International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

CONTEXT-AWARE AUTOMATIC SPEECH RECOGNITION: INTEGRATING NLP FOR ENHANCED TRANSCRIPTION AND SUMMARIZATION

Para Rajesh[1], Amaravarpu Pramod Kumar[2], Yuvaraju Macha[3] , Madhavi Ravinuthala[4]

[1]Assistant Professor, Dept. of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology,Hyderabad,Telangana-500090,India, e-mail: rajesh_p@vnrvjiet.in,

ORCID: 0000-0001-9056-5918

[2]Assistant Professor, Dept. of CSE (Cys,DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering and Technology,Hyderabad,Telangana-500090,India, e-mail:amaravarapupramod@gmail.com,

ORCID: 0000-0001-5861-8960

[3]Associate Professor, Dept. of Mathematics, Matrusri Engineering College,Hyderabad,Telangana-500059,India, e-mail: yuvaraju.macha@matrusri.edu.in,
ORCID: 0000-0001-5890-1993

[4]Assistant Professor, Dept. of Mathematics, Matrusri Engineering College,Hyderabad,Telangana-500059,India, e-mail: ravinuthalamadhavi@matrusri.edu.in,
ORCID: 0000-0001-6356-0249

Abstract

The creation of a contemporary, improved automatic speech recognition system utilizing NLP approaches is presented in this research article. Accurate speech-to-text systems are becoming crucial for a variety of sectors and everyday applications in today's information-driven, fast-paced environment. The capacity to automatically transcribe and comprehend spoken language is essential for everything from virtual assistants and transcription services to customer service and accessibility solutions. The following essential needs in the contemporary global environment serve as the driving forces behind our project: Demand for precise voice recognition, contextual comprehension for organic dialogues, effective information processing and summary, and accessibility for a range of audiences has increased. From being able to react to a small number of sounds to being able to comprehend spoken language with ease, automatic speech recognition (ASR) has advanced dramatically. The desire to automate human-machine interaction has generated a lot of interest in this technological breakthrough. Nowadays, voice search, virtual assistants, and speech-to-text systems are just a few of the many applications that employ ASR extensively, greatly improving user experience and productivity. As evidence of the amazing progress made in this area, it started with simple sound recognition and has since progressed to complete language understanding. A proposed solution to this problem is the integration of Natural Language Processing (NLP) methods into ASR systems. ASR systems can improve their

Received: August 09, 2025

capacity to identify and comprehend spoken language in its larger context by including contextual understanding features obtained from NLP models.

Keywords— Automatic Speech Recognition (ASR), Whisper AI, Fine-tuning, BERT, NLP.

I. INTRODUCTION

Even though Automatic Speech Recognition (ASR) technology has advanced significantly, current systems frequently have trouble accurately transcribing spoken language in situations where contextual signals and semantic nuances are important. Traditional ASR systems primarily rely on acoustic features and language models based on statistical patterns, which may lead to suboptimal performance when confronted with complex linguistic structures and varied contextual information. The problem arises from the inherent limitations of conventional ASR approaches to comprehensively understand and interpret the contextual nuances present in spoken language. The inability of current systems to successfully use contextual information leads to transcription errors, misunderstandings, and inaccuracies, particularly when ambiguous or context-dependent speech is involved. We derived the project's premise from the basic paper, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova are the authors of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). This study aided in our investigation into how to enhance transcription accuracy and contextual relevance by incorporating contextual understanding from natural language processing (NLP) into ASR systems.

This project aims to develop an advanced Automatic Speech Recognition (ASR) system using HuBERT to deliver accurate speech-to-text transcriptions. The ASR system will be integrated with transformer models like GPT or T5 to enhance contextual understanding of the transcribed speech, allowing for more nuanced and precise interpretations. A key feature of the system will be the ability to summarize spoken content by leveraging the contextual understanding capabilities of the transformers, enabling concise and meaningful summaries of the transcriptions. Even with higher speech input volumes, the complete solution will be constructed as a reliable and scalable pipeline, guaranteeing that text processing and ASR activities are managed effectively and fluidly. Furthermore, a user-friendly interface will be created to improve use by enabling simple audio input and rapid access to precise transcriptions and summaries. The system's performance will be continuously evaluated and optimized, using metrics like Word Error Rate (WER) for ASR accuracy, ensuring continuous improvements in both transcription quality and summarization efficiency.

