Study of Various Machine Learning Algorithms for use with Automatic Speech Recognition

T. Sathies Kumar [1], T.Sheela [2], D.Arulselvam[3], S.Premalatha[4], K.Srividya[5]

[2]Associate Professor, Department of ECE, Vinayaka Mission's Kirupananda Variyar Engg. College, Salem

[1,4,5]Assistant Professor, Department of EIE, Sri Sairam Engineering College, Chennai.

[3]Assistant Professor, Department of EEE, Sri Sairam Engineering College, Chennai.

Email id: [1]sathies.ei@sairam.edu.in, [2] sheelamuthu@gmail.com,

[3] arulselvam.eee @sairam.edu.in, [4]premalatha.ei@sairam.edu.in,[5]srividya.ei@sairam.edu.in

2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS) | 978-1-6654-6275-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPECTS56089.2022.10047695

Abstract— **Speech recognition enables which the system recognize and identify the words from aspeech. It aids in balancing technology enabled Human computer interaction (HCI) and Human robot interaction(HRI). Command and Communication of any information is brought down through Speech. Artificial intellingence carried out with the help of circuit involving speech recognition which involved in Controlling and commanding application such as security system, VCR systems. Speech processing make life easier allows user to perform parallel tasks with increased efficiency and effectiveness(i.e., if the user's hands and eyes are occupied elsewhere OR disabled person who are not using their hand). Occasionally the speech signal is not clearly understand by the system which we communicate with them, by solving this issue we made an various investigation. In this paper we give report of investigate the various methods, algorithm which used in the previous speech recognition technique and system to improve the communication between a Human-Computer Interaction and Human Robotic Interaction to obtain a better results. Various algorithms like dynamic time warping (DTW) and Mel frequency Cepstral coefficients (MFCC), Artificial Neural Network (ANN) and Power Normalized Cepstral Coefficients (PNCC) were analyzed.**

Keywords— MFCC, PNCC, ANN, DTW, HRI

I. Introduction

Worldwide smart homes are highly involved with Automatic Speech Recognition [1]. These recognition units perfrom well when the direction comes from closely placed microphones, they give better performance using head-set. The performance slightly decreases when the indication comes from the user from far away microphone. This performance degradation is due to unknown groud noise and unwanted reverberation within the house environment. The unwanted signal distortion caused by reverberation is reduced or treated in recognition system using a acoustic model for the system. It is controlled at the input level by changing the adaptation level of input feature with data trained as taget along with its distortion loss. Systems with equal distortion are trained with data obtained from target for both stationary and non stationary conditions.

The home automation system performance decrease in its performance and has no significant improvemnet even with trained acoustic input models if the Reverberation time is more than 500 ms. Considering the above condition the time of reverberation stays below 500ms for tiny dimension flats, which can be effective in maintaining a close proximity with the talking microphones. Adaptive data recorded system for the test environment is used in the techniques of speech recognition.

Application of SR

A. Digital Assistants

These are voice activated smart sensor mobiles or appliances which includes Cortana, Alexa, Siri [3,4]. All of the above applications are activated using voice command and also it gather information from multiple available databases and other sources to interact for the commands provided They provide better interaction between people and devices.

B. Banking

Voice or speech recognition system [2] indeed provide valuable help to the customer in solving the personalized queries and also responds request such as account balances, payment and transactions. It aids in improving customer care loyalty and satisfaction.

C. Healthcare

Healthcare involves the life of human it requires quick responses. Handling patient with the voice automated system improves quality and speed of the entire health care system. Health records of the patient can be remotely moitored and accessed. It aids in booking an appointment and indicating to the health care professionals and the patients at the right time. AI system improves the bedding allotment administration.

II. CHALLENGES IN SR

A. Advantages/Disadvantages of SR

Voice enabled system performs various task by indicating it to the google home, alexa and other voice recognition process. machine learning and modified algorithms converts speech in text. Acuracy can be improved by modified voice enabled system and software program by reducing errors. The surrounding noise produces false input which includes system in silent room. The next problem

978-1-6654-6275-4/22/$31.00 ©2022 IEEE

associated with the voice enabled system when the sound of a word is same alike and that are differently spelled with different meaning for the alike words hear and hear. The problem can be overcome with the stored contextual information. These stored information require more storage RAM and fast processing processors.

