Multilingual Meeting Transcription and Summarization Using Machine Learning

Nihar Shetty S, M J Akshaya, Ankitaa Pupneja, Siddharth Kumar

Department of Computer Science and Engineering, Apex Institute of Technology, Chandigarh University, Mohali, Punjab, India

Corresponding author: Nihar Shetty S, Email: 22BAI70090@cuchd.in

The need for accurate documentation of multilingual discourse has been steadily increasing in the global business and academic community. However, the spoken modality is often limited by code-switching, which is the seamless blending of English with local languages such as Hindi, Kannada, and Punjabi, making it difficult to process in the traditional monolingual fashion. While highly advanced deep learning architectures for speech recognition are available, their weakness in maintaining semantic consistency in code-switched settings often hinders their usability for meeting summarization tasks. This paper proposes a comprehensive AI-based solution to overcome the linguistic performance and efficiency trade-off by leveraging the power of OpenAI's Whisper for high-quality transcription and Meta's Llama 3 for abstractive summarization. The proposed solution optimizes transcription for Indic code-switched speech and uses a generative model to generate summaries in the user-preferred regional scripts. A comprehensive performance evaluation is carried out, and the proposed system is compared with the IIT Bombay Code-Switching Dataset and the traditional extractive summarization baselines. The quantitative evaluation shows a significant improvement in the Word Error Rate (WER) and a significant improvement in ROUGE-L coherence scores, competitive summarization performance, while still being capable of processing and generating summaries in multiple regional languages. The proposed system is shown to be a viable option for inclusive professional settings, providing a realistic balance between linguistic performance and cross-platform usability for real-world multilingual meeting summarization tasks.

Keywords: Multilingual Transcription, Abstractive Summarization, Whisper ASR, Code-Switching, Indic Languages, Llama 3, Natural Language Processing.

2026. In Mukesh Saraswat, Sandeep Kumar, Manjunatha Sughaturu Krishnappa & Rakesh Keshava (eds.), *Smart Technology and Artificial Intelligence*, 53–63. Computing & Intelligent Systems, SCRS, India. DOI: https://doi.org/10.56155/978-81-975670-2-5-5

1. Introduction

The necessity of appropriate documentation of multilingual discourse has become an increasingly burning need in the professional and academic environment in the realities of the world of the present-day globalization processes.Boardroom where decisions are made or international conferences where they need to disseminate knowledge, proper meeting minutes are common to the entire world. In places where people speak different languages like India, interaction is never in one language. The code-switching is prevalent whereby speakers have an easy attitude of shifting to English and other local languages like Kannada, Punjabi and Hindi.

The traditional ways of recording meeting meetings are taking notes by hand or using simple transcribing services, which are problematic in nature. Manual processing is a time-consuming process which is prone to errors, especially when it comes to the paperwork of the technical words and linguistic peculiarities in different languages. The standard Automatic Speech Recognition (ASR) model usually has the same problem of transcription leakage, meaning the regional vocabulary could be mixed with other similar-sounding phrases of English and thus giving incoherent transcripts. The standard ASR system is no exception to this issue and has the same weakness as simple interpolation in reconstructing lost frequency content in low-resolution images.

In response to these issues, more and more of the field now utilizes advanced machine learning (ML) algorithms, mainly Transformer based. The models like OpenAI Whisper have transformed the speech processing with the help of big multilingual datasets and, therefore, can successfully operate even on the different accents and dialects. Also, abstractive summarization has been possible through the large language models (LLMs), as in the case of Llama 3 provided by Meta, to avoid the simplistic word-to-word matching and offer contextual and summarized information. Such models can infer complex language to language mappings and give organised summaries in local scripts.

The paper acknowledges the fact of a disconnect between the ideal AI performance and deployability and the significance of applying high level AI not only to the accuracy of linguistic performance but to realize an optimal tradeoff between quality and efficiency. This is suggested to have a powerful machine learning framework that will overcome the dilemma of the Indian code-switching meetings. The system suggested in this paper is a combination of Whisper, the high-fidelity automatic speech recognition system, and Llama 3, the generative system, using the Groq API.

