Machine Learning Techniques for Speech Recognition using the Magnitude

C. Gopala Krishnan[1], Y. Harold Robinson[2], Naveen Chilamkurti[3*]

Abstract

Machine learning consists of supervised and unsupervised learning among which supervised learning is used for the speech recognition objectives. Supervised learning is the Data mining task of inferring a function from labeled training data. Speech recognition is the current trend that has gained focus over the decades. Most automation technologies use speech and speech recognition for various perspectives. This paper demonstrates an overview of major technological standpoint and gratitude of the elementary development of speech recognition and provides impression method has been developed in every stage of speech recognition using supervised learning. The project will use DNN to recognize speeches using magnitudes with large datasets.

Key Words: Deep neural network, Magnitude, Speech recognition, Supervised learning.

I. INTRODUCTION

Speech Recognition that otherwise known as automatic speech recognition (ASR) distinguishes the spoken words, persons and transfers them to a machine-readable configure. By transferring spoken audio into text, speech recognition skill enhances the users to organize digital devices. The traditional tools like keystrokes, buttons, keyboards etc the ASR technique can be broadly used in all walks of life [1].

In recent years, neural networks of deep learning have taken force in the field of artificial sensory development such as machine vision [26], [27]. In the emulation of this sense, the developments evolve from the segmentation of objects into images and localization of elements, e.g., localization of pedestrians where more recently, these techniques of usage of neural networks have been applied to voice recognition [2]. The initial developments that made use of neural networks in speech recognition began to implement different types of basic networks such as time-delay neural networks [3]. Due to the low processing capacity that computers had in the 90's, the neural networks did not have great depth, however, thanks to the progress in processing speed, it was possible to start deploying ever deeper neural networks, making them increasingly efficient in pattern recognition for this reason the interest has arisen in the application of these in tasks of speech recognition, even above other recognition techniques [4]. The introduction of Deep Neural Networks (DNN) for speech assignments began in the early 2010's developing DNN in combination with the Hidden Markov Model (HMM) [5], since, it allows modeling the sequential structure of a speech signal.

1.1 Convolution Neural Networks (CNN)

According to the high performance that Convolutional Neural Networks (CNN) have had in pattern recognition it has been begun to apply the concepts of convolution in DNN to speech recognition obtaining hybrids between convolution and fully connected layers [6]. However, at first, they did not exceed 3 convolution layers combined with a higher number of fully connected layers which limits the ability to acquire patterns and makes the performance not really high [7]. Recently, architectures began to be developed deeper and in combination with other types of neural networks like the recurrent neural

Corresponding Author (*): Naveen Chilamkurti, Department of Computer Science and IT, La Trobe University, Melbourne, Australia, N.Chilamkurti@latrobe.edu.au

[1]Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli, INDIA, skywarekrish@gmail.com

[2]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, INDIA, yhrobinphd@gmail.com

[3]Department of Computer Science and IT, La Trobe University, Melbourne, Australia, N.Chilamkurti@latrobe.edu.au

networks which help in the temporal relationship that may exist between signal divisions, improving the reduction of error in phonetic recognition [8].

1.2 Architectures

The architectures that have been developed, according to the state of the art have as input feature maps of phonemes which are used in large vocabulary continuous speech recognition, however for more basic applications, the use of phonemes makes them more complicated for its implementation, so, an alternative to reduce the complexity is the recognition of certain number of words delimited by the application, making the implementation of a CNN more feasible and achieve a better performance by the parameters that have to learn [9]. For this reason, in this work two CNN architectures based on the recognition of complete words are built which have not been developed in the state of the art as a complement to this research area [10].

In simple words speech recognition could be defined as the development of converting speech signal to a progression of words by means of algorithm demonstrated as a computer program [11]. Speech processing is the main areas of signal processing. The contribution of speech recognition is to develop a method for speech input to machine based readable script, which can be very well used in libraries, banks, and various workplaces for effective management [12]. Regular speech recognition today discovers widespread relevance in tasks that necessitate human machine interface like automatic call processing [13]. Fig. 1 illustrates the architectural framework for generating speech recognition.

II. EXISTING SYSTEM

In the existing system, a librarian can use barcode to get a particular book details and its availability in the library [14]. But the exact location where the book is placed cannot be easily identified. It is very difficult and laborious to get a particular book or books when large volume of books, manuals, research papers etc. are stacked in big libraries [15]. Moreover, it is time consuming and there is every possibility for misplacement and wrong issue of books or journals etc. [16]. ASR processing of a multimedia document's soundtrack has proven to be a substantial hindrance for users when searching multimedia digital libraries for sub-document segments (i.e., audio or video clips) [17].