B. *Background noise*

The noise interference in the required signal of interest is known then various technique is employed in removal of known noise. The possible noise sources are captured using a specialized microphone and impulse response of the noise source is calculated in a acoustic environment in order to suppress the noise. Noise impulse response is measured through least mean square. The method shows promising results when the noise consists of music and speech content. Unknown noise sources such as washer or binding noise, Blind Source Separation (BSS) method is more suited. The low frequency signal obtained from the microphone consists of mixture of speech and noise sources. Independent Component Analysis, a type of sub category of BSS, separate various sources through its statistical properties. The above process is well suited for signal that are not Gaussian and it also does not involve position of the emitter or mic. The above consider noise and signal linearly mixed. Separation of noise in real time smart home condition remains an open challenge.

C. *Word spotting*

Detection of words which is uttered has been used extensively over past decades identified from very large speech databases and in continuous speech streams. A wide study is conducted with over 100 samples. The study is governed using spotifying vocabulary in continuous speech. Language models and acoustic adaptation using multisource based recognition were also adopted.

IV Speech Recognition Algorithm

Speech recognition software [15,16,17] is defined as a technology that can process speech uttered in a natural language and convert it into readable text with a high degree of accuracy, using artificial intelligence (AI), machine learning (ML), and natural language (NLP). Two most popular sets of ML algorithms used in the analysis of the speech signal are the Many research studies were now concentrating to improve the ability of hearing aids as hearing loss has been increasing in an alarming rate in the present scenario. Recognizing speech of humans is very difficult if the environment is surrounded by noise back ground, multiple speakers and reverberation. machine learning algorithms is employed to improve human speech recognition considering the signal-to-noise ratio in which a specific number of words specifically fifty percent are recognized, at less time in a cost effective manner.

A. *Mel Frequency Cepstral Coefficients (MFCC)*

The figure 1 shows the overall diagrammatic view of MFCC [6,7,8] algorithm which convert spatial time domain signal into frequency domain signal by involving functional cochlea Mel filters. MFCC is a sort of Discrete Cosine Transform (DCT) which is a logarithmic value of the energy on the Mel frequency scale whose features is extracted through the aid of power spectrum frame and filter banks of MEL.

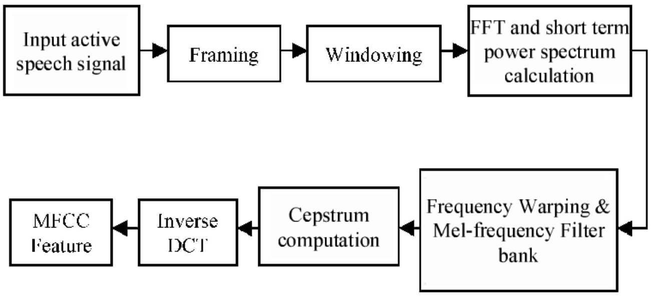
Input active speech signal
Framing
Windowing
FFT and short term power spectrum calculation
Frequency Warping & Mel-frequency Filter bank
Cepstrum computation
Inverse DCT
MFCC Feature

Fig 1. Overall Flow diagram of MFCC Algorithm

Mel scale indicates the received frequency or pitch of a pure tone to its actual frequency measured. Human beings are well structured in discerning very minute changes in pitch over low frequencies than pitch are in high frequencies.

Conversion formula for frequency to Mel scale is:

$M(f)=1125 \ln (1+f/700)$ ....(1)

Mels back to frequency:

$M^{-1}(m) = 700 (\exp (m/1125) - 1)$ .... (2)

Deltas and Delta-Deltas

Deltas and delta-deltas are differential and acceleration coefficients. Feature vectors of MFCC indicate the single frame power spectral envelope. Necessary dynamics is a major factor in speech which is the major trajectories of MFCC coefficient overtime. Calculating the MFCC trajectories and combining them to the original feature vector increases ASR performance (12 MFCC coefficients, get 12 delta coefficients, combine to give a feature vector of length 24).

