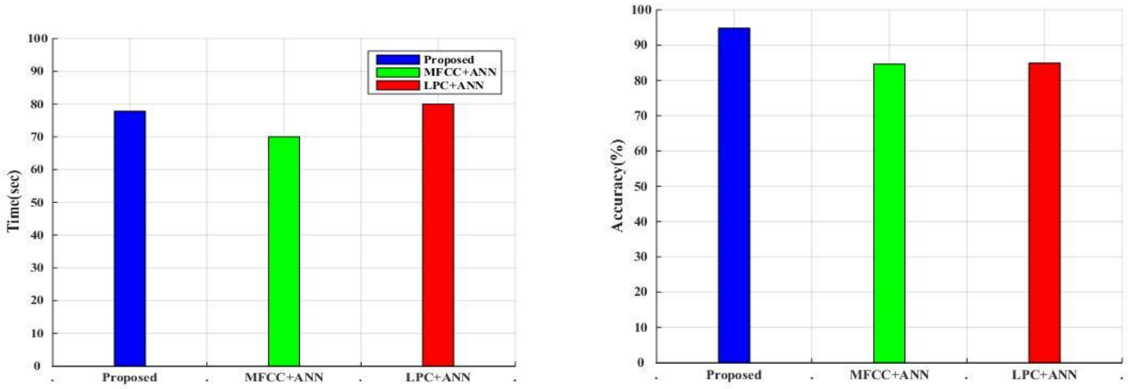
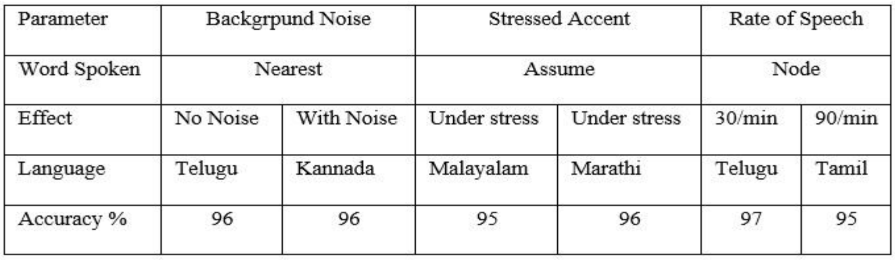


Figure 12: Execution Time and Accuracy Simulated Plots

The execution time with 90/min rate of speech in Tamil accent is 78 sec and accuracy rate is 95 %. The accuracy can be validated as per the simulated data, TP= 98, TN= 66, FN= 4 and FP= 5 for accented word "Node". By substituting the above simulated values in the equation 1 for accuracy, the value comes out to be 94.79 %. (≈95 %). Other than accuracy the other parameters of performance are: Precision 95.14 %, Recall Rate 96.07%, F-Measure 95.60%.

The table below compares the findings and shows accuracy is at or above 95%.

Table I: Finding Comparisons



Parameter	Backgrpund Noise		Stressed Accent		Rate of Speech	
Word Spoken	Nearest		Assume		Node	
Effect	No Noise	With Noise	Under stress	Under stress	30/min	90/min
Language	Telugu	Kannada	Malayalam	Marathi	Telugu	Tamil
Accuracy %	96	96	95	96	97	95

V. CONCLUSION

In present day scenario, human machine interactive systems facilitate communication between the people in real world situations. Natural language of any region has hundreds of accents which creates differences in pronunciations and intonations of speech. The conflict of interest comes when people of native and non-native come together for a cause face to face. For the realization of suggested automated speech recognition and text generation a progressive coding scheme based on Syllable MFCC along with ZCR, & integrated model of HMM speech coding is suggested. The implementation of the suggested work is evaluated over various speech samples with the approach of MFCC & Hidden Markov modeling thereby making the signal stable so as to get accurate features such as frequency, Amplitude, zero crossing rate, peak values etc. This research successfully developed an algorithm for accurate accent recognition using firefly optimized neural network (FO-NN) for accurate system evaluation by considering many parameters of interest. The accuracy evaluated is at or above 95% as compared to other conventional systems.

REFERENCES

[1]. Shaik Riyaz, Bathula Lakshmi Bhavani, S.Venkatrama Phani Kumar, “Automatic Speaker Recognition System in Urdu using MFCC & HMM”, International Journal of Recent Technology and Engineering (IJRTE), Volume 7, Issue-5S4, February 2019.

[2]. Aditay Tripathi, Aanchan Mohan, Saket Anand, Maneesh Singh, “Adversarial Learning of raw Speech Feature for Domain Invariant Speech Recognition”, Research Gate 2018.

[3]. Shubham Toshniwal , Tara N. Sainath, “Multilingual speech recognition with a single end-to-end model” IEEE 2018.

[4]. Ankur Mauryaa, Divya Kumara, R.K.Agarwal, “Speaker Recognition for Hindi Speech Signal using MFCC- GMM Approach” 6th International Conference on Smart Computing and Communications, ICSCC 2017, December 2017, Kurukshetra, India.

[5]. Varuna Shree and T. N. R. Kumar, Identification and Classification of brain tumor MRI Images with feature extraction using DWT and PNN, Brain Informatics (2018) 5:23–30, Springer, 2017.

[6]. Fontaine, V.; Ris, C.; Leich, H. “Comparison between two hybrid HMM/MLP approaches in speech research Acoustics, Speech, and Signal Processing, 2017, ICASSP-17, Conference Proceedings, International Conference on Volume 6, 7-10 May 2017 Page(s):3362 - 3365 vol. 6.

