International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 05, ISSN No. 2455-2143, Pages 151-156
Published Online September 2024 in IJEAST (http://www.ijeast.com)

HUMANOID ROBOT ON MILITARY APPLICATIONS

Sujendra G Bharadwaj, Shruthi B
Department of ECE
SJB Institute of Technology,
Bengaluru, Karnataka, India

Abstract— **The rapid advancement of technology has heightened the demand for developing robots capable of performing complex tasks akin to human abilities. Humanoid robotics is at the forefront of current robotics research, as it aims to enable robots to interact with the environment and people in complex, dynamic ways. This field emphasizes social interactions, like gesture-based communication and collaborative activities, which integrate both physical and social engagement. This integration, essential for three-way interaction, seeks to merge physical and social dynamics at fundamental levels.**
In military contexts, technological innovation has created a need for robots capable of performing tasks akin to those of a soldier. This project focuses on building a humanoid robot designed to execute essential military operations through the use of speech recognition technology.

Keywords— **Humanoid, Robot, Speech Recognition, AI**

I. INTRODUCTION

Humanoid robots are designed to resemble human physiology, with components like a head, torso, arms, and legs, making them well-suited to environments made for humans. As the need grows for machines that are human-friendly and versatile, humanoid robots are increasingly engineered to replicate authentic human expressions, interactions, and movement. This capability allows them to blend more naturally in human- centered settings, providing consistency and precision in ways that humans may find challenging. Notable examples of humanoid robots include AMECA and ALTER3.

Humanoid robots are a subset of professional service robots developed to perform tasks through mimicking human interaction and motion. Their ability to automate processes not only enhances productivity but also leads to significant cost savings, making them a valuable asset across various industries. Although they are relatively new in professional services, humanoid robots are finding use in diverse applications that expand each year.

One major application of humanoid robots is in hazardous environments, such as power plants, where they conduct inspections, maintenance, and respond to emergencies. These robots help alleviate human workers from dangerous and strenuous tasks, thereby improving workplace safety. In space exploration, humanoid robots are being tested for routine

Fig 1.1 Humanoid Robot

functions that would typically require astronaut involvement, supporting tasks that can be time-consuming and labor-intensive

Beyond industrial use, humanoid robots have impactful roles in personal and social domains. They are being used to provide companionship for elderly and ailing individuals, offering social engagement and support. As guides or receptionists, they help facilitate customer interactions in public spaces. There is even potential for humanoid robots to aid in medical research, possibly hosting human tissue for growing transplant organs. Their ability to perform a vast range of roles, from rescue missions to compassionate care, positions humanoid robots as versatile tools for the future

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 05, ISSN No. 2455-2143, Pages 151-156
Published Online September 2024 in IJEAST (http://www.ijeast.com)

II. PROPOSED METHODOLOGY

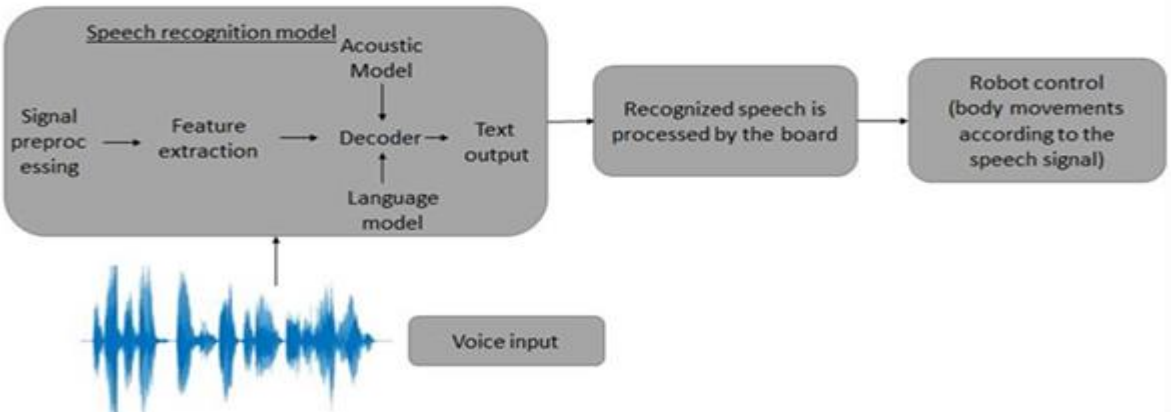


Fig 2.1: Block diagram of Speech to Robot body movement conversion

Methodology Steps:
Step 1: The voice signal is provided as input to the speech recognition system.
Step 2: The speech recognition model incorporates algorithms designed for processing speech signals, involving three primary stages:

Signal Pre-processing:
Signal processing is an area focused on analysing, modifying, and creating signals, including sound, imagery, and scientific data measurements. Techniques within signal processing serve to enhance transmission clarity, optimize digital storage, correct signal distortions, improve video quality, and identify key components in a measured signal.