The Phoneme Matching Algorithm has been developed that represents an early attempt to increase the effect of such speech-to-text transcription errors; this study reports on an attempt to enhance the efficiency [18]. The points-

based scoring system makes it possible to calculate a Phonetic Divergence Score (PDS) when comparing any two given phoneme series with the difference in the selected phonemes' scores indicating the degree of phonetic divergence [19]. For the experiments detailed in this study, PDS scores have been standardized by way of dividing the total PDS score by the number of phonemes in the query term, thus a 10-phoneme query term garnering a gross PDS score of 30 would be assigned a standardized PDS score of 3 [20].
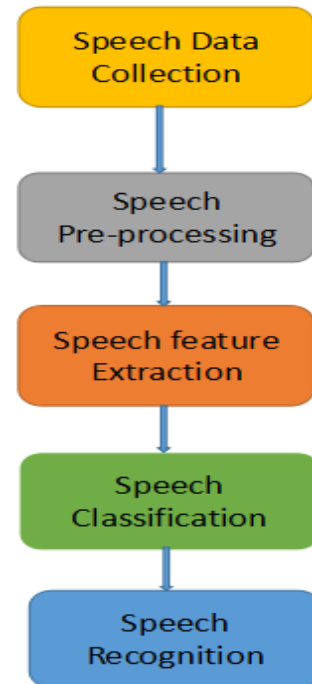

Fig. 1. Architectural framework.

2.1. Speech Technology and Machine Learning

Speech technologies must be developed for decades as a distinctive signal processing region, while the preceding decade has brought an enormous development according to the latest machine learning paradigms. In spite of the intrinsic complexity has been proposed that the relations with cognitive systems, speech recognition methods are the most common interdisciplinary knowledge area. The paper fully based on to develop the machine learning methods that are computational intelligence contributes to provide fields that cover speech construction and auditory observation that demonstrates the speech transmission and language, speech recognition. The spoken dialogue construction is consequently speech synthesis.

2.2. Concepts and recent advances

Moreover, this paper demonstrates the concepts of recent developments in speech signal demonstration, coding that include cognitive coding. The main objective of this paper is to highlight the recent development

according to the latest machine learning methods that has an impact in the development of speech signal processing.

III. PROPOSED SYSTEM

In the existing system various algorithms are used whereas in the proposed system we have implemented only Artificial Neuron Network (ANN) and accordingly input data's have to be fed and output results probably the exact one is expected to be derived. By this means considerable time saving is expected and financial implications too can be minimized.

3.1. Automatic Speech Recognition (ASR)

3.1.1. Basic Principle

ASR demonstrates into a training method that the system learns the reference represents the dissimilar kind of speech recognition that comprises the application of the vocabulary systems. Every reference is learned from spoken application that stored like templates constructed by several methods which illustrate the pattern related properties. The recognizing phase that input pattern is generated by the references. Fig. 2 demonstrates the Automatic Speech Recognition.
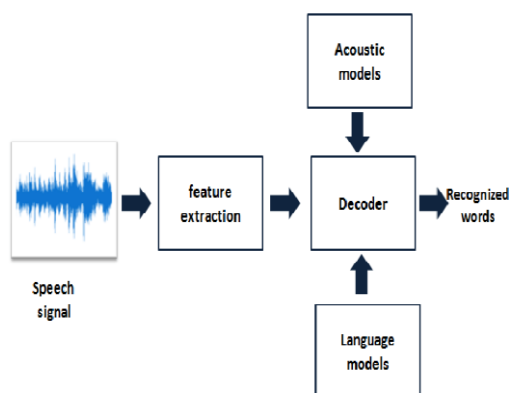

Fig. 2. Automatic Speech Recognition.

3.2. Speech Recognition Techniques

The main objective of speech recognition is too capable of understand the spoken information. The related methods did not execute the expected accuracy, so the proposed methodology has been constructed to implement the speech recognition in an automatic manner with speaker identity. This system could be constructed to implement the feature extraction, modeling, analyzing and also testing process.

3.3. Magnitude Analysis

A suitable feature which can differentiate the "Cholo" and "Thamo" signal is magnitude analysis of the given signals.