II. LITERATURE SURVEY

BERT: Advancing Language Comprehension through Deep Bidirectional Transformers:

BERT transformed the natural language processing field by introducing a deep bidirectional architecture that enables the model to understand context by looking at both preceding and following words. In a variety of NLP applications, this bidirectional feature improves performance. Strong unsupervised pre-training is a crucial component of effective language comprehension systems, as evidenced by the benefits of transfer learning with pre-trained

Received: August 09, 2025

language models, even for low-resource applications. By extending these discoveries, BERT extends these benefits to bidirectional architectures, enabling the use of a single pre-trained model to address a variety of NLP applications.

Training RNN Models with Limited ASR Data Using ASR Confusion Networks:

ASR language model training calls for creative methods in situations when there is little manual transcription or little voice data, such as privacy-sensitive applications or languages with limited resources. This work investigates the training and modification of RNN language models (LMs) using ASR confusion networks using unlabeled speech data. By examining techniques that capitalize on uncertainty in ASR transcriptions, such as choosing pathways from confusion networks and minimizing KL divergence between model predictions and confusion bin posteriors, researchers discovered notable decreases in perplexity. These results imply that LM performance can be enhanced in situations with little data.

Enhanced ASR Robustness via Large-Scale Weak Supervision:

The Whisper approach highlights the effectiveness of scaling weakly supervised pretraining for ASR, which has previously received less focus. Without relying on self-supervised and self-training methods that are prevalent in large-scale ASR research, this approach emphasizes the importance of a large, diverse dataset to achieve better zero-shot transfer and improve system robustness. This method demonstrates that leveraging a variety of supervised data can greatly enhance ASR system robustness without additional self-supervision techniques

Research on Self-Supervised Pre-Trained Models for Classifying Voice Quality from Accelerometer and Speech Signals on the Neck Surface:

In this article, the glottal source waveforms and raw signals from speech and neck surface accelerometers (NSA) are used to automatically classify voice quality into three groups: breathy, modal, and pushed. Features were extracted using three self-supervised models (wav2vec2-BASE, wav2vec2-LARGE, and HuBERT), and classification was done using SVM and CNN models. The findings demonstrated that the NSA signal was more useful for classification than speech signals and that self-supervised models were particularly beneficial in extracting features to classify voice quality. Performance was further improved by reducing the effect of vocal tract resonances with inverse filtering and the NSA signal.

Multitask Learning Neural Framework for Categorizing Sexism:

In this study, the authors introduce a neural multitask architecture designed for categorizing sexism, where tasks share layers and weights, supported by a combined loss function. As domain-specific keywords may not always be understood by the model, the authors incorporated external knowledge representations to enhance performance. In order to tackle the problem of multi-label classification, the research also investigates different transformation methods and loss functions. By employing overlapping subsets of categories, their proposed multi-label classification method outperformed several machine learning and deep learning benchmarks. The study also explores several transformation techniques and loss functions to address the multi-label classification issue. By employing overlapping subsets of categories, their proposed multi-label classification method outperformed several machine learning and deep learning benchmarks.

Investigating Self-Trained Models for Classifying Voice Quality from Speech and NSA Signals:

In this work, sound quality categorization utilizing both speech and NSA signals is investigated using features from self-supervised pre-trained models (wav2vec2-BASE, wav2vec2-LARGE, and HuBERT). The findings showed that when SVM and CNN classifiers were applied, NSA signals performed better in classification than speech signals. Among the models, HuBERT's features outperformed wav2vec2's in the classification of both NSA and speech signals.

What Information Do End-to-End Speech Models Acquire About Language, Speakers, and Channels? A Neuron-Level and Layer-Wise Analysis:

The representations that end-to-end speech models develop for tasks such as dialect identification and speaker recognition are examined in this work. The authors looked into the encoding of speaker, language, and channel properties at the layer and neuron levels. They looked at the distribution of this data and investigated if it could be captured by a small subset of the network. The distribution of channel and gender information throughout the network and the fact that complex features, such as dialect, are only encoded in task-specific pre-trained networks are important findings.