Background of Signal Pre-processing:
According to researchers like Alan V. Oppenheim and Ronald W. Schafer, foundational signal processing concepts can be traced back to numerical analysis techniques of the 17th century. Later, digital adaptations of these methods emerged within digital control systems during the 1940s and 1950s. In 1948, Claude Shannon's landmark paper, "A Mathematical Theory of Communication," published in the Bell System Technical Journal, provided essential theories for information transmission and signal processing.
Throughout the 1960s and 1970s, signal processing developed significantly, and in the 1980s, digital signal processing became widespread due to advancements in dedicated signal processor chips.
Signal pre-processing is critical in establishing a reliable speech or speaker recognition system. This stage separates voiced sounds from unvoiced or silent sections using methods like short-time energy analysis and distribution-based techniques.

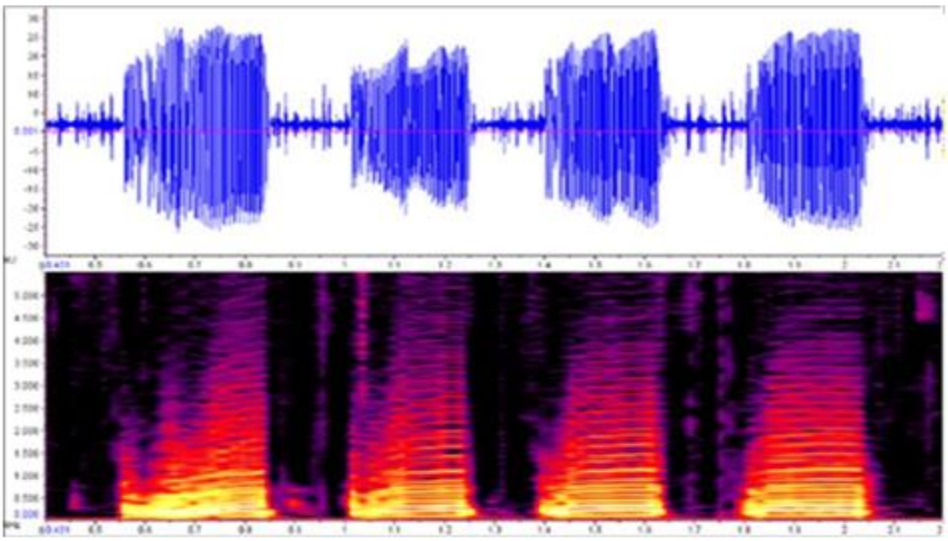


Fig 2.1: Signal Pre-processing

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 05, ISSN No. 2455-2143, Pages 151-156
Published Online September 2024 in IJEAST (http://www.ijeast.com)

Short-time Energy basis :
STE-based methods for separating speech signals operate on the principle that energy levels are typically higher in voiced segments than in silent or unvoiced segments. In this approach, a Gaussian-shaped window is applied to the signal. Using the one-dimensional Mahalanobis distance function, which acts as a linear pattern classifier, the voiced portions can be isolated from the signal.
Initially, a 200ms window is selected to determine the parameters of the Gaussian distribution, as this duration allows for ambient noise capture before the speaker begins talking (often taking more than 200ms to start speaking after recording begins). The Gaussian distribution in a single dimension can then be defined as follows:
Where, μ is the mean and σ is the standard deviation of the distribution. It can be calculated that the probabilities obey:
Based on probability calculations, there's a 99.7% chance that the distance r will be below 3. To apply the Mahala Nobis Distance Method for extracting voiced portions, a 200ms window (corresponding to 4410 samples at a sampling rate of 22,050 samples per second) is used to calculate the average and spread of the distribution. This window length is selected assuming the speaker needs over 200ms to begin speaking after recording starts, thus capturing ambient noise within the window. Following this, the Mahala Nobis distance is computed for each sample beyond the initial window. If the calculated distance exceeds 3, the sample is considered part of the voice and retained; otherwise, it is discarded.

