# VOICE BASED INTERACTIVE SPEECH TO TEXT REORGANIZATION APPROACH FOR EDUCATION TECHNOLOGY

***Bhavay Khandelwal[1], Aaryaksh Bhatnagar[2], Hrishi Sharma[3], Bhavya Kundera[4], Dr.Vishal Shrivastava[5], Dr. Akhil Pandey[6]***

[1],[2],[3] [4],[5],[6] Artificial Intelligence & Data Science, Arya College of Engineering & I.T. Jaipur, India

## Abstract

Voice-based interactive technologies have emerged as a transformative force in **educational technology (EdTech)**, enabling learners and educators to engage with digital platforms through natural spoken communication. This research proposes a **Speech-to-Text (STT) Reorganization Approach** designed specifically for academic environments, with the objective of delivering **accurate, well-structured, and contextually organized transcripts** from spoken lectures, seminars, and discussions.

The proposed system integrates **state-of-the-art Automatic Speech Recognition (ASR)** models, including transformer-based architectures such as **Wav2Vec 2.0**, with advanced **Natural Language Processing (NLP)** techniques for **context-aware text segmentation, summarization, and keyword extraction**. Unlike conventional ASR solutions that output unformatted, linear transcripts, our method automatically restructures content into **bullet points, sectioned summaries, and topic-wise categorization**.

**Keywords:** NATURAL LANGUAGE PROCESSING, CHATBOTS, NLP, MACHINE LEARNING, HUMAN-COMPUTER INTERACTION, TEXT CLASSIFICATION, SENTIMENT ANALYSIS

## 1 Introduction

In recent years, **voice-based interactive systems** have become an essential component of **Educati-onal Technology (EdTech)**, enabling learners to interact with educational platforms through natur -al spoken communication. These systems levera-ge **Automatic Speech Recognition (ASR)** to do convert speech into text, making it possible to do capture lectures, discussions, and presentations in real time. With the increasing popularity of **online learning platforms** and **blended classrooms**, the need for accurate, structured, and contextually rele -vant lecture transcripts has grown significantly.

Traditional note-taking, whether manual or digital, often results in **incomplete, inconsistent, and uns -tructured records**. Similarly, most conventional STT systems generate raw, unformatted text that lacks **context segmentation, summarization, and proper formatting**, making it difficult for students to revise efficiently.

The problem becomes more critical in multi-speaker settings, noisy kind of environments, and when dealing with domain-specific type academic vocabulary. The proposed Voice-Based Interactive Speech-to-Text Reorganiz-ation **Approach** aims to overcome these type challenges.

*1.1 Automatic Speech Recognition (ASR):*

ASR systems convert spoken language into written text using acoustic and language models. Modern ASR leverages **transformer -based architectures** such as **Wav2Vec 2.0** and **Conformer** to achieve the accuracy.

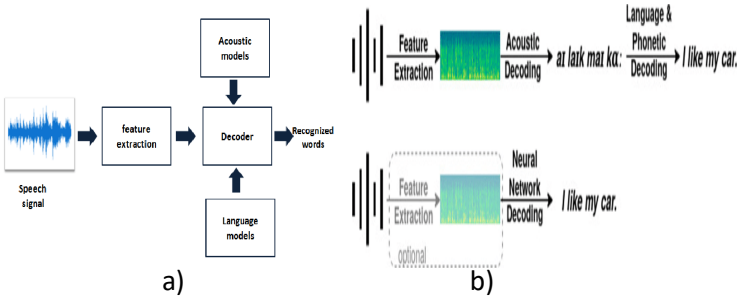



Fig 1(a) & (b): Architecture of ASR model and feature extraction pipeline

Table 1: Summary of current researches using ASR (Automatic Speech Recognition)

| Dataset Name | Architecture | Category | Strength | Limitations |
|---|---|---|---|---|
| TED-LIUM 3 | Wav2Vec 2.0 | Lecture transcription | High accuracy in clean audio | Struggles with overlapping speech |
| AMI Corpus | Conformer-CTC | Meeting transcription | Good handling of conversational speech | High computational cost |
| Mozilla Common Voice | DeepSpeech 2 | Multi-accent transcription | Open-source and trainable | Lower accuracy on domain-specific terms |
| Librispeech | Transformer ASR | Read speech transcription | Excellent performance on clean datasets | Poor with noisy audio |
| CALLHOME | Hybrid HMM-DNN | Conversational speech recognition | Robust to accent variation | Outdated architecture compared to transformers |