A general condition feature can be applied as,

1. Find the max() of the signal and its location.

2. Let apeak range of the duration of the signal is -5000to 5000 location for max().

3. Now, find the max() of the amplitude of the location of before the range and after the range.

4. If both magnitudes before highest and after highest are less than (highest-0.3), then the signal will be 'Thamo', where (highest-0.3) is our minimum countable max.

5. The exception of this condition will be 'Cholo' signal.

Here is common scenario of a Cholo signal where upper figure is original signal and below the figure is pre-processed signal plot. As I can observe that, here the highest max is 1 and if I take another two max before the highest and after the highest those max magnitudes are very near to the highest and differences are in between 0.3 range. Now, the general scenario of the Thamo signal plot which is given besides, after the pre-processing, the maxs of before and after highest of Thamo signals are quite far (in magnitude) at all.

Therefore, we have set a condition of if-else in my MATLAB code as, If before max and after max is less than [highest-0.3], that signal is Thamo signal. Else of the condition is, Thamo signal; where, before max and after max is greater than [highest-0.3], that means in between this range

Analysis process basics

Here, we have got 7 succeed and 5 Failed output. So, there are 3 kinds of failure I have found and the reason behind the failure cases are discussed below,

1. Peak range: In some cases, ± 5000 peak range exceed

2. Magnitude range: Due to peak range violation I got smaller magnitude max

3. Noise problem: In noisy signals, sometimes noise becomes max magnitude

At the end of this experiment, we have obtained proper knowledge about how signal works in moving average filter. I passed the signals through 400 points moving average filter and analyzed the data. I found a pattern for to successfully distinguish between 'cholo' and 'thamo'

audio signals. This can also be done in various other methods as I have seen by the other groups. I found this technique effective and unique and thus used this to solve the problem. For audio signal analysis, this experiment proves as a basic and will help us in future experiments.

3.4. Effectiveness of applying this need to machine learning

The development of speech recognition methods is constructed using the innovative machine learning methodologies to implement the accuracy level. According to the improvement of the machine learning approach has been utilized to provide the excessive impact to this development. HCI systems are having the spoken language understanding and demonstration.

The machine learning concept has automation about the speech recognition system to translate the text to speech technology. It is demonstrated that the system with deep learning approach could be capable of recognizing the speech components in an acoustic surrounding with the capability of analyzing the accuracy with human based interaction system using HCI of style and speaker quality.

The reinforcement related machine learning methodology has been utilized the unsupervised methodology to develop the progress with huge number of datasets. The speech signal is involved to provide the automatic speech recognition in accurate way and signal based synthesis detection to generate the adaptive quantization in view of speech coding approaches.

The innovative methods could be constructed to provide the robustness in a huge amount of dynamic speech signal demonstrations. The frame adaptive approach can indicate that the machine learning tools to increase the popularity and provided the exact solutions. The improved performances of the proposed system have proved by utilizing the predictive coefficients. An increasing amount of progress in speech signal processing with machine learning approach.

3.5. Artificial Neural Networks

Neural networks have incorporated the Markov chain methodology that the statistical methods are demonstrated as a graph. The Markov models have be utilized for providing the probability based state transitions that the neural networks utilize the connection based state transition. The common dissimilarity within the neural network and Markov model is the parallel connectivity. The state transition is the common strength of the connectivity. Artificial Neurons are the common component of Artificial Neural Network that utilizes the function of organic neuron. It is a mathematical formation that is conceived as a representation of natural neuron. Fig. 3 demonstrates the common artificial neuron.
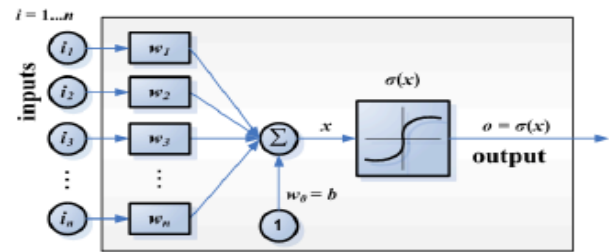

Fig. 3. Structure of Artificial Neural Network.

