Exploration of Machine Learning for Sound and Signal Investigation

Kaustubh Kumar Shukla[1], Sonia Rani[2], Shilpy Gupta[3], C Venkataramanan[4], D Harika[5], A Nivetha[6], Teena Khorwal[7]

[1, 6 & 7]*Department of Electronics and Communication Engineering, Dronacharya Group of Institutions, Greater Noida, Uttar Pradesh, India. dr.kkshuklaece@gmail.com*
[2 & 3]*School of Computing Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India.*
[4]*Department of Electronics and Communication Engineering, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India.*
[5]*Department of Electronics and Communication Engineering, Mohan Babu University, Tirupati, Andhra Pradesh, India.*

Abstract

This study investigates the use of machine learning methods for signal and sound data analysis and interpretation. Its objectives are to create reliable techniques for obtaining features from unprocessed audio signals, find appropriate algorithms for precise classification, regression, and anomaly detection tasks, train models on a variety of datasets, and investigate real-world applications in domains such as speech recognition, noise reduction, and medical diagnosis. Data gathering feature extraction, model selection and training, model evaluation, and practical application are all included in the technique. Future research paths in real-time signal processing and multimodal analysis, practical applications in domains including audio engineering, medical diagnostics, and environmental monitoring, and enhanced machine learning models are among the anticipated results. Speech recognition is crucial for human-computer and human-robot interaction. The Smart-Home Research aims to design a new methodology to reduce speech conversion effort, benefiting visually and physically challenged individuals and those who cannot type. The system uses raw audio signals to extract information from speech and sound, allowing users to connect with relatives, physicians, or caregivers. The study focuses on spotting and Vocabulary Continuous Speech Recognition, using an outsized system to extend robustness and adapt language and acoustic models for multisource-based recognition. This research aims to analyse the challenges of implementing speech recognition technology in various applications, such as classrooms, voice-operated robots, and smart homes. The objectives include evaluating signal performance, analysing speech recognition techniques, identifying problems with SR-mLA, analysing machine learning algorithms, and developing a new framework using the identified efficient algorithms. The research aims to improve pedagogical purposes and enhance lecture transcripts for students. The research evaluates the performance of the speech signal recognition (SR) technique, comparing algorithms like HMM, ANN, PNCC, and DTW. The proposed machine learning framework achieves 98% accuracy for 1-word and 95% for 2-word utterances, but struggles with 2-word recognition due to background noise. Future work includes performance enhancement and multilingual applications.

Keywords: Artificial Intelligence, Machine Learning, Speech recognition, Multilingual Applications, Smart-homes.

I. Introduction

Speech recognition (SR) is a technology that understands words but not their meaning, making it a significant tool for Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI). Advances in signal processing, algorithms, computational architectures, and hardware have led to significant advancements in speech recognition. Research aims to design a novel methodology to reduce effort in speech conversion[1]. SR has applications in various fields, benefiting visually and physically challenged individuals and those who cannot use their hands for typing. Despite progress, the ultimate goal is to comprehend free conversational speech uttered by any speaker in any environment[2]. To achieve fluent speech recognition, research should focus on increasing accuracy rates even with environmental noise. Speech recognition is the process of translating an auditory signal into words. Linguistic analysis is used to achieve speech understanding, as most people connect through voice[3]. Real-time neural network models are used to improve speech recognition performance. A text-to-speech system (TTS) is a computer that converts normal language text into speech, rendering symbolic linguistic representations like phonetic transcriptions into speech[4]. The TTS converts the structural analysis of input text into phoneme conversion and prosody analysis, producing a wave form of the signal that reads the text as a speech. The speech rate is often expressed in words per minute (wpm), which can be calculated by counting the total number of words and dividing it by the number of minutes of speech. Speech recordings can be recorded using smartphones or video cameras, and the required output size for audio to text converted files is determined by the speed memory used[5].

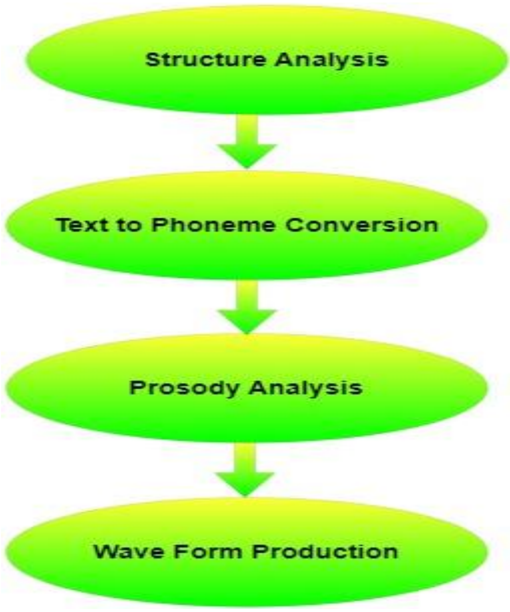

Fig.1 Model of Text-To Speech synthesizer (TTS)

The Smart-Home Research aims to design a replacement smart home system that provides assistance through natural man-machine interaction (voice command), eases social consideration, and provides security by identifying trouble situations[6]. A qualitative user evaluation was conducted on healthy individuals aged 71-88, seven relatives, and three professional careers. The voice-based solution was found to be more accepted than intrusive solutions, suggesting that audio technology has potential to ease daily living for elderly and frail persons. The system uses raw audio signals to extract information from speech and sound, either reacting to orders or modifying the environment without an order. The system can also make it easier for users to connect with relatives, physicians, or caregivers using e-lio1 or Visage2 systems.

