iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** X **Month of publication:** October 2025

DOI: https://doi.org/10.22214/ijraset.2025.74665

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue X Oct 2025- Available at www.ijraset.com

Real-Time AI Note Taker with Contextual Highlights

P. Naveen Raj[1], M. Madesh[2], M. S. Ramesh[3], S. Venkata Lakshmi[4]
[1, 2, 3]*Department of Computer Science Engineering,* [4]*Associate Professor / CSE, K.L.N. College of Engineering, Pottapalayam, Sivagangai*

Abstract: ***This study presents an advanced web-based intelligent note-taking system that enhances participation and productivity in virtual meetings through the use of contextual awareness and real-time transcribing. The proposed system, Real-Time AI Note Taker with Contextual Highlights, uses cutting-edge speech recognition and Natural Language Processing (NLP) techniques to automatically transcribe live audio, identify crucial entities like names, actions, and deadlines, and dynamically highlight key points of discussion. With its interactive and scalable user interface and seamless connection with video conferencing services like Google Meet, the solution was built using the MERN stack architecture (MongoDB, Express.js, React.js, and Node.js). The system makes use of the HuggingFace T5-small NLP model for contextual tagging and summarization, as well as Google Speech-to-Text (STT) for precise speech recognition. The system generates clear summaries and well-structured transcripts that can be exported, saved, and retrieved at a later time. This framework is intended for professionals, educators, and students. It improves focus and memory, minimizes manual note-taking, and lowers cognitive load. The suggested system encourages astute cooperation and effective information handling in both academic and professional settings.***

Keywords: AI Note Taker, Real-Time Transcription, Natural Language Processing, Contextual Highlighting, Summarization, Speech Recognition, MERN Stack, Google Meet Integration, HuggingFace T5-small Model, Google STT, Knowledge Management.

I. INTRODUCTION

Intelligent systems that can automate and improve digital note-taking during meetings, seminars, and conferences are becoming more and more necessary as a result of the quick expansion of online communication and virtual collaboration platforms. Participants frequently find it difficult to stay focused while taking notes in professional and educational settings, which can result in an incomplete or erroneous recording of important topics. Recent developments in artificial intelligence (AI), natural language processing (NLP), and speech-to-text modelling have allowed for the creation of systems that can comprehend context and perform real-time transcription, allowing for automatic note production and summarization [1], [2], [3]. Conventional meeting platforms like Zoom and Google Meet mostly provide simple recording or captioning capabilities, requiring users to manually go through long sessions in order to identify the most crucial information. This procedure takes a long time and is prone to mistakes made by people. Improved accuracy through multi-task learning, multimodal fusion, and transformer-based architectures that increase contextual comprehension and granularity management has been investigated in voice recognition and transcription studies [1], [2], [5]. Creating meaningful and organized notes requires a more nuanced and context-aware transcription process, which is made possible by the incorporation of such sophisticated models. By merging contextual tagging, AI-driven summarization, and real-time speech recognition into a single platform, the suggested solution, Real-Time AI Note Taker with Contextual Highlights, seeks to address the shortcomings of current technologies. By utilizing models like Whisper, GPT, and BERT, the system is able to precisely record live audio, recognize important discussion components like names, due dates, and action items, and produce succinct summaries for easy access [3], [4], [8].

Large language models can be integrated into note-taking systems to increase productivity and accessibility, as shown by similar methods like NoteBar [13] and NexaNota [14]. Real-time collaboration, scalability, and smooth integration with current meeting platforms are, nevertheless, lacking in many of these systems. Through the usage of a MERN stack architecture (MongoDB, Express.js, React, and Node.js), the suggested framework guarantees secure cloud-based data management, effective real-time performance, and user-friendly interfaces. Motivated by the developments in Named Entity Recognition (NER) and speech-to-text generation research, the system additionally integrates entity recognition and contextual highlighting [2], [3], [9]. These elements facilitate efficient task management and decision tracking by enabling the automatic detection and classification of important information.

The significance of multi-modal data handling and adaptive natural language processing (NLP) models for extracting valuable insights from unstructured voice or text has been highlighted in earlier research on AI-based note-taking assistants and educational data extraction [10], [11], [12]. Building upon these frameworks, the current system incorporates intelligent summary strategies that use machine learning techniques based on natural language processing to produce succinct and logical summaries [15]. By showcasing a real-time, context-aware, AI-powered note-taking system that enhances productivity, accessibility, and information retention, this study advances the expanding field of intelligent meeting assistants. The outcomes demonstrate how well the suggested approach converts unstructured speech into searchable, actionable insights, and it has a great deal of promise for use in corporate, academic, and collaborative digital contexts in the future [6], [7], [8], [13], [14].

