iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: https://doi.org/10.22214/ijraset.2026.85356

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 14 Issue X Oct 2026- Available at www.ijraset.com

StudyScribe: An AI-Powered Audio Processing and Smart Learning System for Automated Lecture Note Generation

Mrs. Vanitha D[1], Shreya BM[2], Smitha R[3], Sunaina G Nayak[4], Sushma Raj R[5]
[1]*B.E., M.Tech., Assistant Professor, Dept. of CS&E*
[2, 3, 4, 5]*Department of Computer Science and Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India*

Abstract: ***Digital learning increasingly relies on recorded lectures, but these recordings are long, unstructured, and mixed with content that has no academic value, which makes revision slow and error-prone. This paper presents StudyScribe, a proposed AI-powered pipeline that accepts a lecture as an audio file, video file, or YouTube link and converts it into structured, topic-wise study material. The system combines Transformer-based automatic speech recognition, NLP-based relevance filtering, semantic topic segmentation, and abstractive summarization to remove non-educational content and generate concise notes, key insights, and quiz questions. A review of ten related works spanning speech recognition and text summarization is used to justify the design choices, and the resulting five-stage architecture — input acquisition, transcription, cleaning, segmentation, and summary generation — is described along with directions for implementation and evaluation.***

Keywords: Speech-to-text, Natural Language Processing, Transformer, abstractive summarization, topic segmentation, lecture note generation, e-learning.

I. INTRODUCTION

Recorded lectures, online videos and virtual classrooms have become primary sources of learning in modern engineering education. While this shift has made educational content widely accessible, it has also introduced a new problem: recordings are long, largely unstructured, and mixed with announcements, digressions and casual conversation that carry no academic value. A student revisiting a ninety-minute recording to locate a five-minute explanation of a single concept ends up spending more time searching than learning. The difficulty is more acute for students who miss a class altogether. Manual note-taking, the traditional fallback, is itself unreliable — it depends on the listener's speed, attention span and ability to judge what matters while the lecture is still in progress. Important details are frequently lost simply because the student was writing down the previous point when the next one was spoken.

Recent progress in Automatic Speech Recognition (ASR), Natural Language Processing (NLP) and Transformer-based language models has made it feasible to automate large parts of this workflow. Speech can be converted to text with high accuracy even across accents and noisy audio; irrelevant segments can be filtered using semantic relevance scoring; long transcripts can be split into topic-coherent sections; and abstractive summarization models can condense these sections into readable notes. Individually, each of these capabilities is well established in the literature, but very few systems combine them into a single pipeline aimed specifically at lecture content. This paper presents StudyScribe, a system designed to close that gap. StudyScribe accepts a lecture in the form of an audio file, a video file, or a YouTube link, and produces structured, topic-wise study material — summaries, key points and revision aids — without requiring the student to sit through the entire recording. Section II reviews the transformer-based models and prior systems that motivate the design choices made in StudyScribe. Section III describes the proposed methodology and system architecture. Section IV concludes the paper and outlines the direction of future work.

II. LITERATURE SURVEY

To identify a suitable technical foundation for StudyScribe, ten papers spanning speech recognition, language representation and text summarization were reviewed. The works are summarized below and compared in Table I.

Vaswani et al. [1] introduced the Transformer architecture, replacing recurrent sequence models with self-attention so that a model can weigh every token against every other token in parallel. This removed the sequential bottleneck of RNNs and became the architectural basis for nearly every model discussed in this survey, though at the cost of high memory and compute requirements for training.

©IJRASET: All Rights are Reserved

Building on this architecture, Devlin et al. [2] proposed BERT, which is pre-trained using masked-language modelling and next-sentence prediction to read a sentence bidirectionally rather than left-to-right. This bidirectional context gives BERT strong performance on classification and relevance-filtering tasks, which is directly useful for distinguishing educational from non-educational segments of a transcript, although the model remains large and slow to fine-tune.

For generation-oriented tasks, Lewis et al. [3] combined a bidirectional encoder with an autoregressive decoder in BART, training it to reconstruct text that has been deliberately corrupted. This denoising objective makes BART well suited to cleaning noisy, disfluent lecture transcripts and rewriting them as coherent summaries, though it demands considerable compute at inference time.

Reimers and Gurevych [4] adapted BERT into Sentence-BERT (SBERT) using a Siamese network to produce fixed-length sentence embeddings that can be compared with cosine similarity far more cheaply than pairwise BERT inference. This property makes SBERT a practical choice for grouping semantically related sentences into topic segments across a long transcript, provided the embeddings are fine-tuned for the target domain.

Zhang et al. [5] designed PEGASUS specifically for abstractive summarization by pre-training the model to predict sentences that were deliberately masked out of a document — a gap-sentence objective that teaches the model to recognize which sentences carry the most information. Raffel et al. [6] took a broader approach with T5, casting every NLP task, including summarization, as text-to-text generation within one architecture, trading task-specific tuning for a single flexible model at the cost of a larger resource footprint. Both models report strong summarization quality but are expensive to deploy in resource-constrained settings.

Since lecture transcripts routinely exceed the input length that standard Transformers can handle, Beltagy et al. [7] proposed Longformer, which restricts most attention computation to a local window while still allowing a small number of tokens global attention. This sparse-attention design lets the model process much longer documents without a quadratic increase in cost, although tokens that are not marked for global attention can lose long-range context.

Beyond summarization, Yao et al. [8] applied sequence-to-sequence Transformers to Automatic Question Generation (AQG), producing quiz-style questions directly from educational text — a capability StudyScribe intends to use for self-assessment, though the reviewed system occasionally produces repetitive or shallow questions. On the audio side, Radford et al. [9] introduced Whisper, an encoder-decoder Transformer trained on large-scale multilingual audio that performs transcription and translation robustly even in noisy conditions, at the cost of a large model footprint. Closest in intent to the present work, Albadawi et al. [10] describe an end-to-end pipeline that converts lecture speech to text and applies NLP-based summarization to generate structured notes automatically, reporting improved readability and revision efficiency; this system is treated as the base reference for StudyScribe.

Table I. Summary Of Reviewed Literature

Ref.	Methodology	Key Advantage	Key Limitation	Result
[1]	Self-attention Transformer (Vaswani et al., 2017)	Strong contextual modelling	High compute cost	Basis for modern NLP models
[2]	Bidirectional Transformer, MLM+NSP (Devlin et al., 2018)	High semantic accuracy	Very large model size	State-of-the-art NLP results
[3]	Denoising autoencoder, BART (Lewis et al., 2019)	High-quality abstractive summaries	Resource-intensive	Strong text generation
[4]	Siamese BERT sentence embeddings (Reimers & Gurevych, 2019)	Fast semantic similarity	Needs fine-tuning	Efficient semantic comparison
[5]	Gap-sentence pre-training, PEGASUS (Zhang et al., 2020)	Strong summarization quality	High compute demand	Concise, coherent summaries
[6]	Unified text-to-text, T5 (Raffel et al., 2020)	Multi-task capability	Large resource requirement	Strong benchmark scores
[7]	Sparse local/global attention, Longformer (Beltagy et al., 2020)	Handles long documents	Complex to implement	Efficient long-document processing
[8]	Transformer-based question generation (Yao et al., 2018)	Automates quiz creation	May generate simple questions	Effective assessment support
[9]	Transformer ASR, Whisper (Radford et al., 2022)	Accurate multilingual transcription	High processing cost	Robust speech recognition
[10]	Transformer-based lecture summarization (Albadawi et al., 2025)	Automatic structured notes	High computational requirement	Improved readability, learning efficiency

A. *Research Gap*

Across the reviewed literature, Transformer-based architectures consistently outperform earlier approaches, but three gaps recur. First, most systems are built for one task in isolation — either transcription, or filtering, or summarization — rather than an integrated pipeline. Second, models capable of the highest-quality output (BERT, BART, PEGASUS, Longformer) are computationally heavy, which raises questions about practical deployment for everyday student use. Third, none of the reviewed systems explicitly separate educational from non-educational speech before summarization, so irrelevant classroom conversation risks being carried into the final notes. StudyScribe is proposed to address these three gaps by combining ASR, relevance filtering, topic segmentation and summarization into a single, purpose-built pipeline for lecture content.

III. PROPOSED METHODOLOGY

Based on the gaps identified in Section II, StudyScribe is designed as a five-stage pipeline: input acquisition, speech-to-text conversion, transcript cleaning, topic segmentation, and summary and study-aid generation. Fig. 1 shows the overall system architecture; each stage is described in the subsections below.