Formula is used to calculate delta coefficient

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \ldots\ldots\ldots\ldots(3)$$

Delta-Delta (Acceleration) coefficients are calculated from the deltas, and not from the static coefficients. Where d_t is a delta coefficient, from which frame't' is computed in terms of the static coefficients $C_{t+n} - C_{t-n}$.

ANN algorithm

The multi-layer feed-forward ANN architecture as shown in figure 2.3 MFCC vectors was carried out utilizing the Matlab platform. The ANN is prepared utilizing the back proliferation calculation, which has been shown as compelling in accomplishing minimization of acknowledgment mistake rates. The ANN engineering assumes a basic part in accomplishing a decent and advantageous outcome and this relies to a great extent upon the quantity of neurons at the info and secret layers and the quantity of layers for the errand. To accomplish the necessary ANN architecture, fundamental design of three layer framework, with varying numbers of neurons in each and every layer preparing till the ideal number that gave the best preparing output was accomplished. The biological neural network structure and function are used in designing ANN architecture. In ANN neurons are stacked in layers, similar like neurons in the brain. In FFNN an input layer that receives data for pattern recognition, an output layer solves the problem, and a hidden layer connects the other levels. A cyclic method connects neurons from the feed layer to the output layer. The ANN trains the datasets using advanced training methods that re-alters neuron weights based on the error rate produced between target input and the actual output.

An assortment of expressions from five speakers was utilized for testing the ANN design. The removed element vectors of every one of the expressions introduced in the information layer most likely discourse and speaker distinguished at the yield layer. Since the back spread technique as a rule requires objective worth that is utilized while preparing, which for discourse acknowledgment, is not accessible, we consequently expect the objective yield to '1' for the right speech expression and '0' for other people. This in some way or another supported the right yield and debilitated some unacceptable yield. The network was prepared by changing the upsides of the loads between the network components.

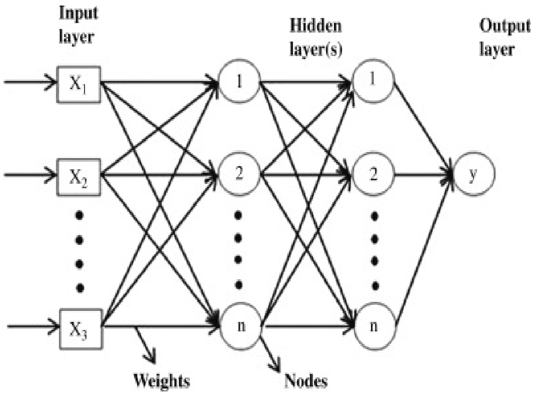

Fig 2 General structure of ANN

Power-Normalized Cepstral Coefficients (PNCC)

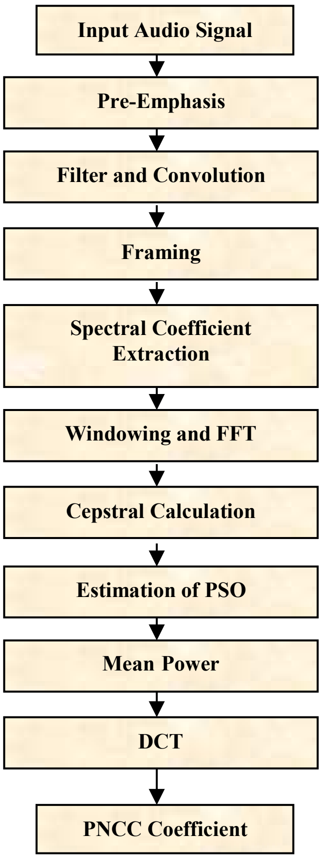

Fig 3 Power-Normalized Cepstral Coefficients (PNCC)

The power normalized Cepstral coefficients is extracted as feature by involving approach inspired by auditory processing is a new feature extraction approach inspired by auditory processing. Power law non linearity swaps the distinctive log non-linearity used in MFCC coefficients, a noise-suppression technique completely based on asymmetric filtering that embodies background excitation, and a temporal masking module are all major new elements of PNCC processing. The figure 3 shows the schematic flow of PNCC algorithm used in the SR processing.

Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a stochastic statistical designed model which is used in machine learning. It is used to explain the evolution of observed events depends on internal factors. A HMM statistical model used in machine learning is illustrated in the figure 4. The input speech signal is analyzed using Gaussian distribution followed by state transition probabilities and later the signals mean and variance is estimated and compared with the mixed weight, as the said processing is not shown it is state to be hidden hence the name.

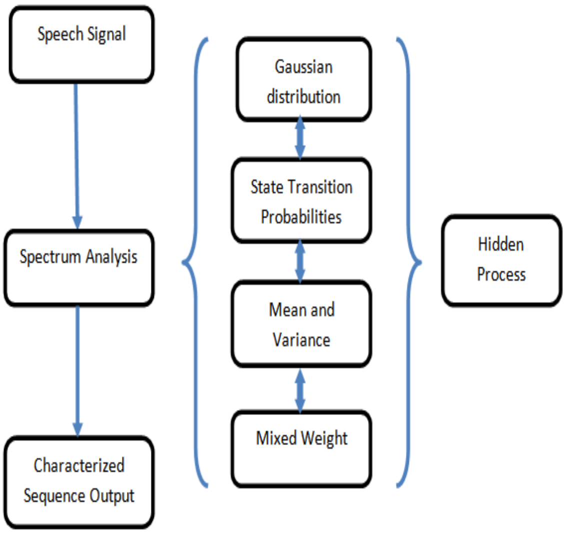

Figure 4 Block diagram of HMM for Speech Recognition

Power Normalized Cepstral Coefficients is spurred by sound-related preparing. Major highlights of PNCC incorporat utilization of force law non linearity replacing the conventional log non-linearity utilized in MFCC coefficients, a clamor concealment calculation dependent on hilter kilter separating that stifles foundation excitation, and a module that achieves fleeting covering. The utilization of medium time power investigation,where natural parameters were assessed over a more extended term than is generally utilized for discourse, just as recurrence smoothing. Trial results show that PNCC handling gives considerable upgrades in acknowledgment exactness contrasted with MFCC and PLP preparing for discourse within the sight of different sorts of added substance commotion and in reverberant conditions, with just somewhat more prominent computational expense than regular MFCC (mel frequency Cepstral coefficients) preparing, and without corrupting the acknowledgment precision that is watched while preparing and testing utilizing clean discourse. PNCC handling likewise gives better acknowledgment precision in boisterous situations than strategies, for example, ETSI advanced front end (AFE) and vector taylor series (VTS) with significantly less calculation. PNCC calculation doesn't utilize a logarithmic non-linearity, every one of the highlights may firmly be impacted by the sign level. So as to lessen this marvel, the PNCC algorithm incorporates a power normalization system which comprises in scaling the intensity of each frame as indicated by a normalization factor assessed on-line utilizing the intensity of past frames.

Frequency transformation

The PNCC is a delegate highlight set that endeavors to incorporate a large number of the sound-related preparing traits in a computationally effective manner. PNCC handling starts in customary design with a brief timeframe Fourier change, with the yields in each casing increased by gammatone recurrence weighting along a force work nonlinearity, and age of cepstral-like coefficients utilizing DCT and mean standardization. Generally, commotion and resonation concealment is practiced by a nonlinear arrangement of activities that perform running clamor concealment and worldly differentiation upgrade, individually working in a medium time setting, with examination interims on the request for 50-15oms.(The outcome this more drawn out term investigation are applied to flag portrayals removed over conventional 20-35ms investigation outlines for discourse acknowledgment). Numerous gatherings have discovered that PNCC preparing gives compelling commotion strength just as concealment of resonation impacts, with minor adjustment, and the calculation required is similar to that utilized in MFCC and PLP include extraction.