Feature Extraction:
Feature extraction is a crucial step in speech recognition as it condenses a large dataset into a simplified yet meaningful form, enhancing its reliability and discriminatory power compared to the original signal. In speech recognition, feature extraction generally produces a multidimensional feature vector for each speech sample. Two primary methods are used for acoustic measurement: temporal domain approaches and parametric approaches, such as linear prediction, which is designed to model the resonant characteristics of the human vocal tract that produces speech sounds.
Linear Prediction Coefficients (LPC) is a technique often used to simulate the human vocal tract, generating a robust set of speech features. This method estimates the speech signal by modeling the formants, then removing these effects from the signal, which enables calculation of the residual frequency and amplitude. Each sample is represented as a linear combination of past samples, and the coefficients of this equation correspond to the formants, which are key frequency components in the speech signal. LPC is widely recognized as an effective tool for analyzing and estimating these formant frequencies.
The formant frequencies, or peaks in resonance, are identified through this technique by calculating LPC coefficients across a sliding window and locating the peaks in the resulting spectrum. This makes LPC particularly useful for encoding high-quality speech at low bit rates. Additionally, LPC analysis can yield other useful features, including Linear Prediction Cepstral Coefficients (LPCC), Log Area Ratio (LAR), Reflection Coefficients (RC), Line Spectral Frequencies (LSF), and Arcus Sine Coefficients (ARCSIN). LPC is commonly used in speech reconstruction, as well as in applications involving musical instruments, telephony, and robotic sound analysis.
The linear prediction approach also identifies the filter coefficients that represent the vocal tract by minimizing the difference, or mean square error, between the actual speech and its predicted form. Essentially, this technique models each speech sample as a weighted sum of previous samples, enabling accurate prediction and analysis of speech over time.

$$\hat{s}(n)=\sum p_k = a_k s(n-k)$$

where $\hat{s}$ is the predicted sample, s is the speech sample, p is the predictor coefficients.
The prediction error can be expressed as:

$$e(n)=s(n)-\hat{s}(n)$$

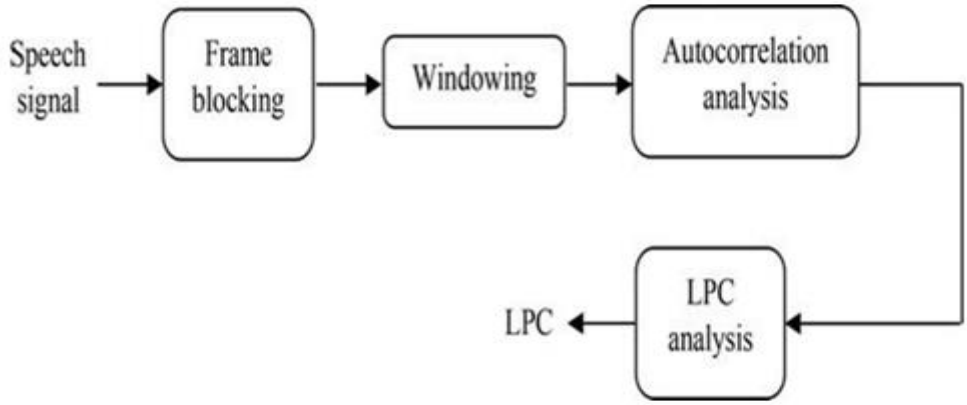
In the next step, each segment of the windowed signal undergoes autocorrelation analysis, where the maximum autocorrelation value indicates the order of the linear prediction analysis. Following this, Linear Prediction Coefficient (LPC) analysis is conducted, transforming each segment of autocorrelation into a set of LPC parameters, which include the LPC coefficients. The relationship between these coefficients and the reflection coefficients can be defined as:

$$am=\log[1-km/1+km]$$

where, am is the linear prediction coefficient, km is the reflection coefficient.

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 05, ISSN No. 2455-2143, Pages 151-156
Published Online September 2024 in IJEAST (http://www.ijeast.com)

Block diagram of LPC processor,



Steps in Linear Prediction Coding (LPC):

- **Frame Blocking:** The input signal is segmented into multiple frames, allowing for the extraction of useful characteristics while processing at a reduced rate.
- **Windowing:** This step involves converting the signal into a time-domain representation of a fixed duration, which helps in analyzing the signal more effectively.
- **Autocorrelation Analysis:** In this analysis, the windowed signal is compared to its delayed counterpart. A specific point on both signals is selected to assess the degree of similarity between them.

LPC Analysis: The analysis focuses on identifying the formants present in the signal.
Linear predictive analysis is adept at extracting vocal tract characteristics from speech, noted for its computational efficiency and accuracy. Additionally, it is utilized in speaker recognition systems, primarily aimed at extracting vocal tract properties. However, traditional linear prediction methods may encounter issues with aliased autocorrelation coefficients. LPC estimates are particularly sensitive to quantization noise, which can limit their generalization capabilities.