Types of Artificial Neural Network
- Feedforward Network
- Recurrent Neural Network
- Modular Neural Network
- Kohonen Self Organizing Maps

A Feedforward Network is a mathematical model according to the construction of biological neural networks. The demonstration that flows within the network influences the architecture of the ANN since a neural network modification based on those functionalities. ANNs are measured as non-linear statistical information modeling components that are the multifaceted relationships within the inputs and outputs are modeled. ANN is also called as a neural network. ANNs have 3 layers that are interconnected with other related systems. The initial layer contains the input neurons. Those neurons forward the information to the middle layer, which in turn forwards the output neurons to the final layer. Fig. 4 illustrates the demonstration of the Feedforward network layer.
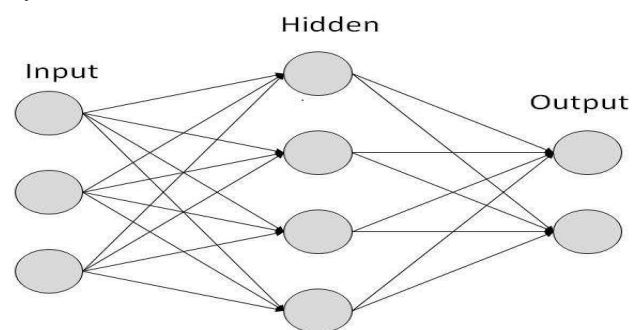

Fig. 4. Feedforward network.

Artificial neural networks are an effort to obtain computers to effort more like the human brain. Your brain doesn't accumulate detailed encoded functions; it has enormous networks of neurons. These change their communications to every other as latest data past throughout them. Speech recognition with this machine learning is completed with big data-based organizations of data to train these networks.

An ANN has a lot of positives but one of the mainly recognized of these is the reality that it preserves

essentially train from scrutinizing the data sets. In this approach, ANN is utilized like a random function estimate component. These types of components assist approximation the nearly all cost-effective and ideal approaches for finding the answers while defining computing distributions. ANN understands data samples rather than whole information to demonstrate at solutions, which has the time utilization. ANNs are measured reasonably simple mathematical elements to augment existing data analysis approaches.

3.6. Working with Audio Files

Speech Recognition utilizes the implementation of audio files with AudioFile class. This class could be started with the way to an audio file and presents a context manager edge.
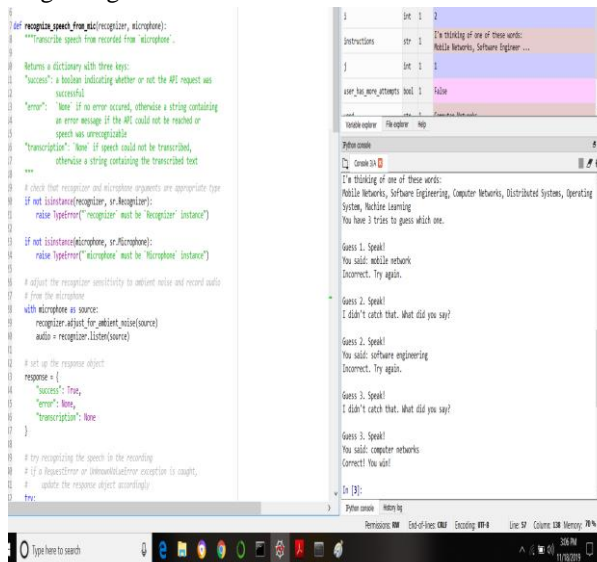

Fig. 5. Retrieving input from audio input.

3.6.1. Supported File Types

Presently, Speech Recognition wires the following file formats:
- WAV: must be in PCM/LPCM format
- AIFF
- AIFF-C
- FLAC: must be native FLAC format; OGG-FLAC is not supported

Recognizing speech needs audio input, and Speech Recognition uses retrieving this input. The Speech Recognition will have you up and operation in just a small number of minutes and it is demonstrated in Fig. 5. The suppleness and ease-of-use of the Speech Recognition package create it as a replacement for any Python coding. Moreover, maintain for every characteristic of every API it wraps is not definite. You will require expending more