II. METHODOLOGY

The suggested Real-Time AI Note Taker with Contextual Highlights takes a methodical, sequential approach, as seen in Figure 1. Prior to creating structured meeting notes with contextual highlights, the process begins with live audio input and proceeds through Natural Language Processing (NLP), speech-to-text conversion, and summarizing. The system's architecture, which combines the HuggingFace T5-small NLP model, the Google Speech-to-Text (STT) API, and Python-based AI modules with the MERN stack (MongoDB, Express.js, React.js, and Node.js), guarantees scalability, responsiveness, and real-time performance. The web application allows users to easily connect to live or recorded meetings by recording audio from camera inputs or conferencing services like Google Meet. To improve the accuracy and clarity of transcription, preprocessing methods including audio normalization and noise reduction are used.

After the audio has been adjusted, the voice-to-text engine accurately converts speech to text for a variety of speakers and accents. After the content has been transcribed, the NLP engine does entity recognition to gain a better understanding of the context. Transformer-based models, such the HuggingFace T5-small, identify and highlight significant features including names, actions, and decisions, allowing users to concentrate on crucial insights without reading the whole transcript. The summary module uses a combination of extractive and abstractive techniques to further distill long discussions into succinct and insightful summaries. Contextual highlighting dynamically highlights important information in real time, increasing comprehension and engagement. With matching timestamps and meeting IDs for convenient management and retrieval, all transcripts, summaries, and highlights are safely kept in MongoDB.Users may read and export findings in formats like PDF or DOCX using the React-based interface. All things considered, the system offers a productive, contextually aware note-taking solution that improves cooperation, communication, and memory of information in both academic and professional settings.

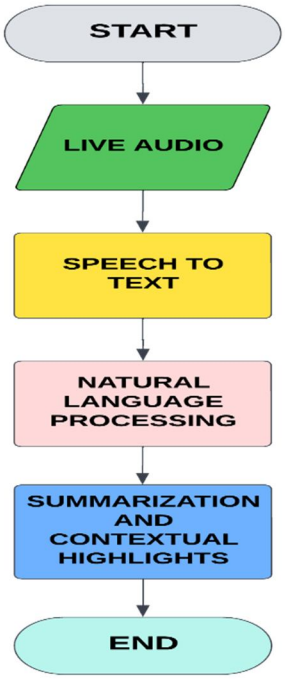


Fig. 1 Flow diagram

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue X Oct 2025- Available at www.ijraset.com

III. AUDIO PREPROCESSING AND SPEECH DATA HANDLING

Audio preprocessing is a crucial initial step in creating a dependable real-time AI note-taking system because it guarantees that incoming sound data is consistent, clear, and tuned for precise speech recognition. The clarity of the incoming audio signal has a significant impact on the caliber of later procedures like transcription, contextual tagging, and summary. In order to facilitate smooth browser-based recording, web APIs or frameworks like WebRTC or MediaStream are first used to capture the audio stream. Python-based modules that allow both live and pre-recorded inputs are then used to process the recorded audio on the backend. The system reduces background noise and improves the accuracy of Google Speech-to-Text (STT) transcription by performing noise reduction, echo removal, and normalization after recording.Techniques like spectral gating and band-pass filtering are essential for improving the intelligibility of human speech in real-world scenarios like online meetings, where background noise from fans, typing, or overlapping speech is frequent. Reducing undesired noise guarantees robustness in a range of acoustic circumstances and greatly increases transcription accuracy.

Following noise reduction, audio normalization modifies amplitude levels to preserve constant loudness, avoiding fluctuations that can compromise the accuracy of recognition. By breaking up continuous audio into digestible pieces, usually lasting 20 to 40 milliseconds each, segmentation further simplifies processing and enables the Google Speech-to-Text (STT) model to handle data effectively while maintaining temporal information. Furthermore, to increase efficiency and guarantee seamless, real-time transcription performance, Voice Activity Detection (VAD) recognizes active speech segments and eliminates background noise or silence.To increase model accuracy, the preprocessed audio is then transformed into feature representations that capture important aspects of human speech, like pitch, energy levels, and frequency. To improve audio quality for accurate transcription and contextual interpretation, the system carries out a number of preprocessing activities, such as segmentation, noise reduction, normalization, and Voice Activity Detection (VAD). With the help of the Google Speech-to-Text (STT) and HuggingFace T5-small models, this structured preprocessing pipeline creates a solid basis for precise real-time note-taking and efficient NLP-based processing in a variety of meeting scenarios.

