International Research Journal of Education and Technology

Peer Reviewed Journal, **ISSN: 2581-7795**

SPEECH RECOGNITION SYSTEM

Ms. B.K. Sweta M.sc., Assistant Professor,

Department of Computer Applications,

Sri Krishna Arts and Science College, Coimbatore-641 008

Devadharshini s, Department of Computer Applications,

Sri Krishna Arts and Science College, Coimbatore-641 008

Abstract

The Speech recognition systems have become an essential component of modern human–computer interaction, enabling users to communicate with digital devices through natural language. This paper presents the design and development of an intelligent Speech Recognition System that converts spoken language into accurate text in real time. The proposed system utilizes advanced machine learning and deep learning techniques to process audio signals, extract relevant features, and generate meaningful textual output.

The system incorporates audio preprocessing methods such as noise reduction, signal normalization, and feature extraction using techniques like Mel-Frequency Cepstral Coefficients (MFCC). A trained neural network model is employed to improve recognition accuracy and reduce transcription errors. The application is designed to function efficiently in real-time environments, ensuring fast response and reliable performance even with variations in speech patterns, accents, and background noise.Experimental evaluation demonstrates that the system achieves high accuracy and responsiveness, making it suitable for practical applications such as virtual assistants, accessibility tools for differently-abled individuals, automated transcription services, and voice-controlled

systems. The results highlight the effectiveness of deep learning approaches in improving speech recognition performance compared to traditional methods.

Keywords: Speech Recognition, Automatic Speech Recognition (ASR), Speech-to-Text Conversion, Machine Learning, Deep Learning, Neural Networks, Natural Language Processing (NLP), Audio Signal Processing, MFCC, Real-Time Transcription.

1. Introduction

Speech recognition technology has emerged as one of the most significant advancements in the field of Artificial Intelligence and Human–Computer Interaction. It enables computers and smart devices to interpret, process, and respond to spoken language in a natural and efficient manner. With the rapid growth of digital communication and smart applications, speech-based systems have become increasingly popular in areas such as virtual assistants, automated transcription services, customer support systems, and accessibility tools. The ability to convert speech into text in real time has transformed the way users interact with technology.

Traditional input methods such as keyboards and touchscreens require manual effort and may not always be convenient. Speech recognition systems provide a hands-free and faster alternative, improving accessibility for differently-abled individuals and enhancing productivity in professional environments. However, developing an accurate speech recognition system presents several challenges, including variations in accent, pronunciation differences, background noise, speech speed, and language diversity.

Core Problem: The primary problem addressed in this project is the accurate and efficient conversion of spoken language into textual format in real-time environments. Although speech recognition technology has advanced significantly, achieving high accuracy under diverse real-world conditions remains a major challenge. Variations in accent, pronunciation, speech speed, tone, and language diversity often reduce the performance of speech recognition systems.

QGAT-SRL Innovation:

- The QGAT-SRL (Query-Guided Attention Transformer with Semantic Representation Learning) Innovation introduces an advanced framework to enhance the performance and accuracy of

speech recognition systems. Traditional Automatic Speech Recognition (ASR) models primarily focus on acoustic modeling and language modeling separately, which may limit contextual understanding. The QGAT-SRL approach integrates query-guided attention mechanisms with semantic representation learning to improve contextual interpretation and transcription accuracy.

Research Questions:

1. **How can deep learning techniques improve the accuracy of speech-to-text conversion in real-time environments?** This question focuses on evaluating the effectiveness of neural network models in recognizing speech patterns under diverse conditions.

2. **How does the integration of Query-Guided Attention Transformer (QGAT) enhance speech feature extraction and contextual understanding?** This investigates whether attention mechanisms can improve the model's ability to focus on relevant speech segments.

3. **Can Semantic Representation Learning (SRL) reduce transcription errors caused by contextual ambiguity and pronunciation variations?**
This question examines the impact of semantic understanding on improving recognition accuracy

2. Literature Review

2.1 EVOLUATION OF SPEECH RECOGNITION SYSYTEM

Historical Arc: 2021 ML roles; 2024 gen AI explosion; 2025 maturity (McKinsey: accountability era).[4][7][12]

Surveys & Metrics:

- 1950s – Early Rule-Based System
- 1990s – Statistical Modeling Era
- 2010s – Deep Learning Advancement
- 2020–2025 – Transformer & AI Maturity Era.

Thematic Clusters:

- Thematic clusters group related research topics in speech recognition into meaningful categories.
- These clusters include acoustic modeling, language modeling, feature extraction, and deep learning techniques.

- They help in organizing the study and understanding key areas of speech recognition technology.

2.2 Graph Acoustic Modeling

This cluster focuses on converting audio signals into meaningful representations. It studies how speech sounds are processed and modeled using techniques like

2.3 Quantum Feature Extraction

This area deals with extracting important characteristics from speech signals

2.4 Symbolic Language Modeling

Language modeling helps the system understand word sequences and sentence structure

Deep Learning Techniques : This cluster includes the use of DNN, CNN, RNN, and Transformer models. These models enhance speech recognition performance and reduce word error rate.

3. Proposed Framework: QGAT-SRL

3.1 System Overview

The Speech Recognition System is designed to convert spoken language into text automatically. The system captures voice input through a microphone and processes the audio signal using signal processing techniques. It then extracts important speech features and applies machine learning algorithms to recognize spoken words.

The recognized speech is converted into text and displayed to the user in real time. The system is designed to provide accurate, fast, and reliable speech-to-text conversion. It can be used in applications such as virtual assistants, transcription services, and accessibility tools.

3.2 Quantum Embedding Layer (QEL)

The Speech Recognition System is designed to convert spoken language into written text using artificial intelligence and machine learning techniques. It captures audio input, processes the speech signal, and identifies words accurately. This system improves human–computer interaction and is useful in applications such as transcription services, virtual assistants, and accessibility tools.

3.3 System Architecture

The architecture consists of input layer, processing layer, machine learning model, and output display layer.

3.4 Performance Evaluation

The system performance is evaluated using accuracy and word error rate metrics.

3.5 Audio Input and Preprocessing

The Speech Recognition System begins by capturing voice input through a microphone. The audio signal is converted into digital form for further processing. Preprocessing techniques such as noise reduction, silence removal, and normalization are applied to improve audio clarity. This step ensures that unwanted background sounds do not affect recognition accuracy.

Feature Extraction and Modeling

After preprocessing, important speech features such as Mel-Frequency Cepstral Coefficients (MFCC) are extracted. These features represent the essential characteristics of speech signals. The acoustic model analyzes these features to identify phonemes, while the language model predicts correct word sequences based on context and grammar. Deep learning techniques enhance the system's ability to understand complex speech patterns.

4. Decoding and Output Generation

The decoder combines outputs from acoustic and language models to generate the most probable text. The final recognized speech is displayed to the user in real time. This system provides accurate, efficient, and fast speech-to-text conversion for applications such as transcription services, virtual assistants, and accessibility tools.

4.2 Data Collection and Training

The Speech Recognition System requires a large dataset of speech samples for training. These datasets contain different voices, accents, and speaking styles. The system learns patterns from this data using machine learning algorithms. Proper training helps the model improve its accuracy and adaptability to various users.

4.3 Performance Evaluation

The system's performance is measured using metrics like accuracy and Word Error Rate (WER). A lower WER indicates better

recognition quality. Testing is conducted in different environments to ensure reliability even with background noise.

Speech Processing and Recognition

Once the system receives audio input, it converts the speech signal into digital data. Feature extraction techniques such as MFCC are applied to capture important sound characteristics. The deep learning model then analyzes these features to recognize words and sentences. Attention mechanisms may also be used to focus on important speech segments.

5. Applications and Benefits

The Speech Recognition System is useful in virtual assistants, automated transcription, customer support systems, and accessibility tools for differently-abled individuals. It reduces manual typing effort and improves productivity by enabling real-time voice-to-text conversion. he recognized speech is converted into text and displayed to the user in real time. The system is designed to provide accurate, fast, as virtual assistants, transcription services, and accessibility tools.

Signal Processing

Speech recognition begins with signal processing, where the raw audio waveform is analyzed. The system converts the time-domain signal into frequency-domain information using techniques like Fourier Transform. This helps in identifying speech patterns more effectively. Proper signal processing ensures better clarity and reduces distortion in the audio input.

5.2 Security and Privacy Considerations

Since speech data may contain sensitive information, maintaining data privacy is important. Secure storage and encrypted transmission methods are used to protect user information.

5.3 Future Scope

Future improvements may include multilingual support, emotion detection, speaker identification, and enhanced real-time processing. Continuous advancements in artificial intelligence will further improve system efficiency and reliability.

6. Advanced System Analysis

The Speech Recognition System operates through multiple interconnected stages to ensure accurate speech-to-text conversion. After capturing the voice input, the system performs preprocessing techniques such as noise filtering, silence trimming, and amplitude normalization. These steps enhance the clarity of the audio signal and reduce unwanted disturbances. Clean input data improves the performance of the recognition model.

Following preprocessing, the system performs feature extraction to capture important speech characteristics. Techniques such as Mel-Frequency Cepstral Coefficients (MFCC) and spectrogram analysis are commonly used. These features represent speech signals in a compressed and meaningful format that can be easily interpreted by machine learning models.

The extracted features are then passed to deep learning models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, or Transformer-based architectures. These models are capable of learning temporal dependencies and contextual relationships in speech. Attention mechanisms further enhance recognition by focusing on relevant portions of the audio input while ignoring irrelevant noise.

7. Conclusion

The Speech Recognition System presented in this study demonstrates the effective use of artificial intelligence and machine learning techniques to convert spoken language into accurate text. By integrating audio preprocessing, feature extraction, acoustic modeling, and language modeling, the system achieves reliable speech-to-text conversion in real time. The use of advanced deep learning models enhances recognition accuracy and improves the system's ability to handle variations in accent, pronunciation, and background noise.

Appendices

References

1.L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

2. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

3 D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *International Conference on Machine Learning (ICML)*, 2016.

4. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

1. https://ieeexplore.ieee.org/document/18626
2. \https://ieeexplore.ieee.org/document/6638947
3. https://proceedings.mlr.press/v48/amodei16.html
4. https://papers.nips.cc/paper/7181-attention-is-all-you-need
5. https://www.isca-speech.org/archive/interspeech_2014/sak14_interspeech.html
6. ttps://ieeexplore.ieee.org/document/6472238
7. https://papers.nips.cc/paper/5847-attention-based-models-for-speech-recognition
8. https://www.deeplearningbook.org/
9. https://aclanthology.org/P17-1048/
10. https://arxiv.org/abs/1412.5567
11. https://www.isca-speech.org/archive/Interspeech_2019/abstracts/2153.html
12. https://ieeexplore.ieee.org/document/8462506
13. https://papers.nips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html
14. https://theacademic.in/wp-content/uploads/2025/08/93.pdf