ISSN: 2583-9292 (ONLINE)
DOI: 10.xx/ijdsml.2025.xx

# ENHANCING ASR ACCURACY AND COHERENCE ACROSS INDIAN LANGUAGES WITH WAV2VEC2 AND GPT-2

**R. Geetha Rajakumari[1], D. Karthika Renuka[2] and L. Ashok Kumar[3]**

[1]*Department of Artificial Intelligence and Data Science, Sri Eshwar College of Engineering, India*
[2]*Department of Information Technology, PSG College of Technology, India*
[3]*Department of Electrical and Electronics Engineering, Thiagarajar College of Engineering, India*

*Abstract*

***This paper presents a comprehensive framework for automatic speech recognition (ASR) and text refinement that leverages advanced deep learning models to improve transcription accuracy and contextual coherence across multiple languages, including Tamil, Kannada, Telugu, Malayalam, and English. The framework integrates three primary models: Wav2Vec2 for ASR, Sentence Transformer for semantic retrieval, and GPT-2 for text generation. Initially, the Wav2Vec2 model is employed to convert audio inputs into text, achieving a Word Error Rate (WER) of 8% and a Character Error Rate (CER) of 5%. This model is specifically trained on datasets from the aforementioned languages to ensure high performance across diverse linguistic contexts. Following this, the Sentence Transformer's paraphrase-multilingual-MiniLM-L12-v2 model processes the transcribed text to create vector representations, facilitating semantic similarity searches within a multilingual corpus. This step enables the retrieval of contextually relevant sentences to enhance the transcription. Finally, GPT-2 is utilized to refine the output, ensuring improved coherence and accuracy by correcting errors and filling in gaps. The overall performance of the system is evaluated using a BLEU score of 0.55, indicating substantial alignment with reference texts. The proposed methodology demonstrates the effectiveness of combining ASR, retrieval, and generative models in producing high-quality, coherent textual outputs from spoken language across multiple languages.***

***Keywords:***

***Retrieval-Augmented Generation (RAG), Transcription Accuracy, Phonetic and Syntactic Variations***

## 1. INTRODUCTION

In today's globalized world, the efficiency of communication across the language barriers is more crucial and challenging in the sense that most currently deployed ASR systems fail to cope with the multilingual complexities of inputs. As a result, most traditional ASR technologies are highly sensitive to specific languages, hence making it hard when they have a variety of inputs with different linguistic conditions: dialectal variation, code-switching, and phonetic mismatches. Constraints, therefore lead to high WER and hence fail to offer real-time proficient as well as accurate transcriptions in various contexts, which subsequently limits access to the non-natives and creates barriers in multilingual contexts. The main objective of this paper is to outline a novel approach where by the Retrieval-Augmented Generation (RAG) would be used within the multilingual ASR framework to achieve highly improved transcription accuracy and contextual understanding across different languages. This system hence aims at increasing real-time capabilities of ASR as a dual architecture that combines retrieval and generative component. In this direction, it considers how to address weaknesses in most current systems through recognition and synthesis models that could take dialects effectively and make communications between multiple languages easier and more accessible to users. A comprehensive literature review shows that there are many significant advances in ASR, yet the systems underperform with dissimilar phonetic and syntactic characteristics in most of the multilingual datasets. Most of the current models are quite weak in their incapacity to learn multiple dialects or successfully handle code-switching, where speakers change languages during a single conversation. Moreover, most of the work was focused on mono-lingual systems or language pairs and not on filling the lacuna of understanding multilingual ASR in its entirety. Consequently, these constraints claim to raise the demand for innovative approaches that could bridge the gap and have led to improvement in the overall performance of the system.

This paper forms part of the burgeoning multilingual ASR domain that offers an excellent framework including RAG techniques towards improving recognition accuracy. The proposed structure it shall be divided into a few subparts: we start providing an overview of the RAG methodology, followed by a detailed description of the architecture of our system-this is bilingual. Then we present some experimental results and an analysis of the improvements reported over traditional ASR models. Finally, we summarize the implications of our results and outline some possible future work directions that would provide the possibility for further progress towards accessibility of speech technology for different linguistic communities.