IV. PROCESS FLOW

The Real-Time AI Note Taker with Contextual Highlights uses a systematic and organized process to accurately and efficiently translate voice into succinct, perceptive text. "Audio Capture," the initial step of the process, involves continuously recording audio input via a microphone or other connected device. This stage ensures that every spoken word is recorded clearly and consistently because the overall correctness of the system is heavily dependent on the quality of the recorded sound. While high-quality recordings minimize recognition errors, noise or echo can distort results and limit clarity. During the Audio Preprocessing step, the gathered data is further enhanced by eliminating noise, altering volume levels, and eliminating distortions or silences.This ensures that only consistent and unambiguous voice signals move on to analysis. Good preparation reduces computational load and increases transcription accuracy in dynamic real-time scenarios where several voices or background noises may clash. Following a defined set of procedures as shown in Figure 2, the system now logically connects each functional block from audio input to summary output.

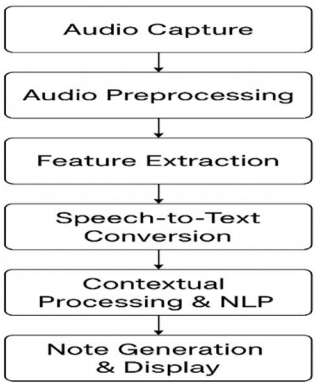


Fig. 2 Process flow

The integration of all the modules to produce real-time notes with contextual highlighting is seen in this Fig 2. During the Feature Extraction stage, the enhanced audio is first converted into measurable patterns, such as pitch, energy levels, and Mel-Frequency Cepstral Coefficients (MFCCs). The accuracy of the ensuing text generation is directly impacted by these extracted qualities, which serve as the foundation for recognizing spoken content and differentiating between words and sounds. The system processes these features using Google Speech-to-Text APIs during the Speech-to-Text Conversion step. For precise and dependable transcriptions for in-the-moment note-taking, these APIs can handle a range of accents, tones, and speaking speeds.Accurate language and high-quality processing are essential since even small mistakes might alter the intended message. The next step, Contextual Processing and NLP, refines the transcribed content by identifying entities, interpreting meaning, and eliminating superfluous utterances. Using NLP approaches with the HuggingFace T5-small models, key point recognition, and the production of simple, context-aware summaries relating to user focus areas are accomplished. The Note Generation and Display stage next arranges this polished content into a logical and aesthetically pleasing output. Contextual highlights are updated as the meeting goes on, and users can view and engage with the notes in real time. You can export, store, or share the finished product, which is an intelligent and structured translation of spoken words into useful notes.

V. SPEECH-TO-TEXT CONVERSION

This phase entails processing and employing automated speech recognition (ASR) algorithms to turn the user's real-time audio input into text. First, the input audio is preprocessed to eliminate background interference and increase clarity using noise reduction, normalization, and feature extraction, including Mel-Frequency Cepstral Coefficients (MFCCs). Following processing, the speech recognition engine receives the features and uses Google Speech-to-Text APIs to correctly recognize words, phonemes, and fully formed phrases. This method guarantees that despite taking into account different accents and speech patterns, the conversion retains contextual accuracy.Subsequent processing steps work on the finalized text, including contextual highlighting, summarization, and keyword extraction. Real-time note-taking is greatly aided by speech-to-text conversion, which accurately and swiftly converts audio input into a text format that can be further examined and improved.

VI. CONTEXTUAL PROCESSING AND NATURAL LANGUAGE UNDERSTANDING

Contextual processing and natural language understanding (NLU) are used to determine the content's meaning and pertinence after voice has been transformed into text. In order to comprehend structure, context, and word-to-phrase relationships, this stage examines the transcribed content. Tokenization, named entity recognition (NER), dependency parsing, and part-of-speech tagging are among the activities the system uses NLP approaches that are implemented with the HuggingFace T5-small model or spaCy-based pipelines. These procedures assist in comprehending linguistic patterns and identifying significant entities (such as names, dates, and themes). Along with capturing the semantic meaning of words within sentences, contextual embeddings make sure that linked keywords and phrases are understood according to their context rather than just their apparent meaning.The system also uses topic modeling and sentiment analysis to determine the underlying tone and intent of the conversation. This stage is essential for improving accuracy in downstream tasks such as action recognition, contextual highlighting, and summarization. By effectively understanding the context, the system ensures that the generated notes are both textually accurate and semantically meaningful, making them useful for immediate comprehension as well as future reference.