Decoder:
The decoder processes the feature-extracted signal and generates the corresponding text output. It comprises two main components:

Acoustic Model
Language Model

Acoustic Model:
The acoustic model processes audio input and generates a probability distribution for each character in the alphabet. It is crucial in automatic speech recognition as it defines the relationship between audio signals and the phonemes or other linguistic units that constitute spoken language. A phoneme is a unique sound unit that differentiates words in a language; it is the smallest component of a word that, when altered, changes the word's meaning. For example, "thumb" and "dumb" differ by the substitution of the "th" phoneme with "d." Phonemes can vary depending on who is speaking. These variations, known as allophones, arise from differences in accent, age, gender, the phoneme's position within the word, or the speaker's emotional state. Phonemes serve as the foundational units that speech recognition algorithms arrange in the correct sequence to form words and sentences. Speech recognition typically applies two methods to achieve this ordering.
Linguistics is the structured study of human language, focusing on accurate and objective analysis of its structure and essence. This field explores both the cognitive and social dimensions of language. Key areas of linguistic analysis align with distinct elements in human language systems, including syntax (sentence structure rules), semantics (meaning), morphology (word structure), phonetics (speech sounds and gestures in sign languages), phonology (the sound system of a language), and pragmatics (the role of social context in meaning). Subfields like biolinguistics, which examines biological factors and the evolution of language, and psycholinguistics, which studies psychological aspects of language, help bridge these areas.
The acoustic model is developed from audio recordings paired with their text transcriptions. These recordings and transcriptions are processed to generate statistical representations of the sounds that compose each word.

Language Model:
The language model transforms these probabilities into coherent words and phrases by assigning likelihoods to them, based on statistical data from training datasets. A language model calculates the probability of an entire sequence for any series of words with length m. Language models are trained on extensive text corpora in one or more languages, which allows them to calculate probabilities. Due to the potential for an infinite number of valid sentences in the language (known as digital infinity), language models must address the

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 05, ISSN No. 2455-2143, Pages 151-156
Published Online September 2024 in IJEAST (http://www.ijeast.com)

challenge of assigning probabilities to legitimate sequences that might not be present in the training data.
Solutions include the use of the Markov assumption and advanced neural architectures like recurrent neural networks or transformers.
Language models have diverse applications in computational linguistics. They originated in speech recognition to prevent low-probability or nonsensical word sequences from being selected. Their applications now span areas such as machine translation (for evaluating possible translations), natural language generation (to produce more human-like text), part-of-speech tagging, parsing, optical character recognition, handwriting recognition, grammar induction, and information retrieval, among others. In information retrieval, language models are applied in the query likelihood model, with each document in a collection linked to its language model.

III. EXPERIMENT AND RESULTS

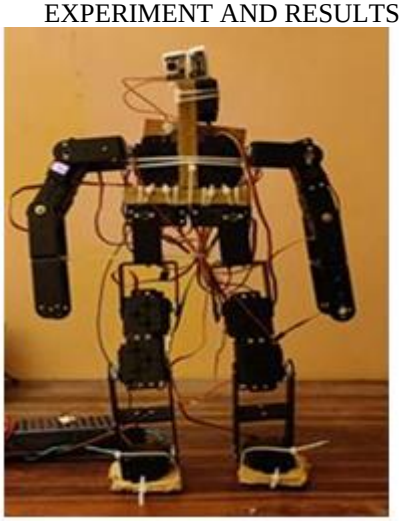


Fig 3.1 Front View

Fig 3.1 shows the front view of the robot model.
All the servo motors are in their initial position. That is zero degrees.
Calibrating the servo motors to make the bot move forward. The angles are as follows: - Right thigh servo - 45 degrees (approximately)

Right thigh knee servo - 32 degrees (approximately) Due to gravity, the right foot drops down.
Left thigh servo - 45 degrees (approximately) Left knee servo 32 degrees(approximately)
Due to gravity, the left foot drops down so that the robot has moved forward.

Fig 3.2 Side View

F**igure 3.2** shows the side view of the robot model and the action performed in shooting.