V- Experiment Results and Discussion

All of the experiments employed a bigram language model having 39-length feature vectors with delta and delta-delta features. Subsets of 1600 clean speech signal as training set and six hundred degraded speech signals for testing in experiments using the DARPA Resource Management (RM1) database were employed. The system is trained on the WSJ0 SI-84 training set and WSJ0 5K test set from the DARPA Wall Street Journal (WSJ) 5000-word database.

The input of speaker speaks: “right”.

Table 1: Comparative Analysis of Various Algorithms

	Accuracy (%)			
	HMM	**DTW**	**ANN**	**PNCC**
Person with clear voice	**92.24%**	**94.22%**	**97.66%**	**97.78%**
Person With infected (cold)voice	**55.56%**	**62.49%**	**88.19%**	**86.33%**
Person in Clear indoor	**64.13%**	**73.15%**	**89.20%**	**87.75%**
Person In noisy indoor	**54.66%**	**54.82%**	**60.38%**	**59.62%**
Person In clear outdoor	**63.66%**	**62.61%**	**85.96%**	**86.32%**
Person In noisy outdoor	**52.86%**	**51.41%**	**55.03%**	**54.52%**

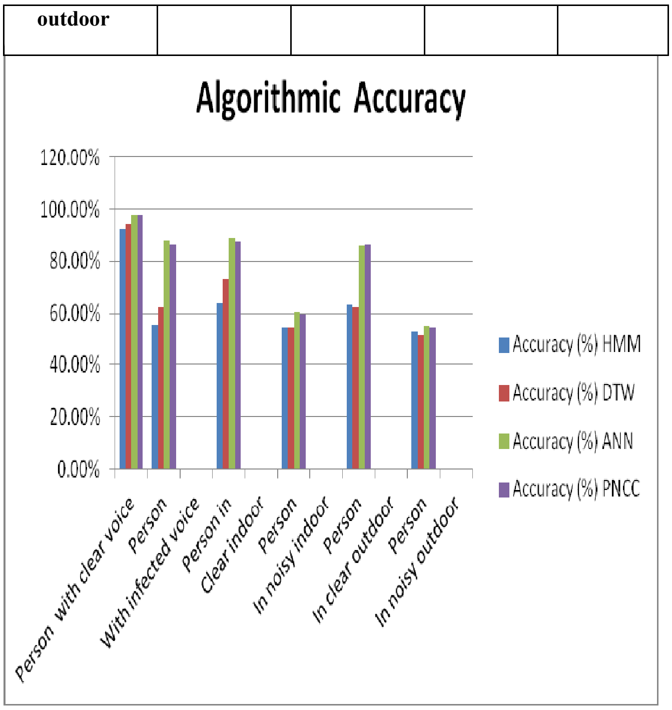

Figure 5: Graphical Representation of Comparative Study of Algorithms

Figure 5 illustrates the test results [18] related to accuracy in voice recognition of speakers with different ambient in various algorithms. The false acceptance ratio will be high during slow speed speaking or unwanted background noise higher than the speaker. Voice Rejection occurs when many cross talks produces in the microphone and output is silent more than 30 seconds during the test process. Table 1 shows artificial neural networks (ANN) and power normalized cepstral coefficients (PNCC) exhibit better efficiency compared with other classical algorithms. In the proposed real time ANN speech recognition system, the accuracy rate of speech recognition increases even with increase environmental noise during human to device or machine communication. All the existing neural networks has over all aurracy of more than 85%, due to the combined models using Power Normalized Cepstral Coefficient (PNCC) and Modified Group Delay Function (ModGDF), SVM or Gaussian models. ANN shows increased accuracy rate in all the noisy environment.

VI . Conclusion

The paper analysis the several Machine Learning Algorithms used for effective Speech Recognition processing that are used in automated voice recognition systems. After analyzing all the algorithms and their problems, it is very clear that the artificial neural network (ANN) and Power normalised cepstral coefficients (PNCC) is preferred for speech recognition for distance communication as compared to other analyzed algorithms. Moreover ANN gives better output while PNCC reduces the noise level in speech recognition.

References

[1] J.C. Segura Ramirez, C. Benitez, A. de la Torre, A. Rubio, "Voice activity detection with noise reduction and long-term spectra divergence estimation", IEEE International Conference on Acoustics Speech and Signal Processing, vol. 2, no. 17-21, pp. 1093-6, May 2004.