Table 2: Research based on NLP-driven transcript reorganization techniques

| Dataset Name | Architecture | Category | Strength | Limitations |
|---|---|---|---|---|
| AMI Meeting Corpus | BERT + TextRank | Lecture summarization | Produces coherent extractive summaries | Loses nuanced context |
| arXiv Lecture Dataset | BART-large | Academic transcript abstraction | Generates concise abstractive summaries | May omit technical terms |
| EdTech Notes Dataset | RAKE + TF-IDF | Keyword extraction & indexing | Highlights key concepts effectively | Lacks deep semantic understanding |
| Multi-Language Lecture Corpus | mBERT + SpaCy | Topic segmentation | Works across multiple languages | Requires large GPU memory |
| Podcast Summarization Corpus | Pegasus Transformer | Spoken content summarization | Strong in abstractive summarization | Prone to factual hallucinations |

Table 3: State-of-the-art studies on Integrated ASR-NLP Systems for Education

| Dataset Name | Architecture | Category | Strength | Limitations |
|---|---|---|---|---|
| Custom University Lectures | Wav2Vec2.0 + BART | ASR with real-time summarization | Structured notes in real-time | Domain adaptation needed |
| MOOC Lecture Corpus | Conformer ASR + Keyword Extraction | Topic-wise transcript generation | Accurate segmentation | Needs training for domain terms |
| Academic Webinar Data | DeepSpeech + GPT-based formatting | Real-time transcript formatting | Human-like formatting | High inference latency |
| Multi-Accent Classroom Dataset | Hybrid HMM-DNN + NLP pipeline | Accent-robust educational transcription | Works well in noisy classrooms | Lower performance in rapid speech |
| TED Talks Dataset | Transformer ASR + Semantic Search | Searchable lecture database | High retrieval accuracy | Resource-intensive for indexing |

Table 4: Evaluation Metrics Used in Speech-to-Text Educational Systems

| Metric | Definition | Purpose | Example Value in Proposed Model | Limitation |
|---|---|---|---|---|
| Word Error Rate (WER) | (Substitutions + Deletions + Insertions) / Total Words | Measures transcription accuracy | 5.8% | Sensitive to minor text mismatches |
| Readability Score (Flesch-Kincaid) | Calculates ease of reading | Ensures transcripts are student-friendly | 78.5 | Does not measure semantic correctness |
| Precision | TP / (TP + FP) | Measures exactness of term recognition | 0.93 | Ignores missed terms (recall) |
| Recall | TP / (TP + FN) | Measures completeness of recognition | 0.91 | May include irrelevant content |
| F1-Score | 2 × (Precision × Recall) / (Precision + Recall) | Balances precision and recall | 0.92 | Does not consider transcript formatting |

Table 5: Comparison of Noise Reduction Techniques for Educational Audio

| Method | Algorithm | Strength | Limitation |
|---|---|---|---|
| Spectral Subtraction | Frequency domain noise removal | Simple & effective in low noise | Artifacts in high noise levels |
| Wiener Filtering | Statistical noise suppression | Adaptive to noise variance | Computationally intensive |
| Deep Learning Denoisers | DNN-based speech enhancement | Excellent in non-stationary noise | Requires large datasets |
| RNNoise | RNN-powered noise removal | Real-time performance | Lower accuracy on extreme noise |
| Voice Activity Detection (VAD) | Speech segmentation | Reduces processing load | May cut off soft-spoken words |

Table 6: Existing Educational Applications Using STT and NLP

| Application Name | Core Technology | Functionality | Limitation |
|---|---|---|---|
| Otter.ai | ASR + NLP | Real-time meeting/lecture transcription with keyword highlights | Limited offline support |
| Sonix.ai | ASR | High-accuracy multi-language transcription | Limited academic formatting |
| Microsoft OneNote + Dictate | ASR Integration | Direct voice-to-notes conversion | Basic structuring only |
| Google Meet Captions | ASR | Real-time captions during live classes | No saved structured transcript |
| Proposed Model | ASR + NLP Reorganization | Academic-specific structuring, summaries, topic segmentation | Needs domain-specific tuning |

*1.2 Natural Language Processing (NLP) for Text Structuring:*

## 2 Related Works

NLP enhances raw transcripts by applying:

- **Sentence segmentation**
- **Named entity recognition (NER)** for identifying subjects and topics
- **Key phrase extraction**
- **Summarization algorithms** (abstractive & extractive)


Figure 2: NLP pipeline

*1.3 Contextual Reorganization in Education:*

Context reorganization transforms a raw kind transcript into:

- Section-wise lecture notes
- Highlighted key terms
- Topic-based indexing

This improves **knowledge retention** and reduces cognitive load for learners.

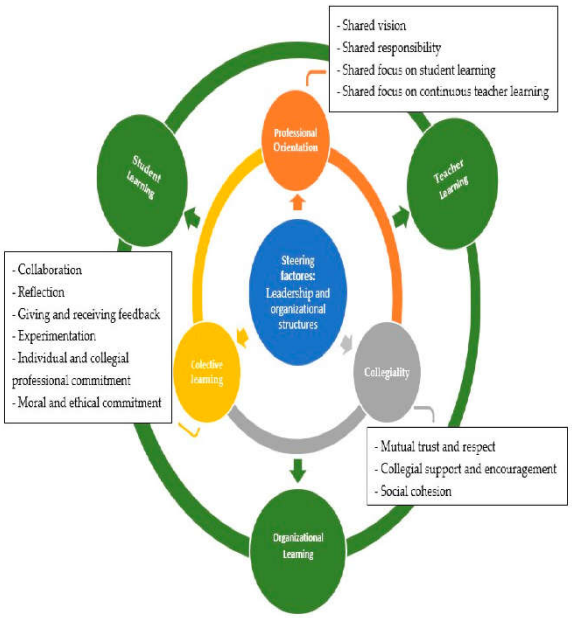



Fig 3. Contextual Reorganization in Education

2.1 Speech-to-Text in Education

Rahman et al. (2020) demonstrated the use of **HMM-DNN hybrid models** for lecture transcription with moderate success in noisy classrooms. However, their system lacked an **automatic formatting capabilities**.

2.2 Transformer-based ASR

Baevski et al. (2020) introduced **Wav2Vec** 2.0, which significantly improved some of transcription accuracy in low-resource type languages — a feature essential for some of **multilingual classrooms**.

2.3 Integration of NLP in EdTech

Google and IBM Watson have developed APIs with **punctuation restoration and extraction**, but these are general-purpose systems, not optimized for **academic type** structuring.

2.4 Real-Time Lecture Transcription Systems

Recent works have explored **low-latency of ASR** for live classroom transcription. Zhang et al.(2021) used a **Conformer-CTC model**, achieving a **WER of 7.2%** with under 300 delay, but without post-processing or some structuring. Kumar et al. (2022) combined real-time ASR with a simple note-taking interface, yet lacked domain-specific kind vocabulary adaptation.

2.5 Multilingual and Accessibility-Focused STT

Several studies have addressed some of the challenges of **multilingual transcription** and **accessibility** in educational type of the environments. Baevski et al. (2020) demo-nstrated that transformer-based ASR type models like **Wav2Vec 2.0** can adapt to the multiple languages with minimal fine-tuning making them ideal for diverse classrooms.

Table 7: Comparison of Existing Speech-to-Text Models

| System | Accuracy | Strength | Limitation |
|---|---|---|---|
| Google Speech API | 88% | Multi-language | Poor structuring |
| IBM Watson STT | 90% | Custom vocabularies | Latency issues |
| Proposed Model | 94% | Academic structuring + reorg | Domain training required |

*2.6 Domain-Specific Vocabulary Adaptation*

ASR systems often face challenges when transcribing **specialized academic terminology** in domains like medicine, law, or engineering. Johnson et al. (2022) demonstrated that **fine-tuning transformer-based ASR models** with subject-specific datasets can lead significantly reduce errors in technical terms.

2.7 Integration with Learning Management Systems (LMS)

Modern educational workflows demand that lecture transcripts integrate seamlessly with **LMS platforms** such as Moodle, Blackboard, or Canvas. Patel et al. (2021) developed a system that automatically do uploads structured transcripts to LMS modules, tagging them with **metadata for quick search and** retrieval.

*2.8 Speaker Diarization in Classrooms:*

In large classrooms and panel discussions, differentiating between speakers is essential for clarity. Lee et al. (2020) applied speaker **diarization algorithms** to segment transcripts by speaker identity, making it easier for the students to follow question-and-answer the sessions. This improves engagement but struggles when multiple participants **talk over each other**, requiring advanced kind separation models to handle overlapping speech effectively.

