sciendo INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS

Research article | DOI: 10.2478/ijssis-2025-0016 Issue 1 | Vol. 18 (2025)

A multi-threaded approach for improved and faster accent transcription of chemical terms

Sonali Kothari*,
Shwetambari Chiwhane,
Shreeja Mehta, Pranav Naranatt,
Md. Asad Ansari and Rithwik Satya

Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra, India

*E-mail: sonali.kothari@sitpune.edu.in

Received for publication February 05, 2025.

Abstract

The proposed research aims to improve real-time transcription of English accents by focusing on transcribing technical terms related to chemistry using an innovative framework that blends multi-threading and asynchronous programming techniques for enhanced speed and accuracy. The method lays the groundwork for real-time transcription by taking into account the nuances of regional accents. Highlighting the significance of embracing language and cultural diversity underscores the importance of advancements in speech-to-text technologies. Continuous learning mechanisms and adaptive language models serve to enhance the quality of transcription. Ensuring performance in time is guaranteed through the use of scalable infrastructure while maintaining operational efficiency through ongoing evaluation and enhancement processes. This innovative approach not only enhances transcription precision but also establishes a system that appreciates and embraces the diverse range of linguistic nuances present in Indian English accents, promoting the development of inclusive and effective communication technologies.

This paper presents a comprehensive framework

- for real-time transcription of English speech in the rich tapestry of regional Indian accents, with a distinctive focus on the intricate classification of chemical terms;
- to navigate through the unique linguistic landscapes, addressing the challenges posed by the diverse accents prevalent across various Indian regions;
- to emphasize the nuanced identification and categorization of chemical terminology within the transcribed content.

Keywords

chemical terms classification, Indian regional-accents of English, multi-threading architecture, chemical entities, speech to text conversion

I. Introduction

In India, the linguistic diversity is like a never-ending tapestry. During this process of synthesis, minor variations in regional pronunciation pose significant difficulties in accurately capturing spoken English [1,3]. However for real time synthesis of speech, the problems with regional speech variation are increased. Hence alongwith identification of pronunciation defects, it is important to tune subject-related subtleties. This paper explores the crux point at which

regional Indian accents meet unbridled chemical terminology in transcribed speech, for instance. With English reshaping itself over the various forms of languages found in India, regional accents are also expressions of exuberant identity. However, usual speech synthesis reduces accuracy of conversion with regional accents. Furthermore, in transcribing, one must identify and classify terms that are particular to specific disciplines, particularly in the realm of chemistry.

To confront this multifaceted challenge, this research aimed to develop an integrated framework that not only handles different regional Indian accents but also separates out and recognizes technical terms in chemistry with ease. The research acknowledged that accurate transcription requires more than just getting the words right in sequence from unknown accents. It needs an understanding of domain-specific jargon, particularly the complexities of chemical terminology. The proposed research seeks to bridge this gap by providing a comprehensive solution that augments both linguistic and domain-specific accuracy. Figure 1 illustrates the high-level implementation approach, where an input audio file is verified, chemical terms are identified within the audio file, and a text transcription for the identified chemical terms is provided.

To enhance the efficiency of the proposed framework, this study introduces the incorporation of multi-threading and asynchronous programming paradigms. The asynchronous nature of the proposed architecture enables parallel processing, a key feature essential for real-time transcription, thus addressing the urgency inherent in spoken communication.

This paper unfolds a novel chapter in the realm of speech transcription, where the celebration of linguistic diversity converges with the precision demanded by specialized domains. The proposed framework not only envisions a more accurate and efficient transcription process but also stands as a testament to the adaptability of technology in embracing the richness of regional linguistic expressions within the broader landscape of English.

II. Methodology

a. Dataset

This section offers a glimpse into the central role that datasets play in designing and testing the performance of proposed transcription systems. The high-quality and diverse nature of the data ensures accuracy and ruggedness in real-world implementations. Using two types of data, each directed at a different aspect under investigation for transcription and discrimination, the most recent study employed one method for research.

For fine-tuning the system, the first set consists of sound recordings that test its ability to transcribe different aspects of Indian-accented English speech, including accents specific to recording studio and whispering. This set of sound recordings represents a wide variety of regional accents in India, with each dialect exhibiting distinct marks or shades. Every detail of how the words sound, including tone and speed, is carefully captured here. The inclusion of whispered speech data enables us to cope with the transcription problems created by low volume or soft speech. The second set comprises text data extracted from various online sources, such as scientific literature, web pages, and other textual material, giving prominence to terms related to chemistry [4].

The reason for including these chemical terms was to assess the system's ability to authentically identify and organize specific domain language within transcribed text without errors or misleading information. The following section discusses dataset composition, dataset source, and dataset utilization in the proposed research for training various models on audio dataset.

a.i Audio dataset

According to research papers presented, the audio dataset used in the proposed research experiment were suitable for testing the accuracy of

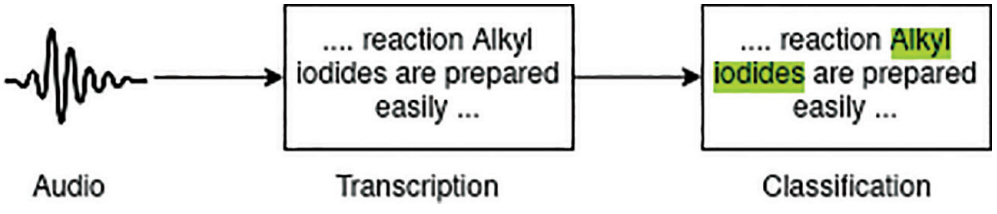

Figure 1: Overview of the proposed work.

transcription systems after being adjusted with whispered speech. The audio dataset is mostly recorded in Indian regional accents. A 50-h collection of such audio material became especially critical for the initial training and validation phases of the proposed research.

The audio dataset used in this dataset was weakly supervised, reflecting the inherent difficulties of whispered speech recognition in real-world environments. This audio dataset spans different Indian regional accents, gathering and rendering different ways of speaking as well as speed and how a person pronounces the words. The current quality level in this area can then be gauged through trial and error by people using proposed standards, until an acceptable version emerges for public release.

The proposed research combines the method described for fine-tuning the system using an audio dataset [5]. Through repeated training and testing, this research produced an initial model that was tuned for whispered speech with a word error rate (WER) of 23%. This initial result signified a massive advance in the system's capability to transcribe whispered speech, standing for the difficulties and achievement of the proposed research under regional Indian accents.

a.ii Chemical dataset

The ChemDataExtractor toolkit was used to identify chemical data from the transcription. It is trained using a collection of 3592 chemical articles from The American Chemical Society (ACS), The Royal Society of Chemistry (RSC), and Springer. It collects information from abstract, text, captions, and tables of the articles for natural language processing using the CHEMDNER corpus [11]. This process involves splitting of transcribed sentences into tokens, normalization of these tokens, clustering of words based on similarities using unsupervised learning algorithms, Part of Speech (POS) tagging, and finally, named chemical entities recognition. The abstracts, texts, and captions of articles are processed separately from the tables, and the results of both are combined using a data interdependency resolver to obtain the resultant chemical terms from the transcription.

b. Initial approach

Figure 2 highlights the initial approach considered for speech recognition. The proposed research aimed to develop a real-time transcription system that handles regional Indian accents and recognizes domain-specific terms, starting with an initial model [Figure 2]. This preliminary model represents a single-pass transcription and classification system, which seamlessly integrates three crucial steps to provide a holistic solution to the proposed objectives.

b.i Audio input and conversion

In the first step of the proposed approach, audio is captured from the user, showcasing the rich variety of regional Indian accents of English. The proposed research saves the user's audio input in a standard file format so that it can be compatible with the proposed transcription system. Since then, the research has utilized the ffmpeg conversion process [6], as it is a necessary part of the first approach. In simple terms, this small step is crucial as it converts the audio dataset of both whispered and normal speech in a respective way that Whisper and the proposed fine-tuned Whisper model can understand. This audio dataset is transcribed using the Whisper model, which is an inherently trained model that understands whispered speech.

