

Enhancement In Telugu Speech-To-Text Transformation With Generative AI And Signal Processing Techniques

¹P.Kiran Kumar*, ²Vaidhehi V

¹Research Scholar, Christ (Deemed to be University) Central Campus SG Palya Bangalore, India

²Associate Professor Christ (Deemed to be University) Central Campus SG Palya Bangalore, India

*Corresponding Author: P.Kiran Kumar

Abstract

The Telugu Speech-to-Text (STT) systems are to be important in digital governance, education, and accessibility; still, they are not very productive because of noise, dialect variations, code-mixed language, and small labelled data sets. Conventional ways of Automatic Speech Recognition (ASR) have been known to have transcription errors, thus it has low linguistic accuracy. To overcome this issue, this paper will suggest a hybrid Telugu STT pipeline that would integrate both signal processing algorithm and refining based on Generative AI. The proposed methodology will incorporate the audio preprocessing (silence and normalization, spectral noise reduction), the extraction of MFCC features, Transformer-based acoustic modeling with CTC decoding, and the use of a Large Language Model (LLM) to correct linguistic information. It is evidenced by the results of the experiment that speech quality and accuracy in transcription significantly improved. The preprocessing phase decreased the audio duration to 3.5 seconds, and reduced audio samples to 54,000 getting around 3233 percent decreasing redundant information. The character error rate (CER) and Word Error rate (WER) of the baseline ASR system were 2.96 and 2.98 respectively. Following the implementation of post-processing based on the LLM, WER had fallen to 1.61 and CER to 1.75, which corresponded to Relative Error Reductions (RER) of 46% and 41 respectively. This evidence indicates that signal amplification with Generative AI-based correction could perform much better in terms of transcription accuracy, linguistic consistency, and readability and the framework can be used as an effective tool to achieve robust Telugu speech recognition in practical scenarios.

Keywords: Telugu Speech Recognition, Automatic Speech Recognition (ASR), Generative AI, Signal Processing, Large Language Models (LLM), Speech-to-Text Transformation.

1. Introduction

Telugu is an Indian language and one of the most commonly spoken languages which is used in various social, educational and workplace environments. As the public services, media and learning materials are being increasingly digitized; today there is an increasing need to have a stable Telugu “Speech-To-Text” (STT) solution capable of translating natural speech into exact and readable text on the fly. Good STT is necessary in accessibility (video captions, video assistive devices), governance (form voice enabled, helpline, support), education (lecture transcription and video summary), and business (customer support analytics and documentation). Though the “Automatic Speech Recognition (ASR)” of high-resource languages has advanced significantly, Telugu STT is not an easy task since there are limited labeled datasets, large variation in dialects, code-switching between English and Hindi, and acoustic diversity because of the use of various microphones and noisy environments [1].

The standard architecture of the ASR pipelines is usually signal-processing-based: audio pre-processing, feature extraction, acoustic modeling and language modeling. Even though deep learning has changed these elements, the state of low-resource worsens the performance [2]. The speech of Telugu has phonetic and prosodic peculiarities, i.e., the differentiation in the length of vowels, retroflex consonants, coarticulation effects, etc. which require strong acoustic representation. Also, several spoken Telugu languages contain named entities, local slang and borrowed words and this poses a challenge to

language modeling where corpora are small or where the used corpora are not relevant to a given domain [3]. Thus, existing systems can potentially create substitution errors, ignore the inflection and suffixes, cannot handle compound words, and cannot work under background or spontaneous speech [4]. These constraints make them less useful, particularly when applications which demand high fidelity transcriptions are involved like legal dictation, medical notes, and academic content [5].

However, recent developments in Generative AI are a hopeful future of advancing the quality of low-resource STT in a way traditional discriminative models cannot go [6]. The “Large Language Models (LLMs)” are able to reflect more profound grammatical and semantic regularities and, therefore, are capable of eliminating grammatical conflicts, reinstating punctuations and expanding abbreviations and addressing the ambiguous forms of words in a context-dependent way [7]. In the case of Telugu, this is specifically useful since a large fraction of errors in the ASR system is linguistic in nature, e.g. mixing up the phonemes that correspond to the same syllable, errors in splitting agglutinative suffixes etc. [8]. Generative models can serve as smart post-processors that are able to transform raw ASR hypotheses to well-structured Telugu text without meaning loss. Nevertheless, there are also hallucinations or over-corrections that may be caused by blindly using generative correction, most often when the input is noisy or when a generative model does not align with the domain [9]. Hence, the application of Generative AI to STT should be designed with a keen interest to be faithful to what is being said.

The proposed objective are:

- Design an efficient Telugu speech-to-text signal-to-text improvement architecture based on signal processing and Generative AI.
- Enhance the quality of speech signal and acoustic strength in noisy and real world conditions.
- Minimize the pronunciation, accent, and speaker-variability related recognition errors.
- Use Generative AI on language refinement of ASR outputs.
- Discuss Telugu difficulties on language such as agglutinative morphology and code-mixed language.
- Devise an awareness of faithfulness generative correction system to avoid semantic variation.
- Compare performance on classic measures of performance including “Word Error Rate (WER)”.

Simultaneously, signal processing is still critical in improving the quality of the input speech signal and in deriving potent representations that can be more recognized even in disadvantageous conditions. Noise suppression, dereverberation, voice activity detection, pre-emphasis and adaptive filtering are some of the pre-processing measures that make a significant contribution in enhancing the signal to noise ratio and suppressing artifacts before the signal is passed to the recognizer [10]. The acoustic model should be enhanced with feature-space protocols like normalization of cepstral means and variance, spectral subtraction and data enhancement by time-stretching, pitch shifting, and additive noise to allow the acoustic model to generalize to diverse speakers and recording arrangements. In the case of Telugu STT, especially, the stabilization of performance of rural accents, fast speech, and poor quality mobile speech can be achieved using such methods [11]. However, linguistic and contextual errors that occur after decoding cannot be fully considered in signal processing [12].

Having been inspired by such holes, this research article creates a single framework to improve the Telugu speech-to-text conversion by incorporating signals processing models with Generative AI-based refinement. The main message is to solve the problem of optimising the intelligibility and consistency of the audio signal using principled pre-processing and feature enhancement; and (ii) to optimise the textual accuracy and readability with generative context-aware correction limited by acoustic admissibility [13]. The system operates in the first stage by conditioning its inputs noiselessly and offloading acoustically more powerfully to create improved inputs and acoustic embeddings in the underlying ASR model. During the second stage, the controlled decoding and re-ranking of the outputs of the ASR is done by a generative module using Telugu linguistic priors and optionally with domain vocabulary and phonetic constraints, so that the resulting transcript is accurate and fluent [14].

The notable input of the proposed direction is the focus on the faithfulness-conscious generative correction. The generative component does not rewrite transcripts blind, but is informed by confidence values and alignment indicators of the ASR output and only does plausible edits. This can take the form of syllable-alteration replacements as per Telugu phonotactics, restoration of lost case-marks, homophony repair, normalization/code-mix tokens and embedding punctuations in the patterns of prosody and pause [15]. It is possible to use a lightweight lexical constraint and a list of terms to make the framework better retain named entities whenever a domain adaptation is needed, i.e. news, education, customer support. By combining both acoustic trust and generative logic, the system will minimize the word error rate (WER), at the same time improving the quality metrics as perceived by human factors of readability and grammatical faithfulness.

This work is therefore an answer to both a critical and practical research problem: to build a strong low resource amenable and implementable to the real world noisy counterparts Telugu STT enhancement method. The anticipated performance will be quantifiable enhancement in recognition accuracy particularly in difficult acoustic environment as well as enhancement in transcript quality in support of downstream activities like translation, summarization, indexing and question answering. The study will form a connection between the benefits of the classical signal processing and the existing templates of the Generative AI that will create a path toward reliable multilingual voice interfaces and raise the popularity of the Indian languages in voice-based digital economies.