[7]. Haruna Chiroma, Ahmad Shukri Mohd Noor, Sameem Abdulkareem, Adamu I. Abubakar, Arief Hermawan, Hongwu Qin, Mukhtar Fatihu Hamza and Tutut Herawan1, Neural Networks Optimization through Genetic Review, Appl. Math. Inf. Sci. 11, No. 6, 1543-1564 (2017).

[8]. Shadiev, R., Hwang, W.Y., Huang, Y.M. and Liu, C.J. Investigating applications of speech-to-text recognition technology for a face-to-face seminar to assist learning of non-native English-speakinng participants, Tech., Pedagogy and Education 25 (2016) 119-34.

[9]. Behravan, H., Hautamaki, V., Siniscalchi, S.M., Kinnunen, T. and Lee, C.H. i-Vector modeling of speech attributes for automatic foreign accent recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (1) (2016) 29-41. [4]

[10]. Chen, N.F., Wee, D., Tong, R., Ma, B. and Li, H. Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL. Speech Communication 84 (1) (2016) 46-56.

[11]. Haizhou Wu, Yongquan Zhou, Qifang, and Mohamed Abdel basset “Training Feedforward Neural Networks Using Symbiotic Organisms Search Algorithm” Computational Intelligence and NEURP Science, 2016.

[12]. Dahl, G.E., Yu, D., Deng, L. and Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 20 (1) (2016) 30-42.

[13]. Thilagar PP, Harikrishnan R. Application of intelligent Firefly Algorithm to solve OPF with STATCOM. Indian Journal of Science and Technology. 2015 Sep; 8(22).

[14]. Sahar E. Bou-Ghazale and John H. L. Hansen, “A Novel Training Approach For Improving Speech Recognition Under Adverse Stressful Conditions”, ISCA achieve, 2014.

[15]. S. King, and M. Hasegawa-Johnson, “Accurate Speech Segmentation by Mimicking Human Auditory Processing”, in International Conference on Acoustic, Speech, and Signal Processing (ICASSP), Vancouver, Canada, 2013, pp. 8096-8100.

[16]. A. H. M. Russell and M. Carey, “Human and computer recognition of regional accents and ethnic groups from British English speech”, Computer Speech and Language, 27(1), pp. 59-74, 2013.

[17]. Li, M., Han, K.J. and Narayanan, S. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. Computer Speech & Language 27 (1) (2013) 151-167.

[18]. Mohit Dua, R. K. Aggarwal, Virender Kadyan, Shelza Dua, “Punjabi Automatic Speech Recognition Using HTK” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012.

[19]. Kumar. K., Aggarwal, R. K., Jain, A, “A Hindi Speech Recognition System for Connected Words using HMM Tool Kit” Int. J. Comput.Syst. Eng. 1(1), 25-32(2012).

[20]. Murat Akbacak, Dimitra Vergyri, Andreas Stolcke, Nicolas Scheffer, and Arindam Mandal Effective Arabic dialect classifcation using diverse phonotactic models. INTERSPEECH’11, pages 737–740, 2011.

[21]. Banati H, Bajaj M. Fire Fly based feature selection approach. IJCSI International Journal of Computer Science Issues. 2011; 8(4):473–80.

[22]. Ibrahim Patel, Dr. Y. Srinivas Rao, “SPEECH RECOGNITION using HMM with MFCC- an analysis using frequency spectra decomposition technique”, Signal & Image Processing: An International Journal (SIPIJ) Vol.1, No.2, December 2010

[23]. A. Ito, T. Konno, M. Ito et al., “Intonation Evaluation of English Utterances Using Synthesized Speech for Computer-assisted Language learning,” International Journal of Innovative Computing Information and Control, vol. 6, no. 3(B), pp. 1501-1514, Mar, 2010.

[24]. Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DWT) Techniques”, Journal of Computing, pp 138-143, vol. 2, Issue 3, March 2010.

[25]. Kittler J. Feature selection and extraction. Handbook of Pattern Recognition and Image Processing, Y. Fu., editor. New York: Academic Press; 1978.

[26]. L.V. Fausett, Fundamentals of Neural Networks, Architectures, Algorithms, and Applications, Prentice Hall Englewood Cliffs.

[27]. K. Woo, T. Yang, K. Park, and C. Lee,“Robust voice activity detection algorithm for estimating noise spectrum,” Electronics Letters, vol.36, no. 2, pp. 180–181, 2000.

[28]. http://cslu.cse.ogi.edu/HLTsurvey/ch1node7.html

ABOUT AUTHORS

Mr. Ashok Shigli, is pursuing Ph.D. (PP-ECE-0021) in the Department of ECE at Rayalaseema University, Kurnool, Andhra Pradesh. He has received B.E (Instrumentation Technology) from Mysore University in 1992, M. Tech (ECE) from IASE University, Sardarshar, Rajasthan, India in 1995. His main research interests are in Instrumentation, Signal and Speech Processing. He has a total experience of 25 years in teaching and research. He has published more than 13 research papers in International referred journals and international Conference proceedings.

.

Dr. K. Srinivasa Rao, is a Research Guide in the Department of ECE at Rayalaseema University, Kurnool, Andhra Pradesh. He has Ph.D. (ECE) from Andhra University, ME (ECE) in digital systems from Osmania University, Hyderabad and BE (ECE) from Andhra University. He is having a vast experience of more than 30 years in teaching and research. He has published more than 80 papers in journals and conferences both at national & International level. He is a member of IEEE and life member of other distinguished societies like FIETE, FIE, ISTE, BMESI and associated with foreign universities. His main research area includes signal Processing, Image Processing, Speech Processing, communication systems, Microwave and Antennas.

Ashok Shigli. “A Stratified Approach for Spoken Word and Accent Recognition: Validation & Analysis.” IOSR Journal of Engineering (IOSRJEN), vol. 09, no. 08, 2019, pp. 01-08.