For such standardization, a few specifications are applied to make the audio transcription ready. First, the audio sample rate is set to 16,000, a baseline standard between audio quality and ease of processing. Moreover, the proposed research confirm whether the audio is single-channel or mono-channel. This reduces the number of processing steps needed to create an algorithm. Since

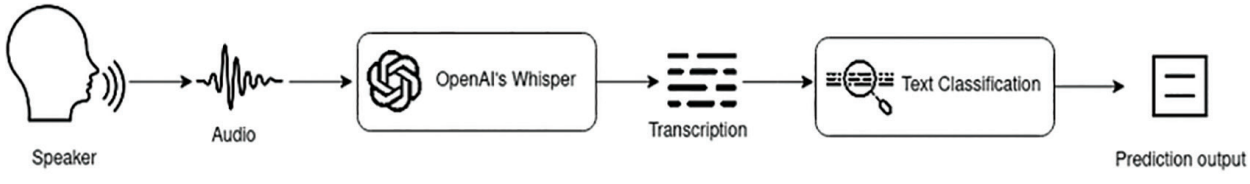

Figure 2: Initial model.

they don't follow general written language patterns, they are better suited for accurate transcription.

b.ii Audio transcription with fine-tuned whisper

As the next step of the proposed research, a fine-tuned Whisper model is used to convert the audio file into word format. Whisper has been custom-trained and tweaked to accommodate whispered speech, a style of talking necessitated by some social settings or environmental restraints [8]. The Whisper model, which is one of the components of the proposed system, acts as a verbal shadow, transcribing the Indian regional accents present in the audio dataset. The output from this stage of transcribed text serves as a basis for the next phase of the classification of domain-specific terms.

b.iii Chemical term classification

The third and final phase of the proposed initial approach focuses on the classification of chemical terms within the transcribed text. Here, the proposed research employs the ChemDataExtractor toolkit [7], a specialized tool designed for extracting and categorizing domain-specific terminology.

This toolkit intently examines transcribed communications, isolating and categorizing technical expressions within the chemical field. This enhances the system's ability to accurately discern and arrange industry-standard terminology. The integration of transcription, identification, and classification within the initial model forms a critical waypoint toward ultimately realizing a transcription framework that functions in real time. Single-pass transcription and simultaneous sorting offer basic features that future improvements can enhance, as explained later in proposed work.

c. Improved approach

In the proposed continuous pursuit of a refined and efficient real-time transcription system that addresses the nuances of regional Indian accents while accurately classifying domain-specific terminology, research presents an evolved approach. Building upon the foundation of the proposed initial two-pass model, the improved approach represents a sophisticated transcription and classification system empowered by multi-threading and asynchronous paradigms.

c.i Overview of the improved approach

The proposed improved approach revolutionizes the transcription and classification process, delivering speed, efficiency, and precision in real-time audio processing. The workflow is meticulously designed to seamlessly handle audio input as the user speaks, ensuring rapid and accurate transcription with the simultaneous classification of chemical terms. The essence of the proposed model is elucidated in Figure 3, offering a holistic depiction of the procedural flow within the enhanced architecture. In real time, audio input, typically captured through a microphone, undergoes a streamlined process of transcription and classification, seamlessly facilitated by the integration of multi-threading and asynchronous paradigms. This orchestration ensures the swift and efficient transformation of spoken words into accurate transcriptions while simultaneously discerning and categorizing domain-specific terms.

c.ii Multi-threading and parallelization

To enable real-time transcription with swiftness, the proposed system leverages multi-threading and parallelization. As shown in Figure 4, the user speaks in real-time through a microphone, and the audio input is divided into segments [13]. These segments are further processed in parallel threads. Each thread is responsible for handling the conversion of audio dataset into a format compatible with the proposed fine-tuned Whisper model [8]. The use of multi-threading optimizes resource allocation and substantially accelerates the transcription process [9, 12]. This approach raises a significant issue regarding the length of the stream to be cut before transcribing it into text. Initially, the proposed findings pointed toward an algorithm named Modified Support Search Tree (MoSST) [10, 11]. However, after rigorous testing with the proposed Whisper model using real-time data from proposed initial implementation, it is concluded that the whisper's built-in normalizer accurately handles streaming. Therefore, adding an extra layer of input stream manager would be redundant.

c.iii Whisper-based transcription

The transcribed text from each thread is gathered and processed through a correction model, which is a critical component that examines the text for errors or inaccuracies. This correction model plays a pivotal role in improving the precision and quality of the

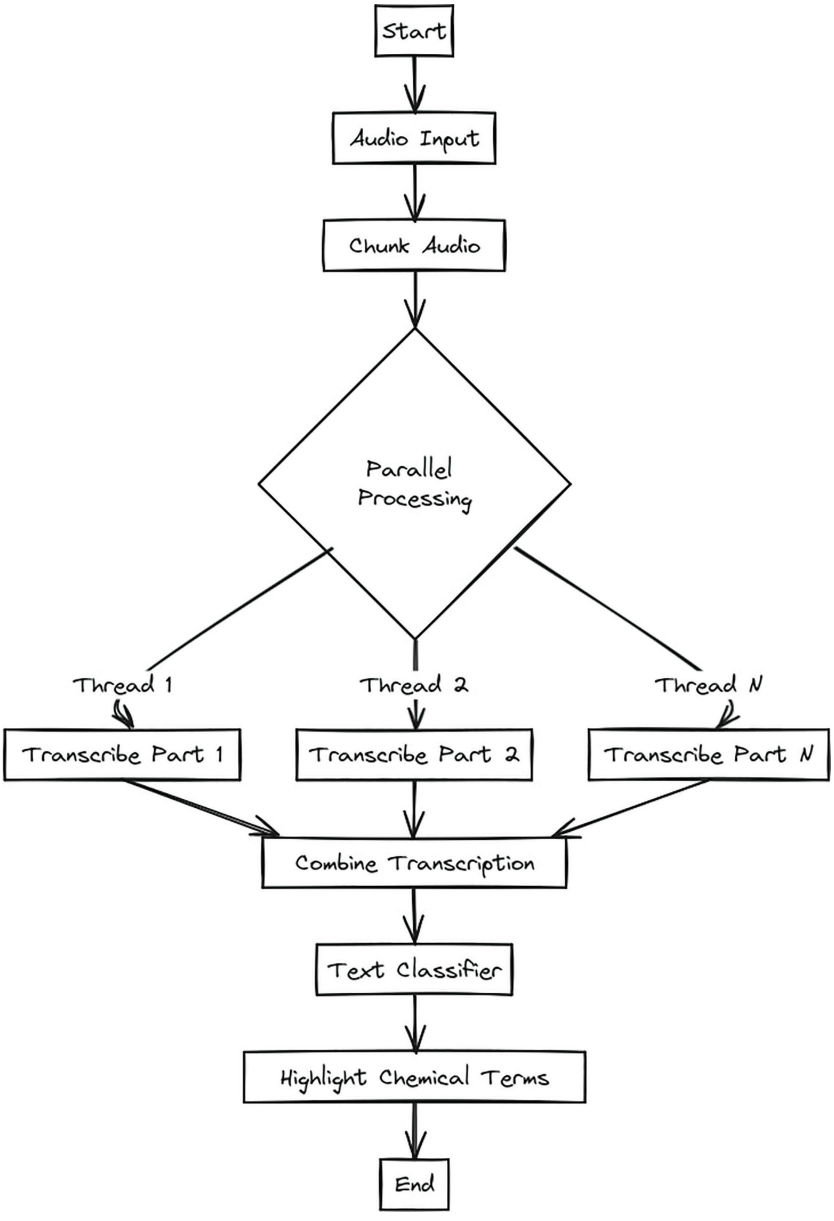

Figure 3: Flow diagram of improved model.

transcription output by addressing common issues such as mispronunciations, noise, or accent-induced transcription errors.

c.iv Chemical term classification

Following the correction model's processing, the transcribed text is then passed on for classification using the ChemDataExtractor toolkit [7]. This toolkit excels in identifying and extracting domain-specific terminology, with a particular emphasis on chemical terms. The classification process improves the system's ability to identify and categorize scientific jargon, a crucial feature for applications involving chemical domain content.

c.v Real-time presentation

Finally, the transcribed text, complete with highlighted chemical terms, is presented to the user in real-time. This real-time presentation not only enhances the user's experience but also ensures immediate access