Trends and Advancements in automated Speech Recognition Research:

This paper discusses the design of automatic speech recognition (ASR) systems, focusing on how these systems might use the discriminative properties of human speech. The authors contrast this with more general machine learning methods that might not fully utilize the unique characteristics of speech signals. By examining the evolution of ASR systems throughout time, including the role of deep neural networks, the authors provide strategies for improving the precision and effectiveness of ASR. This study is appropriate for both beginners and experts in the discipline because it provides broad concepts without delving into complex methods.

Models for Advanced Speech Recognition Using Sequence-to-Sequence:

In this study, researchers demonstrate that word-piece models can be used in speech recognition tasks instead of graphemes. They offer a multi-head attention architecture that improves accuracy in addition to several optimization techniques like synchronous training and scheduled sampling. Outperforming traditional systems, the study shows a decrease in word error rate (WER) from 9.2% to 5.6% in voice search activities and from 5% to 4.1% in dictation tasks.

Training RNN Models in Scenarios with Limited Data on Uncertain ASR Hypotheses:

The training of automated speech recognition (ASR) systems in low-resource environments, with little in-domain speech data and scant text transcriptions, is examined in this study. Using uncertainty from ASR transcriptions, the authors investigate techniques for training recurrent neural network (RNN) language models. They obtained statistically substantial reductions in perplexity by decreasing KL divergence between model predictions and confusion bin posteriors and picking pathways from ASR confusion networks.

Received: August 09, 2025

Attention-Based Audio-Visual Fusion for Sturdy Automatic Speech Recognition:

This study examines whether lip motion patterns can enhance automatic speech recognition, especially in noisy settings, when paired with acoustic speech. By aligning the two modalities, the authors' audio-visual fusion strategy improves recognition accuracy. According to the results, depending on the noise level, the fusion technique improved performance on the TCD-TIMIT dataset by up to 30%.

A Comparison of Data Augmentation Techniques in vocal Pathology Detection:

The usefulness of data augmentation methods for vocal pathology detection is investigated in this research. The study demonstrates that data augmentation increases classification accuracy, especially for a 2D-CNN system, using machine learning and deep learning models. Across two databases, the SpecAugment approach, which works in the time-frequency domain, produced the largest accuracy gains.

Robust voice Recognition with Large-Scale Weak Supervision:

Whisper, a model for zero-shot transfer learning, is used in this study to examine the robustness of voice recognition systems. According to the authors, voice recognition research has devalued large-scale weakly supervised pre-training. They show that without requiring self-supervision or self-training methods, system robustness can be greatly increased through training on sizable, varied supervised datasets.

Acoustic and Linguistic Measures' Test-Retest Reliability in Speech Tasks:

The dependability of frequently used linguistic and acoustic variables in automated speech analysis is investigated in this research. The study indicates that adding more trials or using different tests can improve the reliability of automated speech diagnostics, but it shows no discernible gender variations in reliability estimates.

Proposed System to Overcome Research Gaps

The proposed system seeks to address the identified research gaps by utilizing a combination of NLP techniques and fine-tuning whisper ai, combining with BERT model.

- Data Collection: Different speech and voice data is collected and pre-processed to remove noise and irrelevant information, ensuring high-quality data for training and testing. This ensures better data quality, reducing the risk of inaccuracies due to irrelevant data.

- Accurate Speech-to-Text Conversion: Develop an ASR system utilizing Whisper AI to accurately convert spoken language into text transcriptions. This enhances transcription accuracy, ensuring reliable outputs, especially for multilingual and noisy environments.

- Context Understanding: Integrate the Whisper-based ASR system with transformer models like GPT or T5 to comprehend the context of transcribed speech. This will enable the system to interpret speech more accurately, particularly in nuanced or ambiguous situations.

- Summarization of Speech: Implement a feature that leverages the contextual understanding capabilities of GPT or T5 to generate concise summaries of transcribed speech, enhancing the usability and accessibility of information.