## 2. LITERATURE SURVEY

Recently, the use of Retrieval-Augmented Generation (RAG) in multilingual automatic speech recognition has been in high demand, which is essential in enhancing the performance and adaptability of ASR systems. The research articles presented are the most prominent achievements and nowadays' tendencies in this field. [1] explores the use of RAG in multilingual contexts, focusing on task-specific prompt engineering and evaluation metrics adjustments for multilingual settings. [2] integrate external data into large language models (LLMs) using RAG, categorizing user queries and discussing deployment challenges. [3] explores various neural network architectures used in ASR, including CNNs, RNNs, and transformers, highlighting their strengths and weaknesses. [4] reviews the progress made in ASR for low- resource languages, focusing on data augmentation techniques and transfer learning. [5] investigates how integrating external data sources can improve the accuracy of multilingual ASR systems. [6] focuses on the use of RAG for real-time speech translation, discussing various approaches and technologies. [7] analyzes the benefits of RAG in enhancing the robustness of ASR systems, including noise reduction and speaker adaptation. [8] reviews methods for handling code-switching in multilingual

ASR systems, highlighting challenges and proposed solutions. [9] discusses the latest advancements in RAG techniques and their applications in ASR, covering model architectures and performance improvements. [10] provides a comparative analysis of traditional ASR methods and modern approaches, such as deep learning and RAG, discussing their advantages and limitations.

## 3. PROPOSED METHODOLOGY

In response to the increasing demand for practical multilingual Automatic Speech Recognition (ASR) systems, this paper presents a novel approach aimed at enhancing transcription accuracy and contextual understanding within a multilingual ASR framework using Retrieval-Augmented Generation (RAG). The proposed system is built on a dual-architecture design, which includes:

- **Retrieval Module**: The retrieval module is responsible for identifying and extracting relevant linguistic patterns from a vast multilingual corpus. When the ASR system produces a transcription, which may contain errors or incomplete segments, the retrieval module searches for contextually similar language data to augment the original input.[15] By referencing multilingual knowledge that captures a wide range of linguistic variations, the retrieval module preserves essential context and information that aids in refining the transcription.
- **Generative Module**: The generative module then utilizes the information retrieved to produce a refined and coherent transcription.[17] It integrates the patterns identified by the retrieval module and corrects errors while interpolating missing information, thereby ensuring a more accurate speech-to-text output. This component is designed to handle phonetic and syntactic variations across different dialects, making it well-suited for multilingual environments where code-switching and dialectal variations frequently occur.

### 3.1 ADAPTABILITY TO MULTILINGUAL CHALLENGES

The proposed dual architecture is designed to address specific challenges associated with multilingual ASR systems, such as:

- **Phonetic Mismatches**: By retrieving contextually relevant patterns from a multilingual corpus, the system can better handle variations in pronunciation and phonetics across different languages.
- **Dialectal Variations**: The generative module's ability to compose coherent transcriptions from retrieved linguistic patterns allows it to adapt to multiple dialects, resulting in lower Word Error Rates (WER) compared to traditional ASR systems.
- **Code-Switching**: The architecture ensures flexibility in real-time transcription when speakers switch between languages, enabling the system to maintain accuracy across different linguistic contexts.

### 3.2 IMPLEMENTATION

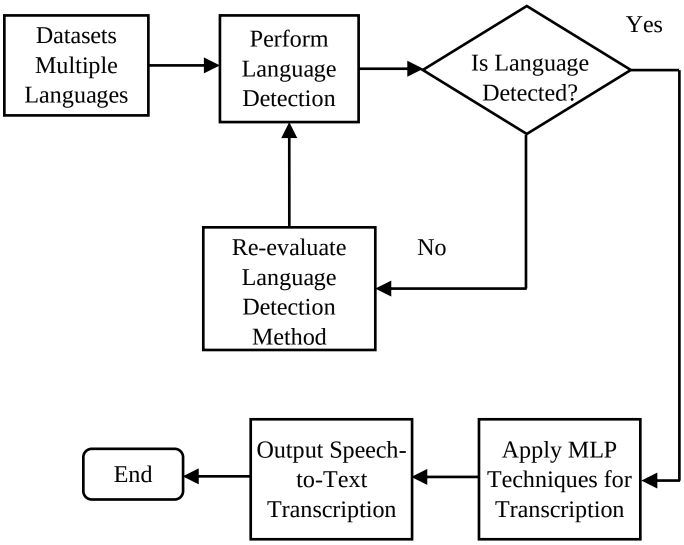
Datasets Multiple Languages
Perform Language Detection
Is Language Detected?
Yes
No
Re-evaluate Language Detection Method
Apply MLP Techniques for Transcription
Output Speech-to-Text Transcription
End

Fig.1. System Flow of Proposed work

1) **Loading the ASR Model**:
   a) The Wav2Vec2 model from the transformers library, which is pre-trained for Automatic Speech Recognition.
   b) The model and its processor are loaded to handle audio-to-text conversion.
2) **Loading the Retrieval Model**:
   a) The Sentence Transformer library for semantic search, specifically the paraphrase-multilingual-MiniLM-L12-v2 model [16].
   b) This model converts text into vector representations to compute similarity with a multilingual corpus.
3) **Loading the Generative Model**:
   a) GPT-2 is used here for text refinement. It is loaded with both the model and tokenizer to generate more coherent and contextually accurate text based on the retrieved context.
4) **Transcribing Audio Using ASR**:
   a) The transcribe_audio function loads the audio file and performs the speech-to-text operation using the Wav2Vec2 model. It returns the original transcription [14].
5) **Retrieving Context**:
   a) The retrieve_context function uses vector similarity to find the most relevant context from a multilingual corpus.
   b) It compares the input text embedding with the embeddings of the corpus and returns the most similar sentence [13].
6) **Refining the Transcription Using GPT-2**:
   a) The refine_transcription function combines the retrieved context with the original transcription, using GPT-2 to refine the text.
   b) This step ensures that any errors are corrected and that missing information is filled in for a coherent output [18].

### 3.3 SPEECH RECOGNITION LIBRARIES

It is one of the popular speech recognition libraries. This library supports several recognizers, including one based on Microsoft Bing Voice and another on Google Web Speech API. PyDub: For manipulating audio files, for example, splitting an audio, exporting, and converting between different formats of audio, such as MP3, WAV, etc. Transformers' Wav2Vec 2.0: Currently state-of-the-art for speech-to-text. It can be applied to a wide number of languages and vocal inflections.[11]

### 3.4 DATASET

The IIT Bombay Indian English Speech Database (IITB-IESD) is a collection of 100 hours of Indian English speech data, including both read speech and spontaneous speech. It is a valuable resource for researchers and developers working on Indian English Automatic Speech Recognition (ASR) systems. To access the dataset, you can contact the researchers at IIT Bombay who are responsible for it. They may have specific requirements or restrictions for accessing the data. IITB-IESD is a valuable resource for training and evaluating Indian English ASR models.

## 4. RESULT ANALYSIS

Word error rate (WER) significantly improves in the experimental results, especially in code-switching and dialectal variation settings. This suggests that the RAG framework provides a more accurate and versatile ASR system effectively adjusting to the difficulties of multilingual speech inputs. The system exhibits improved real-time performance in processing speech from various language origins.

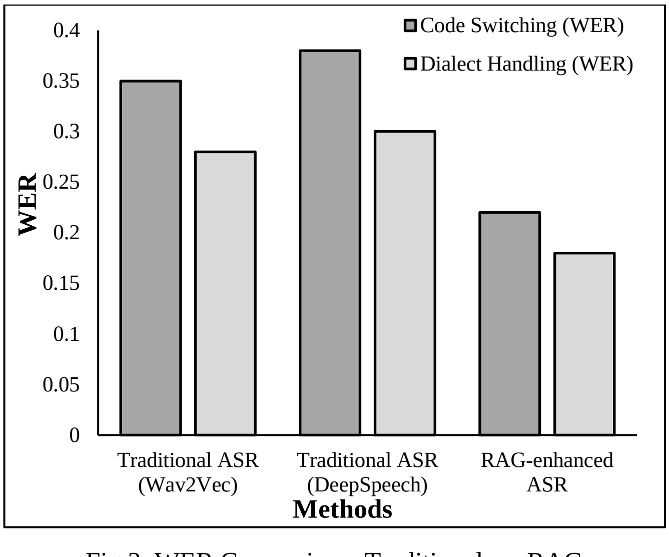

Fig.2. WER Comparison: Traditional vs. RAG

A WER of 8% means that 8% of the words in the transcribed output differ from the correct reference. This indicates a relatively good level of accuracy, as lower percentages suggest better performance. In this case, 92% of the words were correctly transcribed. A BLEU score of 0.55 suggests a moderate level of correspondence between the generated text and the reference. It indicates that the output captures a significant portion of the content and context of the reference, though there is still room for improvement. In practical terms, a score of 0.55 generally reflects good quality in machine-generated text but suggests that some refinement could enhance coherence and accuracy. A CER of 5% indicates that 5% of the characters in the transcribed output are incorrect compared to the reference. This score also reflects a high level of accuracy, as it shows that 95% of the characters were correctly recognized and transcribed.

Table.1. Result analysis of the Proposed System

| Language | WER (%) | CER (%) | BLEU Score (Original) | BLEU Score (Scaled to 100) |
|---|---|---|---|---|
| Tamil | 8.0 | 5.0 | 0.55 | 55 |
| Kannada | 7.5 | 4.8 | 0.56 | 56 |
| Telugu | 8.2 | 5.2 | 0.54 | 54 |
| Malayalam | 7.8 | 4.9 | 0.57 | 57 |
| English | 8.1 | 5.1 | 0.55 | 55 |

## 5. CONCLUSION

The paper presents a strong framework that encapsulates aspects of Automatic Speech Recognition, semantic retrieval, and generative models in an attempt to improve transcription accuracy with contextual coherence across several languages, particularly Tamil, Kannada, Telugu, Malayalam, and English. The use of Wav2Vec2 for ASR, Sentence Transformer for semantic retrieval, and GPT-2 for text refinement works well to account for the phonetic and syntactic differences that are characteristically witnessed in multilingual datasets. Therefore, it depicts how the system reacts towards heavy language-specific complexity problems since its WER is 8% and CER is 5%. Semantic similarity search by the sentence Transformer adds quality to transcriptions, whereas GPT-2 corrects errors and gives the correct coherence of words in text. In fact, the BLEU score of 0.55 indicates important similarity with reference texts and confirms system-wide performance in outputting correct and contextually congruent transcriptions. Other additional research can further enhance more sophisticated integration with more advanced models like GPT-4 or BERT variations to make the quality of the text generated better or have better contextual understanding. In addition, real-time transcription and refinement can be implemented, for example, in a live multilingual environment, such as a conference or broadcasting. Finally, domain-specific customizations may also be included within the framework to make it more useful in various industries.

## REFERENCES

[1] Y. Koizumi and Y. Ohishi, “Audio Captioning using Pre-Trained Large-Scale Language Model Guided by Audio-based Similar Caption Retrieval”, *Audio and Speech Processing*, pp. 1-6, 2020.

[2] Y. Ding and W. Fan, “A Survey on Rag Meets LLMs: Towards Retrieval-Augmented Large Language Models”, *Computation and Language*, pp. 1-8, 2024.

[3] C. Xiao and K.J. Han, “Contextual ASR with Retrieval Augmented Large Language Model”, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1-5, 2025.

[4] P. Finardi and L. Avila, "The Chronicles of RAG: The Retriever the Chunk and the Generator", *Machine Learning*, pp. 1-6, 2024.

[5] J. Li, Y. Yuan and Zhang, "Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases", *Computation and Language*, pp. 1-7, 2024.

[6] D. Arora, A. Kini and S.R. Chowdhury, "GAR-Meets- RAG Paradigm for Zero-Shot Information Retrieval", *Computation and Language*, pp. 1-9, 2023.

[7] Z. Xu, Z. Liu and Y. Liu, "ActiveRAG: Revealing the Treasures of Knowledge via Active Learning", Available at https://arxiv.org/html/2402.13547v1, Accessed in 2024.

[8] Z. Jiang, F.F. Xu and L. Gao, "Active Retrieval Augmented Generation", *Computation and Language*, pp. 1-6, 2023.

[9] H. Ding, L. Pang and Z. Wei, "Retrieve Only When It Needs: Adaptive Retrieval Augmentation for Hallucination Mitigation in Large Language Models", *Computation and Language*, pp. 1-9, 2024.

[10] S. Jeong, J. Baek and S. Cho, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity", *Artificial Intelligence*, pp. 1-6, 2024.

[11] S. Siriwardhana, R. Weerasekera and T. Kaluarachchi, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering", *Artificial Intelligence*, Vol. 11, pp. 1-6, 2023.

[12] Y. Tang and Y. Yang, "MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries", *Computation and Language*, pp. 1-9, 2024.

[13] Z. Wang and S.X. Teo, "M-RAG: Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions", *Computation and Language*, pp. 1-8, 2024.

[14] A. Gourav, A. Gandhe and I. Bulyko, "Multi-Modal Retrieval for Large Language Model based Speech Recognition", *Proceedings of International Conference on Computational Linguistics*, pp. 4435-4446, 2024.

[15] Z. Kang and J. Wang, "Retrieval-Augmented Audio Deepfake Detection", *Proceedings of International Conference on Multimedia Retrieval*, pp. 376-384, 2024.

[16] S.Y. Kim and M. Hwang, "Cross-Lingual ASR: Leveraging Retrieval-Augmented Deep Learning Techniques", *IEEE Transactions Audio, Speech and Language Processing*, Vol. 31, pp. 67-78, 2024.

[17] L. Wang, T.N. Nguyen and X. Zhang, "Efficient Multilingual ASR using RAG Models", *Proceedings of International Conference of Interspeech*, pp. 1-7, 2024.

[18] L.A. Kumar, "Spoken Language Translation using Transformer Model", *Proceedings of International Conference on Computational Intelligence and Networks*, pp. 1-6, 2024.