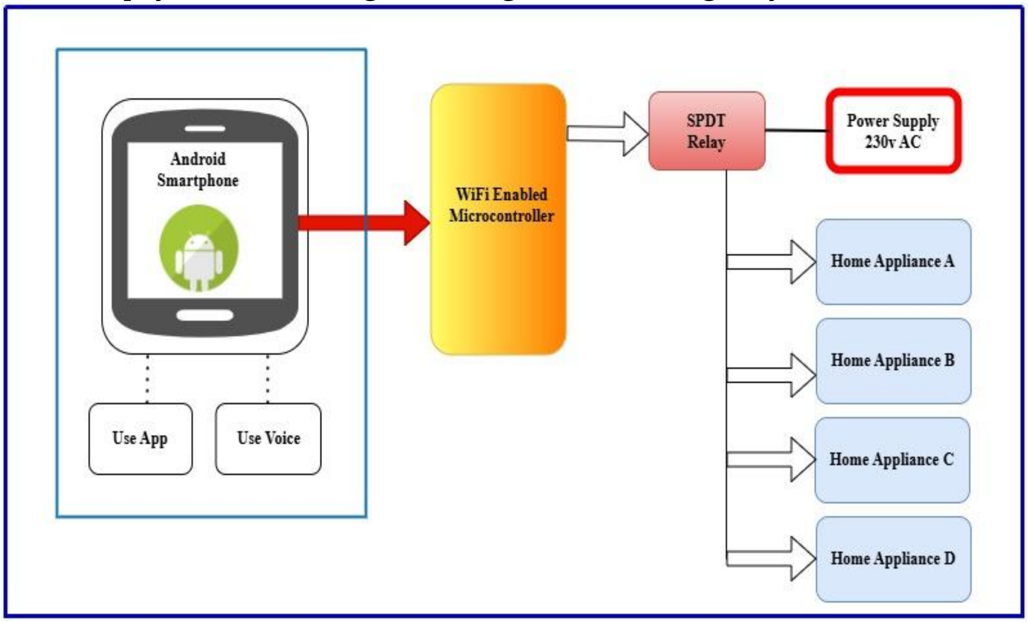

Fig.2 Overall Block diagram of Smart Home System

The research does not focus on the development of latest communication protocols between devices but instead on building communication buses and using standardized technologies and applications. Interoperability of ubiquitous computing elements may be a challenge to address[7]. The Smart-Home Research aims to design a replacement smart home system that provides assistance through natural man-machine interaction (voice command), eases social consideration, and provides security by identifying trouble situations. A qualitative user evaluation was conducted on healthy individuals aged 71-88, seven relatives, and three professional careers. The voice-based solution was found to be more accepted than intrusive solutions, suggesting that audio technology has potential to ease daily living for elderly and frail persons[8]. The system uses raw audio signals to extract information from speech and sound, either reacting to orders or modifying the environment without an order. The system can also make it easier for users to connect with relatives, physicians, or caregivers using e-lio1 or Visage2 systems. The research does not focus on the development of latest communication protocols between devices but instead on building communication buses and using standardized technologies and applications. Interoperability of ubiquitous computing elements may be a challenge to address. Speech recognition relies on noise removal techniques to cancel background noise[9]. Least Mean Square and Recursive Least Square methods are used for speech or classic music noise. Blind Source Separation (BSS) techniques are suitable for unknown noise sources like washer or blender noise. Independent Component Analysis is efficient for non-Gaussian signals but not suitable for realistic recordings[10]. Noise separation in realistic smart home conditions remains a challenge. Spoken word detection has been studied extensively, but performance decreases in noisy or spontaneous speech conditions. In fig.-1 and 2 it has been tried to showcase the entire process of the research[11]. This study focuses on spotting and Vocabulary Continuous Speech Recognition, using an outsized system to extend robustness and adapt language and acoustic models for multisource-based recognition. The approach integrates word matching directly into the ASR system, improving detection rates of demotic order[12].

II. Literature Review

Speech recognition (SR) is a recent technology used in various applications, including smart homes and unmanned military vehicles. It understands modulated sound signals rather than their meanings, making it useful in situations where voice commands may be misunderstood due to environmental noises or infected voices[13]. SR systems are classified into several types, including isolated speech, connected speech, continuous speech, and spontaneous speech. Isolated speech requires one utterance at a time, while connected speech allows separate utterances with minimal pause between them[14]. Continuous speech, also known as computer dictation, allows users to speak almost naturally without stops and pauses[15]. Unintentional speech is natural-sounding, unrehearsed speech, and an ASR system with spontaneous speech capability should be able to handle a wide range of natural speech features. Despite the advantages and benefits of SR, there are limitations to its performance. Speech recognition is an easy process for humans, but it is a difficult task for machines.

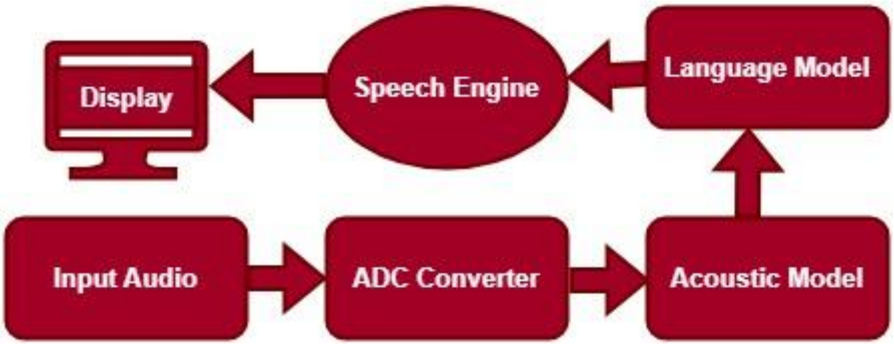

Fig.3 Speech Recognition Process

Machines require training and assistance in separating speech sounds from other sounds[16]. Problems related to SR use include discourse acknowledgment advances, ongoing inscribing (RTC), and post lecture record (PLT). RTC provides immediate presentation of teacher's discourse, while PLT uses client-free SR calculation to create sight and sound class notes with synchronized talk records, teacher sound, and class PowerPoint slides. PLT has shown more prominent word acknowledgment exactness than RTC. Speech is the main barrier in communication between individuals to express feelings, thoughts, emotions, and ideologies. Research is being conducted to improve the efficiency of speech recognition processing in various applications[17]. Some techniques used include acoustic recognition, which simulates human linguistic constraints, and phonetic typewriters, which type in response to words spoken into a microphone[18].

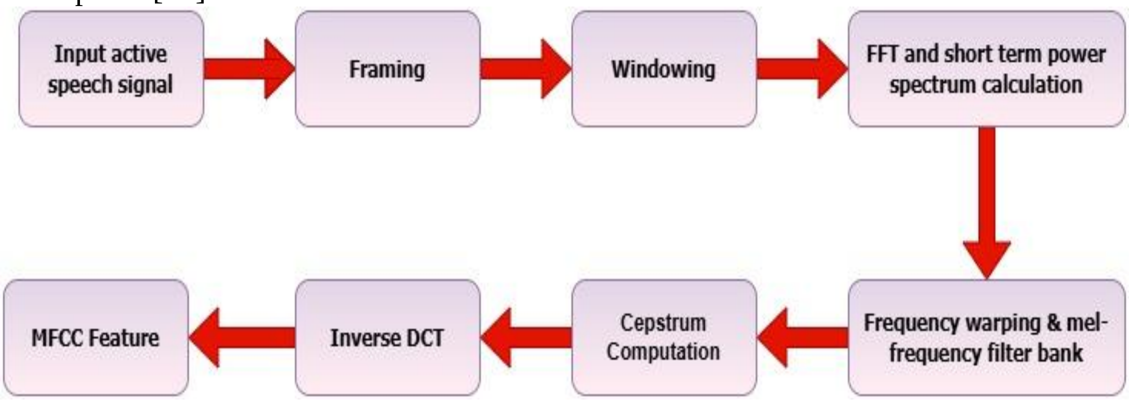

Fig.4 Overall Flow diagram of MFCC Algorithm

The study investigates how machine learning (ML) approaches can be used to enhance automation, efficiency, and accuracy in sound and signal analysis. Each of the three types of machine learning—supervised, unsupervised, and reinforcement learning—has specific uses. In these processes, deep learning—in particular, neural networks—dominates. Real-time processing, interpretability, feature extraction, and data quality are among the difficulties. Healthcare, entertainment, security, and environmental monitoring all employ machine learning (ML) to identify and diagnose problems, assess user preferences, and identify odd noises. With improvements in processing power and the availability of datasets, machine learning in sound and signal analysis has a bright future.

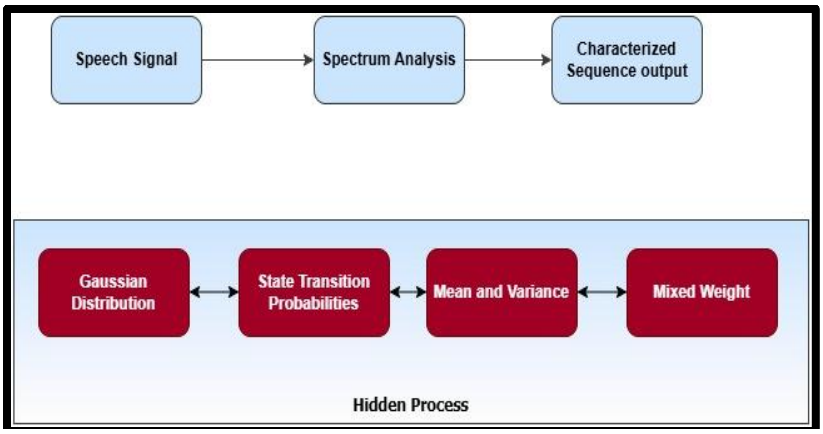

Fig.5 Block diagram of HMM for Speech Recognition

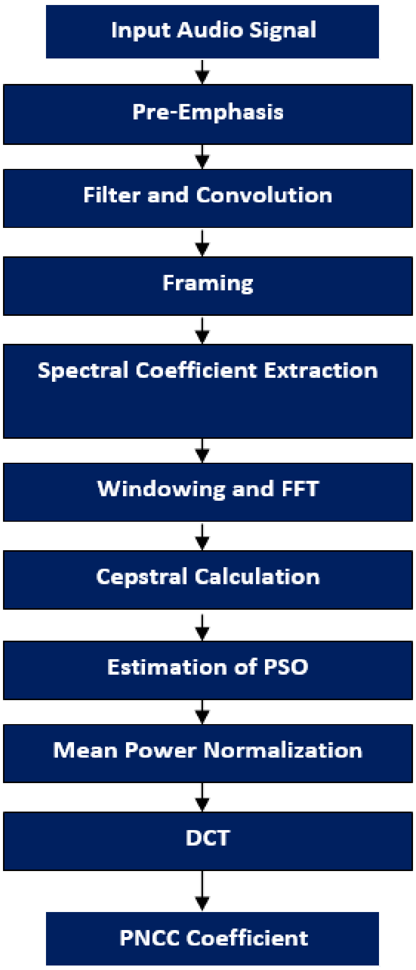

Fig.6 Power-Normalized Cepstral Coefficients (PNCC)

In fig.-3 to fig.-6 it has been tried to showcase the entire process, algorithm and major techniques of the proposed research.

III. Research Methodology

This study aims to analyze the effectiveness of various speech recognition (SR) methods and tools, evaluate signal performance, identify problems with SR-mLA, and analyze machine learning algorithms like ANN, PNCC, VQ, and DTW. Open-ended problems include automatic stoppage due to extra sounds, environmental noise, distance limitations, hands-free mode issues, low voice signal identification, unwanted background noise, and high speech signal range in crowded environments[19].

Table-1 Various Speech Recognition Techniques

Techniques	Merits	Demerits
Linear Predictive Coding (LPC)	Low resources required, easy Implementation.	Failed to distinguish words with same vowel sounds, useful for only single speaker and single language and it is reliable for small vocabulary size.
Perceptual Linear Prediction (PLP)	Discards irrelevant information of the speech and thus improves speech recognition rate.	It gives less recognition rate than MFCC and RASTA techniques.
Relative Spectral Filtering (RASTA)	Useful for multi- speakers and multi languages, and reliable for moderate size vocabulary	It requires moderate hard implementation
Mel Frequency Cepstral Coefficient (MFCC)	Useful for multi-speaker and multi languages, reliable for moderate high size vocabulary and it is easy to implement.	Surrounding noise can influence and obstruct the quality of MFCC results.

The process of turning spoken language into text is known as speech recognition, and numerous methods have been developed to improve efficiency and accuracy. The following are examples of traditional methods: Support Vector Machines (SVM), Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Dynamic Time Warping (DTW), Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Transformer Networks. An early method for voice recognition was template matching, but because it cannot adapt to changes in speech patterns, it is not viable for bigger vocabularies.

Although HMMs are fundamental statistical models for simulating speech's temporal structure, they can be computationally costly and require a lot of training data. Dynamic Time Warping calculates how similar two sequences are, however it isn't appropriate for tasks involving continuous speech or a vast vocabulary. Gaussian Mixture Models (GMM), Support Vector Machines (SVM), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) are examples of statistical approaches. While LSTM networks better handle vanishing gradients and capture long-term dependencies in speech, RNNs are better suited for sequential data.

Transformer Networks provide excellent accuracy, quicker training, and inference by processing input sequences in parallel via self-attention techniques.The goal of end-to-end speech recognition systems is to convert unprocessed audio input into transcriptions directly, bypassing step-by-step procedures like phoneme classification or feature extraction. These systems can handle continuous speech and vast vocabularies, have a streamlined pipeline, and minimize the spread of errors across stages. Although hybrid models can be computationally costly and difficult to create and implement, they integrate the best features of multiple approaches to improve performance.

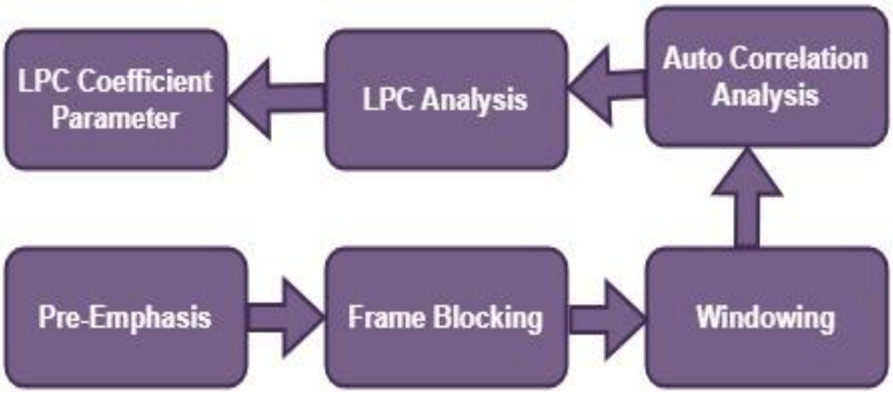

Fig.7 LPC Processing Method Diagram

Hardware/software tools used include microphones, Via Scribe-RTC (IBM), hosted transcription service (HTS)-PLT, and Python-IDLE. The goal is to analyze, highlight, and portray based on voice input signals[20]. Different methodologies and systems are compared for discourse qualities. With the help of table-1, different techniques, merits and demerits have been shown.

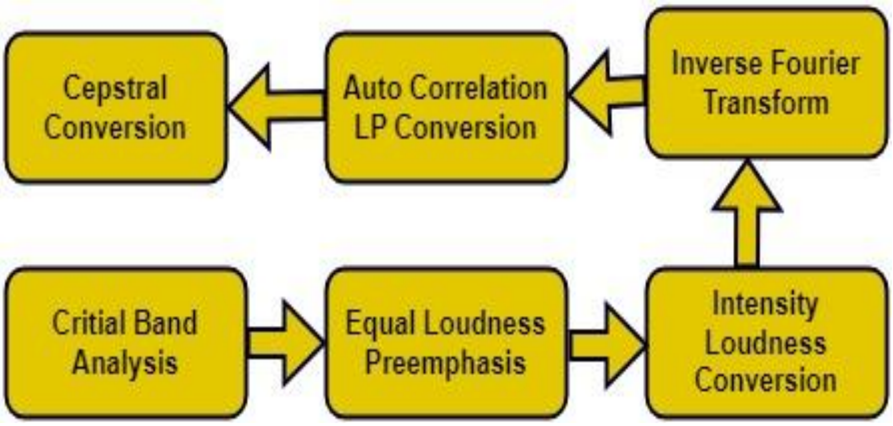

Fig.8 Perceptual Linear Prediction (PLP)

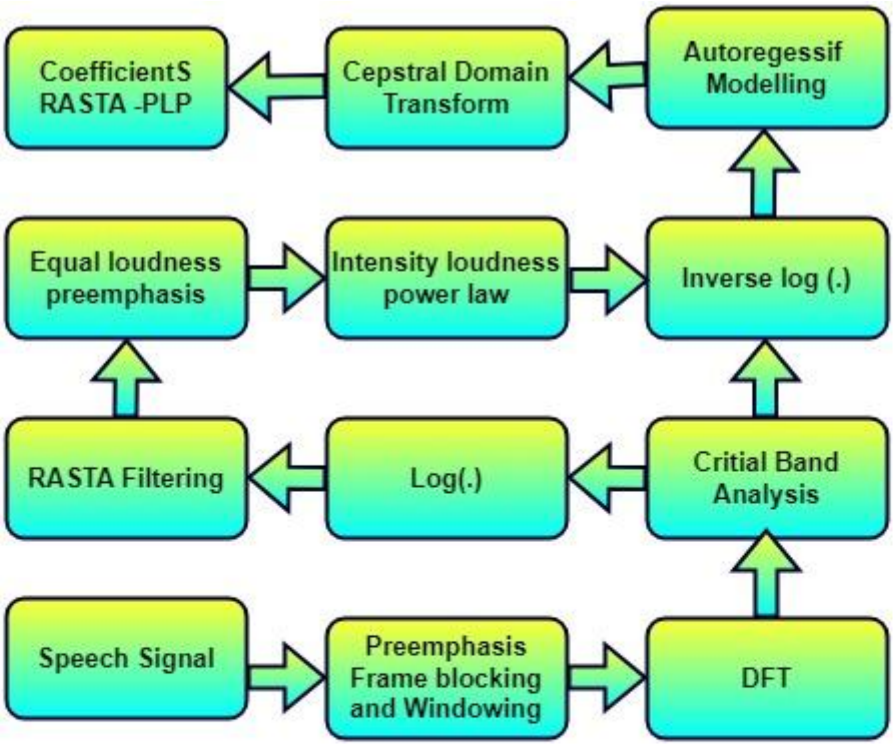

Fig.9 Block Diagram of RASTA algorithm

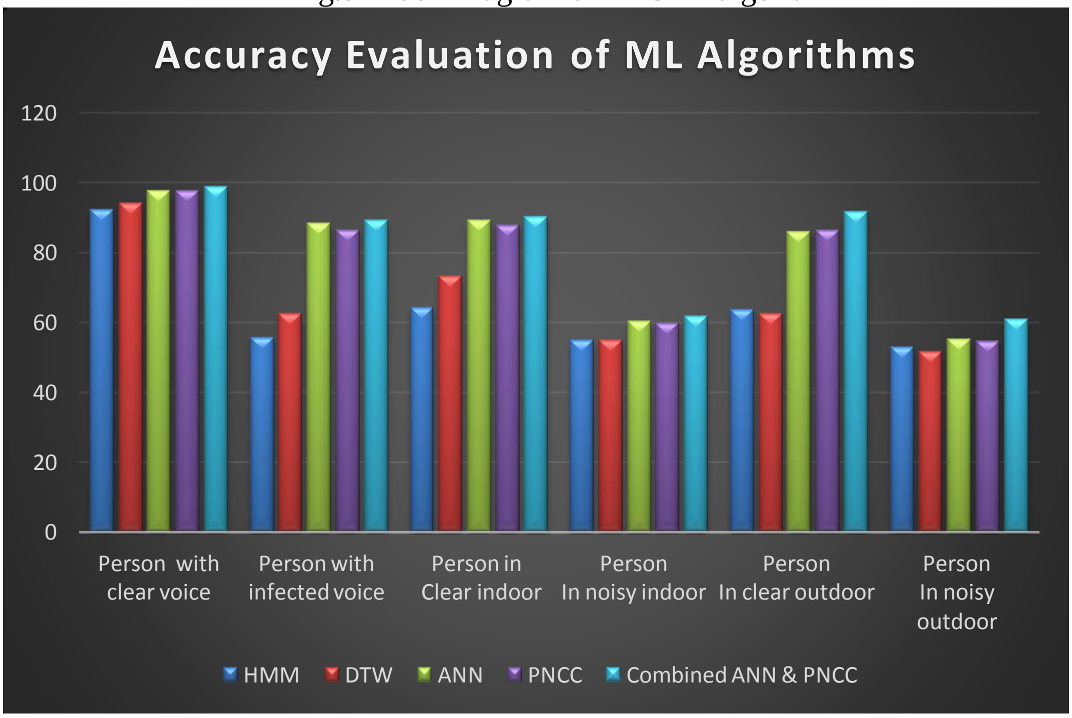

Fig.10 Graphical representation of voice recognition of speakers using various algorithms

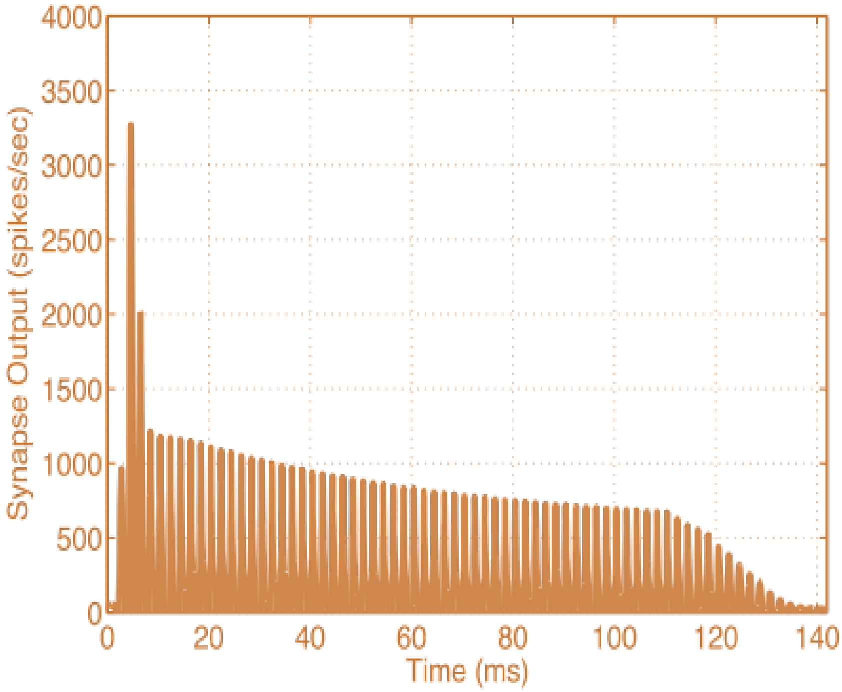

Fig.11 Output of pure tone input with a carrier frequency of 500 Hz at 60 dB SPL.

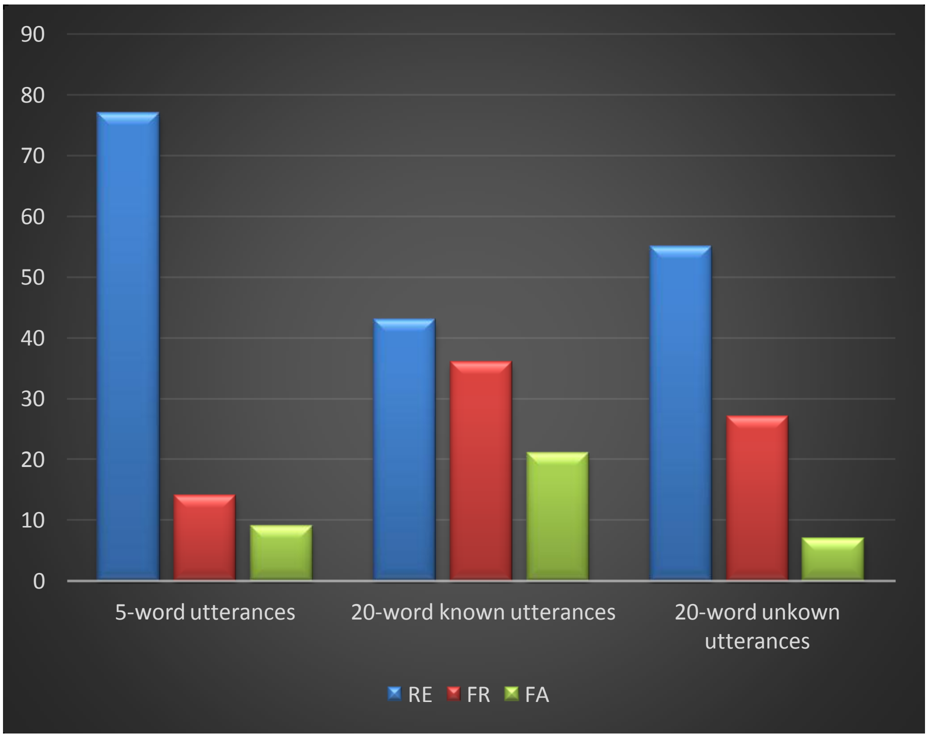

Fig.12 Word utterances analysis

In fig.-7 to fig.-12 it has been tried to showcase the entire analysis, concept and process, algorithm and major techniques of the proposed research.

IV. Results and Discussions

A machine learning pipeline for sound and signal analysis, specifically audio categorization, is demonstrated by this sample of Python code. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from audio files, the dataset is prepared, Convolutional Neural Network (CNN) architecture is defined, and the model is assembled, trained, and evaluated. For more complex audio analysis, other factors to take into account are data augmentation, hyper parameter tuning, model design, transfer learning, attention processes, self-supervised learning, and generative models. For more complicated jobs, the code takes into account strategies like time stretching, pitch shifting, hyper parameter adjustment, and transfer learning; nonetheless, it should be utilized carefully.

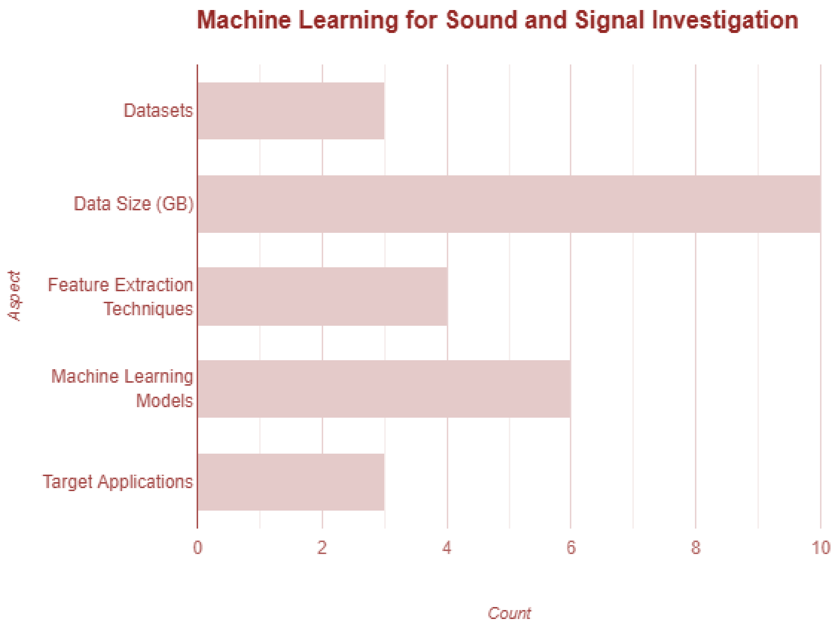

Fig. 13 Investigation of machine learning and sound signal

This research study aims to improve speech recognition performance in various applications, such as voice-operated robots, smart homes, and military vehicles[21]. Various SR methods and tools were analyzed to identify issues with SR-mLA, with Via Scribe (RTC) and HTS (PLT) being well-suited. Machine learning algorithms like Artificial Neural Networks (ANN) and Power-Normalized Cepstral Coefficients (PNCC) were found to be efficient. When combined, the combined framework produced outstanding results.

The combined framework was tested using discourse expressions from 10 unique speakers, with results showing a progress rate of 79% for 5-word utterances and 45% for 20-word utterances. However, a low recognition rate of 17% was possible for 20-word utterances due to obscure discourse expressions. The ability to perceive speakers using consistent discourse waveforms could be valuable for recognition in natural conversation frameworks. Further research is needed to expand the application to local dialects and improve the accuracy of speech recognition in real-time. In fig. 13 to 15 shown the entire analysis of different models.

Python Coding

```
import librosa
import numpy as np
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Dropout

# Load audio files and extract MFCC features
def extract_features(file_name):
   x, sr = librosa.load(file_name, mono=True, sr=22050)
   mfccs = librosa.feature.mfcc(y=x, sr=sr, n_mfcc=40)
   return mfccs.T

# Prepare the dataset
def prepare_dataset(data_path):
   X, y = [], []
   for label in os.listdir(data_path):
      label_path = os.path.join(data_path, label)
      for file in os.listdir(label_path):
         file_path = os.path.join(label_path, file)
         features = extract_features(file_path)
         X.append(features)
         y.append(label)
   return np.array(X), np.array(y)

# Create the model
model = Sequential([
   Conv2D(32, (3, 3), activation='relu', input_shape=(40, 216, 1)),
   MaxPooling2D((2, 2)),
   Dropout(0.25),
   Flatten(),
   Dense(128, activation='relu'),
   Dropout(0.5),
   Dense(num_classes, activation='softmax')
])

# Compile the model
model.compile(loss='sparse_categorical_crossentropy', optimizer='adam', metrics=
['accuracy'])

# Train the model
model.fit(X_train, y_train, epochs=10, batch_size=32, validation_data=(X_val, y_val))

# Evaluate the model
test_loss, test_acc = model.evaluate(X_test, y_test, verbose=2)
print('\nTest accuracy:', test_acc)
```

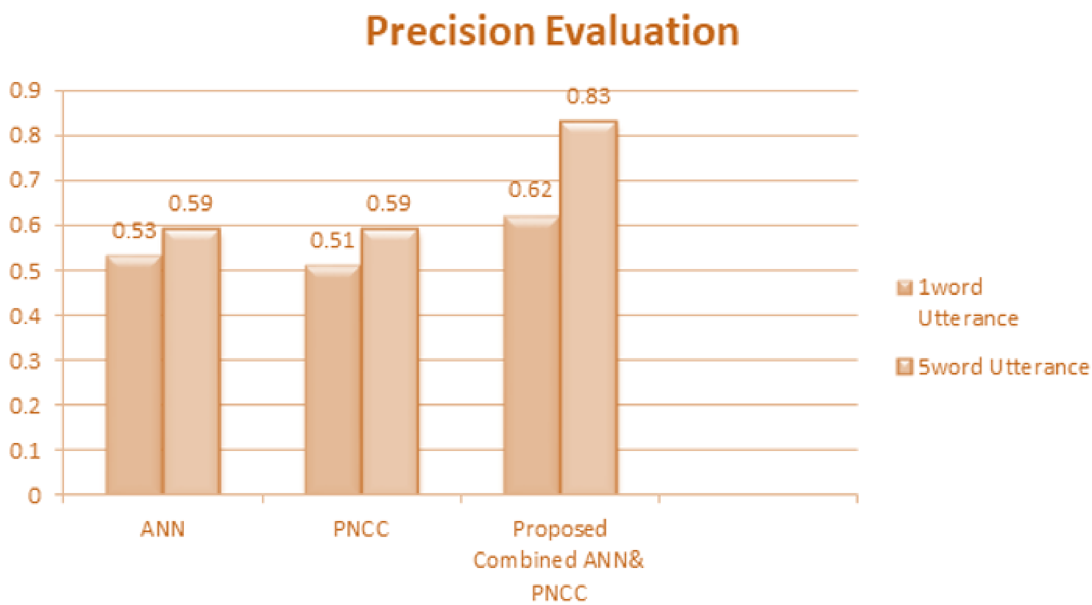

Fig.14 Precision Evaluation Analysis of ANN, PNCC and Proposed Combined ANN & PNNC models

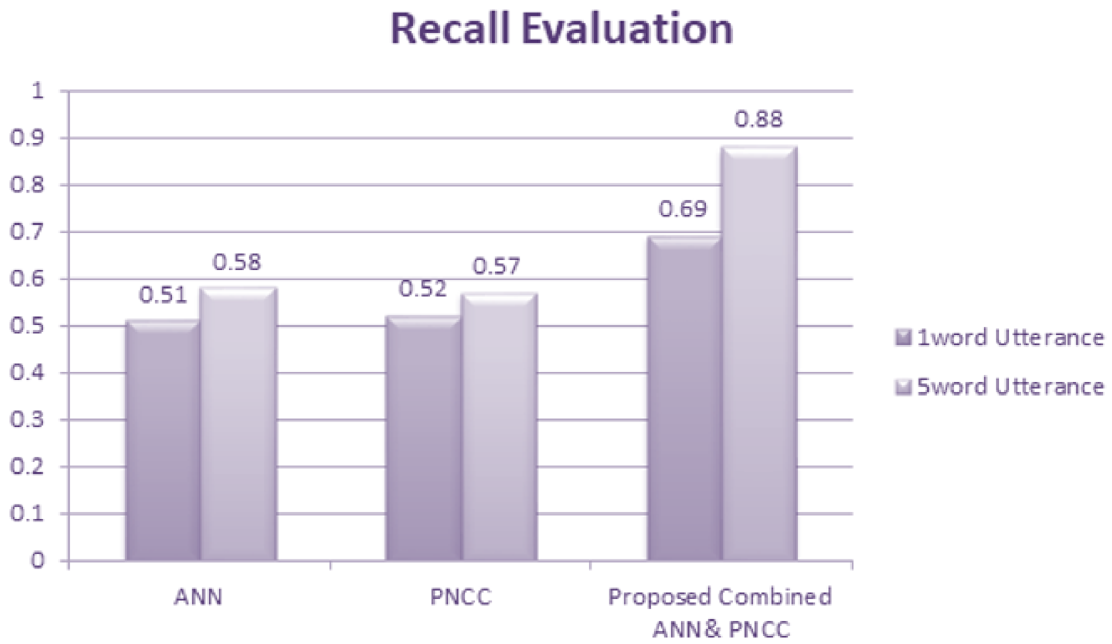

Fig.15 Recall Evaluation Analyses of ANN, PNCC and Proposed Combined ANN & PNNC models

Conclusion

This research evaluates the performance of the SR technique, comparing the accuracy, efficiency, speed, memory requirements, and complexity of existing algorithms and a proposed combined machine learning framework. The study identifies problems with SR-mLA for classroom applications and analyses the performance of machine learning algorithms like HMM, ANN, PNCC, and DTW. The proposed machine learning framework uses ANN and PNCC, achieving a 98% accuracy rate for 1-word utterances and 95% for 2-word utterances. However, a 3% recognition rate is unworkable for 2-word utterances due to high background noise. The framework is considered a novel with reasonable time complexity overheads, maximum efficiency, and better reliability for speech signal recognition. Future work includes performance enhancement, combining machine learning and deep learning models, and developing SR systems for one-word and multiple-word working in noisy environments and multi-lingual applications.

References

[1] A. Asaei, M. Cernak, and H. Bourlard, "Perceptual Information Loss due to Impaired Speech Production," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2433–2443, Dec. 2017, doi: 10.1109/TASLP.2017.2738445.

[2] Z. Lu, P. Heidenreich, and A. M. Zoubir, "Objective quality assessment of speech enhancement algorithms using bootstrap-based multiple hypotheses tests," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 4102–4105. doi: 10.1109/ICASSP.2010.5495741.

[3] L. Pfeifenberger, T. Schrank, M. Zohrer, M. Hagmuller, and F. Pernkopf, "Multi-channel speech processing architectures for noise robust speech recognition: 3rd CHiME challenge results," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, Dec. 2015, pp. 452–459. doi: 10.1109/ASRU.2015.7404830.

[4] S. Amuda, H. Boril, A. Sangwan, and J. H. L. Hansen, "Limited resource speech recognition for Nigerian English," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Mar. 2010, pp. 5090–5093. doi: 10.1109/ICASSP.2010.5495036.

[5] T. Kleinschmidt, M. Mason, E. Wong, and S. Sridharan, "The Australian English Speech Corpus for In-Car Speech processing," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Apr. 2009, pp. 4177–4180. doi: 10.1109/ICASSP.2009.4960549.

[6] S. Mishra, M. Jaiswal, K. K. Shukla, H. Mittal, S. Dubey, and B. K. Sharma, "Design and Analysis of a Novel Microbattery using Multiphysics based on Artificial Intelligence Applications," in *2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, IEEE, Jul. 2024, pp. 1–6. doi: 10.1109/ICSTSN61422.2024.10671311.

[7] H. Kasaudhan, K. K. Shukla, R. Kushwaha, K. Sharma, U. Gupta, and A. Sharma, "Early Detection and Analysis of Lung Cancer Using Artificial Intelligence," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, IEEE, Feb. 2024, pp. 1470–1474. doi: 10.1109/IC2PCT60090.2024.10486335.

[8] S. Mishra, M. M. Kamal, and K. K. Shukla, "An Overview of Artificial Intelligence for Cancer Prognosis and Treatment," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, IEEE, Feb. 2024, pp. 1465–1469. doi: 10.1109/IC2PCT60090.2024.10486792.

[9] K. Ota, E. Duflos, P. Vanheeghe, and M. Yanagida, "Speech Recognitionwith Speech Density Estimation by the Dirichlet Process Mixture," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Mar. 2008, pp. 1553–1556. doi: 10.1109/ICASSP.2008.4517919.

[10] G. Kumar, M. Post, D. Povey, and S. Khudanpur, "Some insights from translating conversational telephone speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2014, pp. 3231–3235. doi: 10.1109/ICASSP.2014.6854197.

[11] M. M. Alsulaiman, G. Muhammad, M. A. Bencherif, A. Mahmood, Z. Ali, and M. Aljabri, "Building a Rich Arabic Speech Database," in *2011 Fifth Asia*

Modelling Symposium, IEEE, May 2011, pp. 100–105. doi: 10.1109/AMS.2011.29.

[12] M. Ressl, J. Prendes, and R. Saint-Nom, "Undergraduate speech processing awareness," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Mar. 2012, pp. 2773–2776. doi: 10.1109/ICASSP.2012.6288492.

[13] Z. Guo, Q. Wang, J. Du, J. Pan, Q.-F. Liu, and C.-H. Lee, "A Variance-Preserving Interpolation Approach for Diffusion Models With Applications to Single Channel Speech Enhancement and Recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 3025–3038, 2024, doi: 10.1109/TASLP.2024.3407533.

[14] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual Alzheimer's Dementia Recognition through Spontaneous Speech: A Signal Processing Grand Challenge," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–2. doi: 10.1109/ICASSP49357.2023.10433923.

[15] J. J. Webber, C. Valentini-Botinhao, E. Williams, G. E. Henter, and S. King, "Autovocoder: Fast Waveform Generation from a Learned Speech Representation Using Differentiable Digital Signal Processing," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095729.

[16] P. Guo, H. Wang, B. Mu, A. Zhang, and P. Chen, "The NPU-ASLP System for Audio-Visual Speech Recognition in MISP 2022 Challenge," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–2. doi: 10.1109/ICASSP49357.2023.10097011.

[17] Y. Ju *et al.*, "TEA-PSE: Tencent-Ethereal-Audio-Lab Personalized Speech Enhancement System for ICASSP 2022 DNS Challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2022, pp. 9291–9295. doi: 10.1109/ICASSP43922.2022.9747765.

[18] A. Hossain, K. Das, P. Khan, and M. F. Kader, "Erratum to 'A BCI system for imagined Bengali speech recognition' [Machine Learning with Applications 13 (2023) 100486]," *Mach. Learn. with Appl.*, vol. 15, p. 100532, Mar. 2024, doi: 10.1016/j.mlwa.2024.100532.

[19] A. Hossain, K. Das, P. Khan, and M. F. Kader, "A BCI system for imagined Bengali speech recognition," *Mach. Learn. with Appl.*, vol. 13, p. 100486, Sep. 2023, doi: 10.1016/j.mlwa.2023.100486.

[20] C. Mansoor, S. K. Chettri, and H. Naleer, "Advancing Heart Disease Prediction through Synergistic Integration of Machine Learning and Deep Learning Techniques," in *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, IEEE, May 2024, pp. 1–6. doi: 10.1109/ACCAI61061.2024.10602447.

[21] S. Murugesan, A. CM, A. Bhuvanesh, S. Manoj, P. E. Darney, and K. S. Rajan S, "Smart Home Automation with Smart Metering using ZigBee Technology and Deep Belief Network," in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, IEEE, Jan. 2024, pp. 988–992. doi: 10.1109/IDCIoT59759.2024.10467342.