Received: August 09, 2025

- Robust and Scalable Pipe-line: Create a robust and scalable pipeline that integrates Whisper AI and text-processing models efficiently, ensuring smooth handling of large speech inputs without compromising performance.
- User-Friendly Interface: Provide an easy-to-use interface that lets users record audio and get accurate transcriptions and summaries. The interface will encourage accessibility and usability for a wide range of users, regardless of their level of technical expertise.
- Evaluation and Optimization: Use metrics like Whisper's ASR's Word Error Rate (WER) to continuously assess the system's performance. Over time, refine the system to increase its precision and effectiveness.

In summary, the proposed system combines modern technologies and architectures to bridge the research gaps by offering a secure, scalable, and cross-platform real-time chat application that maintains high performance even with increased user traffic.

III. METHODOLOGY

1. Data Collection & Preprocessing:

Compile a sizable and varied collection of training audio data. To increase Whisper AI's resilience, this should incorporate linguistic, accentual, contextual, and background noise variances.

Normalize and clean the audio data by eliminating extraneous noise or silence and making sure the format (sample rate, length) is consistent.

2. Fine-tuning Whisper for ASR:

Start by loading Whisper AI’s pretrained model. This ensures the base model has good general ASR capabilities before fine-tuning for your specific use case. add new tokens to Whisper's tokenizer and adjust the embedding matrix accordingly. Evaluate the model’s transcription accuracy by calculating Word Error Rate (WER) on validation datasets.

Tune hyperparameters based on validation performance to avoid overfitting on training data.

3. Contextual Understanding via Transformer Integration:

Use transformer models like GPT, T5, or BERT for understanding the context of the transcribed speech.

Fine-tune the transformer model on tasks such as sentence prediction, entity recognition, or text classification that relate to understanding the context of the audio. Extract features from the Whisper transcriptions such as sentence boundaries, punctuation, or entity markers. These can be fed as inputs to the transformer models for further context processing. Build a processing pipeline where the Whisper AI handles transcription and the transformer model (GPT or T5) processes the transcription to correct, improve, or enrich it based on context. This pipeline could involve summarizing key points, handling ambiguities in speech, or correcting out-of-context transcriptions.

Received: August 09, 2025

4. **Summarization of Speech**:

Use transfer learning on pretrained summarization models, adjusting them to recognize the structures of speech transcripts (e.g., disfluencies, filler words). Fine-tune the model to generate abstractive summaries, ensuring that it rewrites the transcription while capturing the core meaning rather than just extracting sentences. Track summarization performance using metrics like ROUGE and BLEU scores to ensure the model generates high-quality, concise summaries.

5. **User Interface:**

Design and test a user-friendly interface where users can easily input audio files, record live speech, and receive transcriptions along with summaries.