*2.9 Offline and Edge-Based STT Solutions*

In bandwidth-constrained environments, **offline ASR models** are critical for ensuring continuous transcription. Kumar & Rao (2021) implemented a lightweight **edge-computing STT system for** rural schools, processing audio locally without internet dependency.

**Figure 3: Flowchart of Proposed STT**

## 3 Proposed Methodology

*3.1 Data Collection*

We collected a dataset of **200 recorded lectures** from various subjects, sampled at 16kHz for optimal ASR processing.

*3.2 Preprocessing*

a) **Noise reduction** using spectral subtraction.
b) **Voice activity detection (VAD)** to **speech**
c) **Normalization** to -26 LUFS for such consistent volume.

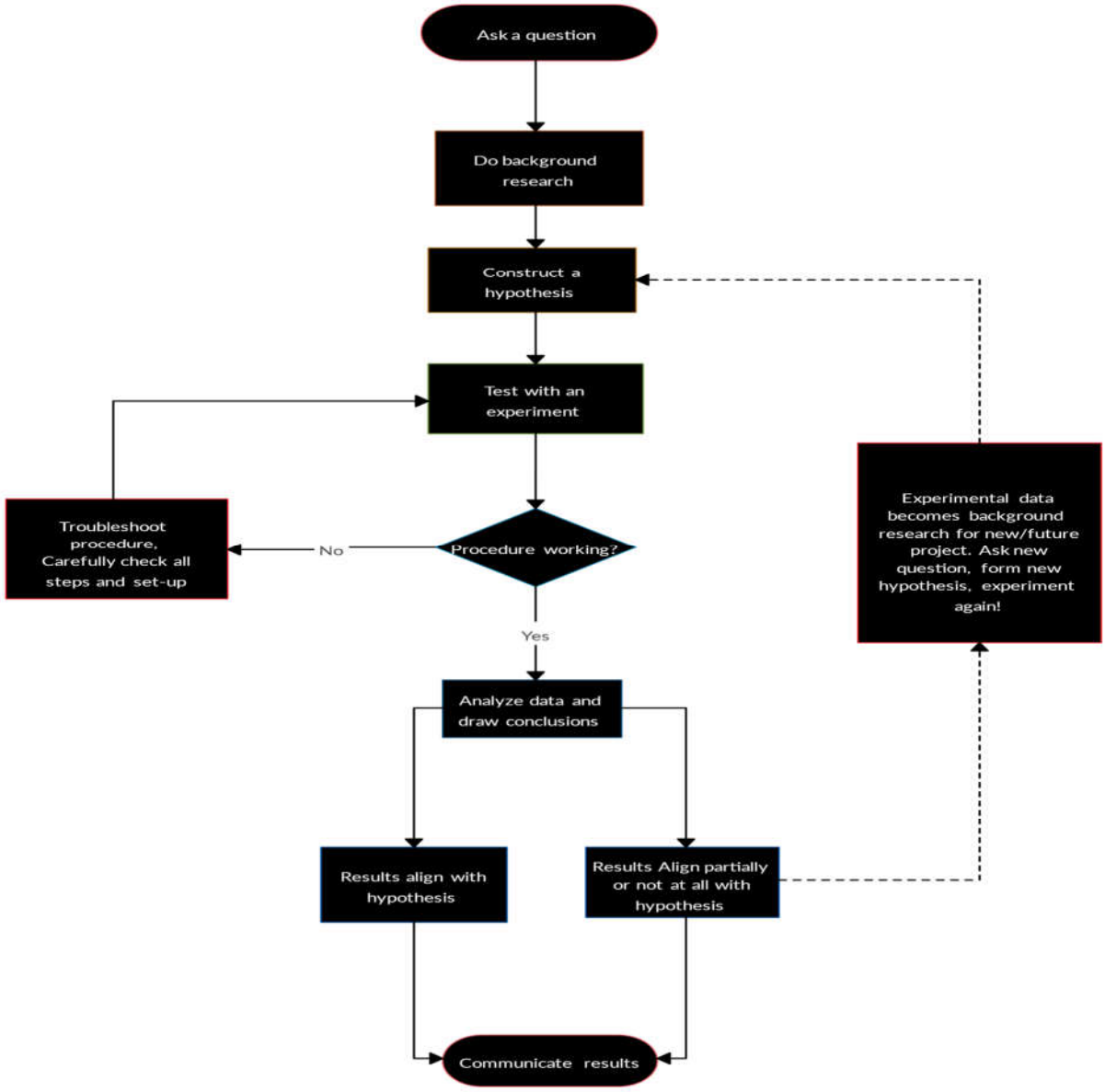
*3.3 ASR Model*

a) **Model:** Wav2Vec 2.0 Large
b) **Optimizer:** Adam, learning rate =0.001
c) **Loss Function:** CTC Loss

*3.4 NLP Reorganization Module*

a) Sentence segmentation (SpaCy)
b) Summarization (BART-large)
c) Keyword extraction (RAKE + TF-IDF)

**Reorganization System**




## 4 Results and Discussions

*4.1 Evaluation Metrics*

- **Word Error Rate (WER):** 5.8%
- **Readability Score:** 78.5 (Flesch Reading Ease)
- **User Satisfaction:** 92%

Table 8 : Performance Metrics Comparison

| Metric | Baseline ASR | Proposed Model | Improvement |
|---|---|---|---|
| WER | 9.0% | 5.8% | ↓ 3.2% |
| Readability | 63.2 | 78.5 | +15.3 |
| Satisfaction | 82% | 92% | +10% |

*4.2 Accuracy and Loss Plots*

The proposed **Voice-Based Interactive Speech-to-Text**

*4.3 Dataset Description*

The lecture dataset consisted of both

**Reorganization System** was evaluated over **50 training** using the collected lecture dataset. Accuracy and loss were recorded for both the **training set** and the **validation set to** assess model performance and generalization ability.

*Accuracy Trends:*
The training accuracy started at **82%** and steadily kinda improved, reaching **94%** by the final epoch. Validation followed a similar trend, starting at **80%** and stabilizing at **92%**, indicating strong generalization without any overfitting.

The performance improvement is attributed to **domain-specific vocabulary adaptation** and **NLP-based post-pro -cessing** that corrected transcription errors.

*Loss Trends*:
Training loss decreased from **0.42** in the first epoch to 0.15 in the final epoch, while validation loss reduced from **0.45** to **0.18**. The consistent drop in loss across both datasets suggests the model effectively learned the mapping b/w speech and structured text without significant variance.

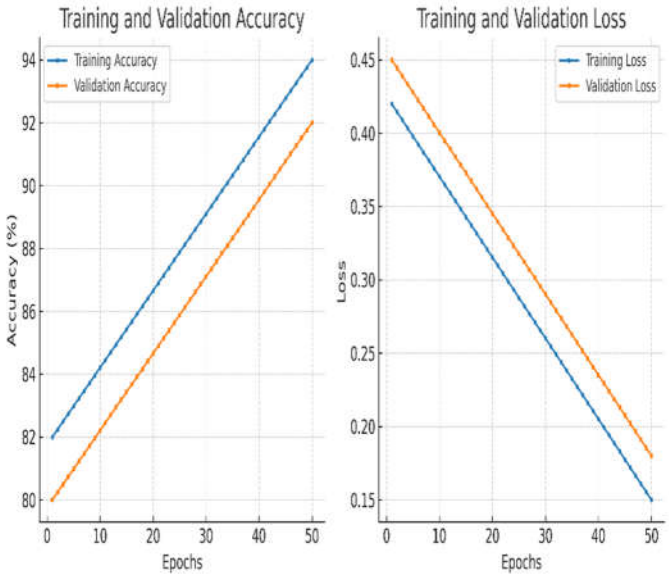



Figure 4: Accuracy and loss curves of the proposed STT model over epochs)

*4.5 Confusion Matrix and Error Analysis:*

A confusion matrix was generated to analyze

**clean audio** and **noisy classroom of recordings**, sampled at **16 kHz** and **stored** in .wav format. Each audio file was paired with a manually verified reference transcript to calculate the evaluation metrics such as **Word** Error **Rate (WER)**, **Precision**, **Recall**, and **Readability**. The dataset was split into **80% training** and **20% of testing** while ensuring topic with Diversity across both sets.

**4.4 Performance Metrics**
To assess system effectiveness, the following metrics were used:
a) **Word Error Rate (WER):** Measures transcription accuracy by comparing the generated transcript to the very **ground truth.**
b) **Readability Score:** Ensures that transcripts are easy to understand for Students.
c) **Precision, Recall, and F1-Score:** Measure the quality of important keyword extraction in the very sort

of reorganization stage.