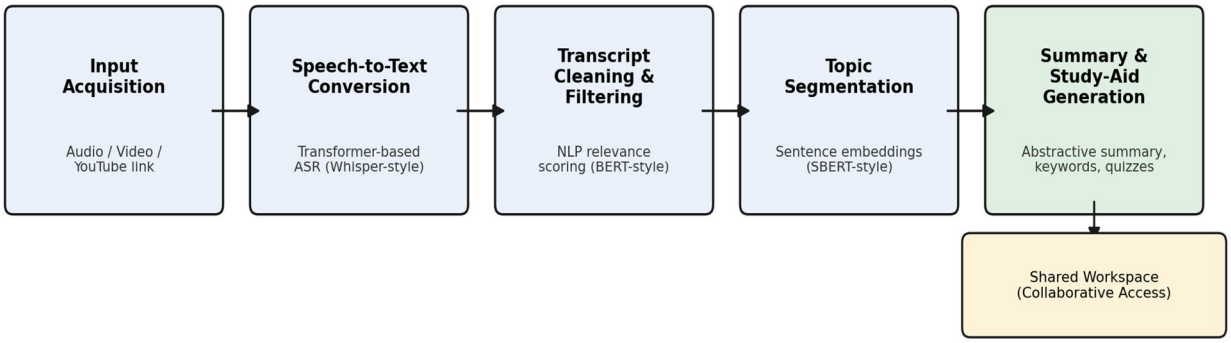


Fig. 1. Proposed system architecture of StudyScribe.

A. *Input Acquisition*

The system accepts three input types — audio files, video files, and YouTube links — so that a student can submit whatever format the lecture happens to be available in. Video inputs are demuxed to extract the audio track, and YouTube links are resolved to a downloadable audio stream, so that every input converges to a single normalized audio format before the next stage.

B. *Speech-to-Text Conversion*

The normalized audio is transcribed using a Transformer-based ASR model in the style of Whisper [9], chosen for its demonstrated robustness to accents, background noise and mixed-language speech — all common in a classroom recording. The output at this stage is a raw, timestamped transcript.

C. *Transcript Cleaning and Relevance Filtering*

The raw transcript is passed through an NLP filtering stage that scores each segment for topical relevance, informed by the contextual classification approach of BERT [2]. Segments corresponding to administrative announcements, small talk, or off-topic discussion are flagged and excluded, leaving a cleaned transcript that contains only academically relevant content.

D. *Topic Segmentation*

The cleaned transcript is divided into topic-coherent sections using sentence-level embeddings, following the Sentence-BERT approach of Reimers and Gurevych [4]. Cosine similarity between consecutive sentence embeddings is used to detect topic boundaries, so that the transcript is broken into segments that each correspond to a single concept rather than an arbitrary block of time.

E. Summary and Study-Aid Generation

Each topic segment is summarized abstractively using a pre-trained sequence-to-sequence model in the family of BART [3] or PEGASUS [5], producing concise notes rather than merely extracted sentences. From the same segment, keywords and important concepts are identified to support customizable note formats — bullet points, paragraphs, or mind maps — as well as quiz questions generated using a sequence-to-sequence question-generation approach similar to [8], allowing the student to self-test on the material covered.

F. Collaborative Access

The generated notes and study aids are intended to be exposed through a shared workspace so that multiple students working on the same course can access, review, and jointly refine the processed material, supporting group revision rather than a purely individual workflow.

G. Expected Outcomes

The pipeline is expected to reduce the time students spend revising a lecture by replacing full-length playback with structured, topic-wise notes; to improve retention by pairing summaries with self-generated quiz questions; and to make missed lectures easier to catch up on since the output no longer depends on someone else's handwritten notes. At the present stage, the project has completed the requirements analysis and the architectural design shown in Fig. 1. Module-level implementation, integration, and evaluation against metrics such as transcription accuracy, ROUGE-based summary quality, and topic-segmentation precision are planned for the next phase, after which these expected outcomes will be verified against measured results.