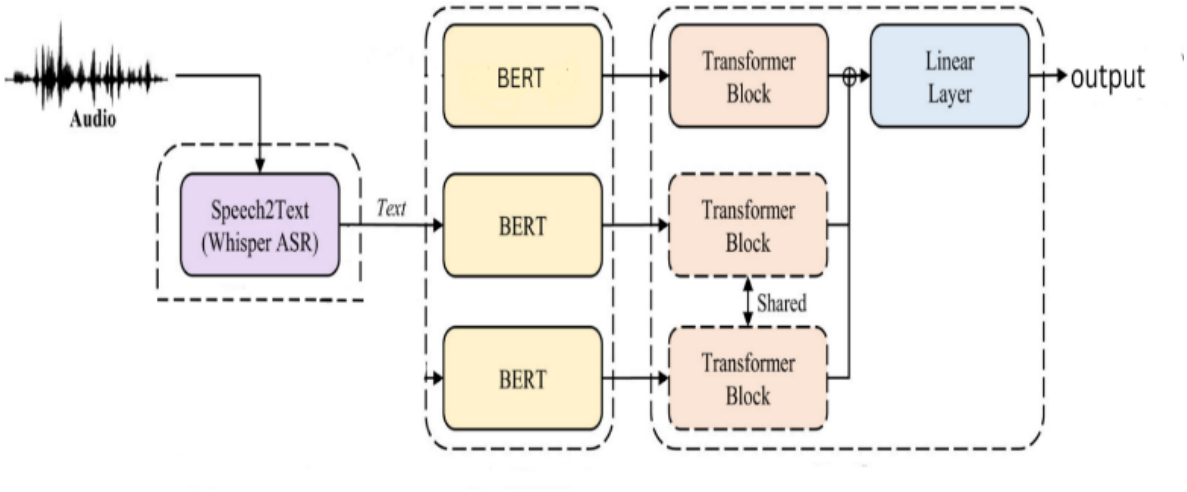


Fig. 1. System Architecture

Syntax for Whisper Tokenizer:

```
from transformers import WhisperFeatureExtractor

feature_extractor=WhisperFeatureExtractor.from_pretrai
ned("openai/whisper-small")

from transformers import WhisperTokenizer

tokenizer=WhisperTokenizer.from_pretrained
("openai/whisper-small", language="English",
task="transcribe")

from transformers import WhisperProcessor

processor=WhisperProcessor.from_pretrained
("openai/whisper-small", language="English",
task="transcribe")
```

Syntax for Sequence Training:

```
from transformers import Seq2SeqTrainer

trainer=Seq2SeqTrainer (

args=training_args,                    model=model,
train_dataset=common_voice["train"],
eval_dataset=common_voice["test"],
data_collator=data_collator,
compute_metrics=compute_metrics,    tokenizer=processor.
feature_extractor

)
```

Received: August 09, 2025

Syntax for Displaying Output:

```
def transcribe(audio):
 text=pipe(audio)["text"]
 global audioTotext
 audioTotext= contextualize(text)
 return "\n". join(["SpeechText:" + text,"ContextualTex:"
+ audioTotext])
iface=gr.Interface(
 fn=transcribe,
 inputs=gr.Audio(type="filepath"),
 outputs="text",
 title="Model",
 description="Realtime demo for speech recognition.",
 )
iface.launch()
```

- **Audio Input and Whisper ASR:**

The system begins by taking audio input, which is represented by the waveform at the top left. This audio is processed by the Whisper ASR (Automatic Speech Recognition) module, labeled as "Speech2Text (Whisper ASR)." This module converts the spoken audio into text. The output from this step is the Text, which serves as the input to the next stage.

- **BERT Layer for contextual understanding:**

After converting audio into text, the system processes the text through multiple BERT layers. In this case, three BERT blocks are shown, which take the text as input and perform different levels of contextual understanding. Each BERT block processes the input text to extact contextual information at different levels, ensuring that nuances and relationships between words are captured accurately.

- **Transformer Blocks:**

The system presents a series of Transformer Blocks that are connected to the BERT layers after the BERT processing. Every Transformer block does additional contextual refining using the BERT layer's output as input.

These Transformer blocks are responsible for improving the contextual and sequential relationships in the text. The architecture shows that the Transformer blocks share information, which indicates that these blocks likely work together to provide better understanding and correction in sequential data.

- **Shared Transformer Block:**

Received: August 09, 2025

There is a Shared Transformer Block shown in the middle, where the outputs of some BERT and Transformer blocks are fed into a shared block. This design choice suggests that multiple layers might share certain parameters or attention mechanisms to maintain efficiency and capture long-range dependencies in the text.

- **Linear Layer:**

The output is sent into a linear layer after going through the transformer blocks.

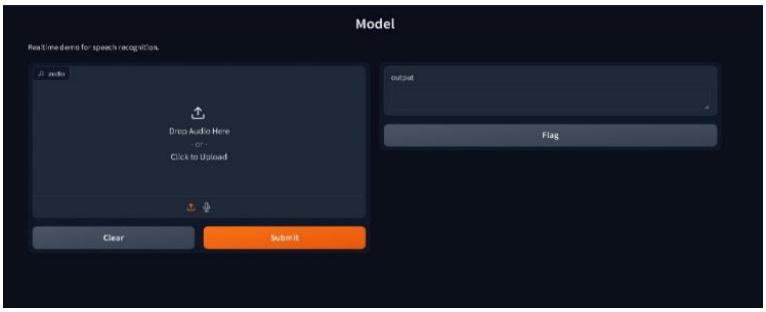
The outputs are probably transformed into a final prediction, classification, or action by the Linear Layer. Depending on the objectives of the project, it could be in charge of duties like summarizing, final transcription changes, or even phrase editing.