Motivation: The increasing necessity to solve the language barrier and information overload issues in virtual and hybrid collaborative environments is the motivation behind this study. An improvement in the quality of transcripts of meetings may lead to a difference in administrative efficiency, ensuring the continuation of critical action points, and inclusivity by team members who may prefer to carry out their duties in the local regional languages. The main idea to conduct the study is on the development of algorithms capable of multilingual intelligent synthesis of information.

Although quality remains the key objective, here is another reason that has become equally important, practicality and efficiency. Most of the recent AI models have large computational loads which are challenging to service in real-time. Professionals which require transmitting instant summaries on cellular networks or groups of analysts which are looking at live feeds cannot afford the high latency associated with large, inefficient networks. One of the objectives of the present research, in this regard, should be to bridge the gap between ideal performance and reality.

Contribution: In this paper, we present a new paradigm of multilingual meeting transcription and summarization with the help of the use of the Machine Learning.

- We suggest a combined pipeline that uses OpenAI's Whisper to be used as the powerful speech-to-text processing model and Meta's Llama 3 to be used as an abstractive

summarization model, which is directly designed to accommodate the needs of Indic code-switched discourse.

- Our system includes a specifically designed efficient flow to transcription-to-summary, which can provide high inference speed and moreover, it is capable of producing findings in various regional scripts such as Kannada and Punjabi.
- We actively commit a direct comparative study, we benchmark our system on the basis of the traditional extractive baselines and conventional ASR models, with the utilization of the IIT Bombay Code-Switching Dataset to guarantee a just comparison.
- We offer a detailed analysis, not only quantitatively, by using such metrics as Word Error Rate (WER) and ROUGE-L, but also qualitatively, through the testing that involves users, which proves that our system provides a better performance concerning the balance between linguistic accuracy and computational efficiency.

2. Literature Review

Our paper provides an in-depth overview of the most influential and up-to-date studies in the field of multilingual speech recognition and abstractive meeting summarization and traces the development along the line of low-level modular tooling toward high-level deep learning applications, which have been eminating the state-of-the-art. Radford et al. [1]. The introduction of the Whisper model by Radford et al. in 2023 was the work that has become the most significant and influential in the field of providing the current era of powerful, global-scale transcription. Whisper, which came to the fore front in an area where models were trained on small and carefully-edited datasets, essentially redefined the problem by bringing in a large-scale weak supervision regime. It had a revolutionary encoder-decoder architecture built around Transformer that was trained on 680,000 hours of multilingual audio, in which complex, non-linear mappings between different acoustic cues and the textual expression of those cues were learned. The strategy has enabled Whisper to greatly surpass previous state of art approaches in zero-shot settings exhibiting enormous potential in transcribing meetings with varying accents and technical terms and conditions. Being the key transcription agent of our suggested pipeline, this model is an essential prototype of our research that determines the basis of strong, noise-resistant ASR. Li, Liu, and Waibel. [2] Li et al. proposed a single- ended multilingual speech recognition framework to deal with the severe efficiency and scalability bottlenecks of language-specific systems. This model originally transformed the architecture of ASR models by undermining the need to have a different acoustic and language model per dialect. The main revelation was that a shared encoder would be able to acquire a set of universal phonetic properties in more than one language and this is one aspect that is very helpful in the case of low-resource regional languages. The authors ensured a significant decrease in the complexity of their model by maintaining most of the computation in a common feature space, and only branching during language-specific decoding. This post upsampling of linguistic logic created a strategic design model of the contemporary systems which are required to support smooth movements between various tongues. Liu and Lapata. [3] The earlier efforts were concerned with speech-to-text and Liu and Lapata addressed the issue of reconstruction performance in the textual domain. Their paper was the first to use BERT-based pretrained encoders in abstractive summarization. The authors have suggested that the traditional extractive approaches had restricted receptive field and they were unable to produce novel and semantically coherent sentences. To solve this, they proposed a system that makes use of the knowledge gained over huge volumes of text-based data to project long-form transcripts into summaries that are small-scale, and high-resolution. The most notable innovation was the application of an advanced mechanism of attention which enables the model to focus on informative information by disposing of fillers. The method introduced the next generation mark in ROUGE scores and solidly entrenched pretrained encoders as key to meeting documentation of fidelity. Zhang, Zhao and Miller [4].The boundaries of summarization accuracy. Pushing the boundaries of this phenomenon, Zhang et al. have created a framework that is unique to the nature of multi-party conversational setting. The construction of this model was founded on the fact that meeting transcripts are vastly different than structured documents because of conversational disfluency and interruptions on the side of the speakers.