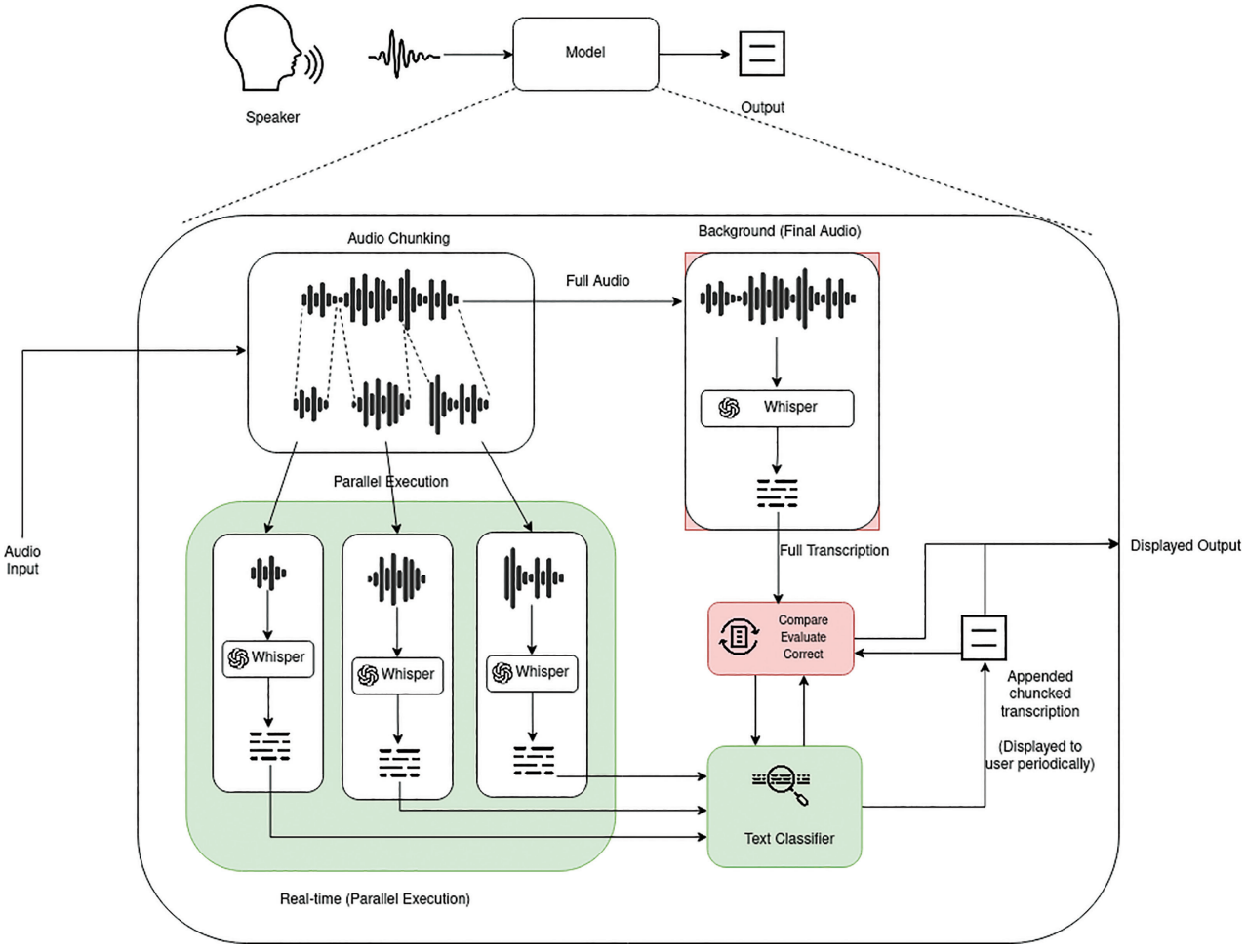

Figure 4: Improved model.

to the transcription results, allowing for quick feedback and communication.

The improved approach marks a significant step forward in the development of a versatile, real-time transcription system tailored to the unique linguistic landscape of regional Indian accents while excelling in the classification of domain-specific terminology. The implementation's design, enabled by multi-threading and asynchronous paradigms, holds the promise of streamlined and efficient real-time transcription and classification, significantly contributing to the broader domain of speech recognition and natural language processing.

III. Results

In a proposed comparative evaluation of the initial single-pass model and the enhanced multi-threaded approach, the research aimed to measure the impact of incorporating parallel processing on real-time transcription and classification. The initial model, with its sequential processing, is contrasted with the improved model that harnesses the power of multi-threading for parallelization, enabling swift and efficient transcription of live audio inputs from real users. The comparison between these two models consists of two different tests: real-time environment-based tests, simulating live user interactions, and stress tests, assessing the system's robustness under challenging conditions. It is noteworthy that in both test categories, the computation time includes the actual audio duration, providing a holistic measurement of the system's efficiency. These results not only offer insights into the system's responsiveness during real-time interactions but also gauge its resilience when subjected to intensified workloads, providing a comprehensive understanding of its practical utility and scalability.

a. Experimental setup

Both models used the fine-tuned base model of Whisper [8] for transcription. The evaluation involved taking live input from users of various accents of English in real-time, simulating the unpredictability of live user input. The proposed metrics focused on the total time taken for transcription and classification, emphasizing the efficiency gains achieved by the improved model. All computations were performed locally to avoid extra time incurred during the transmission of the audio input.

b. Recorder audio test

The initial model, which used a single forward procedure for transcription and classification, demonstrated competent performance. Even when subjected to the rigorous demands of real-time input, its performance was approaching the improved model. However, the improved model, leveraging parallel processing, showcased notable advancements in terms of speed, responsiveness, and overall accuracy.

The comparative results are succinctly summarized in Table 1 and depicted graphically in Figure 5, illustrating the tangible improvements achieved by the enhanced model across various scenarios. Figure 5 graphically presents a comparison of the performance of 10 audio file samples using both the initial and improved models. Notably, the improved model demonstrated efficacy in handling real-time audio inputs, marking a significant stride toward the practical implementation of the proposed transcription system. Table 1 discusses the performance of the initial and improved models for converting the same audio files into text and identifying chemical terms available within them.

In the planned real-time audio test, the research is focused to measure both overall performance and the time taken for the first transcription, which is important for showing how fast and efficient the system. While the initial model exhibited competent

Table 1: Comparative results (in seconds)

Audio file	Audio duration	Initial model	Improved model
audio 001	38.15	44.80	40.83
audio 002	70.97	79.53	79.83
audio 003	80.69	87.78	82.72
audio 004	54.86	62.19	59.21
audio 005	33.25	38.09	39.40
audio 006	40.93	58.66	53.68
audio 007	48.13	53.85	51.81
audio 008	33.49	38.68	35.13
audio 009	33.94	38.55	33.82
audio 010	48.95	54.28	50.15

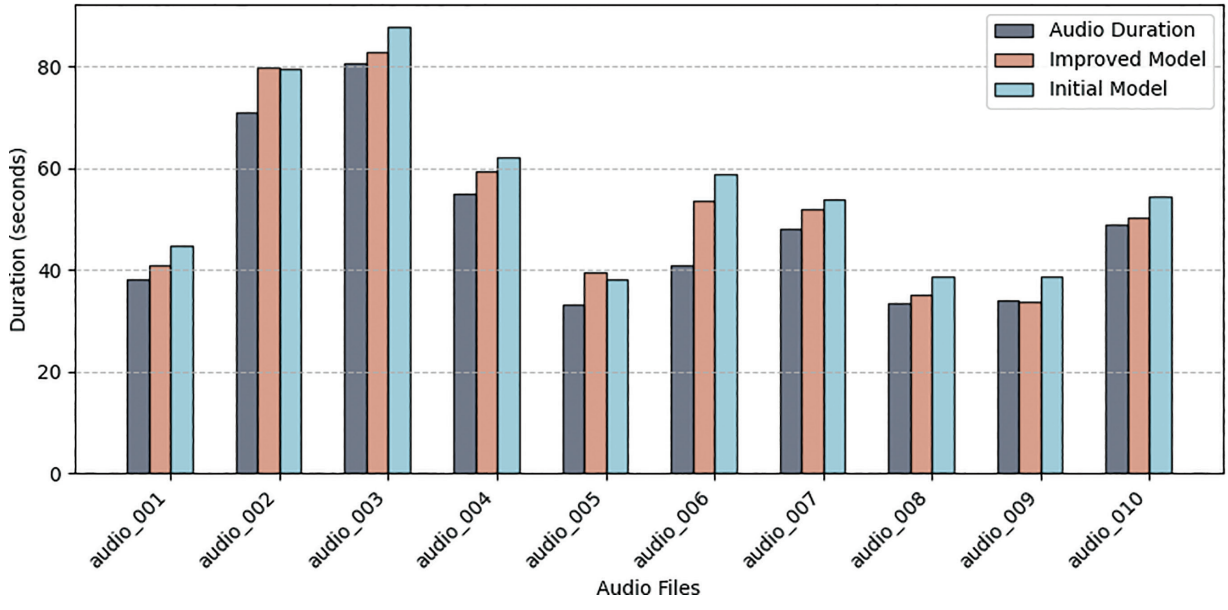

Figure 5: Comparison performance (in seconds).

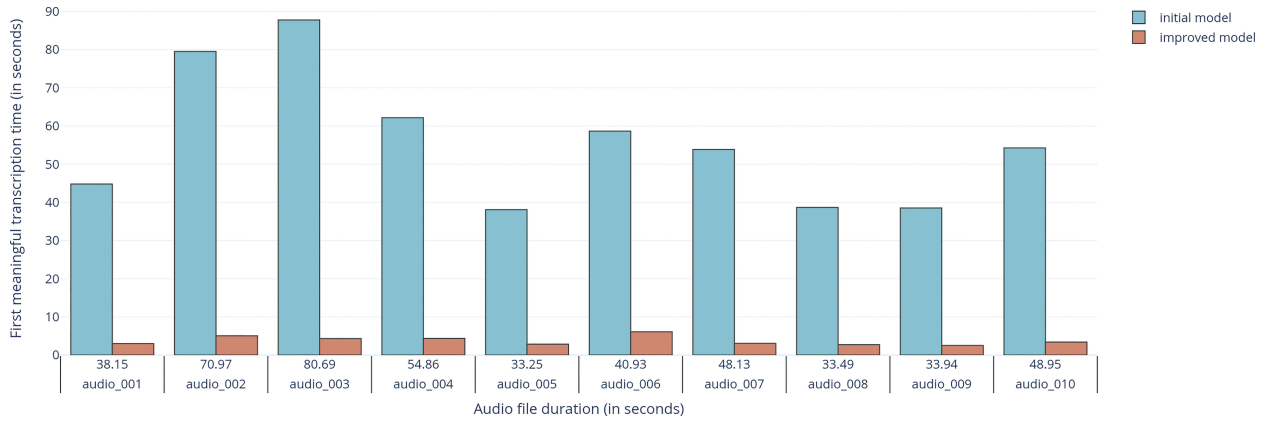

Figure 6: First meaningful transcription time.

performance, particularly in the face of real-time input demands, it notably lagged behind the improved model in terms of the time taken for the first meaningful transcription output. This disparity underscores the transformative impact of parallel processing in the proposed improved model, wherein the audio input is processed simultaneously in real-time while taking the input, significantly reducing the time required to generate the initial transcription output. This enhancement is pivotal for real-time transcription environments.

In Figure 6, it is evident that the improved model achieves a notably faster first transcription time compared to the initial model. This enhanced speed is attributed to the improved model's real-time transcription capability, which allows for instantaneous output without the need to wait for all data to be received, a key distinction from the initial model's sequential processing approach.

Table 2 highlights the time taken by the initial and improved models to provide a meaningful speech-to-text conversion for the first time after the completion of their training.

c. Stress test

In a proposed evaluation of system resilience and processing efficiency under stress, research subjected both the initial and improved models to a rigorous stress test involving long-duration audio inputs (as shown in Figure 7 and Table 3). The intrinsic challenge here lies in the sustained processing of extensive audio dataset, revealing crucial insights into each model's capacity to manage prolonged inputs.

The initial model, characterized by a sequential processing approach, demonstrated notable limitations as it processed the entirety of the input data after its completion. In contrast, the improved model, with its real-time, multi-threaded processing, exhibited a remarkable reduction in computing time. By efficiently handling audio input in parallel and providing results in real-time, the improved model showcased its adeptness in managing stress scenarios. These findings underscore the effectiveness of the proposed improved model, validating its suitability for real-time transcription scenarios. The integration of multi-threading has proven instrumental in enhancing the efficiency and accuracy of the proposed transcription system, marking a substantial leap forward in the realm of real-time speech processing.

To fine-tune models, the proposed research considered a custom dataset, and for testing purposes, the proposed research handpicked four distinct accents from that dataset, calculated the WER and

Table 2: First meaningful transcription time (in seconds)

Audio	Duration	Initial model	Improved model
audio 001	38.15	44.80	3.00
audio 002	70.97	79.53	5.05
audio 003	80.69	87.78	4.33
audio 004	54.86	62.19	4.35
audio 005	33.25	38.09	2.87
audio 006	40.93	58.66	6.10
audio 007	48.13	53.85	3.05
audio 008	33.49	38.68	2.73
audio 009	33.94	38.55	2.51
audio 010	48.95	54.28	3.40

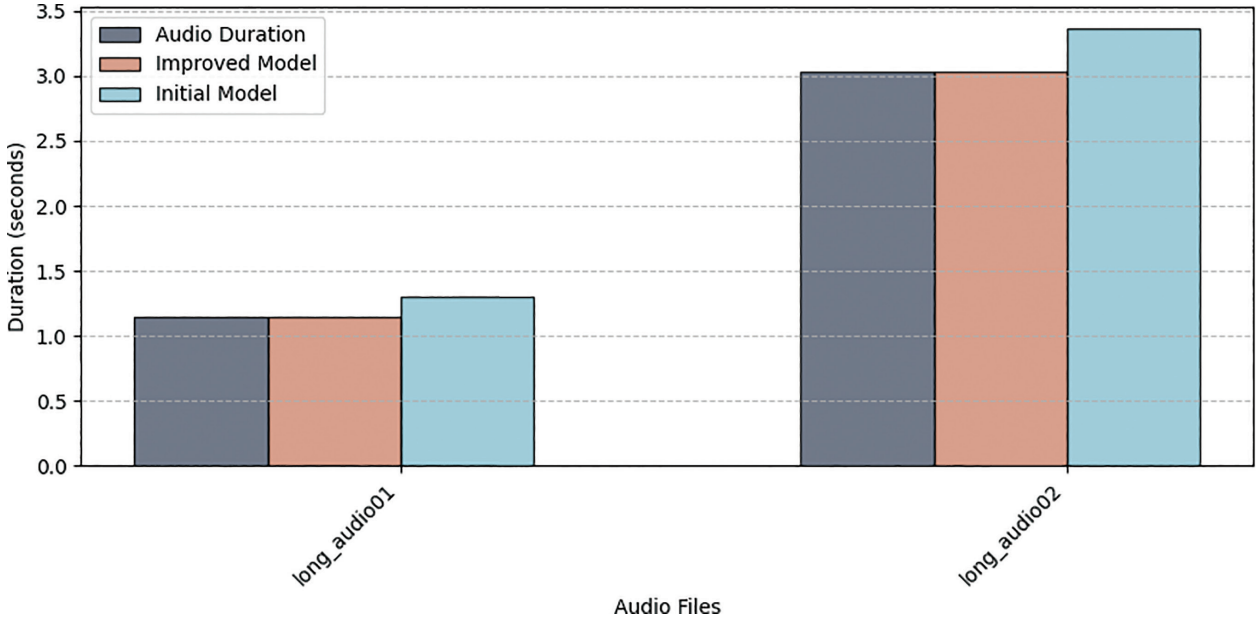

Figure 7: Stress testing (hours).

Table 3: Stress testing (hours)

Audio	Duration	Initial model	Improved model
long audio01	1.144	1.299	1.144
long audio02	3.027	3.363	3.029

time taken for three different automatic speech recognition (ASR) models, which include DeepSpeech, Wav2Vec2, and Whisper in noise and noise-free settings.

The evaluation of three ASR models, DeepSpeech, Wav2Vec2, and Whisper, using WER and transcription time metrics, was conducted on a validation set. For clean audio, Whisper achieved the lowest WER at 22.00 (as shown in Figure 8), surpassing Wav2Vec2 (36.04) and DeepSpeech (68.18). In the presence of augmented background noise, Whisper maintained the lowest WER at 28.00, followed by Wav2Vec2 (39.04) and DeepSpeech (75.28) (as shown in Figure 9). Whisper consistently outperformed the other models across all audio categories.

The study also assessed the time taken for transcription, with Whisper being the fastest at 9.55 s (as shown in Figure 10). Overall, Whisper proved most effective for accurately transcribing speech with chemical terms in Indian regional accents, while Wav2Vec2 showed promise but may need further refinement. Furthermore, a comparison between fine-tuned Whisper, base Whisper, and Google's Speech-to-Text revealed that fine-tuned Whisper marginally outperformed the base Whisper by 2.15% in WER. Both versions of Whisper outperformed Google's Speech-to-Text. Fine-tuning Whisper on Indian accents notably improved chemical term detection accuracy in Indian-accented English, achieving a 7.1% WER, while the standard Whisper had 8.6%, and Google Speech-to-Text had 9.4% WER. Whisper emerged as the most accurate ASR model for this specific task (as shown in Figures 11 and 12).

In the final implementation of the project, the test data set was run again in the completed implementation with faster-whisper for transcription and ChemDataExtractor for classification. Figure 13 shows that the data contained 100 chemical elements consisting of accented speech from multiple regions. The results were as follows: 88 of the total chemical elements were perfectly transcribed and predicted, while 2 of them were not picked up by the ChemDataExtractor classifier. The remaining 10 were falsely classified by the ChemDataExtractor as general terms like "Oh," which is a chemical term implying "alcohol," but "Oh" is also a word from the English dictionary. These are a few of the problems that could be investigated further under the study of noise reduction and better transcription.

As the above section and results are focused on DeepSpeech, Google Speech-To-Text (Google), Whisper (OpenAI), and Wav2Vec2 (Meta), a

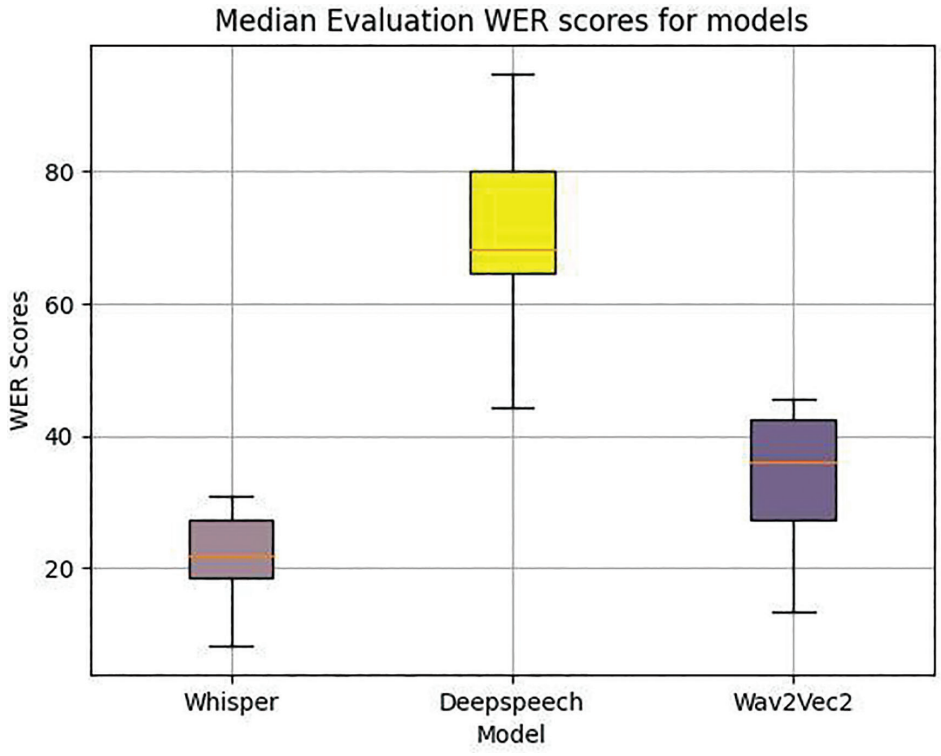

Figure 8: WER scores without noise. WER, word error rate.

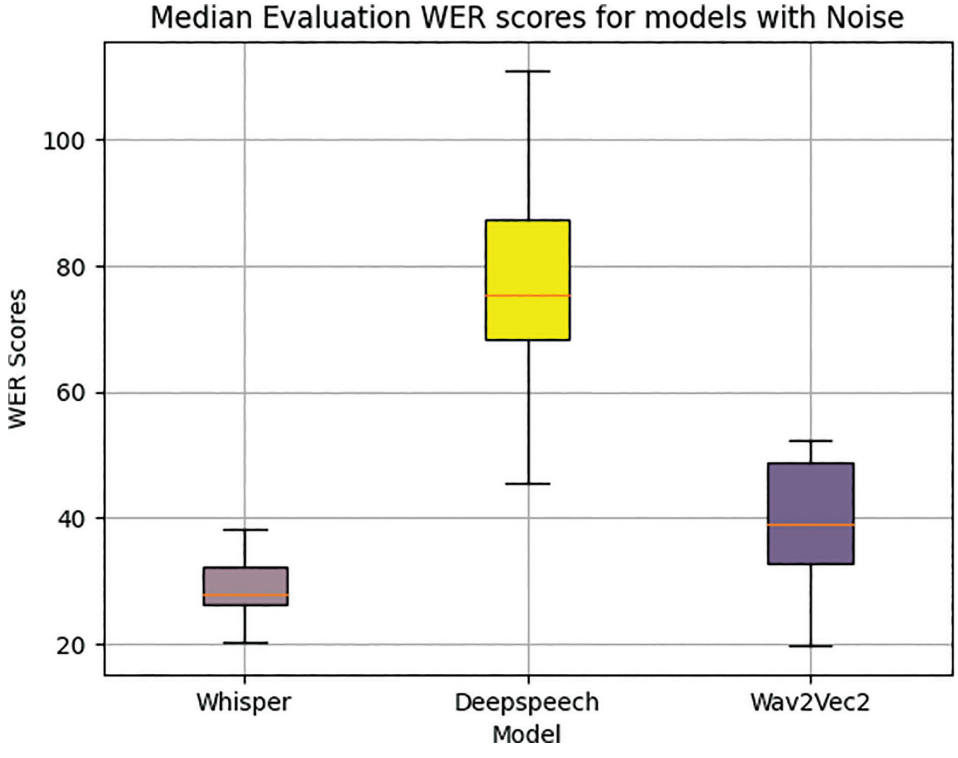

Figure 9: WER scores with noise. WER, word error rate.

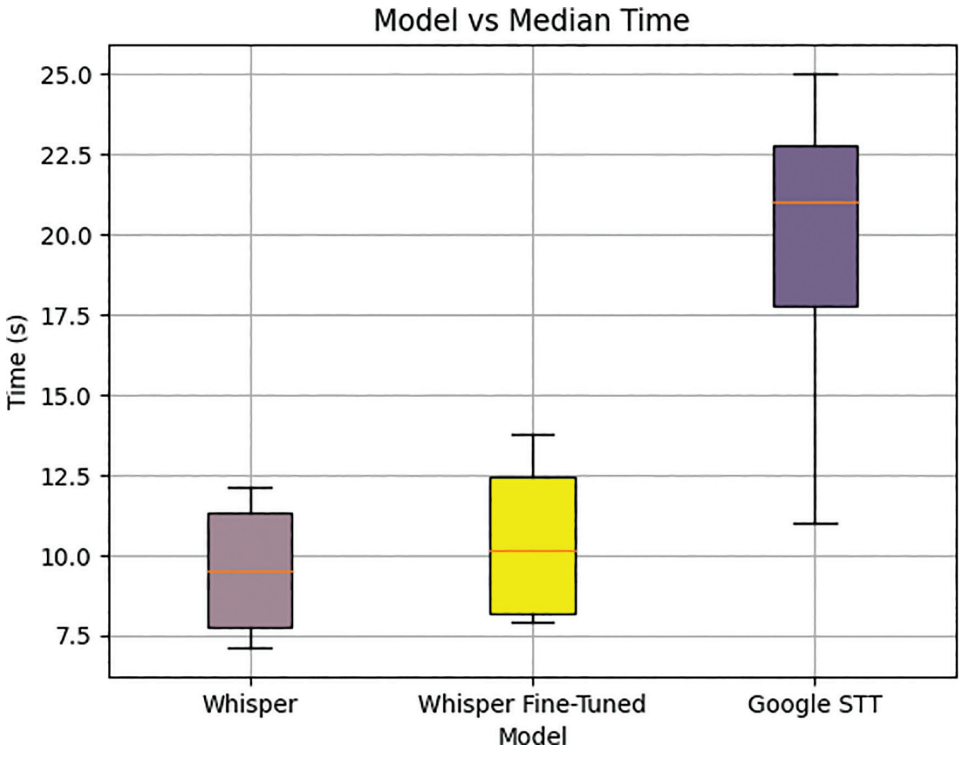

Figure 10: Time taken for transcription.

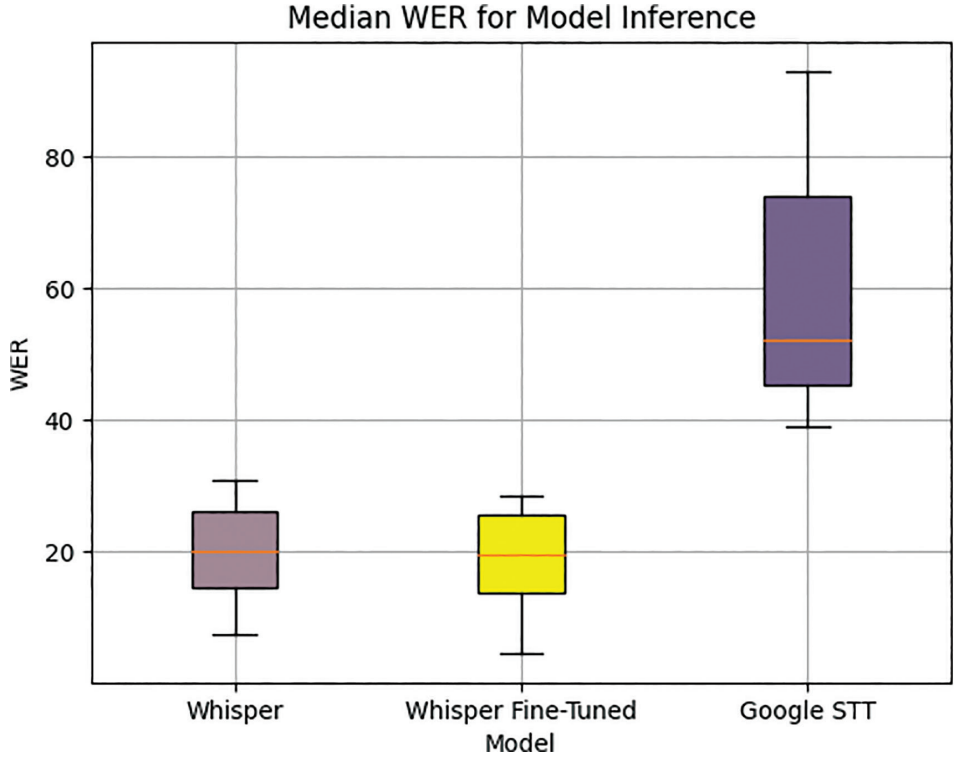

Figure 11: WER comparison with Google-STT. WER, word error rate; STT, Speech-to-Text.

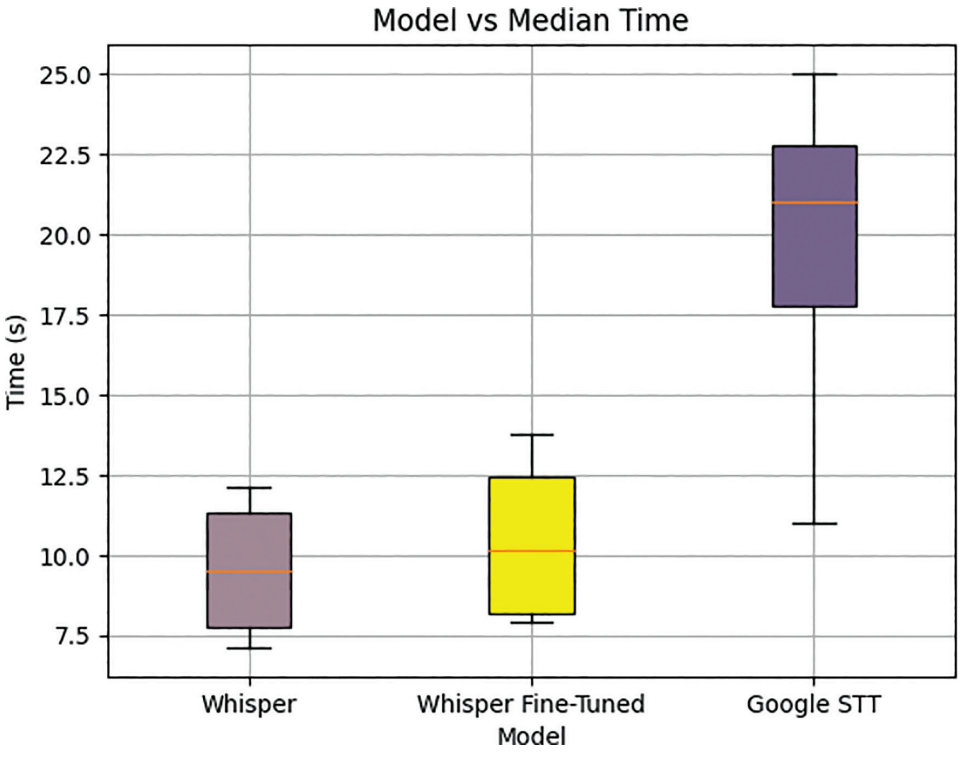

Figure 12: Time taken for transcription comparison with Google STT. STT, Speech-to-Text.

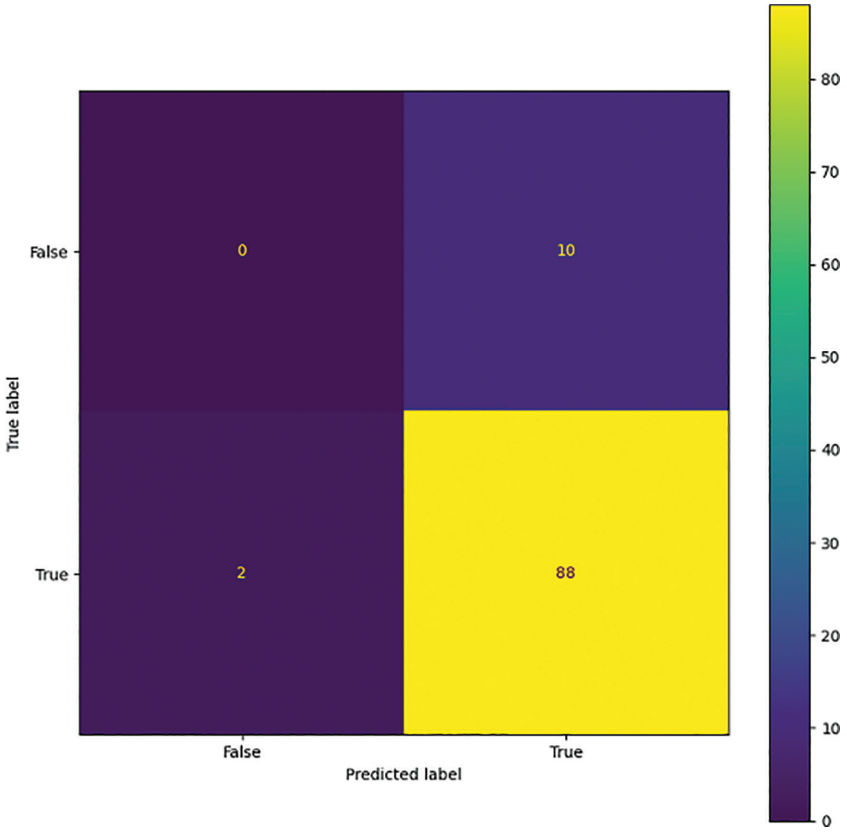

Figure 13: Confusion matrix for classification of chemical elements from text.

Table 4: Performance of existing AER systems over Indian accents

Feature	Whisper (OpenAI) [16]	Wav2Vec2 (Meta) [17]	Google STT [18]
Indian Accent Support	Strong (multilingual model trained on diverse accents) [19,20]	Varies (depends on fine-tuned dataset) [20]	Good (Google has extensive Indian English training data) [21]
Regional Variants (Hindi-English, Tamil-English, etc.)	Handles code-switching well [22]	Requires specific fine-tuning for mixed languages [23]	Decent but struggles with heavy accents [18]
Noise Robustness	Strong (performs well in real-world noisy environments) [16]	Moderate (depends on fine-tuned model) [17]	Good (handles background noise effectively) [18]
Spoken Speed Adaptability	Good (handles fast speech well) [22]	Varies (pre-trained models sometimes struggle) [23]	Good (adjusts well to fast-paced speech) [18]

AER, Assembly-enhanced recognition; STT, Speech-to-Text.

Table 5: Performance of existing AER systems for chemical term recognition

Feature	Whisper (OpenAI)	Wav2Vec2 (Meta)	Google STT
Chemical Terms Recognition	Limited (depends on general training data, not domain-specific) [16]	Can be fine-tuned for better accuracy [17]	Good (Google's general corpus covers some scientific terms) [18]
Adaptability to Scientific Jargon	Poor without custom fine-tuning [19]	Can be trained on specialized datasets [20]	Better but not perfect [21]
Handling of Long & Complex Terms	Struggles with rare chemical names [16]	Can be improved with domain-specific training [17]	Sometimes recognizes common scientific terms but struggles with rare ones [18]

AER, Assembly-enhanced recognition; STT, Speech-to-Text.

comparison of these techniques is provided with respect to Indian accents and chemical terms as discussed in Table 4 and Table 5.

c.i Indian accents performance

Indian English and regional accents present unique challenges for ASR systems due to phonetic variations, code-mixing (mixing English with Hindi, Tamil, etc.), and fast speech patterns.

c.ii Domain-specific chemical terms

Scientific and chemical terminology is challenging for ASR systems because it requires familiarity with complex, rarely spoken words.

Here are a few sneak peaks of the application created using these tools as backend (Figure 14). The web application offers live speech-to-text transcription of user-spoken words. It identifies chemical terms in the transcription and presents the titles of the top five research articles related to these chemicals. Additionally, users can listen to the titles by clicking on the "speak" button. Furthermore, users can also access information about the chemical's structure. In addition to the other features, the web application provides users with the option to email the list of articles. Users can easily do so by entering their email address and clicking on the "send" button.

IV. Conclusion

In the realm of real-time speech transcription, proposed research has significantly advanced the capabilities of transcription systems by addressing regional Indian accents and the precise classification of domain-specific terminology. The research began with an initial model that adeptly handled whispered speech and accent variations, setting the foundation for the proposed journey toward an

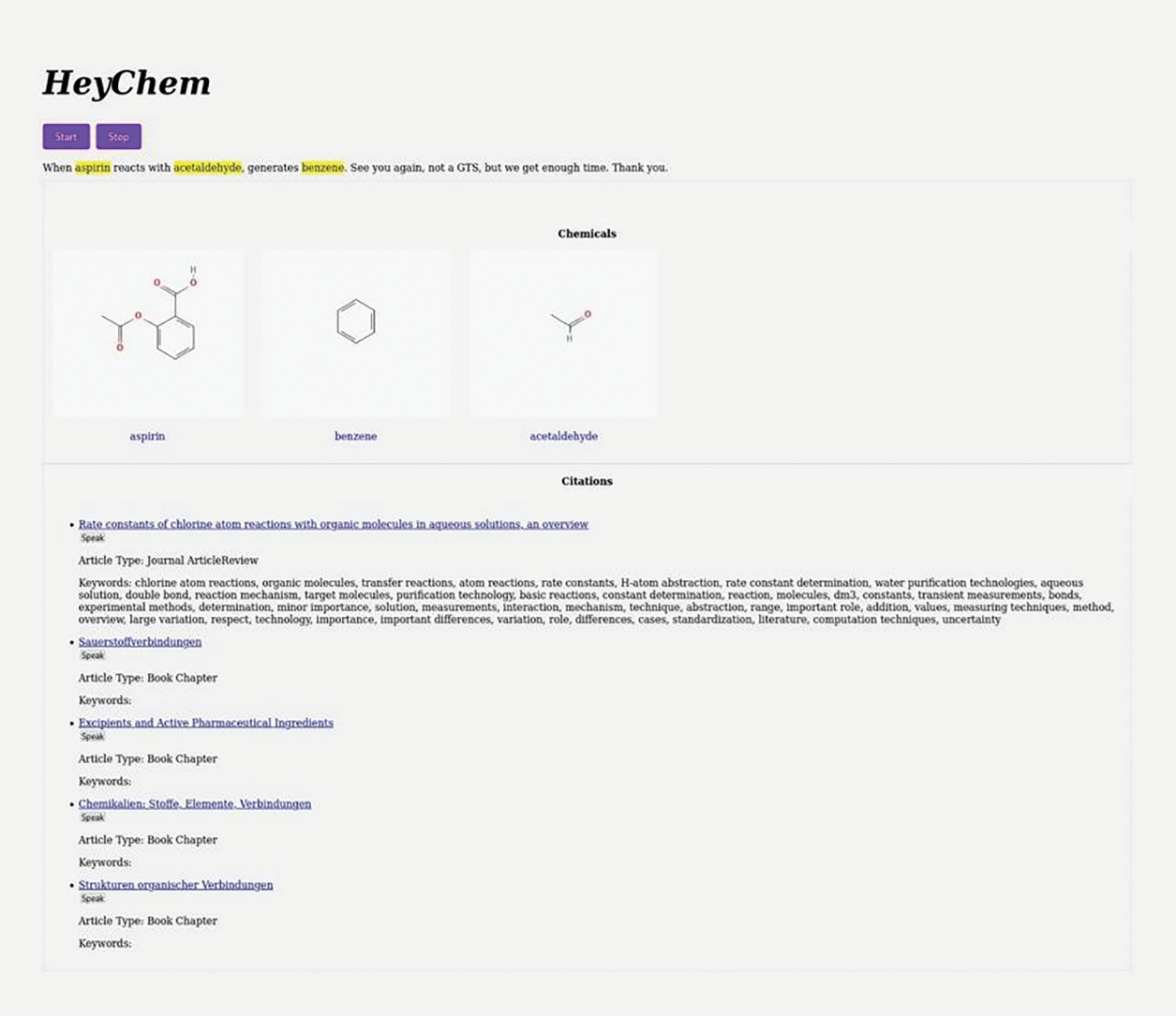

Figure 14: Web application.

enhanced approach. The improved model, featuring multi-threading and asynchronous paradigms, has ushered in a new era of real-time transcription, characterized by swiftness, accuracy, and real-time presentation of transcribed content. The integration of the fine-tuned Whisper, the correction module, and the ChemDataExtractor toolkit underlines the proposed commitment to accuracy and domain-specific precision. Real-time presentation of transcribed text with highlighted chemical terms offers a user-friendly, informative, and interactive experience. This research embodies a substantial stride toward a real-time transcription system that embraces linguistic diversity, domain specificity, and efficiency, with a vision for further enhancements and adaptability to the evolving needs of real-time transcription in diverse linguistic landscapes and domain-specific content, highlighting the transformative potential of technology in effective communication. In conclusion, the advancements achieved through the proposed improved transcription model, notably the integration of multi-threading and real-time processing, present a transformative leap in the domain of real-time audio transcription. The efficiencies gained are not only apparent in scenarios of diverse regional accents and domain-specific jargon but also hold significant promise for real-world applications such as streaming services [14] and live-telecasting videos [15]. The reduced latency and swift responsiveness of the proposed model make

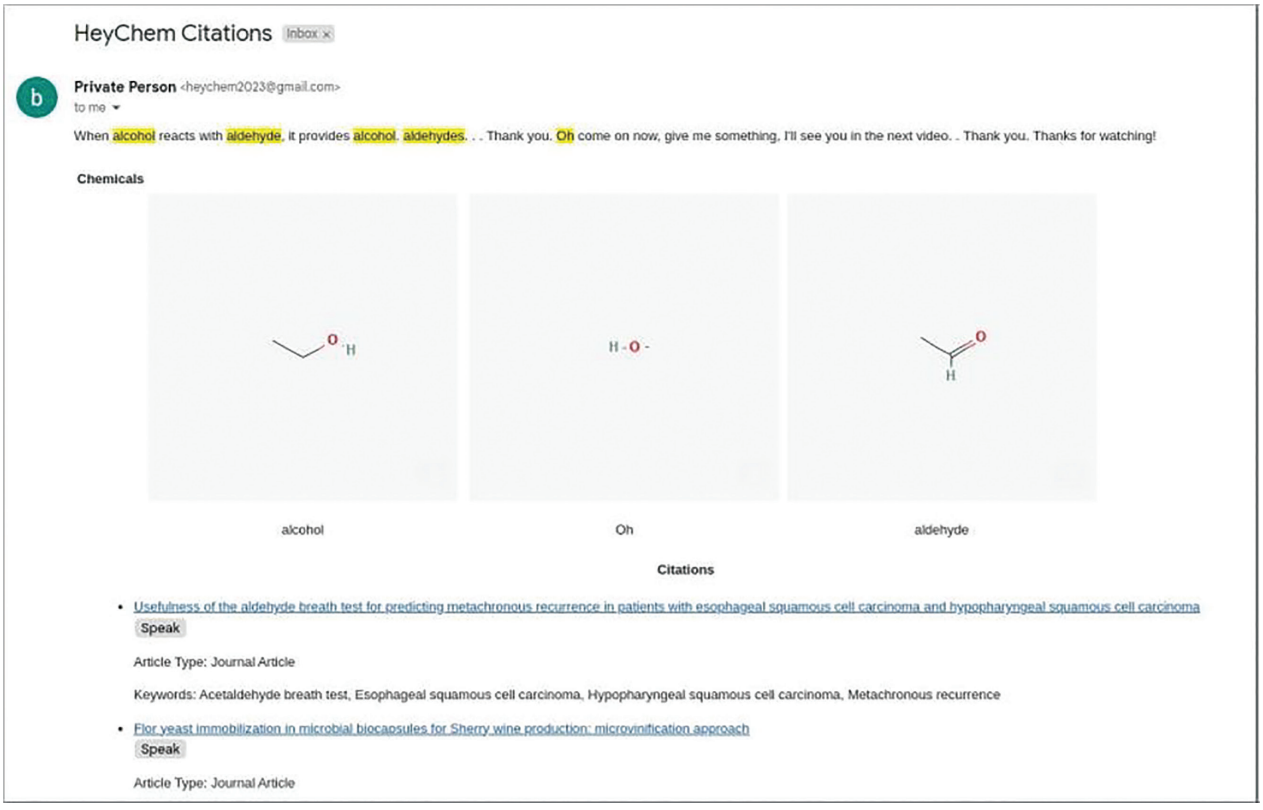

Figure 15: Email details.

it an invaluable tool for providing subtitles in real-time during live broadcasts, enhancing accessibility and user experience in dynamic, time-sensitive environments. This research lays the groundwork for a more seamless integration of speech-to-text technology into various media platforms, promising a future where transcription is not just accurate but practically instantaneous.

With the specific requirement to identify chemical terms, the proposed research is specifically focused on identifying chemical terms. But for future work, a dataset of any other domain can be provided, and the model can be trained over it without making changes in the various parameters. This can make the proposed research useful for all such domains, which need domain-specific knowledge for speech-to-text conversion and identifying specific domain-related terms.

Author Contribution

Sonali Kothari: Project administration, Funding acquisition, Supervision, Writing—Review & Editing, Conceptualization. Shwetambari Chiwhane: Project administration, Writing—Review & Editing, Conceptualization, Supervision. Shreeja Mehta: Data Curation, Writing—Original Draft, Methodology, Visualization. Pranav Naranatt: Writing—Original Draft, Methodology, Visualization, Software. Rithwik Satya: Data Curation, Methodology, Software, Writing—Original Draft. Md. Asad Ansari: Data Curation, Methodology, Software, Writing—Original Draft.

References

[1] Hinsvark, Arthur, et al. Accented Speech Recognition: A Survey. arXiv:2104.10747, arXiv, 2 June 2021. arXiv.org, DOI: 10.48550/arXiv.2104.10747.

[2] Droua-Hamdani, G., Selouani, S. A., & Boudraa, M. (2012, June 6). Speaker-independent ASR for Modern Standard Arabic: effect of regional accents. International Journal of Speech Technology, 15(4), 487–493. DOI: 10.1007/s10772-012-9146-4

[3] Vergyri, Dimitra & Lamel, Lori & Gauvain, Jean-Luc. (2010). Auto- matic speech recognition of

multiple accented English data. 1652-1655. 10.21437/Interspeech.2010-477.

[4] Lin, Zhaofeng, Tanvina Patel, and Odette Scharenborg. “Improving Whis- pered Speech Recognition Performance Using Pseudo-Whispered Based Data Augmentation.” 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2023.

[5] Chang, Jungwon, and Hosung Nam. “Exploring the feasibility of fine- tuning large-scale speech recognition models for domain-specific appli- cations: A case study on Whisper model and KsponSpeech dataset.” Phonetics and Speech Sciences 15.3 (2023): 83-88.

[6] Hock, KIA Siang, and L. I. Lingxia. “Automated processing of massive audio/video content using FFmpeg.” Code4Lib Journal 23 (2014).

[7] Swain, M. C., & Cole, J. M. (2016, October 6). ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. Journal of Chemical Information and Modeling, 56(10), 1894–1904. DOI: 10.1021/acs.jcim.6b00207

[8] Kothari, S.., Chiwhane, S.., Satya, R.., Ansari, M. A.., Mehta, S.., Naranatt, P.., & Karthikeyan, M.. (2023). Fine-tuning ASR Model Performance on Indian Regional Accents for Accurate Chemi- cal Term Prediction in Audio. International Journal of Intelligent Sys- tems and Applications in Engineering, 11(4), 485–494. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3583

[9] Phillips, S., & Rogers, A. (1999). International Journal of Parallel Programming, 27(4), 257–288. doi:10.1023/a:1018741730355

[10] Dong, Qianqian, et al. “Learning When to Translate for Streaming Speech.” ArXiv (Cornell University), 15 Sept. 2021, DOI: 10.48550/arxiv.2109.07368. Accessed 4 Feb. 2024.

[11] Krallinger, M., Rabal, O., Leitner, F. et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 7 (Suppl 1), S2 (2015). DOI: 10.1186/1758-2946-7-S1-S2

[12] Chong, Jike & Friedland, Gerald & Janin, Adam & Morgan, Nelson & Oei, Chris. (2010). Opportunities and challenges of parallelizing speech recognition. 2-2.

[13] Saito, Takashi. “A framework of human-based speech transcription with a speech chunking front-end.” 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 2015.

[14] Jorge, Javier, et al. “Live streaming speech recognition using deep bidirectional LSTM acoustic models and interpolated language models.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 148-161.

[15] Perero-Codosero, Juan M., et al. “Exploring Open-Source Deep Learn- ing ASR for Speech-to-Text TV program transcription.” IberSPEECH. 2018.

[16] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” arXiv preprint arXiv:2212.04356, 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[17] Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Advances in Neural Information Processing Systems, vol. 33, pp. 12449-12460, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[18] Google Cloud, “Speech-to-Text API,” 2023. [Online]. Available: https://cloud.google.com/speech-to-text

[19] P. Jyothi and M. Hasegawa-Johnson, “Acoustic Model Adaptation for Indian English Speech Recognition,” in Proc. Interspeech, 2015, pp. 1565-1569.

[20] A. Gupta, P. K. Ghosh, and H. A. Murthy, “Automatic Speech Recognition for Indian Accents: A Survey,” in IEEE Access, vol. 10, pp. 59347-59365, 2022. [Online]. Available: DOI: 10.1109/ACCESS.2022.3179123

[21] S. Manjunath and K. R. Ramakrishnan, “Domain-Specific Speech Recognition: Challenges and Solutions,” in IEEE Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 431-445, 2021.

[22] A. Rao, R. Patel, and M. S. Deshpande, “Performance Evaluation of ASR Systems for Indian Accents Using Deep Learning Techniques,” in 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 678-683, IEEE, 2021.

[23] S. Setty, S. R. Patil, and N. Gupta, “Speech Recognition for Chemistry Terminology Using Deep Learning,” in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 7, pp. 3456-3465, 2022.