VII. NOTE GENERATION, SUMMARIZATION, AND HIGHLIGHTING

The content that has been contextually evaluated and processed is now converted into insightful and organized notes. Only the most crucial information required for understanding is left after the algorithm cuts out unnecessary, redundant, or filler material from the transcribed data. Providing a succinct yet thorough synthesis that precisely captures the main points of the speech is the aim of this phase. In the process of summarizing, both extractive and abstractive processes are used. Analogous to human-written summaries, abstractive summarizing rewords and reconstructs the information into a more organic and fluid shape, whereas extractive summarizing chooses key portions straight from the transcript. This guarantees that the notes are not only more concise but also more logical and easily comprehensible.

To increase readability and interest, the notes are highlighted after the summary has been created. It automatically recognizes and highlights important words, phrases, and context-driven elements such as decisions, actions, or emphasized points. This allows users to quickly identify important information without having to read the entire note. In order to make sure that the highlighted material actually represents significant concepts rather than meaningless terms, the method makes use of contextual understanding from earlier phases.

Quick reference and easy navigation are made possible by the final output's tidy structure and clear appearance. In a condensed version, each note only contains the most important information while preserving the original content's logical flow.By ensuring that the notes are accurate, relevant, and contextually rich, this step provides users with a dependable and effective method of recording speech in real time and turning it into structured written information.

VIII. SYSTEM OPTIMIZATION AND PERFORMANCE IMPROVEMENT

At this stage, a number of tactics are employed to raise the general effectiveness, precision, and performance of the AI note-taking system. By improving audio preprocessing and feature extraction methods, real-time transcription latency is reduced, increasing the processing speed of speech-to-text conversion. Sophisticated algorithms are modified to handle a range of accents, background noise, and speech fluctuations in order to increase recognition accuracy. Furthermore, to improve contextual comprehension, the natural language processing modules have been upgraded. To better capture significant information and relationships from the transcribed text, this involves enhancing entity recognition, refining language models, and optimizing tokenization and parsing techniques. The efficiency of note-generation and summarization processes is increased by the use of optimized algorithms that reduce computing time while maintaining coherence and relevance.

Another way to improve performance is to implement memory and resource management strategies, which allow the system to operate correctly even on hardware with limited computing power. Highlighted content's correctness, latency, and efficacy are regularly tested and assessed, and adjustments are made as needed. Overall, these optimization and performance improvement techniques enable the system to provide fast, reliable, and accurate real-time note-taking, making it suitable for real-world applications in lectures, meetings, and discussions without compromising quality or user experience.

IX. RESULT AND DISCUSSION

The system effectively integrates speech recognition**,** natural language processing**,** and text summarization techniques to transform real-time speech into accurate and meaningful meeting notes. The Speech-to-Text interface records live audio during the conference and accurately translates it into text, as seen in Fig. 3. Participants' spoken content will be shown on the interface instantly thanks to this real-time transcription. The generated text is accurate and clear since the model performs consistently across a range of speakers, accents, and contextual factors.