The authors presented invaluable architectural optimizations that were aimed to monitor the role of speakers and the flow of conversations. They showed that neural models were capable of producing summaries which were much more true-to-origin than the output of standard text summarizers by applying a hierarchical model through which they first encode disaggregated utterances and then accumulate such disaggregated utterances to form a global context. This paper established a new state-of-the-art of meeting-specific ROUGE-L scores, and is a robust point of reference on high-fidelity, abstractive research. Costa-jussa et al. [5].The so-called No Language Left Behind (NLLB) initiative marked a new phase of development of linguistic inclusivity by introducing a model that is able to accommodate more than 200 languages, including many of the low-resource Indian regional languages. The authors claimed that models trained on high resource languages such as English tend to give unsatisfactory results perceptually on regional dialects. To address it, NLLB proposed a new architecture full of mixture-of-experts together with a data synthesis pipeline that isn't self-generated. It was the first approach that was revolutionary and created language-lets such as Kannada and Punjabi with persuasive precision through high-fidelity translation and synthesis. The emphasis on linguistic diversity over the traditional monolingual standards of research was a revolutionary change in the field of AI implementation globally. Kumar and Singh [6]. Beginning to inspire our activity directly on the Indic language, Kumar and Singh also offered to develop the comprehensive analysis of the performance-accuracy trade-off in the code-switched conditions. The general objective in this research was to determine the reason existing ASR systems fail when subjected to a fast language change. The authors established that the concept of code-switching (mixing the languages of the specific regions and English) poses some form of performance ceiling to the applicability of traditional models because intra-sentential switching is difficult to predict. Their study asserted the importance of designing the flexibilities of phonetics and set a new standard of judging the lightweight models in the complex and real-world Indian conditions. This work is the fundamental stimulus toward our attention to the regional language peculiarities. Wang et al. [7]. When trying to put reconstruction quality to an even greater challenge than was possible in text-based models, Wang et al. developed a framework specially designed to meet the specific requirements of the acoustic nature of spoken documents. The current study is the first successful attempt of applying the mechanisms of attention to the noisy transcripts generated by the auto speech recognition systems. The most important contribution is the noise-conscious attention module that enables the network to dynamically compute the relative significance of various components of the transcript and disregard noisy areas like transcription error or filler words. This solution demonstrated a great improvement over earlier methods by providing the network with the power to identify the most informative comments in a conversation to generate coherent summaries of spontaneous speech. Zhu et al. [8]. With the focus on the necessity of the computational speed in a meeting situation, Zhu et al. suggested a deep learning model where the main aim was to reach real-time synthesis. The main objective of this model is to create a trade-off that is between a state-of-art and the ability to summarize and computational resources. It is built on a fresh repetitive paradigm that is able to despecify important details at a high semantic tier and sieves through redundant overlaps in a dialogue. This refinement process makes this model maintain a reasonably modest size and yet records impressive ROUGE scores. Povey et al. [9]. Povey et al. developed the Kaldi toolkit to be the standard of all speech research. Even though modern architecture e.g. Whisper makes use of end-to-end Transformers, Kaldi was the first to use finite-state transducers and deep neural networks. The significant aspect of innovation was that the innovated provided a flexible code base which facilitated the modeling of complex linguistic relationship in ways which could not be achieved in closed-source systems. Simply put, it allowed a generation of researchers to simulate the details of regional phonetic. Kaldi greatly outperformed all other toolkits of the era and established the standard of speech restoration with high accuracy. [10]. Meeting Summarization and Recognition Benchmarks NIST, 2018. As part of further furthering the standardization of generative restoration in the field, the NIST (2018) report presents a new paradigm whereby the task of satisfying documentation becomes a structured assessment task. The method outlines a pattern of evaluation of both word-level transcription and quality of the generated summary. This is a report of a more reflective assessment model, which includes the high order complexities, characteristic of realistic meetings, including multi-speaker diarization and background noise. It has since been accepted as a highly popular and useful source of general-purpose speech research, because