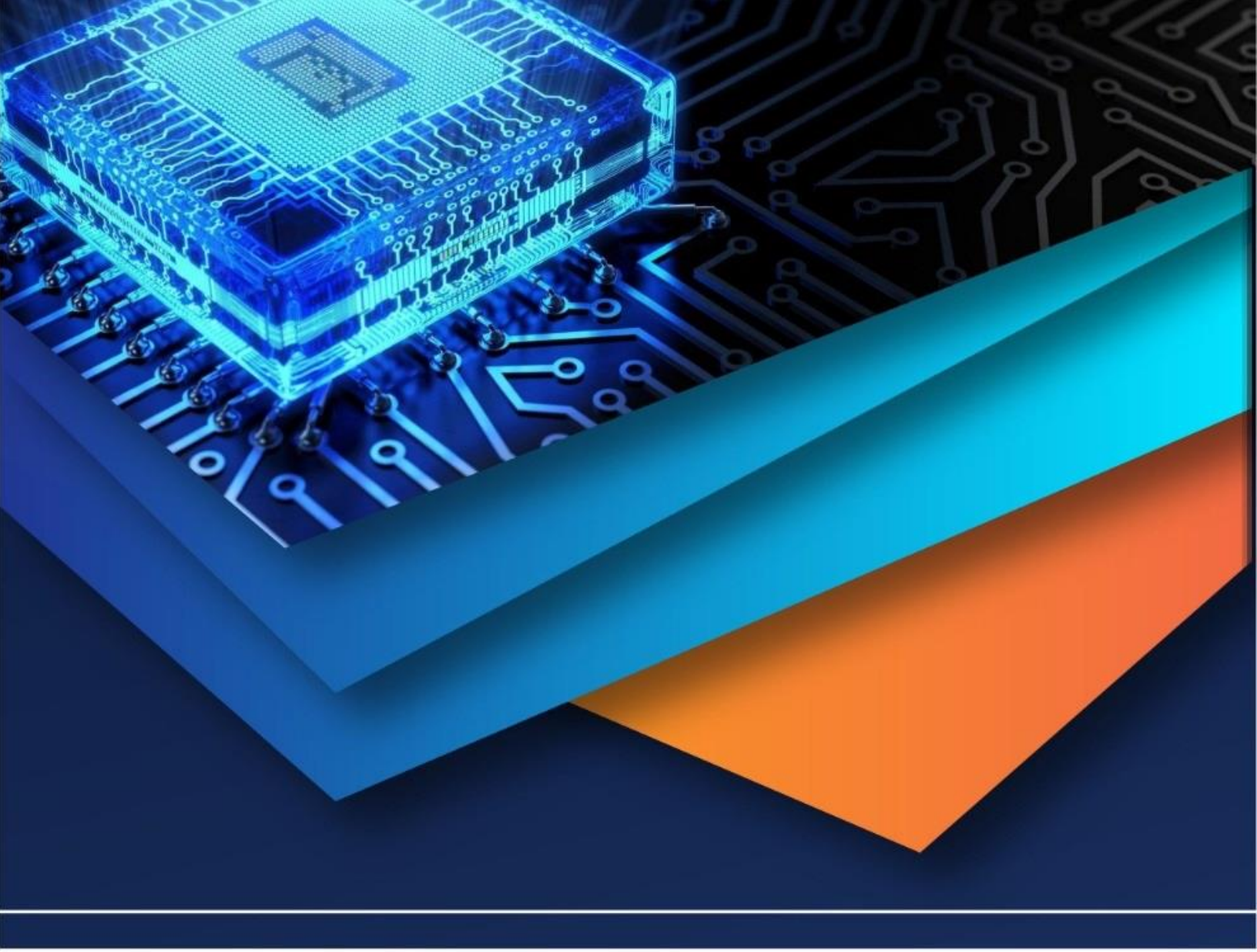
IV. CONCLUSION AND FUTURE WORK

This paper reviewed ten Transformer-based approaches to speech recognition, language representation and text summarization, and used the identified gaps — task isolation, high computational cost, and the absence of educational-relevance filtering — to motivate the design of StudyScribe, a proposed pipeline that combines automatic speech recognition, NLP-based content filtering, semantic topic segmentation, and abstractive summarization into a single system for converting raw lecture recordings into structured study material.

Future work will focus on implementing each pipeline stage, integrating them into a working prototype, and evaluating the system on real lecture recordings against metrics such as transcription accuracy, summary quality (e.g., ROUGE scores), and topic-segmentation precision. Subsequent phases will also explore lighter-weight model variants to reduce the computational cost flagged as a limitation across the reviewed literature, so that StudyScribe remains practical to deploy for everyday student use.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.
[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.
[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in Proc. ACL, 2020.
[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, 2019.
[5] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in Proc. ICML, 2020.
[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, 2020.
[7] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," arXiv:2004.05150, 2020.
[8] X. Yao, D. Zhao, Y. Gao, B. Zhang, and N. Duan, "Neural question generation from text: A preliminary study," in Proc. NLPCC, 2018.
[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv:2212.04356, 2022.
[10] M. A. Albadawi, H. Alshammari, A. Alqahtani, et al., "AI-based automatic lecture summarization and note generation system using transformer models," IEEE Access, vol. 13, pp. 45678–45692, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)