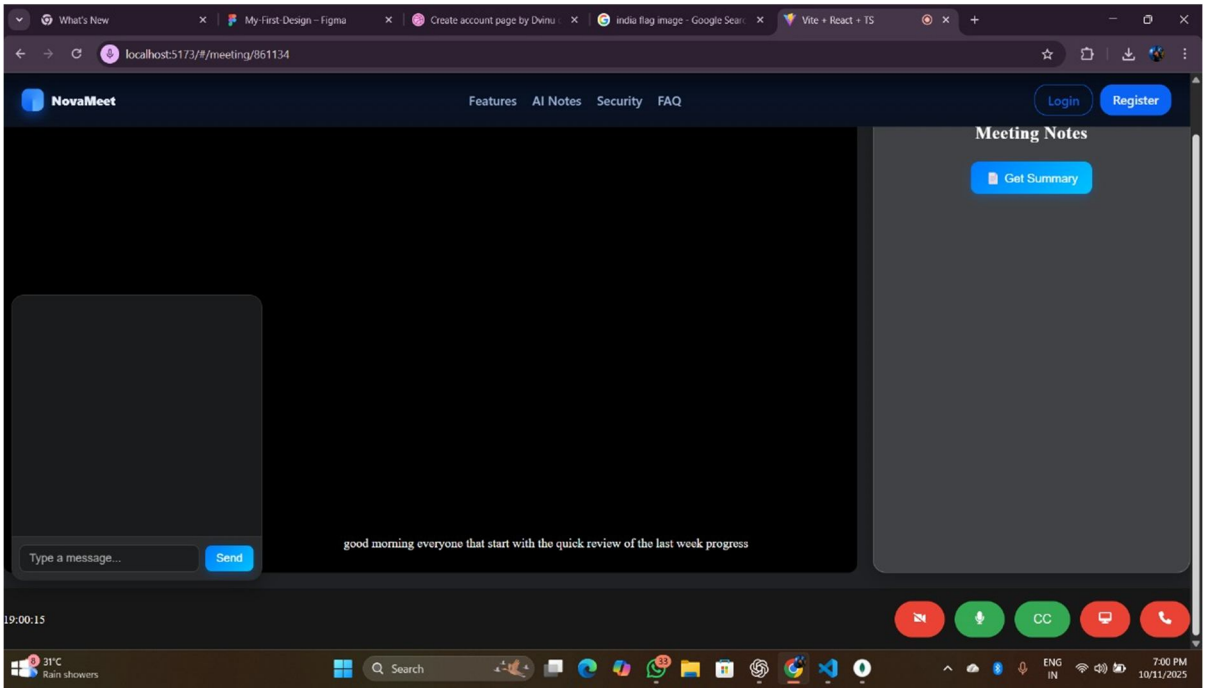


Fig. 3 Speech to Text

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue X Oct 2025- Available at www.ijraset.com

The transcribed data is safely kept in the MongoDB database in Figure 4, where each statement is associated with its respective timestamp and meeting ID. This makes it possible to efficiently arrange and retrieve meeting transcripts for processing at a later time. Scalability is supported by the defined database design, which also makes it possible for the summarization module to integrate seamlessly.

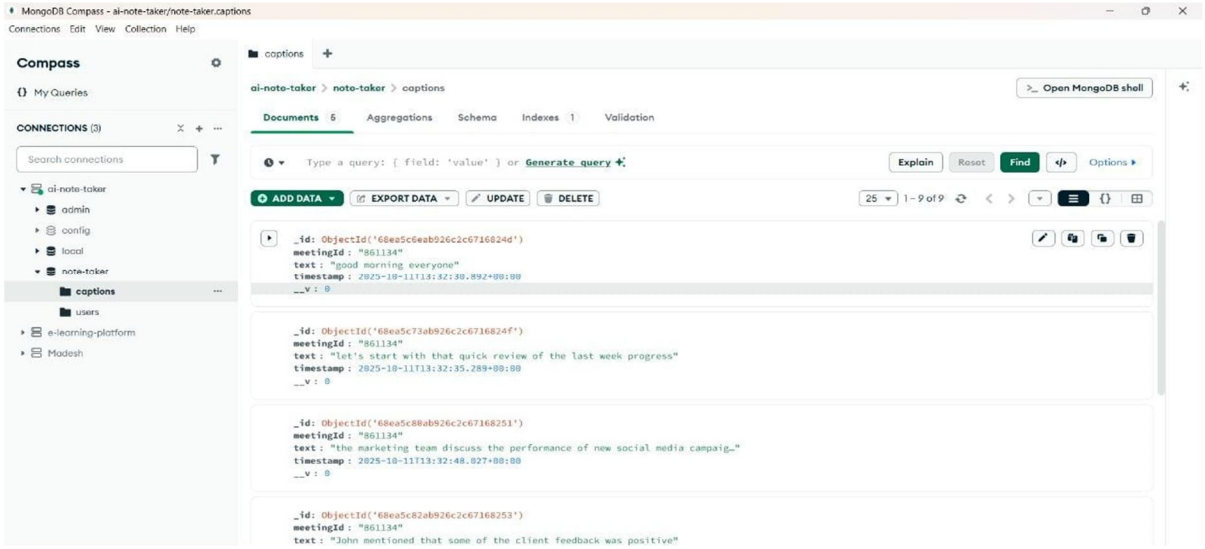


Fig. 4 Database

The Meeting Notes dashboard is shown in Fig. 5, where the system emphasizes the most important action items that were taken from the transcripts that were saved and automatically creates a summary. By removing unnecessary details and highlighting the most important information, the summary module makes it possible for users to swiftly go over the main points of debate and decisions made during the session.