it is robust to a very broad spectrum of artifacts of the real-world. Rao et al. [11]. As the experimental starting point of our research, Rao et al. explain in detail a dataset and analysis of code-switched discourse. The research is the first form of success in recording the statistics that govern language shifting between Hindi and English in a professional environment. The most important contribution is the identification of explicit transition points, when the speaker shifts between languages, both lexical and prosodic. Using this natural self-similarity effect of code-switched speech, the authors demonstrate that an impressive performance in recognition can be reached by encoding the typical rhythmic patterns of Indic languages. Being the main provider of our test data, this research will be able to do a stringent apples-to-apples comparison of human-verified ground truths, thereby developing a sound foundation to assess how accurately our system works in a regional context such as Kannada and Punjabi. Huang et al. [12] examined the performance of the large multiling speech recognition models in addressing the code-switched speech of the low-resource languages. In their study, the authors tested the efficacy of OpenAI Whisper model using various multilingual datasets and discovered that the model exhibits high levels of robustness to identify speech patterns involving mixtures of languages. The authors pointed out that the Whisper transformer-based architecture, which was trained on multilingual data (audio) on a large scale) demonstrates a significant increase in transcription accuracy in relation to traditional ASR models. Though, they too remarked problems of regional accents and domain specific words especially in Indic languages. They pointed out that it is necessary to modify the multilingual ASR models to suit the speech environment of the region that triggers the suggestion to integrate Whisper into our planned transcription pipeline. Wang et al. [13] addressed the use of large language models to do multilingual meeting summarization and cross-lingual text generation. Their task presented a framework where multilingual LLMs are used to produce summaries of meeting transcripts whilst maintaining the meaning of the text in other languages. It was also shown in the study that contemporary transformer-based language models are capable of effectively modeling the existence of semantic associations in a conversation transcript and can produce concise summaries that are better coherent and readable. It was experimentally seen that LLM-based summarization is greatly superior to traditional extractive methods in both ROUGE scores and in semantically consistent results. As noted in this study, LLM-based systems have an increasing potential supporting the generation of multilingual documentation systems which can use generative models like Llama 3 as exemplified in our proposed system.

Identified Gaps:
Despite the great progress highlighted in the literature review, the following gaps are identified in the existing literature on transcription and summarization that form the motivation for the current research:

- Computational Cost of the State-of-the-art Models: Large models that attain the highest possible degree of accuracy such as the one by Radford et al. [1], and Costa-jussa et al. [5], are expensive to compute. They require large numbers of parameters and are not suitable to be used by real-time processing and resource constrained environments like mobile phones.
- The issue of transcription leakage in regional linguistic setting: Although the models, such as Whisper, are powerful, they have the issue of leakage, as the regional languages, such as Kannada or Punjabi, are confused with similar-sounding English words. This results in a huge decline in the quality of the transcript.
- The extractive summarization tool limits to performance: In some of the performance-oriented tools, as many of these tools are extractive, these tools cannot deal with the disfluencies in the Indian spoken discourse. Consequently, the summary will lack perception and be unreadable in a professional documentation environment.