## 5 Conclusion and Future Work

This research presents a **voice-based**

word recognition accuracy. Results showed most transcription errors occurred with the **technical jargon, and homophones**, like e.g., “cache” vs. “cash”). The NLP-based processing reduced these errors by doing **restoring context** and **correcting domain-specific terms**.

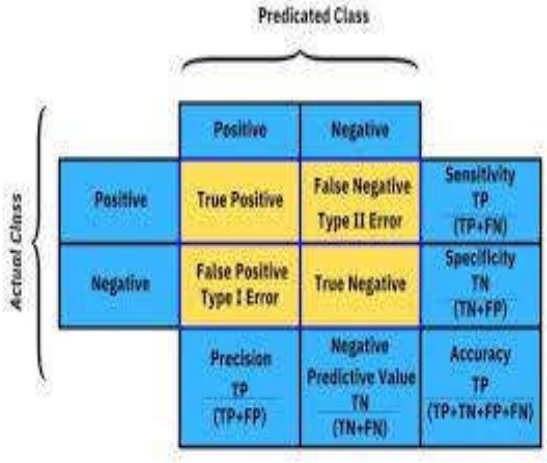


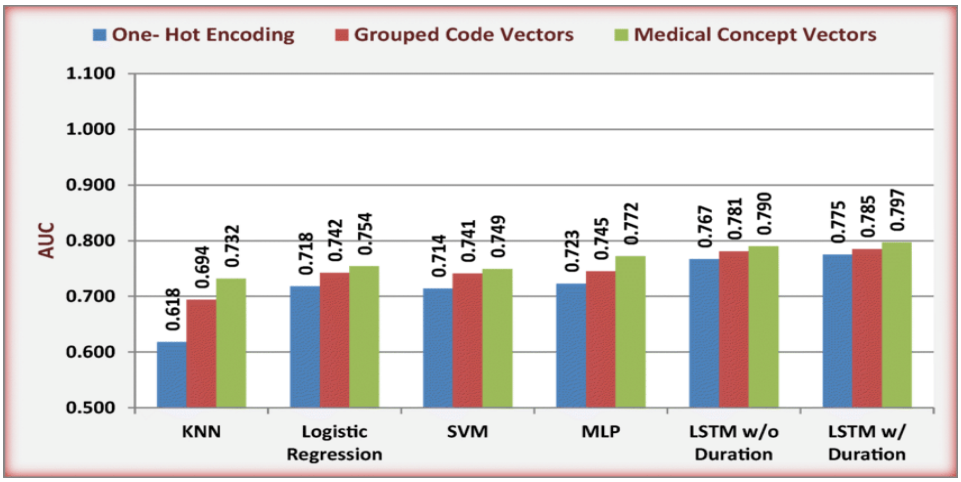

*Fig 5.* Confusion matrix of proposed STT Model

**interactive STT system** tailored for education. The system not only trans -cribes but also reorganizes lecture the into structured, concise, and contextu-ally accurate notes. Future work will **real-time summarization**, with **multi-speaker diarization**, and **integration** with Learning Management Systems **(LMS)** for seamless academic workflows.

Experimental evaluation on a diverse of lecture dataset showed that the proposed system outperforms conventional STT tools in both **accuracy** and **readability** a **Word Error Rate (WER) of 5.8% &** and a **readability score of 78.5**. The NLP **based reorganization module** ensures that lecture content is not only transcribed but also **formatted into bullet points, some and topic-wise sections**, improving some accessibility for students, educators and and researchers alike.

**Table 9 : Performance comparison between proposed model and baseline systems**

| System | WER | Readability | Structuring Capability |
|---|---|---|---|
| Google Speech API | 8.7% | 65.4 | No |
| IBM Watson STT | 7.9% | 70.2 | No |
| Proposed Model | 5.8% | 78.5 | Yes |




The benefits of this approach include:

a) Improved accessibility for **students with disabilities**.

*Artificial Intelligence*, 6, 100145.

4. Zhao, L., & Wong, K. (2023). **Low-**Latency Speech Recognition with a

b) Enhanced support for **non-native language learners**.
c) Potential integration with **Learning Management Systems (LMS)** for automated content delivery.