time for identifying the obtainable options to find out if Speech Recognition may utilize in your scrupulous case and it is illustrated in Fig. 6 and Fig. 7.
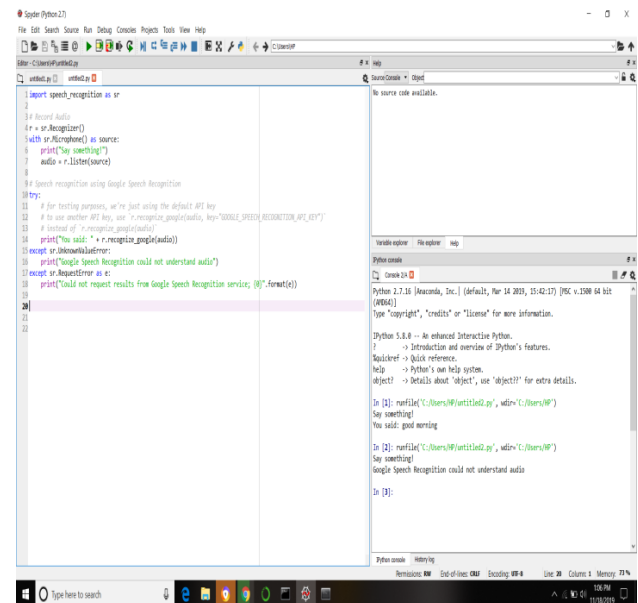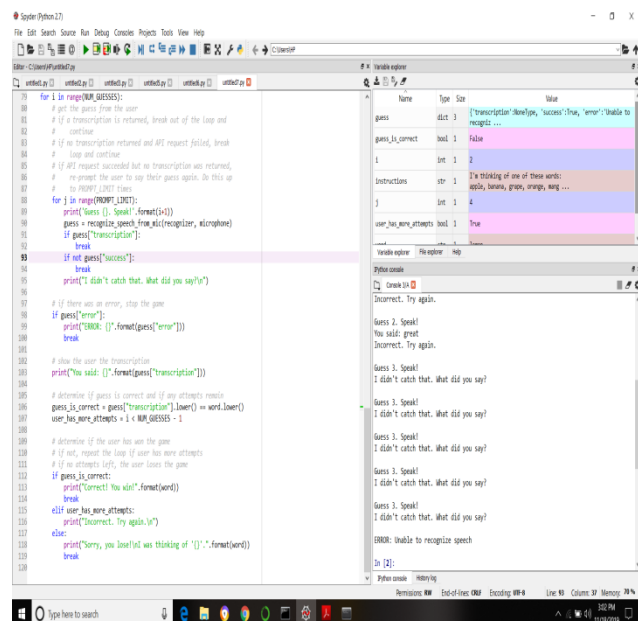

Fig. 6. API formation.

Fig. 7. Output values.

3.7. Training an ANN

ANNs are comparatively rudimentary electronic networks of "neurons" according to the neural organization of the brain. They procedure has been constructed to store data, analyzing their classification of the system with the mostly used classification of the system. The errors from the original classification of the initial record is fed back into the network, and utilized to adapt the networks algorithm the next time.
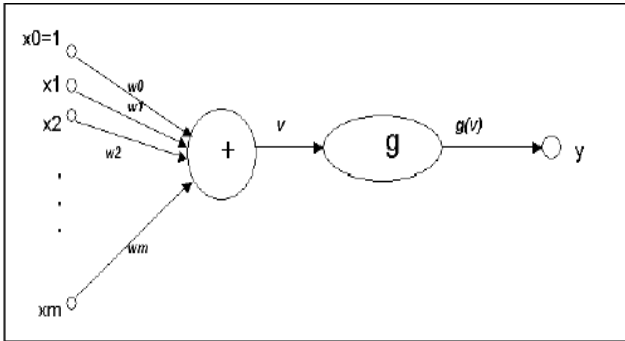
Fig. 8. Training of ANN.

In the training phase, the correct class for each record is known (this is termed supervised training), and the output nodes can therefore be assigned "correct" values -- "1" for the node corresponding to the correct class, and "0" for the others and it is illustrated in Fig. 8.

3.8. Coding Source in Matlab

```
clc
clear all
close all
signals = ['E:\nufaer\test_data\V-2.mp3'];
audo = audioread(signals);
figure;
subplot(2,1,1);
plot(audo);
title('Original audio signal');
m=400;
b=(1/m)*ones(1,m);
pstv=abs(audo);
audof=filter(b,1,pstv);
audof=audof / max(audof);
subplot(2,1,2);
plot(audof);
title('Pre-processed signal');
maxmag=max(audof)
maxloc=find(audof==maxmag);
nextmag=max(audof(maxloc+5000:end))
prevmag=max(audof(1:maxloc-5000))
if((nextmag<(maxmag-0.3))&&(prevmag<(maxmag-0.3)))
disp('signal is thamo');
else
disp('signal is cholo');
end
```

