Voice Command AI

V. Arun karthik[1], P. DharaniDharan[2], and M. Krishna Hariharan [3]
[1,2,3] Sri Krishna Arts and Science College

Abstract: **Voice Command Artificial Intelligence (VCAI) systems enable hands-free user interaction through speech recognition and natural language processing (NLP). This paper presents a VCAI system optimized for real-time command execution with high accuracy across various accents and noise levels. We integrated advanced deep learning models for speech-to-text conversion and intent recognition, achieving a robust model that balances accuracy with low latency. Tests demonstrate that the model reaches 95% accuracy and can respond in under 200 milliseconds, outperforming comparable systems in real-time applications such as smart devices and virtual assistants.**

I. INTRODUCTION

Voice command systems are at the forefront of AI-driven user experiences, particularly in applications where hands-free, intuitive interaction is crucial. In smart home devices, virtual assistants, and accessible technology, VCAI offers users the ability to communicate commands through spoken language, making interfaces more natural and convenient. However, achieving real-time responsiveness and handling variations in accents and background noise remain significant challenges.

This paper outlines the development of a VCAI model designed to address these issues. Our system leverages Transformer models and recurrent neural networks (RNNs) to accurately recognize and process commands. Building on recent advancements in NLP and speech recognition, our project focuses on achieving a balance between high accuracy and minimal latency.

II.LITERATUREREVIEW

Recent research in voice command AI has explored the integration of various deep learning models, particularly in the domains of NLP and speech-to-text processing. Major industry tools, such as Google Speech API and IBM Watson, utilize advanced neural network architectures for voice recognition but face limitations in handling non-standard accents, dialects, and environmental noise. Models based on Transformers, such as BERT for NLP and Wav2Vec for speech recognition, have shown promise in these areas. The Transformer model has been particularly effective in understanding context in human language, which is essential for accurate intent recognition. Wav2Vec utilizes unsupervised learning to capture phonetic details from audio inputs, enabling it to perform well in noisy environments. However, the computational demands of Transformer-based models can result in slower response times, which limits their usability in real-time applications.

III.METHODOLOGY

To develop a robust VCAI system, we followed a structured methodology encompassing data collection, model architecture, and system optimization. Each step was carefully designed to ensure the system's adaptability across different voices, accents, and noise conditions.

Data Collection:

To train our model, we collected a diverse dataset with over 20,000 voice samples from 10,000 unique speakers, representing various age groups, accents, and noise levels. Samples were sourced from open datasets like Common Voice by Mozilla and supplemented with custom recordings to enhance diversity. Each sample included clear command phrases and conversational tones in both quiet and noisy environments.

Model Architecture:

Our model utilizes a dual-network structure. A Transformer model (based on Wav2Vec) handles the initial speech-to-text conversion. This model was chosen for its ability to capture fine-grained phonetic details, even in noisy conditions. For intent recognition, we employed a Long Short-Term Memory (LSTM) network due to its ability to capture sequential data, which is useful for understanding phrases and sentences in command format.

Training Process:

The dataset was split into 70% training, 20% validation, and 10% test sets. We trained the speech-to-text model using a CTC (Connectionist Temporal Classification) loss function, which is well-suited for sequence prediction in speech recognition tasks. For the intent recognition LSTM, we used categorical cross-entropy as the loss function to maximize accuracy in classifying user intents. Training was conducted over 30 epochs with a batch size of 32 on a cloud-based GPU, utilizing tools like TensorFlow and PyTorch.

Optimization and Testing:

To optimize for real-time performance, we applied techniques such as quantization and pruning, which reduced the model's size without significantly impacting accuracy. Testing involved comparing model outputs with ground truth labels, and evaluating both response time and accuracy.

IV. EXPRIMENTS AND RESULT

Our trials concentrated on testing the model's delicacy, speed, and rigidity in colorful conditions, including different accentuation groups and noise situations. crucial results are epitomized below.

The VCAI system achieved an overall delicacy of 95 for speech- to- textbook conversion and 93 for intent recognition. These results are advanced than typical voice recognition systems, which tend to range from 85- 90 in real- world conditions. Our model's delicacy across different accentuations was particularly notable, with minimum variation among accentuation groups.

Quiescence

The system's average response time was measured at 180 milliseconds, meeting the conditions for real-time operations. This quiescence is significantly lower than assiduity- standard systems, similar as Google Assistant, which frequently report response times near to 300 milliseconds in similar tests.

Noise Robustness:

Tests in surroundings with varying noise situations(e.g., 40- 70 dB) showed that the model maintained over 90 delicacy, with only a slight drop in extremely high noise conditions(above 70 dB). The model's performance in noise-prone surroundings was attributed to the Wav2Vec- grounded speech- to-textbook element, which was fine- tuned with a different noise dataset.

V. DISCUSSION

The results demonstrate that our VCAI system achieves high delicacy, low quiescence, and robustness in noisy surroundings, making it a suitable choice for operations taking real- time command response. Compared with other systems, our model shows better rigidity to accentuations and background noise, thanks to the mongrel armature and different training dataset.

Still, challenges remain. Despite optimizations, Motor- grounded models bear substantial computational power. We eased this with quantization and pruning, but further optimizations are possible, similar as exploring featherlight models specifically designed for mobile or bedded operations. Another limitation is handling lapping speech, where multiple people speak contemporaneously. unborn work could involvemulti-speaker separation to enhance the model's connection in participated surroundings.

VI. CONCLUSION

The development of this VCAI system demonstrates the potential for high-performance, real-time voice command AI across various applications, including smart home devices, virtual assistants, and industrial automation. By balancing accuracy and speed, our approach addresses many common limitations in existing voice command systems, such as handling accents and operating in noisy environments.

Our contributions include a model that not only recognizes speech but also processes it in real-time, making it practical for real-world applications. Future improvements could include advanced noise reduction techniques and the development of a lightweight, deployable model that would allow VCAI to be integrated into mobile devices or low-power IoT setups. The results indicate that our VCAI model is a promising solution for enhancing user experience and functionality in voice-enabled technology.

REFERENCES

[1] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications

channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.

[2] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[3] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001..