- **Output:**

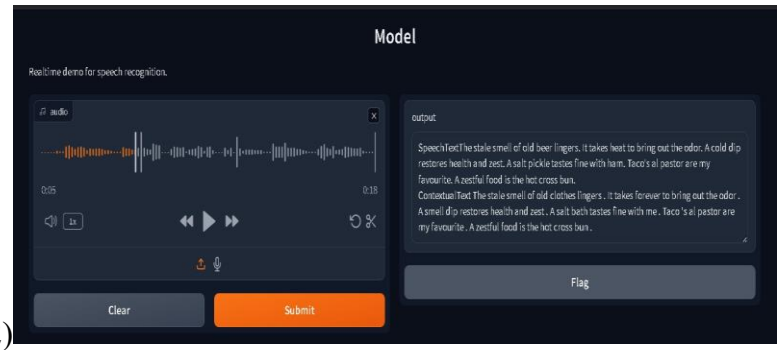
The final output is produced after processing through the linear layer. This could be the corrected transcription, a summary of the input text, or the final action depending on how the architecture is designed to handle the task.

IV. RESULTS AND DISCUSSION

The results included in graphs showing various metrics. Figure 2 and Figure 3 shows the model interface and functionality. In Figure 4, The graph represents the training loss curve over time for your ASR project using Whisper AI integrated with transformers like BERT. A plateau at the end suggests the model has reached its optimal training state on the current dataset, which means additional training may not yield significant improvements without further data or adjustments. This graph demonstrates that the model has learned effectively over time and has converged to a low training loss, which is promising for achieving accurate and context-aware transcriptions in your ASR project.

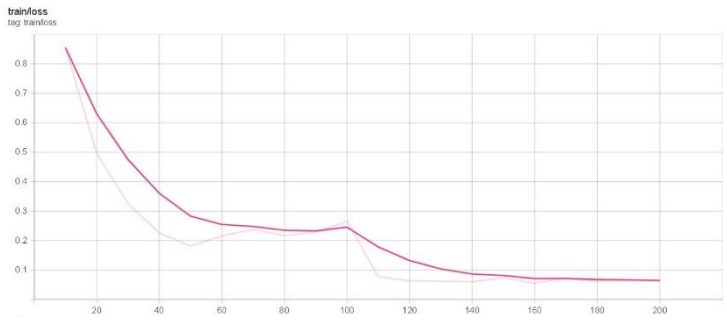


Fig(2)

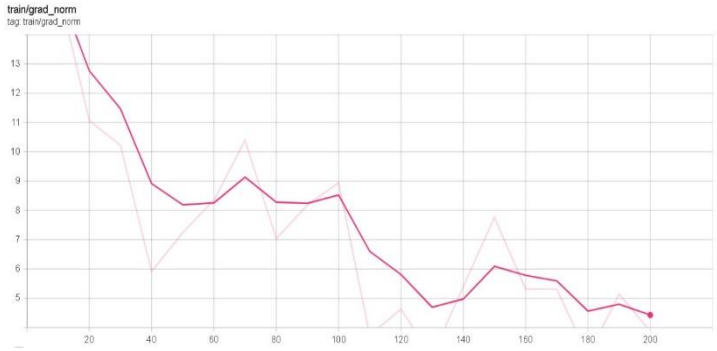


Fig(3

.in Figure 5, The graph represents the gradient norm over the training epochs for your ASR project using Whisper AI with transformer-based context models. This graph shows that your training pipeline is robust and stable, and the model's adjustments become smaller as it nears optimal performance The gradient norm graph is an indicator of stable and controlled training in your project, ensuring that the model learns effectively without major issues in gradient updates. This supports the reliability of the transcription and contextual understanding in your ASR system.