IV. CONCLUSION

In this paper, we have tried to introduce a simple approach which could be used to recognize connected speech and the person concerned. The speech features extorted are compared with related speeches in the database for identification. This approach utilizes it possible by the speech of the broadcaster and it will be simple to authenticate their independence. It generates control access to various   applications like window

speech recognition. Indeed, the application of this technique will certainly enhance smooth and perfect administrative innovations in the day today activates wherever manpower is entertained in multiples such as libraries banks and various workplaces etc. In the present study I have tried to develop a device which will enable to find the presence of a particular data from the cluster of datasets using python. In future this device can be taken to the next level by using Artificial Neural Network (ANN) which will lead us to work with incomplete knowledge on information related to the speech and the person concerned. Further in the new application the network layers will be built and trained to show the pictorial representation of in-built data.

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 9099, 2018.

[2] Balaji, S., Golden Julie, E., Harold Robinson, Y.: Development of Fuzzy based Energy Efficient Cluster Routing Protocol to Increase the Lifetime of Wireless Sensor Networks, *Mobile Networks and Applications*, vol. 24, no. 2, pp. 394-406, 2019.

[3] Hsu, W.N., Y. Zhang, A Lee and J.R Glass, "Exploiting depth and highway connections in convolutional recurrent deep neural networks for speech recognition," in *Proceedings of the International Conference on Interspeech*, University of California, San Francisco, California, USA, pp. 395-399. 2016

[4] Pahini A. Trivedi, "Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Networks," *International Journal of Engineering Development and Research*, Volume 2, Issue 4, 2014.

[5] M. S. Hossain and G. Muhammad, ``Emotion recognition using deep learning approach from audiovisual emotional big data," *Inf. Fusion*, vol. 49, pp. 6978, Sep. 2019.

[6] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," *IEEE Access*, vol. 5, pp. 326337, 2016.

[7] Ondruska P., J. Dequaire, D. Z. Wang and Posner, End-to-end tracking and semantic segmentation using recwrent neural networks. *Master Thesis*, Cornell University, Ithaca, New York, USA, 2016.

[8] N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?" in *Proc. ACM 16th Int. Workshop Mobile Comput. Syst. Appl.*, 2015, pp. 117122.

[9] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, "Speech emotion recognition

in emotional feedback for human-robot interaction," *Int. J. Adv. Res. Artif. Intell.*, vol. 4, no. 2, pp. 2027, 2015.

[10] Subramanian Balaji, Yesudhas Harold Robinson, Enoch Golden Julie, "GBMS: A New Centralized Graph Based Mirror System Approach to Prevent Evaders for Data Handling with Arithmetic Coding in Wireless Sensor Networks," *Ingénierie des Systèmes d'Information*, vol. 24, no. 5, pp. 481-490, 2019.

[11] Orozco, I., M.E. Buemi and J.J. Berlles, "A study on pedestrian detection using a deep convolutional neural network," in *Proceedings of the International Conference on Pattern Recognition Systems (ICPRS-16)*, April 20-22, 2016, IET, Talca, Chile, ISBN: 978-l-78561-283-l, pp. 1-15, 2016.

[12] P. S. Apirajitha, C. Gopala Krishnan, G. Aravind Swaminathan, E. Manohar, "Enhanced Secure User Data on Cloud using Cloud Data Centre Computing and Decoy Technique," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 9, pp. 1436-1439, July 2019.

[13] C. G. Krishnan, K. Sivakumar and E. Manohar, "An Enhanced Method to Secure and Energy Effective Data Transfer in WSN using Hierarchical and Dynamic Elliptic Curve Cryptosystem," in *Proceeding of 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, pp. 1-7, 2018.

[14] C. Gopala Krishnan, A. Renga Rajan, R. Manikandan, "Delay Reduction by Providing Location Based Services using Hybrid Cache in peer to peer Networks," *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, vol. 9, no. 6, pp. 2078-2094, Jun. 2015

[15] Harold Robinson, Y, Balaji, S., Golden Julie, E.: FPSOEE: Fuzzy-enabled particle swarm optimization-based energy-efficient algorithm in mobile ad-hoc networks, Journal of Intelligent & Fuzzy Systems, vol. 36, no. 4, pp. 3541–3553, 2019.