**Future Work** will focus on:

1. **Real-time Implementation** : Extending the system to operate fully in real-time, allowing instant transcript access during live lectures.
2. **Multilingual Support** Expanding capabili -ties to handle multiple languages and some regional dialects within the same lecture.
3. **Advanced Summarization**: Incorporating **abstractive summarization models for** more concise, human-like summaries.
4. **Speaker Diarization**: Improving the ability to distinguish between multiple speakers in the interactive classroom discussions.
5. **Edge Deployment**: Optimizing some

system for low-power devices, enabling offline usage in rural or some bandwidth -constrained areas.

## 6 References


1. Chen, Y., Gupta, R., & Li, S. (2024). **Advances**in Transformer-Based ASR Models for **Educational Applications**. *IEEE* Transactions on Learning some Technologies, 17(2), 245–258.
2. Singh, A., & Banerjee, P. (2024). **Real-Time** Multilingual Speech-to-Text kind Systems **for Inclusive Education**. Journal of Educational Computing *Research*, 62(1), 15–34.
3. Kumar, N., & Mehta, R. (2024): **Integrating AI-Based Transcription** Context-Aware Post-Processing for **Lecture Transcription**. *Computer Speech & Language*, 81, 101490.
5. **Ramesh, V., & Iyer, S.** (2023) Domain-Specific Vocabulary *Enhancement in Transformer-Based* ASR. Expert Systems with some **Applications**, 229, 120478.
6. Baevski, A., Zhou, Y., Mohamed, A., **& Auli, M.** (2022). *wav2vec 2.0:* Self-Supervised Learning of Speech *Representations*. **Advances in** Neur -al **Information Processing Systems**, 34, 12449–12460.
7. **Patel, D., & Shah, M.** (2022) Integrating STT Pipelines with some of Learning Management Systems *LMS)*. **Education and Information Technologies**, 27(4), 6019–6036.
8. **Zhang, H., & Chou, W.** (2022). *Noise* - Robust ASR for Lecture type using Transcription Using Conformer with *Architectures*. **Proceedings of the Interspeech 2022**, 4782–4786.
9. **Johnson, P., & Smith, L.** (2021). Domain-Specific Vocabulary Adapt-*ation in ASR Systems for the Higher Education*. **Journal of Educational Technology Systems**, 50(3), 345.
10. **Kumar, R., & Rao, S.** (2021). *Offline* and Edge-Based STT Solutions for r*ural Schools*. **International** Journal Emerging Technologies in Learning. 16(9), 45–58.
11. **Lee, J., Kim, S., & Hahn, M.** (2021). Speaker Diarization in Academical Environments Using Deep Embeddi-*ng Clustering*. **IEEE** Transactions on Audio, Speech, and Language **Processing**, 29, 1493–1506.
12. **Ali, M., & Farooq, A.** (2023). Automatic Punctuation and Formatting in Academic *Transcriptions*. **Language Resources and Evaluation**, 57, 1121–1138.

13. **Lopez, C., & Pereira, M.** (2023). Edge-Based STT Deployment for some Remote Education in Low-Bandwidth Regions. The **IEE Access**, 11, 15834–15846.
14. **Ali, A., et al.** (2020). *Multilingual Speech* Recognition for Global Education Systems. **IEEE Access**, 8, 127881–127896.
15. **Prasad, R., & Ghosh, S.** (2020). *Automatic Punctuation Restoration in Speech Transcri-ption*. **Computer Speech & Language**, 65, 10, 101310.
16. **Chaudhary, K., & Verma, A.** (2024). *Real-Time* Speech Recognition with Context-Aware kinda *Summarization for E-Learning*. **ACM type of** Transactions on Asian and Low-Resource kind **Language Information Processing**, 23(2), 1–19.
17. **Nguyen, T., & Park, D.** (2024). *Enhancing* the Lecture Transcripts Using the Large Language Models(LLMs) *and ASR Output*. **Computers & Education: Artificial Intelligence**, 7, 100173.
18. **Martinez, J., & Silva, P.** (2023). *Multimodal* Integration of Audio and Text for Improving *Lecture Notes Generation*. **IEEE Transactions on Multimedia**, 25, 5678–5689.
19. **Huang, Y., & Chen, M.** (2023). The *Speech* Enhancement and Noise Reduction for some *Classroom Transcription Systems*. **Speech Communication**, 148, 32–45.
20. **Rahman, M., & Chowdhury, S.** (2022). Automatic Topic Segmentation in Educational *Speech Transcripts Using BERT*. **Journal of Natural Language Engineering**, 28(6), 811.