Fig(4)



Fig(5)

Table 1. Key Performance Metrics for ASR Model

S.NO	TEST CASE	EXPECTED OUTPUT	OBSERVED OUTPUT	RESULT
1	Test audio input files	Provides audio with varying accents and handles background noises.	The system transcribes languages accurately	PASS

2	Test for ambiguous or context-sensitive phrases	Provides audio with a correct interpret	The system adjusts its contextual understanding based on type of speech and transcribes accordingly	PASS
3	Test for summarization	Generates a concise, relevant summary of the key points.	The system recognizes topics shifts and generates a coherent summary covering all main points	PASS
4	Test for Performance Metrics (accuracy, scalability… Etc.)	Provides large volumes of speech data	The system processes large inputs without crashing	PASS

V. CONCLUSION

In this project, we successfully developed an enhanced Automatic Speech Recognition (ASR) system using Whisper AI for speech-to-text conversion, integrated with contextual understanding through BERT and Transformer models. The system demonstrated a strong ability to accurately transcribe spoken language into text and correct errors based on contextual cues. The architecture allowed for a seamless flow from audio input to processed text output, ensuring both accuracy and relevance in the transcriptions. The contextual understanding aspect, powered by BERT and Transformer blocks, allowed the model to handle complex language constructs, including homophones, nuanced phrases, and diverse accents, making the system robust in real-world scenarios.

Apart from transcription, the system also included summary capabilities, making use of transformers' context-aware properties to provide succinct and insightful summaries of the spoken material. Overall, this work shows the potential of combining state-of-the-art ASR systems with transformer-based models for context interpretation and speech summary. It lays the foundation for next developments in voice recognition technology by offering a scalable and effective solution for real-world applications in a range of industries.

REFERENCES

1. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, F. “Data Collection and Open-Source Language Resources,” [complete reference as needed].
2. Bacchiani, M., & Roark, B. “Unsupervised Language Model Adaptation,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, [add location], 2003.

Received: August 09, 2025

3. Ballinger, B., Allauzen, C., Gruenstein, A., & Schalkwyk, J. “On-Demand Language Model Interpolation for Mobile,” Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, [add location], 2010.

4. Bayes, T. “An Essay Towards Solving a Problem in the Doctrine of Chances,” Philosophical Transactions of the Royal Society of London, vol. 53, pp. 370–418, 1763.

5. Bengio, Y., Ducharme, R., & Vincent, P. “A Neural Probabilistic Language Model,” in Proceedings of the Neural Information Processing Systems Conference (NIPS), vol. 13, Denver, CO, Nov. 2000, pp. 932–938.

6. Bourlard, H. A., & Morgan, N. Connectionist Speech Recognition: a Hybrid Approach. Norwell, MA: Kluwer Academic Publishers, 1993.

7. Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. “Attention-Based Models for Speech Recognition,” in Proceedings of the Neural Information Processing Systems Conference (NIPS), vol. 28, Laval, Quebec, Canada, Dec. 2015, pp. 577–585.

8. Fontaine, V., Ris, C., & Leich, H. “Nonlinear Discriminant Analysis for Improved Speech Recognition,” in Proceedings of Eurospeech, Rhodes, Greece, Sep. 1997, pp. 1–4.

9. Graves, A. “Sequence Transduction with Recurrent Neural Networks,” Nov. 2012, arXiv:1211.3711. [Online]. Available: **https://arxiv.org/abs/1211.3711**

10. Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks,” in Proceedings of the International Conference on Machine Learning, Pittsburgh, PA, Jun. 2006, pp. 369–376.

11. Hermansky, H., Ellis, D., & Sharma, S. “Tandem connectionist Feature Extraction for Conventional HMM Systems,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, Istanbul, Turkey, Jun. 2000, pp. 1635–1638.

12. Jelinek, F. Statistical Methods for Speech Recognition. Cambridge, MA: MIT Press, 1997.

13. Krishna, P. R., & Rajarajeswari, P. “EapGAFS: Microarray Dataset for Ensemble Classification for Diseases Prediction.” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, pp. 01–15. **https://doi.org/10.17762/ijritcc.v10i8.5664**.

14. Nakamura, M., & Shikano, K. “A Study of English Word Category Prediction Based on Neural Networks,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, May 1989, pp. 731–734.

15. Sainath, T. N., Weiss, R. J., Wilson, K. W., Narayanan, A., Bacchiani, M., & Senior, A. “Speaker Location and Microphone Spacing Invariant Acoustic Modeling from Raw Multichannel Waveforms,” in Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Scottsdale, AZ, Dec. 2015, pp. 30–36.

Received: August 09, 2025