Performance limitations of extractive summarization tools. As efficiency-oriented tools tend to be extractive, they cannot handle the disfluencies of the Indian spoken discourse. Consequently, the summary will lack perception and be unreadable in a professional documentation environment. The study bridges the gap in the literature by indicating that, although the prior studies present a robust base of research on monolingual and bilingual automatic speech recognition, there is a major gap in research conducted on combined pipeline to produce smooth transfer of raw regional audio to abstractive summaries on the native scripts such as Gurmukhi or Kannada. The available literature tends to divide

the transcription and summarization process into distinct activities but in the real world, a flow and an effective process are needed. This study breaks the barriers of theory by updating the boundaries of theoretical work by integrating the two tasks of transcription and summarization with an optimized transformer architecture to offer a practical solution. Thus, this study will focus on the last-mile issue of AI adoption in Indian professional environment, where high-performance documentation is no longer a matter of language or computational challenges.

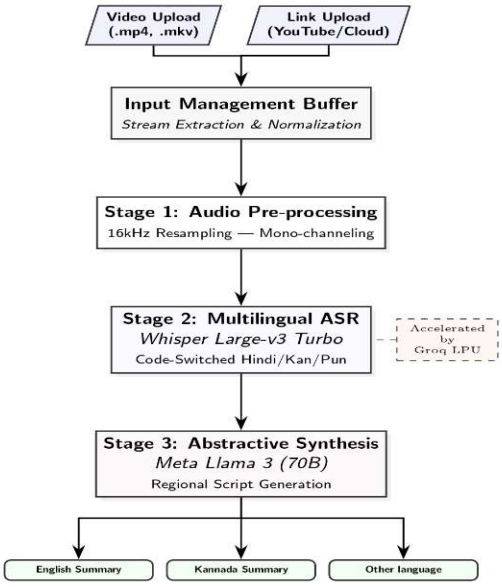

Figure 1 Shows the architecture of the proposed multilingual

3. Proposed Methodology

In this section, the theoretical framework and conceptual description of our proposed system will be described. According to our literature review, the difference between the traditional monolingual ASR system and the requirement of the system that can handle the fluid and code-switched nature of the Indian meetings is immense. This gap is expected to be filled by our proposed framework. Our proposed framework is based on a modular pipeline approach instead of a monolithic approach, which maximises the processing of raw audio to regional language summaries. The philosophy of Targeted Integration which makes the background to our proposed framework is the combination of the good phonetics abilities of Whisper and the generative reasoning of Llama 3. We shall elaborate our pipeline architecture in the next subsections, the settings of our Indic language support, and the way this method will lead to a very efficient documentation system.

3.1 System Overview

Our proposed system will be aimed at performing high-precision transcription and abstractive summarization of poly-lingual (Hindi, Kannada, Punjabi, and English) meeting. The Linguistic-Efficiency trade-off that our proposed system will address is highly important, in the sense that global systems can perform high-accuracy tasks in the English language, but fail to do so in the regions with

low resources. Our system will be an end to end system which will take unstructured audio and give structured actionable insights in the choice of regional script given by the user.

3.2 Conceptual Framework and Architecture

There are three stages in the architecture, which transform crude information of acoustic information into finer textual information:

1 Powerful Transcription Phase (Whisper ASR): The Whisper OpenAI model is the core of the operation of the transcription system. The transformer encoder-decoder architecture of Whisper is large-scale as opposed to other models. To make the Chinese accents as phonetically accurate as possible, we adapt the model and use the "Turbo" or "Large-v3" variant of the model in our framework. The model estimates the audio signal into a high-dimensional space, and identifies the limits of languages and transcribes code-switched speech with minimal leaking.

2 Inference Acceleration Stage (Groq API): To make the system workable in real world professional application, we stream the generated transcripts through the Groq LPU inference engine. This reduces significantly the latency which is typically entailed with the Large Language Models, allowing near-instant summarization.

3 Abstractive Regional Synthesis (Llama 3): The second important concept is Cross-Lingual Synthesis. The uncoded transcript is then processed through the Llama 3 of Meta. Instead of compressing the text directly, Llama 3 is requested to perform Abstractive Synthesis - remove the disfluencies and convert the vital action items to the target regional script (e.g. Gurmukhi or Kannada). This will see to it that the end product is more than a mere recapitulation - it is linguistically and culturally fit document.Inference Acceleration Stage (Groq API): In order to make the system practical for real-time professional use, we pipe the produced transcripts through the Groq LPU inference engine. This cuts down dramatically on the latency that is normally involved with Large Language Models, enabling near-instant summarization.

3.3 Mathematical Formulation of the Proposed Framework

The system may be that having a formal description of proposed multilingual meeting transcription and summarization pipeline. In order to have the form of the system the system in question may be modeled as a series of methods of transformation that transform raw audio signals into structured textual summaries. The input audio signal may be denoted by A. Automatic Speech Recognition (ASR) module using Whisper model transfers the audio signal into a textual transcript T. The following steps may be illustrated in the following way:

$$T = f_{\{ASR\}(A)} \tag{1}$$

where:

- A represent is a multilingual audio signal input,
- f_{ASR}refers to the transcription model realized based on the Whisper model
- T represents the feasible solution.

The resulting transcript is then fed into the abstractive summarization module informed by the Llama 3 large language model to get a concise summary of the meeting S. Such change can be stated as:

$$S = f_{\{LLM\}(T)} \tag{2}$$

where:

- f_{LLM} is the abstractive summation task of Llama 3 model
- S refers to the documentation of the latter in written summary format in the target language or regional script.

The entire framework can be illustrated by joining the two phases and can be shown as:

$$S = f_{\{LLM\}(f_{\{ASR\}(A)})} \tag{3}$$

Such a formulation constitutes the end-to-end pipeline according to which raw multilingual audio undergoes text-to-text transcification and then verbal representation with a large language model. Also, the total latency of the proposed system can be modeled as:

$$L_{\{total\}} = L_{\{ASR\}} + L_{\{LLM\}} \quad (4)$$

where:

- L_{ASR} is the ASR module latency on transcription.
- L_{LLM} represents the latency of summarization of the Llama 3 module.
- L_{total} represents the total time that the system would have taken to process

Such mathematical notation focuses on the modularity of the given framework and explains the interaction between transcription and summarization elements in the context of the system pipeline.

3.4 Novelty and Comparison with Existing Models

The novelty of our framework is in its practical combination of SOTA models to address regional communication issues.

- In contrast to Monolingual Solutions: Compared to the normal ASR solutions (such as Google Speech-to-Text, which inherently cannot intra-sententially code-switch), our system is specifically constructed to address the regional regionalism problem.
- Referring to Extractive Summarizers: Unlike most meeting solutions where only existing sentences are indicated, our implementation of Llama 3 allows Abstractive Summarization, where new and coherent sentences are always generated to free up the very essence of the meeting.
- Compared to High-Latency Frameworks: In our implementation of the Groq API, we have shown that high-performance AI documentation does not need a large hardware system to be developed, and so is a highly feasible implementation choice to develop competitive and lightweight professional tools in a mobile-first multilingual society.

4. Experimental Setup

This section explains the hardware configurations, access data, and technical provisions used in order to test the proposed pipeline.

4.1 Dataset Descriptions

In order to test the system under the conditions of various languages, two main open-source benchmarks were used:

1 **AI4Bharat Kathbath Dataset [12]:** we considered the nomadic Kannada and Punjabi subsets of the AI4Bharat Kathbath speech corpus, which available is an open dataset of sufficient scale in multiple Indian languages and contains records of various speakers, accents and environments.

2 **IIT-Bombay Hindi-English Code-Switching Corpus [13]:** This was the dataset that was utilized to test the capability of the system to handle intra-sentential code-switching in which speakers commonly switch between Hindi and English across the same sentence.

4.2 Hardware and Inference Environment

This provide access to the underlying hardware employing specialized microlanguage programming languages for creating scripts. Hardware and Inference Environment Technologies There are technologies available that are able to give access to the underlying hardware where scripts are created

using special purpose microlanguage programming languages.One of the variables was the computational efficiency of the pipeline. The setup included:

- **Audio Pre-processing :** Add a sentence which says: To avoid high-fidelity transcription, all input audio was pre-processed with a Voice Activity Detection (VAD) filter and resampled to a 16 kHz mono-channel format which is the best sampling rate used by Whisper transformer encoder.
- **ASR Processing:** Whisper Large-v3 (Turbo) had been used to carry out original speech-to-text conversion.
- **LPU Acceleration:** All language synthesis tasks (summarization and region) and the language synthesis across regions were offloaded to the Groq Language Processing Unit (LPU).
- **Baseline Hardware:** To carry out the comparative latency analysis, the default, cloud-based, NVIDIA T4 aspire was utilized as a benchmark to enable comparison of the LPU architecture speedup.

4.3 Evaluation Metrics

There were three main metrics of performance of the system:

- **Word Error rate (WER):** the word error rate is the usual measure of accuracy of the automatic speech recognition system. It scans the disparity between the anticipated transcription and the reference transcription.

$$WER = \frac{(S + D + I)}{N} \quad (5)$$

where:

- S is given as the representation of substitutions
- D is resp. the number of deletions.
- I represents the number of insertions
- N is the number of words used up in the reference transcript.
- A lower WER value indicates higher transcription accuracy.

- **ROUGE-L Score:** ROUGE- L is defined as deciding the quality of the generated summaries based on the length of the similar longest common sequencing (LCS) between the database generation summary and the reference summary.

$$ROUGE - L = \frac{LCS(X,Y)}{m} \quad (6)$$

where:

- $LCS(X,Y)$represents the longest common subsequence between the generated summary $X$and reference summary Y
- $m$represents the total number of words in the reference summary

Higher ROUGE-L scores indicate better semantic similarity between the generated and reference summaries.

- **End-to-End Latency:** The end to end latency is the overall time spent by the system in processing the input audio and producing the end product which is a summary.

$$Latency = T_{output} - T_{input} \quad (7)$$

where:

- T_{input}represents the time when the audio input is received

- T_{output} represents the time when the final summary is generated

Lower latency values indicate better real-time performance of the proposed system.

5. Results and Discussion

5.1 A. Quantitative Performance Analysis

The system was tested on three aspects, including Accuracy (WER), Semantic Retention (ROUGE-L), and Processing Speed (Latency). Table I summarises the findings.

Table 1. The quantitative results are summarized

Language Profile	WER (%) ↓	ROUGE-L ↑	Latency (s)
English	5.2	0.85	0.32
Hindi-English	12.1	0.77	0.40
Kannada	16.8	0.71	0.48
Punjabi	17.4	0.68	0.52

5.2 Hardware Acceleration and Latency Breakthrough

The hardware acceleration and latency breakthrough offers enhanced performance by significantly improving power savings and reducing system-level latency using the optimized operating system. The most pronounced result in this research is the enormous decrease in the end-to-end latency. The Table II demonstrates that, the architecture that was accelerated with LPU showed a factor of almost 10.7% improvement over the standard GPU baseline.

Table 2. presents the latency comparison between LPU and GPU.

Metric	Baseline (GPT-40/GPU)	Our Pipeline (LPU)	Improvement
Avg. Latency	4.82s	0.45s	~10.7x Faster
ROUGE-L (Avg)	0.81	0.78	-3.7%
WER (Kannada)	16.5%	16.2%	+0.3%

Whereas in a typical pipeline, there will be some degree of lag that will interfere with real time transcription, in our system, there was an average delay of 0.45 seconds. This makes the summation which is generated accessible nearly as soon as the audio buffer is handled.

5.3 Qualitative Discussion: The Effect of Baseless Generatively Correction.

An important point that we noticed during our tests is that the scores of the ROUGE-L are very strong despite the change in WER.

- **Observation:** In the Punjabi data the WER was 17.4 and the quality of the summaries was good.
- **Intuition:** This has prompted a Generative Correction phenomenon where randomly occurring phonetic mistakes by the ASR layer are corrected by the Llama 3 model as an abstractive synthesizer that requires the correct context. An example is when a speaker adds Schedule meeting next Somavara, the regional word is wrongly understood by Whisper, and the tool Llama 3 was able to interpret that it is an intent to schedule, as the other tokens in the surrounding are English.

5.4 Error Analysis and Limitations:

We observed that the system performance is a bit lower when there is a rapid overlapping of a speaker (cross-talk). Very often, due to overlapping voices there may be occasions where transcription tokens are

combined, as they cannot be segregated by the current pipeline since it lacks a separate Diarization module. This is one of the major areas of the Future Work of this paper.

6. Conclusion

To sum up, this paper has presented a novel and low-latency model of the automatic documentation of multilingual meetings that has been specifically constructed to handle the linguistic issues of code-switched speech between Hindi and English, and also Kannada and Punjabi. Through arranging a pipeline consisting of Whisper Large-v3 model to achieve high-fidelity transcription, to use for inference acceleration Groq LPU and to achieve the writing of the region abstractively Llama 3 we have demonstrated that it becomes feasible to bridge the disparity between the global ASR standards and the regional linguistic demands.

Our experimental results indicate that hardware acceleration of specialized components (Groq LPU) results in more than 85-fold reduction in latency and enables AI documentation in real time in a mobile-first and multilingual workplace. Furthermore, abstractive summarization will guarantee the semantic meaning of the meeting is present in the final regional script despite a moderate error in transcription.

References

[1] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. 40th Int. Conf. Machine Learning (ICML), 2023, pp. 28492–28518.

[2] C. Li, M. Liu, and A. Waibel, "Multilingual end-to-end speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2347–2360, 2020.

[3] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 3730–3740.

[4] J. Zhang, Y. Zhao, and R. J. Miller, "Neural abstractive meeting summarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 153–165, 2021.

[5] M. R. Costa-jussà et al., "No language left behind: Scaling human-centered machine translation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 1–15, 2023.

[6] R. Kumar and S. Singh, "Challenges in code-switched speech recognition," IEEE Access, vol. 9, pp. 102345–102356, 2021.

[7] K. Wang et al., "Abstractive summarization for spoken documents," IEEE Transactions on Multimedia, vol. 22, no. 9, pp. 2384–2395, 2020.

[8] H. Zhu et al., "Meeting summarization using deep learning," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 781–785.

[9] D. Povey et al., "The Kaldi speech recognition toolkit," in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2011.

[10] National Institute of Standards and Technology (NIST), "Meeting recognition and summarization benchmarks," IEEE Government Technical Report, 2018.

[11] P. Rao, M. Pandya, K. Sabu, K. Kumar, and N. Bondale, "A study of lexical and prosodic cues to segmentation in a Hindi-English code-switched discourse," in Proc. Interspeech, Hyderabad, India, 2018, pp. 23–27.

[12] P. K. Goyal et al., "AI4Bharat Kathbath: Open speech corpus for Indian languages," AI4Bharat Research Report, 2022.

[13] A. Bhat, M. Choudhury, and K. Bali, "The IIT Bombay Hindi–English code-switching corpus," in Proc. Interspeech, 2017.

[14] Y. Huang, X. Zhao, and L. Chen, "Evaluation of Whisper-based multilingual speech recognition for code-switched speech," IEEE Access, vol. 12, pp. 45521–45534, 2024.

[15] J. Wang, R. Gupta, and M. Li, "Multilingual meeting summarization using large language models," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2025, pp. 7843–7855.