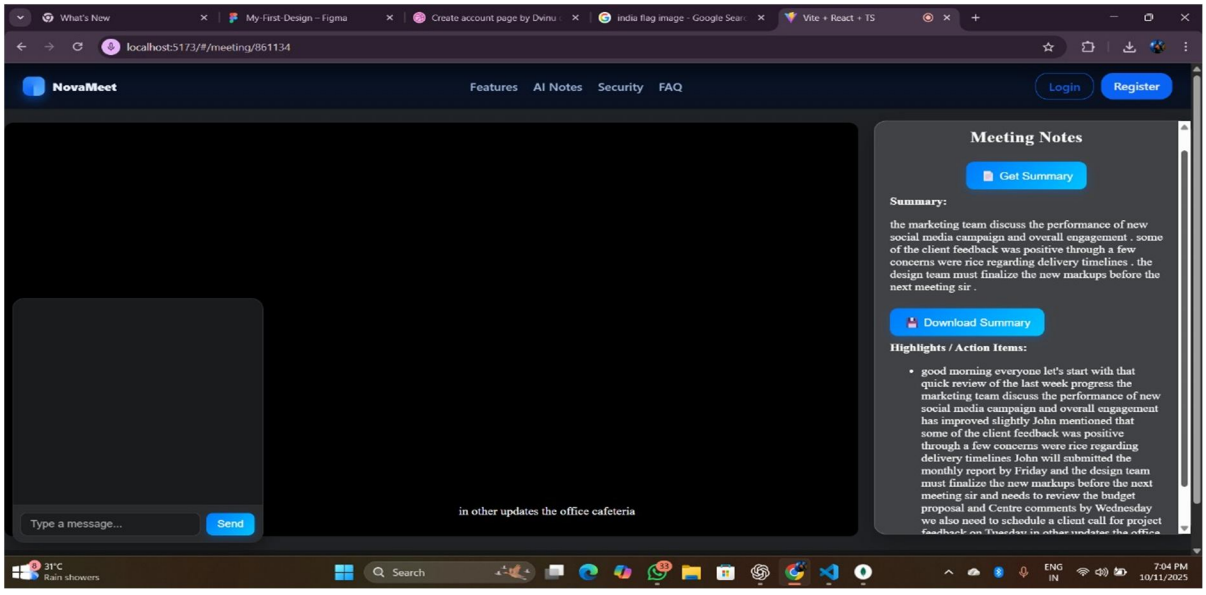


Fig. 5 Summary & Highlights

Lastly, the condensed report is available for download in a properly prepared document format, as seen in Fig. 6. With the help of this functionality, users can easily save and distribute succinct, contextualized notes that include important meeting highlights. The system maintains minimal latency, high transcription accuracy, and effective resource consumption, allowing for smooth real-time operation, according to the overall performance study.