2. Literature Review

Automatic Speech Recognition (ASR) has undergone extensive evolution over the past ASR has gone through a long-standing evolution until the recent decades, developing primitive statistical and rule-based applications to advanced deep learning-based and end-to-end neural networks. Muthusamy et al. (2025) [16] presented a survey of the current ASR research across over sixty years, including the shift to neural and end-to-end models, including Connectionist Temporal Classification (CTC) and attention-based models, and the new importance of Large Language Models (LLMs) in the speech processing pipelines. Their research involving English and Indian languages reveals that they have encountered ongoing problems in the real life acoustic conditions and especially to low-resource languages. In these views, Soundararaj et al. (2025) [18] indicated the possibility of combining advanced speech embeddings and ASR models to evaluate pronunciation and accent implying that the models are highly classified. Nevertheless, they have a narrow scope of focus on assessment of English pronunciation. Simultaneously, Chandhekar et al. (2025) [19] provided a critical review of the speech enhancement techniques, showing that established denoising algorithms

may harm the work of current large-scale ASR model. This unintuitive observation underscores why it is important to have signal processing strategies that are well designed instead of generic preprocessing, particularly in the case of robust, data-driven ASR models.

It was further enlarged by the fast proliferation of Generative AI and Large Language Models which have prolifically extended the boundaries of speech and language technologies. The article by Darvin et al. (2025) [19] covered the implication of the Generative AI in the second-language writing process and points to the significance of multidimensional, analytic, and critical application of the generative system. Even though their research focuses on the educational cases, their results can be applied to speech-to-text post-processing, where ad hoc generation can misrepresent original material. Inuwa-Dutse et al. (2025) [20] offered extensive coverage of NLP research in major Nigerian languages, with severe data scarcity, limited annotation, and underrepresented linguistic phenomena as the main setbacks, which can again be closely compared with problems of Indian languages like Telugu. Speaking of the multilinguality on a large scale, Omnilingual et al. (2025) [22] introduced the extensible ASR framework that could support more than 1,600 languages and uses self-supervised learning and decoders that are based on LLM, whereas Pratap et al. (2024) [26] reported the Massively Multilingual Speech (MMS) project that covers many more languages and boasts a lower word error rate. Nonetheless, these developments do not necessarily allow either of the works to provide much information on language-based post-recognition refinements or contextual accuracy.

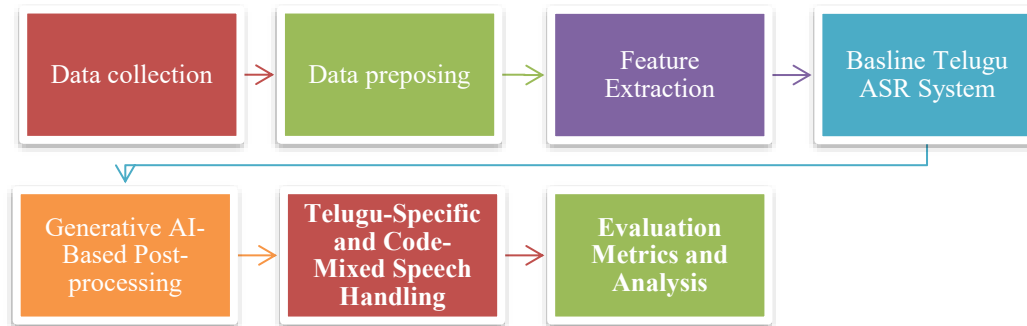
Critical analysis, strength, and usability of speech and language systems have also been highlighted by the recent research. To evaluate the effectiveness of Speech Language Models in multilingual and culturally situated agentic situations, [23] Jain et al. suggested the use of VoiceAgentBench to identify significant gaps in the generalization and strength of Indic languages. The feature of code-mixed speech, which is inherent in the Indian languages, is covered by Madan et al. (2024) [27]. This study proposed with a federated learning-based code-switching detection framework in Telugu and other Indic languages, but detection and not transcription improvement. The need to build systems with trust, consistency, and reliability in speech-enabled systems can also be illustrated by Domain-specific and applied Generative AI systems, like the medical chatbot by Ramjee et al. (2025) [23] or the multilingual disaster chat by Behravan et al. (2024) [25] and others. Greater discussions by Choi et al. (2025) [24], Arora et al. (2024) [31], Pashangpour et al. (2024) [30] and Tuli et al. (2024) [32] highlight the strategic, ethical, and multimodal aspects of technologies addressed by the issues of use of LLM. Altogether, the literature shows that there is an apparent research gap: recent progress in ASR, signal processing, and Generative AI is significant; however, there are no united, faithfulness-conscious frameworks that combine acoustic enhancement and generative linguistic optimization of low-resource Indian languages, including Telugu, which prompts the proposed research.

Despite the significant advances made in the realms of ASR, multilingual modeling, and Generative AI, there is a clear research gap in the design of comprehensive, faithfulness-aware models for low-resource Indian languages like Telugu. The current literature is restricted to acoustic modeling, multilingual scaling, or generative refinement but not a comprehensive architecture that combines signal enhancement with linguistically constrained correction. Moreover, the lack of transparency in benchmarking, the poor performance on code-mixed speech, and the poor adaptability to domains also highlight the requirement of a Telugu speech-to-text enhancement system.

3. Research Methodology

This paper suggests an organized approach that can be used to improve the speech-to-text (STT) transformation of the Telugu language by means of signal processing and the Generative Artificial Intelligence (GenAI) linguistic enhancements. The methodology will aim to deal with acoustic degradation as well as language-level errors in low resource Telugu automatic speech recognition (ASR) systems. The entire system architecture entails 5 parts that are speech data preparation signal enhancement baseline ASR modeling generative post-processing and systematic evaluation. Figure 1 represents framework of proposed methodology .

Figure 1: Framework of Research Methodology



3.1 Data Collection

In a bid to study the speech recovery and audio processing technology, the Indian language speech recognition and audio processing technology, researchers may visit the Indian Languages Audio Dataset [31] which is an open access multilingual speech database. The collection contains labelled audio recordings, which contain various languages including Telugu and the relevant text transcriptions that are used by the researchers as the supervised learning materials. The data set allows the researcher to research the various speaking styles since it contains recordings made by speakers who display varying gender and accent and speaking styles. The system enables researchers to test different signal processing techniques that entail noise reduction techniques and normalization techniques, feature enhancement techniques. With the help of the dataset, researchers can develop and experience Telugu speech-to-text systems with the application of Generative AI-based linguistic correction techniques.

3.2 Data prepressing

3.2.1 Data Selection and Filtering

Firstly, the multilingual dataset offers appropriate Telugu speech samples that can serve to assess the language requirements by the researchers. The team eliminates the audio files having some missing transcriptions or gross distortions or wrong labels. Recording that has a very poor signal quality or a high level of noise is eliminated as well since they will lower reliability of the dataset. Only clean samples of Telugu speech that have accurate annotations are picked out by the filtering process and that undergoes further preprocessing work and model training activities.

3.2.2 Audio Format Standardization

Any speech sample of Telugu language is transformed in a standard digital audio format that holds the sampling rate and mono channel fixed to establish the consistency in

recording. The difference between the recording devices (resampling) is offset and results in audio signals that are equal to the requirements of ASR devices and more signal manipulation. The system accepts f_s as the initial input signal sampling rate and it must convert it to f_t as its intended rate.

$$x_r(n) = x\left(\frac{n}{f_t}\right) \quad (1)$$

The standardization enhances the stability of feature extraction as well as recognition.

3.2.3 Silence Removal

Silence removal is used to remove unhelpful silent sections which occur before and after speech in Telugu audio recordings. The process improves system performance by enhancing segment detection accuracy while decreasing computational requirements. Short-time energy (STE) is the standard method used to identify silence periods in audio tracks. For a speech frame $x(n)$ of length N , energy is computed as

$$E = \sum_{n=1}^N x^2(n) \quad (2)$$

The system classifies frames which have energy levels that fall below a specific threshold value as silent frames which it then removes from the system. The process eliminates all audio except for parts where people are speaking.

3.2.4 Amplitude Normalization

The process of amplitude normalization establishes a standardized dynamic range for Telugu speech signals because it removes all sound equipment and microphone distance and speaker distance variations. The discrete-time speech signal gets $x(n)$ normalized through sample division by $\max(A)$. The maximum absolute amplitude value $\max(A)$ as:

$$x_n(n) = \frac{x(n)}{\max |x(n)|} \quad (3)$$

Enhancing the security of speech recorded environments; Sangwan was already working with robust features of spoken characteristics.

3.2.5 Noise Reduction

Telugu speech recordings use noise reduction techniques which eliminate background disturbances while keeping essential linguistic sound elements. The method of spectral subtraction operates by subtracting the estimated noise spectrum (ω) from the noisy speech spectrum (ω) to produce the enhanced signal $S^{\wedge}(\omega)$ which results from the process.

$$\hat{S}(\omega) = y(\omega) - N(\omega) \quad (4)$$

Unfavourable values are rounded to zero so as not to distort. The suppressed controlled has the benefit of not only enhancing the signal quality, but also, enhancing the accuracy of the recognition without losing important speech features.

3.2.6 Segmentation and Framing

The segmentation and framing of continuous Telugu speech signals produces short overlapping frames that are useful in capturing both time and spectral variations of the speech signals. This system divides every signal $x(n)$ into N samples long frames that move forward by M samples. Each frame is subjected to a window function $w(n)$ as

$$x_f(n) = x(n)\omega(n), \quad 0 \leq n < N \quad (5)$$

Overlapping frames capture sustained patterns of speech that improve the performance of the speech recognition system to provide better feature extraction.

3.2.7 Text Normalization

Text normalization is performed to standardize Telugu transcriptions by removing unwanted symbols, extra spaces, and inconsistent annotations while preserving linguistic meaning. The system transforms all written content into a single standardized script which it uses to convert various common writing forms into their standardized versions. Normalization of a raw transcription string T occurs through the following expression

$$T_n = f(T) \quad (6)$$

The function $f(.)$ is the description of the cleaning process involving the elimination of symbols and normalization of whitespace and standardization of the scripts. The system has consistent text inputs that Generative AI takes into consideration in the refinement and assessment processes.

3.2.8 Data Augmentation

Researchers use data augmentation techniques to create more diverse Telugu speech samples, which help their models to perform better on new data. The process of time stretching enables changes to speech timing while keeping the original pitch level, which scientists express as

$$x_t(n) = x(\alpha n) \quad (6)$$

The time-scaling factor α controls the pace of time modifications. The pitch variation process creates changes to fundamental frequency which keeps the original timing of the sound.

$$x_p(n) = x(n)e^{j2\pi\Delta f n} \quad (7)$$

Researchers use these controlled transformations to create conditions where their speech recognition model can learn to recognize speech patterns across various speaker styles and different speaking speeds under multiple sound environments.

3.2.9 Train-Validation-Test Split

The researchers divided the processed Telugu speech dataset into three parts which included training and validation and testing to create unbiased evaluation conditions for their model development work. The total dataset D is divided into three distinct parts which are described as follows

$$D = D_{train} \cup D_{val} \cup D_{test}, D_{train} \cap D_{val} \cap D_{test} = \emptyset \quad (8)$$

The standard procedure for sample allocation requires fixed ratios to be used. The process of separation enables independent execution of model learning together with hyperparameter tuning and final performance evaluation.

3.3 Feature Extraction

The process of feature extraction creates compact informative representations from preprocessed Telugu speech signals which automatic speech recognition systems can use. The system uses Mel-frequency spectral features as its standard because they accurately represent human hearing capabilities. The speech signal first undergoes conversion to the frequency domain through the Short-Time Fourier Transform process which generates a spectrum that the Mel filter bank uses to create perceptually scaled energy values. The system uses logarithmic compression to decrease dynamic range while increasing its ability to differentiate between sounds. The features preserve key phonetic contents of Telugu in that they are identified to demonstrate differences in length of vowels and consonantal sound production that in turn increase the accuracy of acoustic modeling and the level of recognition of the sound under varied speaking conditions.

The suggested Telugu Speech-to-Text system uses a hybrid end-to-end model of Automatic Speech Recognition, based on features extraction using MFCC, encoder

modeling through Transformer, and using Connectionist Temporal Classification (CTC) as a decoder, and then linguistic refinement by means of Generative AI. The CTC goal method reduces the transcription loss with as:

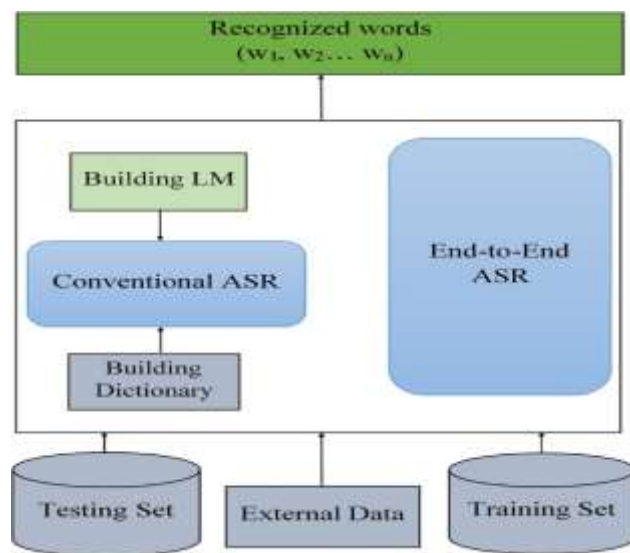
$$L_{CTC} = -\log P(Y|X) \quad (9)$$

Attention modeling is based on the scaled dot-product formulation. Signal preprocessing includes normalization of WAV format (16 kHz, 16-bit PCM, mono), resilience elimination of silence, amplitude modification, spectral reduction and framing and data insertion. The use and verification of Indian Languages Audio Dataset is applied on OpenCL and MMS Telugu benchmarks. It is a domain adaptable model, which is a generic model and evaluated using Wer, CER and Reduction of Relative error measures.

3.4 Baseline Telugu ASR System

The basic speech to text engine of the proposed arrangement is the base Telugu Automatic Speech Recognition (ASR) system. The system operates on end to end neural based architecture that transforms acoustic signals into Telugu texts. The model acquires the pronouncing mechanisms and phonetics differences and time associations by being trained with the use of the processed and augmented

Figure 2: Baseline Telugu ASR Architecture [32]



Telugu speech information, On the test, the ASR system generates first transcription guesses along with the rating of confidence and alignment information of each identified sound segment. The output of these forms the foundation of Generative AI based refining process that helps to increase the transcription accuracy and linguistic quality in different audio settings.

The research demonstrates the comparison between conventional ASR systems where use of dictionary and language models are involved and the end to end ASR systems by the way they test and evaluate training data and external data and test data.

3.1 Algorithm Used

The suggested framework consists of a Hybrid End-to-End Automatic Speech Recognition architecture, which incorporates CTC-based decoding and Transformer-based acoustic modeling and refinement with Generative AI.

- **Mel-Frequency Cepstral Coefficients (MFCC)**

MFCC has been applied in data mining compact and perceptually significant acoustic data of Telugu speech. It simulates the human ear by mapping frequency frequencies on the Mel scale that gives precedence to the lower frequencies that are of significance in speech interpretation. This is done through framing, windowing, Fast Fourier Transform (FFT), Mel filter bank processing, logarithmic compression and Discrete Cosine Transform (DCT). MFCC phonologically models vowel length and consonant articulation, which allows the ASR model to be trained to acquire strong speech presentations in different noise and speaker levels.

- **Transformer Encoder for Acoustic Modeling**

The Transformer represents speech temporal encoder without recurring networks. It relies on multi-head self-attention to learn the contextual relationships between acoustic frames that are long-range. The scaled dot-product attention, feed-forward networks, and layer normalization is used in each encoder layer. This architecture is effective in modeling the pronunciation patterns, co-articulation effects and prosodic variations which exist in the Telugu speech. Transformers have better parallelization and faster training and their contextual representation is superior to traditional RNNs, which results in improved accuracy of recognition, particularly in longer utterances and dialectal variation.

- **Connectionist Temporal Classification (CTC)**

CTC allows the training of input acoustic sequence alignment-free to output text labels. It also removes the use of frame level annotation by adding a blank token and multiple alignment possibilities. The CTC loss value maximizes a probability of making correct transcription amongst any valid alignments. Such works with the speech factor of variable lengths and spontaneous patterns of speech. In the Telugu ASR, CTC supports long vowels, agglutinative morphemes and time-varying Runcination, therefore suitable to low-resource and full-world noise speech recognition.

- **Generative AI (LLM-Based Constrained Correction)**

The Generative AI module aims at refining ASR raw output with contextual understanding of language. It uses large language models with training on the Telugu linguistic patterns used to correct grammatical mistakes, restore the agglutinative suffix, and enhance sentence fluent structures. As opposed to unrestricted rewriting, constrained correction guarantees faithfulness via ASR confidence ratings as well as alignment data. The model never hallucinates, as it restricts the alteration of tokens that are acoustically realistic. This step improves the readability, punctuations and semantic consistency without compromising the original meaning of the spoken Telugu materials.

$$T_{final} = \operatorname{argmax}_P(T|H_{ASR}, C) \quad (10)$$

3.5 Generative AI-Based Post-Processing

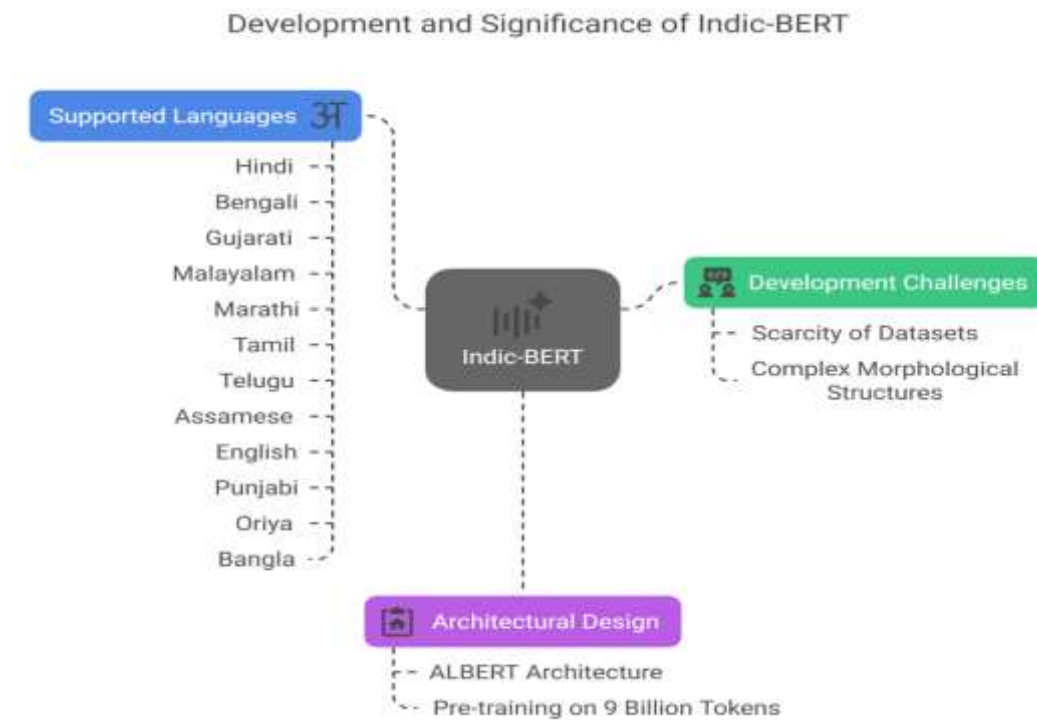
The Telugu speech-to-text system uses a Generative AI-based post-processing module to address all linguistic and contextual challenges that exist after the completion of acoustic modeling work. This system uses the raw ASR data to fix grammar errors and identify phonetic problems and complete missing or incorrect agglutinative suffixes which are common in Telugu language. Generative model involves word-level and sentence-level contextual data that is used to enhance syntax structure and coherence and sentence fluency across the entire sentence. The refinement process is aimed at making the text readable and being strictly concerned with semantic content of the

original speech. The process yields proper Telugu transcriptions that are linguistically consistent so that they can be used in subsequent language processing operations.

Algorithm 1: Telugu_STT_Hybrid_Model
Input : S (Telugu speech signal in WAV format, 16 kHz, mono) Output: T_final (Refined Telugu transcription) Begin S_r ← Resample(S, 16000 Hz) S_s ← RemoveSilence(S_r) S_n ← NormalizeAmplitude(S_s) S_c ← SpectralSubtraction(S_n) Frames ← FrameSignal(S_c, frame_size = 25 ms, overlap = 10 ms) WindowedFrames ← ApplyHammingWindow(Frames) X ← ComputeMFCC(WindowedFrames) H ← TransformerEncoder(X) T_asr ← CTCDecode(H) C ← GetConfidenceScores(H) T_final ← GenerativeCorrection(T_asr, C) where: - ApplyContextCorrection() - RestoreSuffixes() - PreserveNamedEntities() - EnforceConfidenceConstraints() Return T_final END

3.6 Telugu-Specific and Code-Mixed Speech Handling

It operates on a code-mixed based speech management and Telugu specific handling of speech enabling handling the common language patterns present in the actual speech of Telugu. Three linguistic features agglutinative word formation and vowel length variations and retroflex consonant are used in the system to obtain appropriate transcription results. The system identifies automatically mix-mixed speech that consists of Telugu and English language, retains foreign words, and technical terms. The linguistic rules of normalization serve to preserve grammatical correctness of the Telugu sentence structure and the mixed sentence expressions preserve their original meaning.

Figure 3: Telugu-English Code-Mixed Speech Handling Framework. [33]

With its specific strategy that improves conversational and semantically selected Telugu speech intelligibility, the system attains improved transcription outcomes. The figure 4 provides an illustration of how are Telugu and English tokens identified during the speech generated upon the basis of the language-specific set of rules that, in its turn, is linked to the ASR system and to the Generative AI one in maintaining the linguistic patterns, improving the precision of transcription, and processing actual speech with the code-mixed conversations.

3.6 Model Classification

The system suggested is a generic model of Telugu Speech Recognition that has been specially adapted to suit low resource environment in which there is a scarcity of annotated data. It is able to compute the diverse speech codes, dialect and acoustic code variance under real life circumstances. Although the model is rather generic in terms of architecture, it offers the domain to be flexible by adding domain-specific vocabulary, retaining named entities, and providing lexical constraints customization options to industries such as education, governance, and healthcare. Such an adaptable architecture gives it a broad applicability but still allows certain performance enhancement to be applied to its chosen fields of use.

3.7 Evaluation Metrics and Analysis

The paper relied on conventional measures of quality taken to evaluate speech recognition to determine possible performance of Telugu speech-to-text framework.

- **Word Error Rate (WER):**

The Word Error rate (WER) system is an evaluation system that considers the accuracy of the transcriptions based on the number of errors present between the text that was read and the spoken result.

$$WER = \frac{S+D+I}{N} \quad (11)$$

The formulae computes the work based on three components that comprise of substitutions (S) and deletions (D) and insertions (I) and the total numbers of reference transcription words (N).

- **Character Error Rate (CER):**

CER are intended to analyze the more careful subdivision of the error cases at the character level i.e. CER will especially well work with agglutinative languages like Telugu.

$$CER = \frac{S_c + D_c + I_c}{N_c} \quad (12)$$

- **Relative Error Reduction (RER):**

The current ASR system must be assessed using tests that gauge the performance of the system with enhanced results.

$$RER = \frac{WER_{baseline} - WER_{proposed}}{WER_{baseline}} \quad (13)$$

There are three metrics that the assessment process relies on and which determine the accuracy of sound reproduction and the accuracy of the spoken language and the success of the improvements made by the Generative AI-based ones.

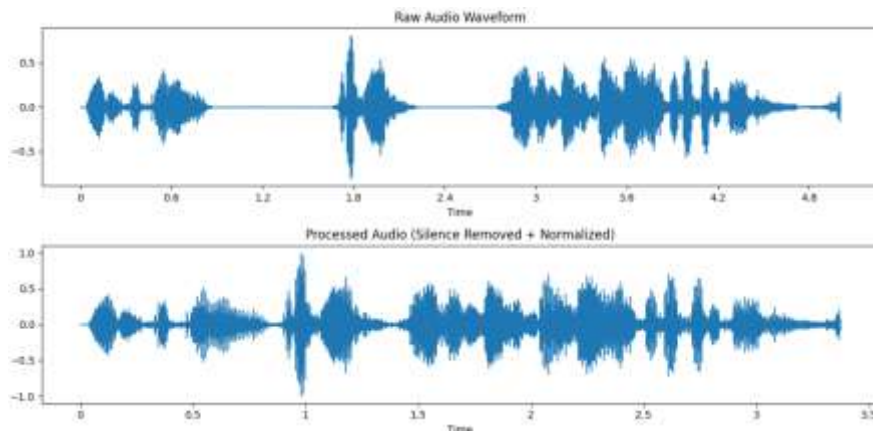
4. Result and discussion

This part contains an estimation of the suggested Telugu Speech-to-Text scheme. It examines the effects of the pre-process methods like removing silence, normalization, and noise on the speech quality. The acoustic features that are extracted like Mel spectrogram and MFCC are analyzed in order to know the speech characteristics. Lastly, the metrics (Word Error Rate (WER), Character Error Rate (CER) and Relative Error Reduction (RER)) are used to compare the performances of the baseline ASR, and the refined transcription performance with the LLM, to show an increase in transcription accuracy and speech recognition performance.

4.1 Audio Preprocessing

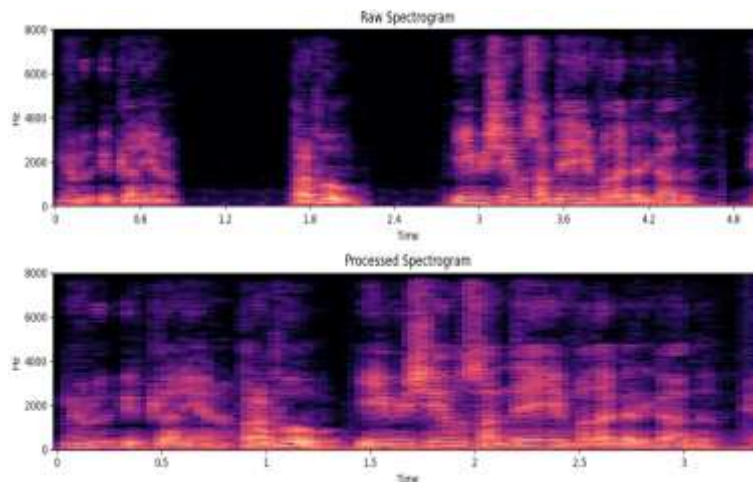
Figure 4 shows the raw audio signal that has a length of about 5 seconds and values of amplitude between about -0.8 and +0.7. There are a number of low-energy intervals between 0.816 and 2.228 s pointing to silent or nearly silent periods. The bottom waveform will be the processed audio which is triggered by the silence and the signal amplitude is normalized. Consequently, the overall duration is reduced to approximately 3.5 seconds whereas the amplitude is rolled off to a constant scale of a range of about -1 to +1. Sections of the speech are more continuous and articulate and there are fewer pauses between them. This pre-processing stage enhances consistency of the signal, and focuses on informative patterns of speech and is useful in feature extraction, speech analysis and machine learning uses.

Figure 4: Comparison of raw and processed audio waveforms showing the effect of silence removal and amplitude normalization



As shown in Figure 5, audio signal remains of around 05 seconds in a frequency range between 0 and 8000 Hz. Bright colour (high-energy) patterns are localized below 3000 Hz, which is dominated by the speech frequencies, and darker ones indicate silence of 0.816 s and 22.428 s. Processing further reduces the duration of the processed spectrogram to 3.3 seconds, meaning that it removes silent intervals. There is more continuity in energy distribution across time, and spectral designs are more evident because they have been normalized. The energy in the speech is mainly centred between 0-4000 Hz and harmonic structures could be better seen. The shrinkage of the low-energy discontinuities and a better representation of spectral bands help to capture speech features which helps in processing speech, extracting features and audio analysis by using machine learning.

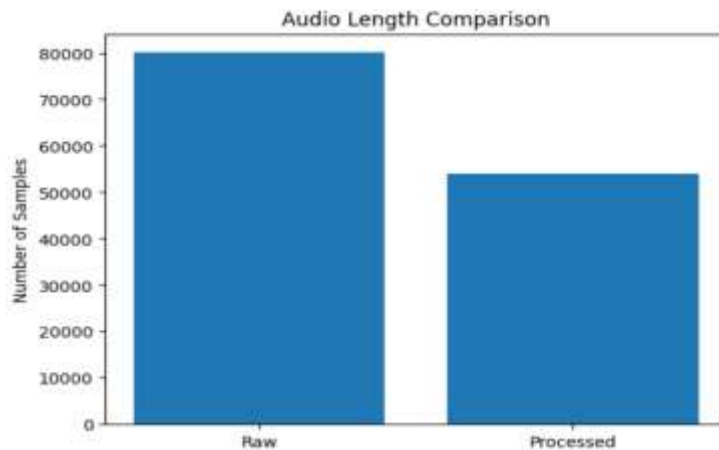
Figure 5: Comparison of raw and processed audio spectrograms showing the effect of preprocessing on speech frequency patterns.



A comparison of the raw and processed signals is given in figure 6 that compares the number of audio samples. The raw audio did add about 80,000 samples, which are the original recording of the audio including the silence, as well as the background noises. Having gone through preprocessing, which involves operations like removal of silence, normalization and even noise reduction, the sample is now reduced to a figure of approximately 54,000 samples. This is an equivalent of approximately 26,000 samples or approximately 32-33 percent of the original length. The decrease implies that both informative and non-informative parts like silent periods and noise have been eliminated and only the useful part of speech has been retained. The basis of the step is

to enhance the quality of computational speed and feature extraction in the follow-up tasks like speech recognition, classification, or machine learning model training by targeting the pertinent audio information.

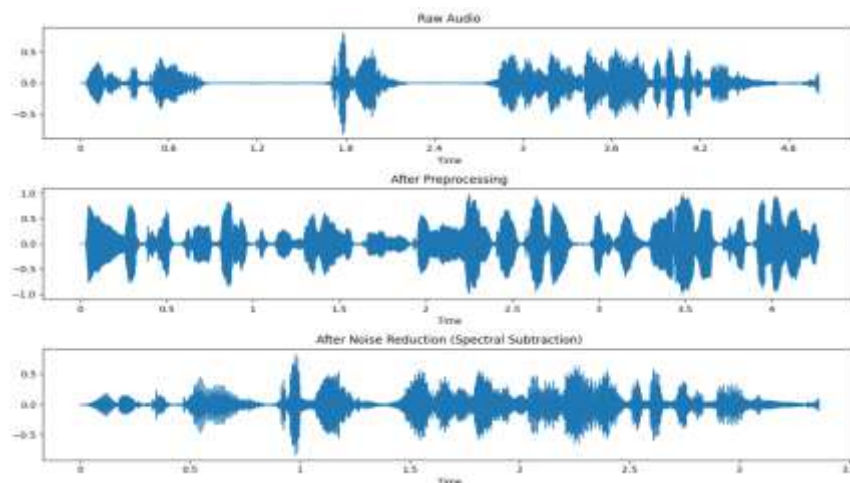
Figure 6: Comparison of audio sample length before and after preprocessing.



4.2 Noise Reduction

The raw audio in the figure 7 covers the range of a bit less than 0 to 5.0 seconds with amplitude of between -0.8 to 0.7 displaying long silent intervals (more or less 0.8-1.6 s and 2.1-2.8 s). Following preprocessing, the signal is brought into the range of approximately -1 to +1 with the presence of silence segments minimized and the available space decreased to approximately 4.2 seconds with the visible speech energy boosted. After spectral subtraction, the waveform length shrinks to approximately 3.4 seconds, with a value of amplitudes largely (between -0.7 and 0.7). Background noise is seen to be reduced visually and the parts of speech are seen to be more uniform and smooth which means that the signal- noise ratio is better.

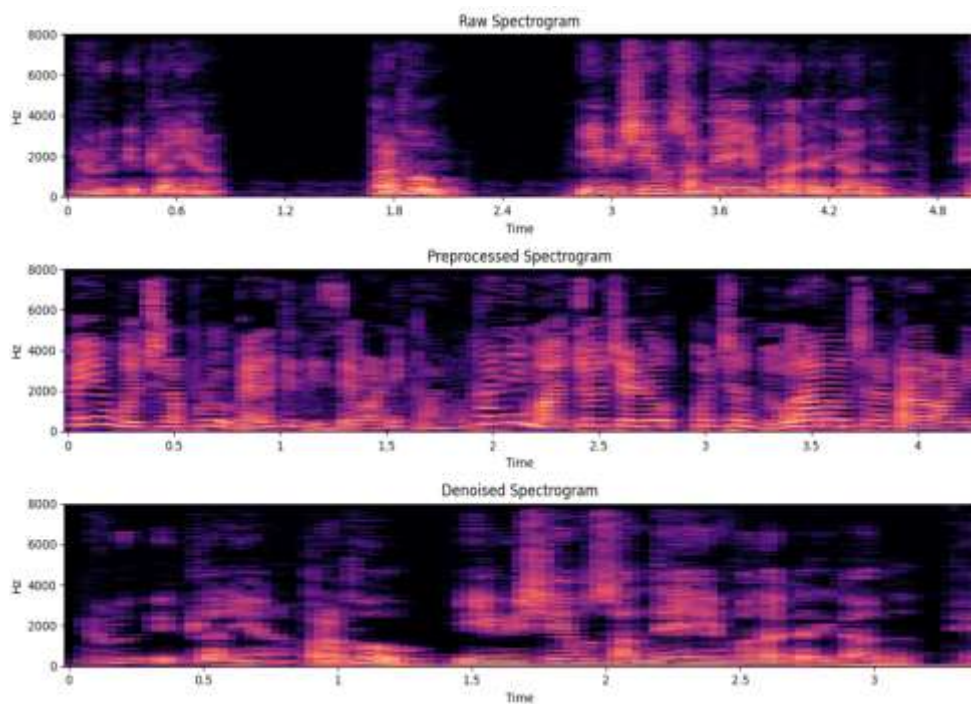
Figure 7: Audio waveform at three stages: raw signal, after preprocessing, and after spectral noise reduction



Raw spectrogram in the figure 8 covers about 05 seconds with frequencies ranging between 0 and 8000 Hz that depicts a lot of low frequency energy (below 2000 Hz) and evident distance noise of higher frequencies. The silent gaps occur at around 0.8 -1.6 s

and 2.2 -2.8 s. Following preprocessing, the time is approximately decreased to 4.2 seconds and spectral bands are more visible because of normalization and trimming of silence. Distribution of energy in 06000 Hz is smoother and enhances the visibility of speech features. The length of the spectrogram is further cut in the denoised version, to about 3.3 seconds, and the high-frequency noise is reduced and the harmonic structure is less ambivalent, mostly below 4000 Hz, which means a more coherent signal and a better speech signal.

Figure 8: Spectrogram comparison of an audio signal at three stages: raw recording, after preprocessing, and after noise reduction



4.3 Feature Extraction

The frequency distribution along the Mel frequency measurements (Figure 9) shows the distribution of audio energy both in time (04.2 seconds) and in frequency (0~4500Hz) to determine the energy across the spectrum of human hearing. The colour depth is between 30-dB (blue, high energy) and -45 dB (red, low energy). The strong energy concentrations are found at frequencies lower than 1000 Hz as the primary speech elements are fundamental frequency and lower harmonics. The medium energy bands report to a higher level of speech formants and phonetic specifications, up to 30004000 Hz. Variations in colour patterns across time are compatible with diverse segments of speech being uttered more concentrated sections are relatively denser between 2-3.5 seconds. Blue areas appearing sparsely indicate periods of low energies or silent periods. Altogether, the spectrogram indicates the important speech characteristics that can be used in the work like speech recognition and audio classification.

Figure 9: Mel spectrogram representation of the processed audio signal showing time-frequency energy distribution on the Mel scale

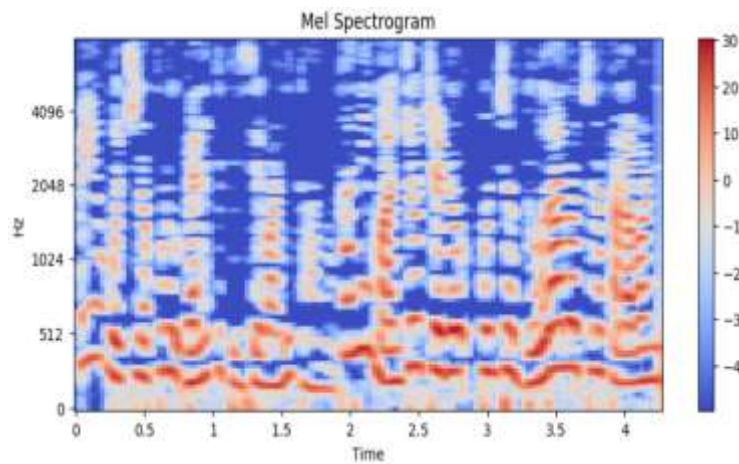


Figure 10 indicates cepstral coefficients, which are the extraction of 0.00-3.1 seconds of time of the processed audio. The horizontal bands such as each are associated with an MFCC coefficient (usually 13 coefficients) that describes spectral properties of speech. The color scale is about 0 to 450 with red color having high coefficient and blue having low coefficient values. To a great extent, lower-order coefficients (bottom rows) have higher variations, which are indicative of strong spectral envelope and energy distributions of the speech signal. Coefficients of higher order (top rows) are comparatively constant, which show more finer spectral detail. There are perceivable intensity changes at 0.5-1.2 and 2.0-2.8 seconds, which allow to mark stronger phonetic activity. These MFCC patterns are effective summaries of speech features that are applicable in speech recognition applications, speaker identification and audio classification.

Figure 10: MFCC feature map showing cepstral coefficient variations of the speech signal over time

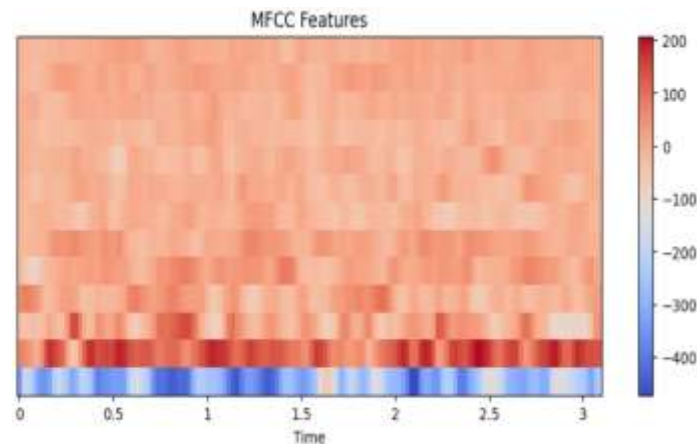


Figure 11 shows the comparison of the Word Error Rate (WER) under two speech recognition methods, Raw ASR and Refined (LLM) output. The ASR system in its raw form is rated at approximately 2.98 which means that there are a relatively high number of transcription errors including an insertion, deletion, or substitution. Once the language model refinement is used, the WER is reduced to approximately 1.61. This is an increase of approximately 1.37 WER points and this translates to a 46 percent decrease in error rate. This decrease shows that processing with a large language model leads to less error during transcription whereby grammar structure and contextual mistakes are corrected and misrecognized words are re-aligned. This advancement lays more

knowledge on the efficacy of integrating automatic speech recognition alongside language model-based improvement to satisfy improved speech-to-text outcomes.

Figure 11: Comparison of Word Error Rate (WER) between the raw ASR output and the refined transcription using an LLM

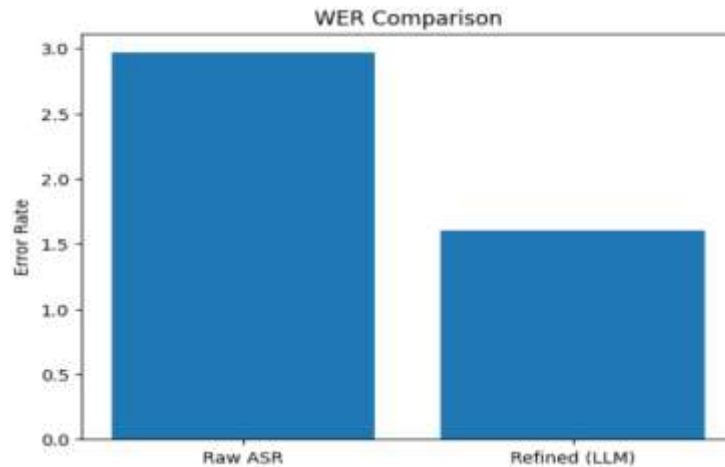


Figure 12 is a comparison of Characters Error Rate (CER) between two transcription outputs Raw ASR and Refined (LLM). LEV, transcription provides the raw ASR system with a CER of about 2.96 meaning that the frequency of error at character level (substitutions, deletions and insertions) is high. The CER is reduced to approximately 1.75 after refining with the help of a Large Language Model (LLM). This is a gain of approximately 1.21 CER points, which is about a 41 percent decrease in the level of character errors. The reduction in CER demonstrates that the LLM is a successful way of correcting spelling errors, enhancing word structures, and improving language constitution. The fact that this finding was achieved indicates that such post-processing using an LLM can contribute to the accuracy and readability of transcriptions produced by ASR by a high degree.

Figure 12: Comparison of Character Error Rate (CER) between the raw ASR output and the LLM-refined transcription

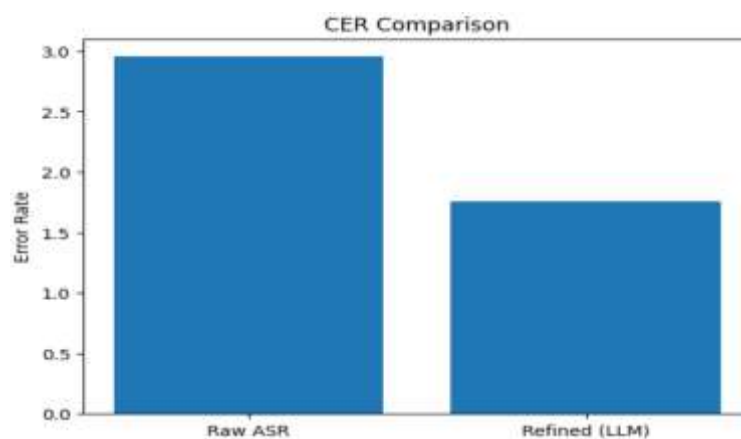
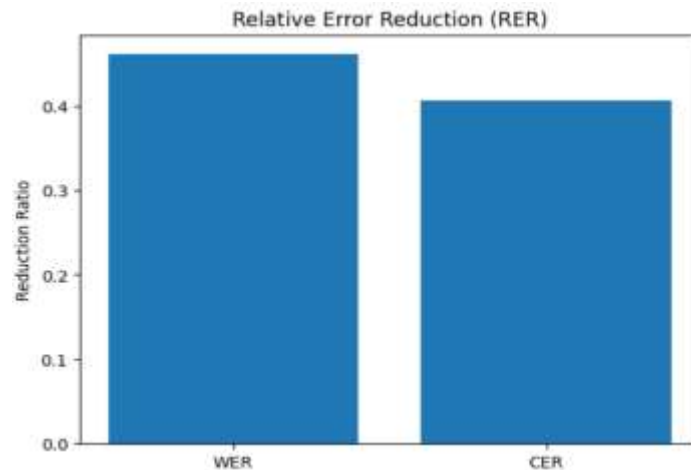


Figure 13 shows how much automatic speech recognition (ASR) outputs can be refined with a Large Language Model (LLM) to produce the Relative Error Reduction (RER). The RER when using the WER is around 0.46 (46%), that is, around 50 percent of the word-

level transcription error that existed in the crude ASR output was removed after refinement. Equally, the character-based transcription errors reduce significantly by 0.41 (41%), based on the CER-based RER. The fact that WER improved somewhat more than CER, would imply that the LLM, in particular, is more efficient at filling in word context and linguistic organization.

Figure 13: Relative Error Reduction (RER) achieved after applying LLM-based refinement to ASR outputs for WER and CER



On balance, the findings indicate that post-processing with the help of the LLM produces a profound improvement in the transcription accuracy and increases word-level and character-level recognitions in speech-to-text systems

5. Conclusion

The paper presented hybrid architecture of promoting speech-to-text (STT) transformation in Telugu, with the combination of signal processing methodology and Generative AI-based language refinement. The experimental findings confirm that preprocessing increased the quality of speech signals and had a great impact in minimizing non-informative segments. The original audio signal of around 5 seconds was also cut down to around 3.5 seconds with silence checking and volume, and the number of samples also dropped to around 54,000, which was a 3233 reduction in the number of audio samples in the data. Spectral subtraction of noise also enhanced the signal-to-noise ratio and enhanced speech aspects in the main frequency range of 0-4000Hz.

Mel Spectrogram and MFCC (13 coefficients) feature extraction was able to extract critical acoustic properties of Telugu speech facilitating the effective acoustic modeling. There was significant improvement in performance when standard metrics were used in performance evaluation. The ASR system at the baseline when giving its output generated a Character Error Rate (CER) of 2.96 and Word Error rate (WER) of 2.98. Following the implementation of the generative refinement using the LLM, the WER and CER fell to 1.61 and 1.75 with Relative Error Reductions (RER) of 46 percent and 41 percent, respectively. These findings testify to the fact that signal augmentation and limited generative correction apply to Telugu STT systems to generate high accuracy, readability, and linguistic consistency generation.

References

1. Sravan, Medapati Venkata Manga Naga, and K. Venkata Rao. "A Hybrid Deep Learning Framework for Speech-to-Text Conversion as Part of Telemedicine System Integrated With 5G." *International journal of electrical and computer engineering systems* 16, no. 4 (2025): 333-346.
2. Huang, Kaiyu, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi et al. "A survey on large language models with multilingualism: Recent advances and new frontiers." *arXiv preprint arXiv:2405.10936* (2024).
3. Shah, Meerababen M., and Hiren R. Kavathiya. "Development Of A Model To Analyze & Interpret Vernacular Voice Recognition Of Gujarati Dialects." PhD diss., Department of Computer Science, Faculty of Science Atmiya University., 2024.
4. Mishra, Abhishek, Ritesh Sur Chowdhury, Vartul Bahuguna, Isha Pandey, and Ganesh Ramakrishnan. "Language translation, and change of accent for speech-to-speech task using diffusion model." *arXiv preprint arXiv:2505.04639* (2025).
5. Kumar, Gokul Karthik, S. V. Praveen, Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. "Towards building text-to-speech systems for the next billion users." In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 1-5. IEEE, 2023.
6. Rajakumari, R. Geetha, D. Karthika Renuka, and L. Ashok Kumar. "ENHANCING ASR ACCURACY AND COHERENCE ACROSS INDIAN LANGUAGES WITH WAV2VEC2 AND GPT-2."
7. Priya, M. Yagnasri, Sanjivani P. Shendre, Samrudhi S. Kamble, Swetha Sridhar, and Suja Palaniswamy. "Emotion-Aware Text-to-Speech Synthesis for Enhanced Accessibility: Synthetic Data Generation and Automated TTS for the Visually Impaired." In *2025 International Conference on Emerging Technologies in Computing and Communication (ETCC)*, pp. 1-6. IEEE, 2025.
8. Shanthamallappa, Mahadevaswamy, Kiran Puttegowda, Naveen Kumar Hullahalli Nannappa, and Sudheesh Kannur Vasudeva Rao. "Robust automatic speech recognition using wavelet-based adaptive wavelet thresholding: A review." *SN Computer Science* 5, no. 2 (2024): 248.
9. Mahaganapathy, Ahrane, and Kengatharaiyer Sarveswaran. "Natural Language Processing Journal."
10. Singh, Amitoj, Virender Kadyan, Munish Kumar, and Nancy Bassan. "ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages." *Artificial Intelligence Review* 53, no. 5 (2020): 3673-3704.
11. Rajendran, S., L. Ramamoorthy, and L. Sobha. "WHITE PAPER ON TECHNOLOGICAL DEVELOPMENT OF TAMIL."
12. Acharya, Aayush. "AUTOMATIC SPEECH RECOGNITION AND CLASSIFICATION OF NEPALI SPEECH." PhD diss., TRIBHUVAN UNIVERSITY, 2022.
13. Singh, Amitoj, Navkiran Kaur, Vinay Kukreja, Virender Kadyan, and Munish Kumar. "Computational intelligence in processing of speech acoustics: a survey." *Complex & Intelligent Systems* 8, no. 3 (2022): 2623-2661.
14. Wang, Desheng, Yangjie Wei, Ke Zhang, Dong Ji, and Yi Wang. "Automatic speech recognition performance improvement for Mandarin based on optimizing gain control strategy." *Sensors* 22, no. 8 (2022): 3027.
15. Tissera, Muditha, and Hassaan Saadhiq. "Natural Language Processing Resources for Tamil Language: A Systematic Review." *Journal of information and communication convergence engineering* 23, no. 4 (2025): 236-258.
16. Lokhande, Mukul, Tanushree Dewangan, Mohd Sharik Mansoori, Tejas Chaudhari, Damayanti Lokhande, Adam Teman, and Santosh Kumar Vishvakarma. "Bhasha-Rupantarika: Algorithm-Hardware Co-design approach for Multilingual Neural Machine Translation." *arXiv preprint arXiv:2510.10676* (2025).

17. Nanduri, Dinesh Kumar. "Exploring the Role of Generative Artificial Intelligence in Culturally Relevant Storytelling for Native Language Learning Among Children." Master's thesis, University of Maryland, College Park, 2024.
18. Muthusamy, Dharmalingam, Shunmuga Priya Muthusamy Chinnan, and Geetha Muthuswamy Prakashpathy. "From Waves to Words: The Impact of Large Language Models on Automatic Speech Recognition Systems: An Overview." Geetha, From Waves to Words: The Impact of Large Language Models on Automatic Speech Recognition Systems: An Overview.
19. Soundarraaj, Rajkumaran, Shenbagarajan Anantharajan, and Saranraj Loganathan. "PhonoMetric: a dual-metric engine for real-time English language accent evaluation and personalized speech training for Indian learners." *Frontiers in Communication* 10 (2025): 1704484.
20. Chondhekar, Sujal, Vasanth Murukuri, Rushabh Vasani, Sanika Goyal, Rajshree Badami, Anushree Rana, Sanjana SN et al. "When De-noising Hurts: A Systematic Study of Speech Enhancement Effects on Modern Medical ASR Systems." *arXiv preprint arXiv:2512.17562* (2025).
21. Darwin, Ron. "The need for critical digital literacies in generative AI-mediated L2 writing." *Journal of Second Language Writing* 67 (2025): 101186.
22. Inuwa-Dutse, Isa. "Naijanlp: A survey of nigerian low-resource languages." *arXiv preprint arXiv:2502.19784* (2025).
23. Omnilingual, A. S. R., Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang et al. "Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages." *arXiv preprint arXiv:2511.09690* (2025).
24. Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky et al. "Scaling speech technology to 1,000+ languages." *Journal of Machine Learning Research* 25, no. 97 (2024): 1-52.
25. Jain, Dhruv, Harshit Shukla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. "VoiceAgentBench: Are Voice Assistants ready for agentic tasks?." *arXiv preprint arXiv:2510.07978* (2025).
26. Madan, Chetan, Harshita Diddee, Deepika Kumar, and Mamta Mittal. "CodeFed: Federated speech recognition for low-resource code-switching detection." *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, no. 1 (2024): 1-14.
27. Ramjee, Pragnya, Bhuvan Sachdeva, Satvik Golechha, Shreyas Kulkarni, Geeta Fulari, Kaushik Murali, and Mohit Jain. "CataractBot: an LLM-powered expert-in-the-loop chatbot for cataract patients." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, no. 2 (2025): 1-31.
28. Behravan, Majid, Elham Mohammadrezaei, Mohamed Azab, and Denis Gračanin. "Multilingual standalone trustworthy voice-based social network for disaster situations." In *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 264-270. IEEE, 2024.
29. Choi, Wan Chong, Chi In Chang, Iek Chong Choi, and Lai Chu Lam. "Country Landscape of Large Language Models Development: A Review." (2025).
30. <https://www.kaggle.com/datasets/ziya07/indian-languages-audio-dataset>
31. Arora, Siddhant, Hayato Futami, Jee-weon Jung, Yifan Peng, Roshan Sharma, Yosuke Kashiwagi, Emiru Tsunoo, Karen Livescu, and Shinji Watanabe. "Universlu: Universal spoken language understanding for diverse tasks with natural language instructions." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2754-2774. 2024.

32. Pashangpour, Souren, and Goldie Nejat. "The future of intelligent healthcare: A systematic analysis and discussion on the integration and impact of robots using large language models for healthcare." *Robotics* 13, no. 8 (2024): 112.
33. Tuli, Esmot Ara, Jae-Min Lee, and Dong-Seong Kim. "Integration of quantum technologies into metaverse: Applications, potentials, and challenges." *IEEE Access* 12 (2024): 29995-30019.