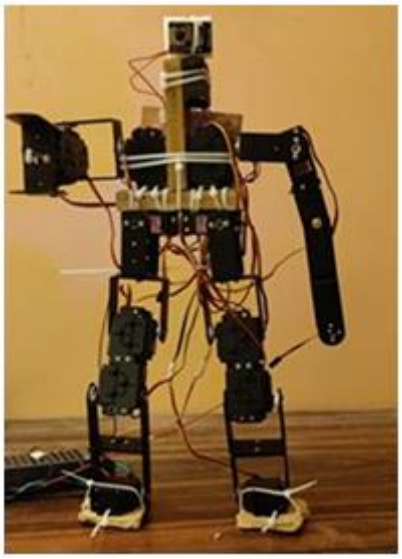


Fig 3.3 Shooting view

Figure 3.3 shows the shooting view of the robot.
Calibrating the servo motors to make the bot shoot. The angles are as follows: - Right shoulder servo 90 degrees (approximately)
Aim the gun and shoot.

IV. CONCLUSION

In conclusion, this project demonstrates the design and implementation of a voice-controlled humanoid robot that can perform basic actions like walking, firing, and standing, making it particularly useful for military applications. By employing an Arduino Blue controller with a speech-to-text converter and an ESP32 board for command processing, this setup enables the robot to respond directly to user commands in real time. Such

developments in humanoid robotics are pivotal, as they offer a more natural and intuitive way of interaction, opening doors for numerous applications across fields like manufacturing, healthcare, and agriculture.

The robot's capacity for precise, deterministic behavior also addresses the increasing demand for reliable and efficient automation solutions, particularly in environments where human productivity and safety are priorities. As advancements in robotic technology continue, this approach underscores the potential of humanoid robots to contribute meaningfully to both industrial and social contexts, reinforcing their role as valuable companions and collaborators in modern society. Future work could further explore expanding the robot's capabilities and refining its interactions, ultimately enhancing its adaptability and usefulness in diverse applications.

V. REFERENCE

[1]. Banerjee, A., & Roy, M. (2023). "Automatic Speech Recognition for Affective Computing in Robots." Pattern Recognition Letters, 169 (pp. 12-20).

[2]. Brown, T., & Kim, J. (2023). "Speech-Driven Multimodal Interaction for Social Robots." Robotics and Automation Letters, 9(3), (pp. 540-549).

[3]. Chen, Y., & Li, Z. (2023). "Advancements in Real-Time Speech Recognition for Humanoid Robots." Robotics and Autonomous Systems, 150, (pp. 32-40).

[4]. Garcia, F., & Huang, Y. (2023). "Advances in Voice Command Recognition for Personal Robots." IEEE Access, 12, (pp. 10456-10467).

[5]. Garcia, L., & Huang, K. (2023). "Speech Recognition for Autonomous Robots in Industrial Environments." Robotics and Computer-Integrated Manufacturing, 78, (pp. 201-210).

[6]. Kim, D., & Choi, H. (2023). "End-to-End Speech Recognition Models for Robotics Applications." IEEE Robotics and Automation Magazine, 30(2), (pp. 79-87).

[7]. Kumar, A., & Mehta, S. (2023). "Speech and Emotion Detection for Healthcare Robots." Biomedical Signal Processing and Control, 89, (pp. 21-30).

[8]. Lee, M., & Park, S. (2023). "Robust Speech Recognition for Smart Home Applications Using RNNs." Journal of Speech Technologies, 18(1), (pp. 52-62).

[9]. Li, H., & Yang, Z. (2023). "Noise-Resilient Speech Recognition for Smart Assistive Robots." IEEE Transactions on Neural Networks and Learning Systems, 35(5), (pp. 1130-1140).

[10]. Nguyen, T., & Tran, V. (2023). "Speech-Driven Control Systems for Assistive Robots." IEEE Transactions on Cognitive and Developmental Systems, 15(3), (pp. 203-213).

[11]. Park, Y., & Lee, S. (2023). "Improving Speech Recognition Accuracy in Multilingual Humanoid Robots." Speech Communication, 142, (pp. 56-65).

[12]. Patel, R., & Kumar, S. (2023). "Deep Learning Approaches for Speech Recognition in Noisy Environments." Speech Communication, 135, (pp. 45-55).

[13]. Shen, T., & Li, Y. (2023). "Machine Learning Approaches to Speech Recognition for Enhanced HRI." Journal of Human-Robot Interaction, 16(4), (pp. 345-354).

[14]. Singh, P., & Rao, T. (2023). "Speech Recognition-Based Human-Robot Interaction Models." International Journal of Robotics Research, 42(3), (pp. 211-219).

[15]. Wilson, G., & Anderson, J. (2023). "Automatic Speech Recognition Models for Low-Resource Languages." IEEE Transactions on Audio, Speech, and Language Processing, 31, (pp. 123-130).