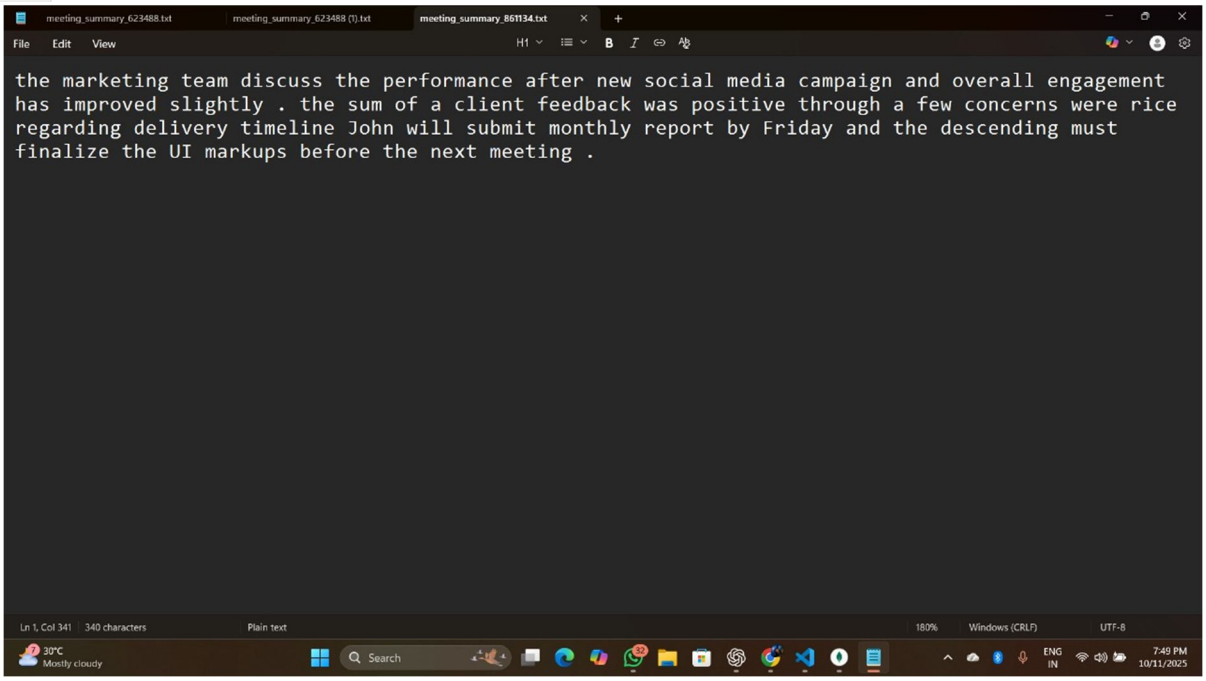


Fig. 6 Downloaded Document

X. CONCLUSION

The research effectively demonstrates an AI-based note-taking system that converts speech into coherent and intelligible text in real-time. The technology first records live audio, then uses advanced speech-to-text algorithms to process and accurately translate spoken input into text. The system is guaranteed to comprehend the semantic meaning, identify significant entities, and document the relationships between concepts using natural language comprehension and contextual processing following transcription. This comprehension serves as the cornerstone for crafting concise and coherent notes.

Contextual highlights are used for important points, action items, and decisions, and the note generating and summarizing module efficiently condenses lengthy transcriptions into information that is clear and easy to grasp. Because of this, people can quickly get important information without reading the entire document. By combining deep learning with natural language processing models, the system can adapt to different accents, speech patterns, and situations, increasing its accuracy and utility.

All things considered, the method bridges the gap between useful written notes and spoken content, significantly increasing productivity, understanding, and information retention. Because of its real-time processing capability, which provides users with rapid, reliable, and contextually relevant notes, it is particularly well-suited for lectures, meetings, and conversations. The experiment shows how note-taking applications with AI capabilities could simplify knowledge management in both professional and academic contexts.

REFERENCES

[1] Xu, Chen, Xiaoqian Liu, Yuhao Zhang, Anxiang Ma, Tong Xiao, Jingbo Zhu, Dapeng Man, and Wu Yang. "Unveiling the Fundamental Obstacle in Speech-to-Text Modeling: Understanding and Mitigating the Granularity Challenge." IEEE Transactions on Audio, Speech and Language Processing 33 (2025): 1719-1729.

[2] Ning, Jinzhong, Yuanyuan Sun, Zhihao Yang, Zhijun Wang, Ling Luo, Hongfei Lin, and Yijia Zhang. "GenEn-MNER: Enhancing Nested Chinese NER With Multimodal Fusion and Alignment via Speech-to-Text Generation." IEEE Transactions on Audio, Speech and Language Processing (2025).

[3] Mo, Ying, Jiahao Liu, Hongyin Tang, Qifan Wang, Zenglin Xu, Jingang Wang, Xiaojun Quan, Wei Wu, and Zhoujun Li. "Multi-Task Multi-Attention Transformer for Generative Named Entity Recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024).