[16] M. Poongodi, Gopalakrishnan, Vijayakumar and I mad Fakhri Al-Shaikhli, "An Effective Energy Ba sed Location Optimization in Wireless Sensor Net works in Structural Health Monitoring," *Internatio nal Journal of Pure and Applied Mathematics*, v ol. 116, no. 22, pp. 275-288, 2017.

[17] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden MARKOV models with deep belief networks," in *Proceedings of IEEE Workshop Autom. Speech Recognit. Understand.*, pp. 216221, 2013.

[18] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 6068, Aug. 2017.

[19] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, "Learning emotion discriminative and domain-invariant features for domain adaptation in speech emotion recognition,"' *Speech Commun.*, vol. 93, pp. 110, Oct. 2017.

[20] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,"' *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 15761590, Oct. 2017.

[21] Yoshioka, T., K. Ohnishi, F. Fang and T. Nakatani, "Noise robust speech recognition using recent developments in neural networks for computer vision," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, ISBN:978-l-4799-9988-0, pp. 5730-5734, 2016.

[22] Sercu T., C. Puhrsch, B. Kingsbury and Y. LeCun, "Very deep multilingual convolutional neural networks for LVCSR," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, ISBN:978-l-4799-9988-0, pp. 4955-4959, 2016.

[23] J. Deng, X. Xu, Z. Zhang, and S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500504, 2017.

[24] Qian, Y. and P.C. Woodland, "Very deep convolutional neural networks for robust speech recognition," in *Proceedings of the 2016 IEEE International Workshop on Spoken Language Technology (SLT)*, San Diego, USA, ISBN:978-l-5090-4903-5, pp. 481-488, 2016.

[25] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, pp. 2227-2231, 2017.

[26] Ji-Hae Kim, Byung-Gyu Kim, Partha Pratim Roy, Da-Mi Jeong "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," *IEEE Access*, vol. 7, pp. 41273-41285, 2019.

[27] Ji-Hae Kim, Gwang-Soo Hong, Byung-Gyu Kim, Debi P. Dogra, "deepGesture: Deep Learning-based Gesture Recognition Scheme using Motion Sensors," *Displays*, vol. 55, pp. 38-45, 2018.

Authors

Dr. C. Gopala Krishnan received his PhD in Computer science and Engineering in 2018 from St. Peter's University, Chennai, India. He received his M.E. degree in Computer science and Engineering in 2010 from Anna University Tirunelveli, India and B.E. degree in Computer Science and Engineering in 2002 from Madurai Kamaraj University, India. Currently he is working as an Associate Professor in the department of CSE at Francis Xavier Engineering College. His areas of interest are Mobile computing, Operating systems, Computer Networks, Computer Graphics and Multi Core Architecture. He has published many papers in International Conferences and International journals in various fields of advanced technologies. He is working on developing communication protocols for wireless networks and protocols optimized for wireless and mobility that can support file and database access. He is also investigating operating systems support for mobile hosts. He is a member of various Professional bodies.

Dr. Y. Harold Robinson is currently working in the School of Information Technology and Engineering, Vellore Institute of Technology, Vellore. He has received Ph. D. degree in Information and Communication Engineering from Anna University, Chennai in the year 2016. He is having more than 12 years of experience in teaching. He has published more than 50 papers in various International Journals and presented more than 45 papers in both national and International Conferences. He has written 4 book chapters by Springer, IGI global Publication. He is acting as an editor for a book Title as "Successful Implementation and Deployment of IoT Projects in Smart Cities "IGI Global in the Advances in Environmental Engineering and Green Technologies (AEEGT)" book series. He is one of the editor for the book "Handbook of Research on Block chain Technology: Trend and Technologies" publish by Elsevier. He is acting as a Reviewer of many journals like Multimedia Tools and Applications, Wireless Personal Communication by Springer Publication.

Dr. Naveen Chilamkurti obtained his Ph.D. from La Trobe University and currently is the Acting Head of Department, Computer Science and Computer Engineering, La Trobe University, Australia. He is the Inaugural Editor-in-Chief for International Journal of Wireless Networks and Broadband Technologies launched in July 2011. He has published about 300 Journal and conference papers. His current research areas include intelligent transport systems (ITS), wireless multimedia and wireless sensor networks. He currently serves on the editorial boards of several international journals including Wiley IJCS, SCN, Inderscience JETWI, and IJIPT. He is a Senior Member of IEEE.