[2] J. Poruba, "Speech Enhancement based on non-linear Spectral subtraction", Proceeding of the Fourth IEEE International Conference on Devices Circuit and System, pp. T031-1-T031-4, April 2002.

[3] Nils Westerlund, Mattia Dahl, Ingvar Claesson, "Speech Enhancement using on adaptive gain equalizer with frequency-dependent parameter settings", Proceeding of the *IEEE*, vol. 7, pp. 3718-3722, 2004.

[4] Lawrence R. Rabiner, "A tutorial on Hidden Markov Model and selected applications in speech recognition", Proceedings of the IEEE, vol. 77, no. 2, pp. 172-175, February 1989.

[5] C. Ganesh Babu, R. Hari Kumar, P.T. Vanathi, "Performance Analysis of Hybrid Robust Automatic Speech Recognition System", IEEE International Conference on Signal Processing Computing and Control (ISPCC), pp. 162-165, 2012.

[6] Rangachari Sundarrajan, C. Loizou Philipos, "A noise-estimation algorithm for highly non-stationary environments", Speech Communication, vol. 48, pp. 220-231, August 2005.

[7] M.T. Balamuragan, M. Balaji, "SOPC- Based Speech to Text Conversion", *NIOS-II* Embedded processors design contest-outstanding, pp. 83-108, 2006.

[8] C.Ganesh Babu, P.T. Vanathi, "Performance Analysis of Voice Activity Detection Algorithm for Robust Speech Recognition System under Different Noisy Environment", Journal of Scientific and Industrial Research, vol. 69, no. 7, pp. 515-522, 2010.

[9] Saleh Albelwi and Ausif Mahmood ”A Framework for Designing the Architectures of Deep Convolutional Neural Networks” Entropy 2017, 19, 242; doi:10.3390/e19060242.

[10] Varghese M P, T Muthumanickam, "Review Paper on Implementation of Neural Network for FPGA System Design" , Indian Journal of Natural Sciences, Vol.11,February 2021.

[11] Z. Zhang, Z. He, G. Cao and W. Cao, "Animal Detection From Highly Cluttered Natural Scenes Using Spatiotemporal Object Region Proposals and Patch Verification," in *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 2079-2092, Oct. 2016, doi: 10.1109/TMM.2016.2594138.

[12] Francisco J. Quiles-Latorre, Andrés Gersnoviez, Manuel Ortiz-López, Francisco J. Jiménez-Álvarez, Francisco J. Montoro-García, María Brox, "Active Electronic Egg for Breeding of Endangered Birds", IEEE Sensors Journal, vol.21, no.22, pp.26086-26103, 2021.

[13] A. Singh, M. Pietrasik, G. Natha, N. Ghouaiel, K. Brizel and N. Ray, "Animal Detection in Man-made Environments," *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1427-1438, doi: 10.1109/WACV45572.2020.9093504.

[14] R. Vera-Amaro, M. E. R. Angeles and A. Luviano-Juarez, "Design and Analysis of Wireless Sensor Networks for Animal Tracking in Large Monitoring Polar Regions Using Phase-Type Distributions and Single Sensor Model," in *IEEE Access*, vol. 7, pp. 45911-45929, 2019, doi: 10.1109/ACCESS.2019.2908308.

[15] Varghese M P, T Muthumanickam," Review Paper On Energy Model For Vlsi Circuits Using Neural Networks" , Advanced Science Letters, Volume 26,Page 216-219, May 2020.

[16] A. Kamilaris, F.X. Prenafeta-Boldu, Deep learning in agriculture: a survey. Compute Electron. Agric. 147, 70–90 (2018).

[17] Sajid Shaikh, Mayur Jadhav, Naveen Nehe and Prof. Usha Verma, “Automatic Animal Detection And Warning System” International Journal of Advance Foundation and Research in Computer (IJAFRC) Volume 2, Special Issue (NCRTIT 2015), January 2015.

[18] https://www.googlecolab.com