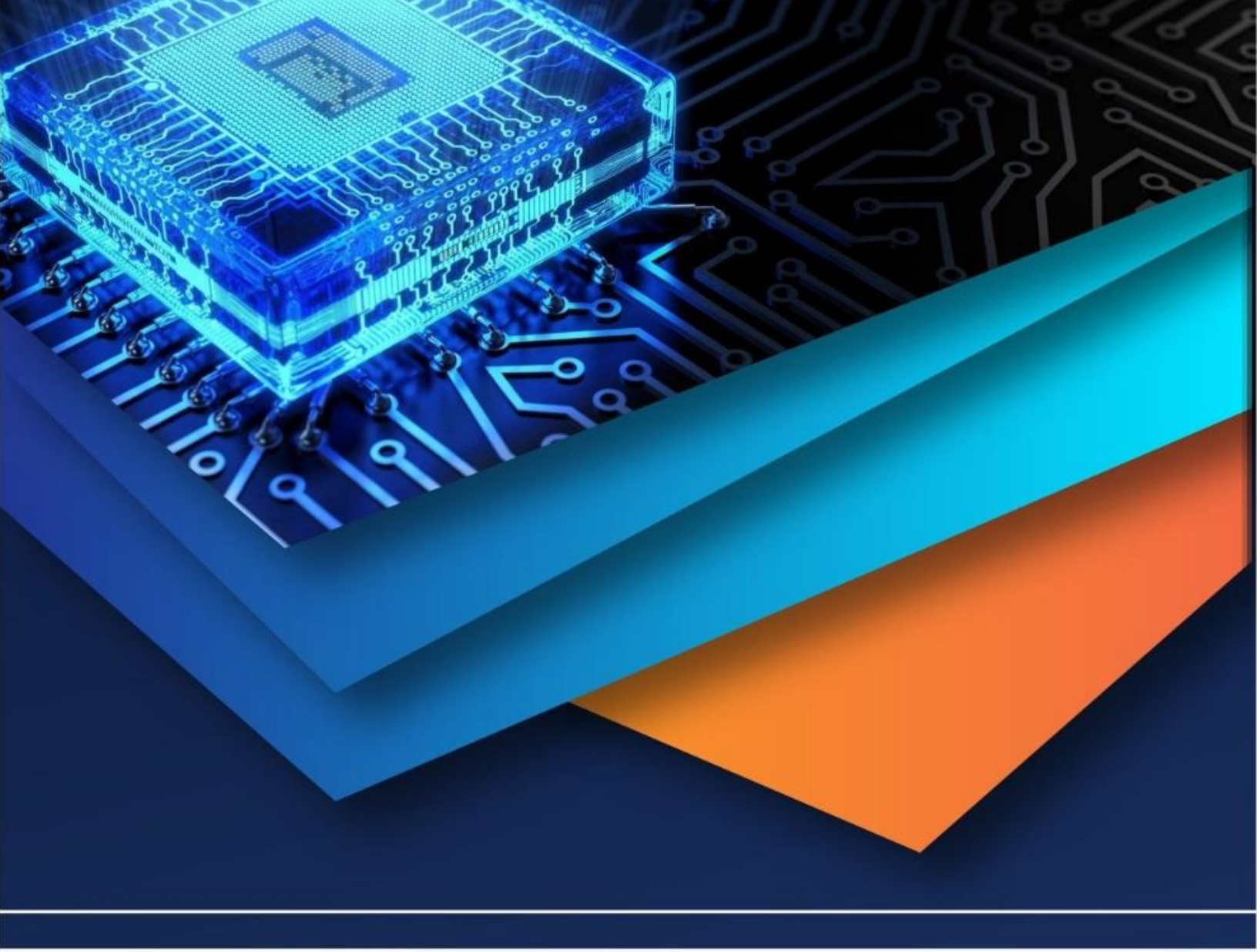
[4] Chatterjee, Sheshadri, Ranjan Chaudhuri, and Patrick Mikalef. "Examining the dimensions of adopting natural language processing and big data analytics applications in firms." IEEE Transactions on Engineering Management 71 (2024): 3001-3015.

[5] Feng, Xin, Yue Zhao, Wei Zong, and Xiaona Xu. "Adaptive multi-task learning for speech to text translation." EURASIP Journal on Audio, Speech, and Music Processing 2024, no. 1 (2024): 36.

[6] Zhang, Ying. "A Study on the Translation of Spoken English from Speech to Text." Journal of ICT Standardization 12, no. 4 (2024): 429-441.

[7] Arriaga, Carlos, Alejandro Pozo, Javier Conde, and Alvaro Alonso. "Evaluation of real-time transcriptions using end-to-end ASR models." arXiv preprint arXiv:2409.05674 (2024).

[8] MUHZINA, MA, P. M. Sulfath, and K. M. Sheena. "Smart Note Taker: A Digital Assistant for Efficient Note-taking." Authorea Preprints (2025).
[9] Feng, Xin, Yue Zhao, Wei Zong, and Xiaona Xu. "Adaptive multi-task learning for speech to text translation." EURASIP Journal on Audio, Speech, and Music Processing 2024, no. 1 (2024): 36.
[10] Zhou, YunYu, Cheng Tang, and Atsushi Shimada. "Extracting Learning Data From Handwritten Notes: A New Approach to Educational Data Analysis Based on Image Segmentation and Generative AI." IEEE Access (2025).
[11] Zhou, YunYu, Cheng Tang, and Atsushi Shimada. "A Novel Approach: Enhancing Data Extraction from Student Handwritten Notes Using Multi-Task U-net and GPT-4." In 2024 7th International Symposium on Autonomous Systems (ISAS), pp. 1-6. IEEE, 2024.
[12] Tang, Yun, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. "Improving speech translation by understanding and learning from the auxiliary text translation task." arXiv preprint arXiv:2107.05782 (2021).
[13] Wisoff, Josh, Yao Tang, Zhengyu Fang, Jordan Guzman, YuTang Wang, and Alex Yu. "NoteBar: An AI-Assisted Note-Taking System for Personal Knowledge Management." arXiv preprint arXiv:2509.03610 (2025).
[14] Jiang, Mi, Junran Gao, Zeyu Pan, Yue Wu, and Zile Wang. "NexaNota: An AI-Powered Smart Linked Lecture Note-Taking System Leveraging Large Language Models." In Proceedings of the 2025 International Conference on Big Data and Informatization Education, pp. 242-248. 2025.
[15] Adhikari, Surabhi. "Nlp based machine learning approaches for text summarization." In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 535-538. IEEE, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY