ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

ROME BUSINESS SCHOOL

Bert-Based Speech-to-Text Notes Generator for Educational Content Accessibility

(Data Science Management)

Isaac, Onoriode, Oshevire[1]; Caleb, Ande[2]; Oluwatosin, Oluwaseun, Babatunde[3]; Grace, Jesutola, Ajayi[4]; Chigozie, David, Eze[5]; Timilehin Ilupeju[6]; Oluwatobi Balogun[7]

[1](25ADS1002OI87); [2](25ADS2805AO83); [3](25ADS1009BO97); [4](25ADS0911AG96); [5](25ADS2312EC89)

[1;2;3;4;5;6;7]Rome Business School Nigeria

Publication Date: 2026/03/23

How to Cite: Isaac, Onoriode, Oshevire; Caleb, Ande; Oluwatosin, Oluwaseun, Babatunde; Grace, Jesutola, Ajayi; Chigozie, David, Eze; Timilehin Ilupeju; Oluwatobi Balogun (2026) Bert-Based Speech-to-Text Notes Generator for Educational Content Accessibility. *International Journal of Innovative Science and Research Technology*, 11(3), 1818-1851. https://doi.org/10.38124/ijisrt/26mar883

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

CERTIFICATION

This project work (Research) titled, BERT-BASED SPEECH-TO-TEXT NOTES GENERATOR FOR EDUCATIONAL CONTENT ACCESSIBILTY, prepared and submitted by ISAAC, ONORIODE, OSHEVIRE, CALEB, ANDE, OLUWATOSIN, OLUWASEUN, BABATUNDE, GRACE, JESUTOLA, AJAYI, CHIGOZIE, DAVID, EZE in partial fulfilment of the award of the degree of MASTER IN DATA SCIENCE MANAGEMENT is hereby accepted.

Date Defended / Approved: February 22, 2026

Dr. Tobi Balogun
Project Supervisor

Samuel Igwe
Head of Academics, RBSN

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

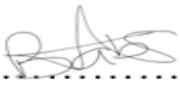
DECLARATION STATEMENT

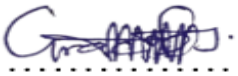
We, ISAAC, ONORIODE, OSHEVIRE, CALEB, ANDE, OLUWATOSIN, OLUWASEUN, BABATUNDE, GRACE, JESUTOLA, AJAYI, CHIGOZIE, DAVID, EZE students of the Master in Data Science Management at Rome Business School, carried out this research work and hereby declare that the (dissertation) is our own work.

We have not copied the work of others in any way, as we have maintained the required ethical standards.

Signed:  Date: 09/03/2026

Signed:  Date: 09/03/2026

Signed:  Date: 09/03/2026

Signed:  Date: 09/03/2026

Signed:  Date: 09/03/2026

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

DEDICATION

This project is dedicated to God, our families, lecturers, and all individuals who have contributed to our academic growth and success.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

ACKNOWLEDGEMENT

We express our sincere gratitude to God Almighty for wisdom, strength, and guidance throughout this research. Our profound appreciation goes to our project supervisor, Dr. Tobi Balogun, for his invaluable guidance, constructive criticism, and academic mentorship. We also thank the lecturers of the faculty of Data Science Management for their support and knowledge imparted during this program. Special thanks to our families and friends for their encouragement and support throughout the period of this study.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

ABSTRACT

Students often find it difficult to take accurate and complete notes during lectures due to fast-paced speech, unfamiliar accents, background noise, and the pressure of multitasking. These challenges are even more pronounced for students with learning difficulties, disabilities, or those who are non-native English speakers. Traditional note-taking methods do not always guarantee clarity or completeness, which affects comprehension and academic performance. With advancements in artificial intelligence (AI), it is now possible to explore automated tools that can transcribe and summarize lectures to support more effective learning.

This study addresses the problem of limited access to accurate and real-time lecture notes. Existing speech-to-text systems are often trained on clean, studio-quality datasets and struggle to perform well in real-world classroom environments with noise, diverse accents, and technical terms. Most available solutions are not tailored for Nigerian contexts and fail to meet the academic needs of students. To solve this problem, a solution that integrates advanced AI models was developed to improve transcription accuracy and automatically summarize educational content.

The system combines Wav2Vec 2.0 for speech recognition and BERT for extractive summarization. Publicly available datasets such as LJ Speech and CNN/DailyMail were used for training and testing. The audio was preprocessed using noise reduction and segmentation, while the text data underwent tokenization and lemmatization. The models were fine-tuned and integrated into a single application with a graphical interface. The system achieved a Word Error Rate (WER) of 0.2 and a ROUGE-1 score of 0.8, indicating strong performance. The interface allows users to upload or record audio, generate full transcripts, produce summaries, and export the output in readable formats.

In conclusion, this project demonstrates that combining transformer-based models like Wav2Vec 2.0 and BERT can provide an efficient and accessible solution for lecture note generation. It enhances learning for all students, particularly those with special needs, and supports inclusive education through AI-based tools.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

TABLE OF CONTENT

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

LIST OF TABLES

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

LIST OF FIGURES

Figures **Page**

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

ABBREVIATIONS

RBSN	Rome Business School Nigeria
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CER	Character Error Rate
CNN	Convolutional Neural Network
CSR	Continuous Speech Recognition
CTC	Connectionist Temporal Classification
DTW	Dynamic Time Warping
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
GUI	Graphical User Interface
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RTF	Real-Time Factor
STT	Speech-to-Text
WER	Word Error Rate

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

CHAPTER ONE
INTRODUCTION

- *Background to Study*

Note-taking is a critical aspect of academic success, yet the traditional method presents significant challenges. During lectures, students are expected to listen, process information, and write simultaneously, which often leads to incomplete or disorganized notes. This multitasking reduces attention and comprehension, especially in fast-paced or technical courses. For example, a student may miss the explanation of a data structures algorithm while attempting to document the previous point. Such inefficiencies can hinder learning, particularly during revision or exam preparation.

To address these challenges, Speech-to-Text (STT) technology, also known as Automatic Speech Recognition (ASR), has emerged as an innovative solution. STT systems convert spoken language into written text in real-time using natural language processing and machine learning techniques (Tan et al., 2024). These tools allow students to focus more on understanding the lecture content rather than on writing, thereby improving engagement and retention (Le et al., 2024). The generated transcripts are editable, allowing students to highlight, add notes, and personalize their notes based on individual learning styles.

Beyond improving note-taking efficiency, STT tools also promote accessibility and inclusivity in education. Students with hearing impairments can follow lectures through real-time captions, as seen in tools like Google Meet or Microsoft Teams (Zhao et al., 2025). Those with motor skill challenges, such as students with cerebral palsy, can use STT to automatically generate notes without having to write manually, encouraging independent learning (Chen et al., 2024). Additionally, non-native English speakers benefit from the ability to replay or reread transcripts at their own pace, helping them overcome difficulties with fast speech or unfamiliar accents.

Recent advances in artificial intelligence, particularly in transformer-based models, have significantly improved STT systems. The Transformer architecture, introduced by Vaswani et al. in 2017, processes entire sequences of data simultaneously using self-attention mechanisms (Karmakar et al., 2024). This allows for better handling of long-range dependencies in both speech and text compared to older models like Recurrent Neural Networks (RNNs).

Notable Transformer-based models like BERT, GPT-3, and T5 have significantly advanced speech-to-text and NLP systems. BERT offers strong contextual understanding for summarization tasks, GPT-3 excels at generating fluent, human-like text across various domains, and T5 unifies language tasks into a simple text-to-text format for greater flexibility (Chang et al., 2024; Karmakar et al., 2024). Unlike earlier models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memories (LSTMs), which struggle with long-range dependencies, transformers use attention mechanisms to process sequences more efficiently and accurately, an approach further detailed in Chapter Two (Fan et al., 2022; Soydaner, 2022). These models enable modern STT systems not only to transcribe audio accurately but also to summarize and refine the resulting text for better usability.

In Nigeria, although university enrollment has grown, many students still struggle with academic performance. Research shows that about 30% of students fail to graduate on time, with factors such as poor study habits and ineffective note-taking contributing to this trend (Ahmadu et al., 2024). While existing STT models have been explored, many lack the ability to summarize content or adapt to noisy and varied speech inputs. For instance, the CNN-based system by Dua et al. (2022) performed well with tonal speech but failed to manage long-term speech patterns. Similarly, Hussein and Mahmood's (2022) hybrid CNN-GRU model improved accuracy but struggled with longer inputs and noise interference.

Therefore, this study proposes the development of a Transformer-Based Speech-to-Text Notes Generator that combines Wav2Vec 2.0 and BERT. Wav2Vec, a self-supervised model developed by Facebook AI, learns rich representations from raw audio and is effective even with limited labeled data. It is robust to variations in accent and background noise. BERT is then applied to summarize the transcribed text, making the notes more concise and contextually accurate. This integrated system aims to improve lecture engagement, reduce cognitive load, and provide accessible, high-quality learning support for students.

- *Statement of the Problem*

University education in Nigeria has witnessed significant expansion over the past few decades, with an increasing number of students enrolling in tertiary institutions. Despite this growth, a concerning issue persists as the high rate of student failure in Nigerian universities. Recent statistics indicate that approximately 30% of university students in Nigeria fail to graduate within the stipulated time frame or drop out entirely before completing their programs (Ahmadu et al., 2024).

Existing speech-to-text tools offer some assistance, but they frequently fall short in accuracy, especially with accents, background noise, and specialized terminology. As seen in the works of Dua et al. (2022), their CNN-based speech-to-text system showed an accuracy of 89.15% using tonal speech signals, with a Word Error Rate (WER) of 10.56%. However, the model lacked the ability to handle long-range speech dependencies. While these models contribute to advancement of speech recognition, they do not fully address the need for contextual understanding and summarization in educational settings. Therefore, in this study, a BERT-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

based speech-to-text note generator will be proposed to mitigate these issues by using Wav2Vec 2.0 for accurate speech recognition and integrating BERT for automatic summarization of transcripts.

➢ *Aim and Objective*

The aim of this project is to implement a speech recognition and summarization system that transcribes and summarizes audio content.

- *The Objectives are:*

✓ To collect and preprocess audio and text summarization data.
✓ To implement wav2vec and BERT-based text summarization model.
✓ To evaluate the performance of the model using Word Error Rate (WER) and ROUGE.

➢ *Research Methodology*

This research follows a structured approach to achieve the stated objectives, involving the collection and preprocessing of audio and text for summarization, implementation, and evaluation. The methodology includes the following phases:

To achieve the first objective, the process begins with data collection, utilizing open-source datasets from Kaggle, for speech recognition and CNN/DailyMail for text summarization. Additional lecture audio data is gathered to ensure diversity in accents and academic content. The collected speech data undergoes preprocessing, including noise reduction through spectral subtraction, segmentation of long lectures into smaller, manageable chunks, and normalization to standard formats like 16kHz mono-channel audio. For text summarization, preprocessing involves tokenization, stopword removal, and lemmatization or stemming to refine textual content for better summarization accuracy.

To implement the wav2vec and BERT-based text summarization model, the process begins with fine-tuning a pre-trained wav2vec 2.0 model on collected lecture datasets. This model converts audio into latent speech representations and transcribes it into text, using techniques such as Viterbi decoding or Connectionist Temporal Classification (CTC) to align audio features with text sequences.

To evaluate the model's performance, transcription accuracy is assessed using Word Error Rate (WER), which quantifies errors by measuring insertions, deletions, and substitutions in the transcribed text compared to ground truth transcripts. Summarization quality is evaluated with ROUGE scores, which compare machine-generated summaries to human-written references, analyzing overlaps of unigrams, bigrams, and longest common subsequences. Additionally, the BLEU score may be used as an optional metric to assess precision in sentence-level summarization.

➢ *Scope of the Study*

The study focuses on developing a speech-to-text Notes Generator within the domains of speech recognition and natural language processing for educational use. The system uses Wav2Vec 2.0 to convert spoken lectures into text and BERT to summarize the transcribed content into clear, readable notes. Python would be used for the implementation, along with libraries such as Hugging Face Transformers, NLTK, and Librosa. The project also involves preparing the data, training the models, and evaluating their performance using Word Error Rate (WER) and ROUGE scores and a GUI for easy access to transcribed and summarized lecture notes.

➢ *Significance of the Study:*

The Speech-to-text Notes Generator project enhances education by simplifying note-taking and making learning more efficient for student and educators. It enables students to keep up lecturers during class dictation, especially when the lecturer speaks too quickly for students to write everything down. The project improves efficiency by automating transcription and summarization, saving time, and providing organized notes.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

CHAPTER TWO
LITERATURE REVIEW

This chapter provides an overview of the existing literature and research methods related to Speech-to-Text. The main aim is to build a theoretical framework for the study and point out any gaps in the current research that this study tries to fill.

A. Speech-to-Text Recognition System

Speech-to-text recognition is the systematic process of converting audio signal into text that can be used in a variety of communications technologies, from virtual assistants to transcription services. In speech-to-text recognition systems, the conversion is often considered the most important step because it directly affects the accuracy and reliability of the final transcription. This stage involves mapping raw audio signals to meaningful textual output, which requires precise interpretation of phonemes, speech patterns, and linguistic structures. Any errors in this process such as misrecognizing words due to accent, speed, or noise can distort the intended message, making subsequent tasks like summarization or analysis ineffective, as there are several stages in the process which include Acoustic models that trace the relationship between the audio signal and the linguistic units of speech (such as phonemes) of the conversation (Zhao et al., 2025). By studying aspects such as frequency and amplitude, models discover patterns matching specific sounds. Language models estimate the probability of word sequences to choose the most likely interpretation for ambiguous sounds. They rely on statistical patterns in large text corpora to predict which words are most likely to follow others, improving the accuracy of transcriptions in context (Benzirar et al., 2025). The language model involves statistical analysis of a number of large text corpora, which are used to understand common word pairings and grammatical representations of language. Before the processing stage, the system extracts relevant features from the raw audio input. Feature extraction is a key step in speech recognition where important characteristics of the audio are identified. Methods such as Mel-frequency cepstral coefficients (MFCCs) are used to capture the characteristics of the speech signal that can help to accurately recognize the speech (Avro et al., 2025). MFCCs represent the short-term power spectrum of sound and are widely used in STT systems due to their effectiveness in modeling the human ear's response. Decoding and Recognition involve translating audio features into meaningful text. Using algorithms such as the Hidden Markov Model (HMM) or neural networks, the system encodes the audio input into text using an acoustic and language model. HMM is a statistical model that assumes the system the system being modeled is Markov process with hidden states, suitable for time-series data like speech. The encoded text is then combined with the language model to generate the most likely transcription.

Advanced systems typically employ deep learning methods, which use multi-layered neural networks to increase recognition rate. These networks learn complex representations from large datasets, improving the system's ability to handle varied speech inputs.

End-to-end modeling is a modern approach that tries to eliminate intermediate coding steps in order to reduce processing time. In this approach, output data can be directly mapped from the audio to the text allowing for more efficient and accurate transcriptions (Du et al., 2024; Orosoo et al., 2025; Niyozmatova et al., 2025).

➢ *Evolution of Speech-to-Text Technology*

Speech recognition has a rich history dating back to the 1950s with the introduction of basic recognition systems, such as Bell Labs' "Audrey" system, which recognized digits spoken in English. Subsequent decades witnessed significant advancements:

The development of Hidden Markov Models (HMMs) in the 1970s laid the foundation for statistical modeling of speech signals. Early speech recognition systems, such as "Harpy" developed by Carnegie Mellon University in the 1970s had the capacity to recognize about 1,000 words due to the limited computational power and memory resources available at the time. Additionally, the algorithms used were based on rule-based and template matching methods, which could not efficiently scale to larger vocabularies (Valencia-Angulo et al., 2025).

The emergence of dynamic time warping (DTW) and continuous speech recognition (CSR) techniques in the 1990s, real-time speech recognition became feasible. DTW is an algorithm used to measure the similarity between two temporal sequences that may vary in speed. It allowed early systems to align spoke input with stored templates even if the speech was not perfectly timed, improving recognition accuracy. CSR, on the other hand, enabled systems to recognize natural, uninterrupted speech instead of isolated words, making it possible to process spoken language as it is typically used in real conversations or lectures. These advancements reduced the processing delays and rigid input requirements seen in earlier systems, paving the way for real-time STT applications (Valencia-Angulo et al., 2025; Benzirar et al., 2025). The integration of deep learning methods, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks in the 2010s, revolutionized speech-to-text systems. This progress was made possible by key advancements in computational power (notably GPUs), the availability of large-scale labelled speech datasets, and improved training algorithms for deep neural networks. These factors enabled RNNs and LSTMs to process long sequences of speech data more effectively, improving transcription accuracy and allowing real-time speech recognition (Choi et al., 2024; Fang et al., 2024).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

The integration of deep learning methods, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, has revolutionized speech-to-text systems. More recently, innovations such as Transformer-based models (e.g., OpenAI's Whisper and Google's Wav2Vec) have significantly enhanced accuracy by handling diverse accents and noisy environments. These models leverage self-attention mechanisms to process entire speech sequences efficiently, resulting in higher precision, better contextual understanding, and real-time transcription capabilities in modern STT systems.

B. Introduction to Machine Learning and Deep Learning

➢ *Machine Learning*

Machine Learning (ML) is a branch of artificial intelligence that enables systems to learn from data and improve their performance on specific tasks without being explicitly programmed. It involves algorithms that identify patterns and make decisions based on training data (Karmakar et al., 2024). ML has become a fundamental tool in speech recognition, powering systems that can process and interpret spoken language effectively.

Machine learning is generally categorized into:

- Supervised learning, which uses labeled data to train models.
- Unsupervised learning**,** which detects patterns in unlabeled data.
- Reinforcement learning, where models learn optimal behavior through feedback from interactions with an environment.

These learning paradigms are essential in designing intelligent systems that adapt to diverse data, such as varying accents, noise levels, or speaking speeds in STT applications.

➢ *Deep Learning*

Deep Learning (DL) is a specialized subfield of ML that utilizes multi-layered neural networks to model complex patterns in large datasets. These deep neural networks are capable of learning both low-level and high-level features, making them particularly powerful for audio and language processing tasks (Fang et al., 2024).

Deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are especially suited for handling sequential data like speech. They maintain memory of previous inputs, which helps in understanding context and predicting future inputs in audio streams. This capability makes them integral to modern speech recognition pipelines.

More recently, transformer-based models like Wav2Vec 2.0 and BERT have emerged as powerful alternatives, capable of processing sequences more efficiently by using attention mechanisms instead of relying on recurrence. These models have significantly improved the accuracy and scalability of STT systems (Chang et al., 2024).

C. Methods of Developing Speech-to-Text Recognition System

Developing Speech-to-Text Recognition systems involves complex methodologies that integrate machine learning technics. This article outlines the primary methods used in the development of speech-to-text recognition systems.

➢ *Recurrent Neural Networks (RNN)*

RNNs have become integral to the development of advanced speech-to-text recognition systems. These networks are designed to process sequential data, making them particularly suitable for interpreting human speech, which is inherently temporal and sequential. This article examines the role of RNNs in speech-to-text recognition, exploring their architecture, functionality, and impact on the field. Speech-to-text recognition involves converting spoken language into written text, a complex task due to the variability and nuances of human speech. Variations in accent, speed, intonation, and background noise are just a few examples. Traditional models like Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) failed to capture temporal dependencies and contextual information over long sequences. In addition, RNNs introduce loops as part of the network architecture and allow for information to persist and thereby allowing the modeling of temporal dynamic. This enables RNNs to store previous inputs in their own state which makes them a memory of the sequence, which is essential for understanding context in speech. The standard architecture of RNNs causes problem of vanishing and explosion of gradients in training, further constraining learning of long-term dependencies (Fang et al., 2024).

To overcome these limitations, advanced variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) were developed. LSTMs include memory cells and gating mechanisms that regulate the flow of information to prevent the problem of gradient vanishing and explosion (Jbene et al., 2025; Quayum et al., 2025; Patil et al., 2025). GRUs reduce the complexity and performance of an LSTM architecture, but provides a quicker alternative. In speech-to-text speech recognition systems, RNNs process input features that are extracted from audio signals, such as Mel-frequency cepstral coefficients (MFCCs) (Jbene et al., 2025; Quayum et al., 2025; Patil et al., 2025). Because RNNs process features sequentially, the system can take into account the temporal context of speech, improving the accuracy of phoneme and word recognition. Integration of connectionist temporal classification (CTC) allows RNNs to align input audio sequences with corresponding text transcriptions without having to

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

segment the speech, thus speeding up the training process. The use of RNNs has led to significant performance improvements in speech recognition. Deep Speech is an end-to-end speech recognition system developed by Baidu that utilizes deep RNNs to achieve excellent accuracy levels. Google's speech recognition systems have also adopted RNNs to improve their performance. More especially, their speech recognition systems have been trained with RNNs, improving accuracy in speech recognition tasks that involve continuous and spontaneous speech (Khan et al., 2024; Quayum et al., 2025; Patil et al., 2025).

Though recurrent neural networks have many advantages, they are computationally expensive. They require large amounts of computational resources and time to train. This problem is especially so for large-scale datasets. To address this problem, techniques such as parallel computing, optimization algorithms, and hardware accelerators such as Graphics Processing Units (GPUs) are employed. Other approaches such as hybrid models that combine RNNs with Convolutional Neural Networks (CNNs) or incorporate attention mechanisms have also been proposed to improve their efficiency and accuracy (Kazemi and Alvanchi, 2025). Recurrent neural networks have significantly advanced speech-to-text recognition systems by capturing temporal sequences and contextual information within the model. The ability to capture sequential data has led to more accurate and robust systems, contributing to the growth of human-machine communication interfaces and applications such as virtual assistants, transcription services, and language translation tools.

- *Long Short-Term Memory (LSTM)*

The purpose of recurrent neural networks is to identify patterns in data sequences, including spoken language, handwriting, genomes, and text. However, problems like vanishing or exploding gradients during training make it difficult for traditional RNNs to handle long-term dependencies. By adding memory cells and gating mechanisms that control the information flow, LSTM networks overcome these difficulties (Shao et al., 2025). Because of their architecture, LSTMs can learn and retain information over long input sequences, which makes them appropriate for tasks involving temporal dependencies. The system must correctly translate audio signals, which are time-series data, to their corresponding textual representations in order to perform speech-to-text recognition (Dolaeva et al., 2025). Because LSTM networks can capture context and long-range dependencies, they are well-suited to handle such sequential data. The model can retain contextual information over time by processing speech signals through LSTM layers, which improves word and phrase recognition (Khan et al., 2025; Hasan et al., 2025). Significant gains have been made when LSTM networks are integrated into the language and acoustic modeling parts of speech recognition systems. In language modeling, LSTMs aid in word sequence prediction by taking into account the context given by preceding words, while in acoustic modeling, they aid in learning the relationship between audio features and phonetic units (Pradhan and Yajnik, 2024).

The ability of LSTM networks to handle sequential data and temporal dependencies is their main benefit when it comes to speech-to-text systems. In order to comprehend context and distinguish between homophones—words that sound the same but have different meanings. LSTMs must be able to retain prior inputs for extended periods of time. Furthermore, it has been demonstrated that LSTM networks perform more accurately than conventional models. Studies show that in continuous speech recognition tasks, LSTMs lower word error rates. They are a potent tool for enhancing the resilience of speech recognition systems against speech alterations like accents, speaking rates, and background noise because of their capacity to model intricate patterns in data (Dekel et al., 2024). An important development in machine learning and artificial intelligence applications is the use of LSTM networks in speech-to-text recognition systems. LSTMs improve the system's capacity to accurately transcribe spoken language into text by efficiently modeling temporal dependencies and preserving contextual information. This advancement not only makes speech recognition technologies more useful in a wider range of applications, but it also makes human-computer interactions more efficient and natural.

- *Bidirectional Long Short-Term Memory (BI-LSTM)*

In the field of speech-to-text recognition systems, BI-LSTM networks have become a significant breakthrough. These networks are a particular kind of Recurrent Neural Network (RNN) architecture that takes dependencies into account in both forward and backward temporal directions when processing sequential data. By capturing the rich contextual information present in human language, this bidirectional processing capability improves speech recognition task performance. Accurately translating spoken language into written text is the goal of speech-to-text recognition systems. The sequential processing of data from the past to the future by traditional RNNs restricts their capacity to use future context in prediction. Understanding the subtleties of speech, where a word's meaning can be greatly influenced by both its preceding and succeeding words, may be beyond the scope of this unidirectional approach (Choi et al., 2024).

In order to overcome the vanishing gradient issue and enable the model to learn long-term dependencies, Long Short-Term Memory (LSTM) networks introduce memory cells and gating mechanisms that control information flow. Standard LSTMs, however, continue to function in a unidirectional fashion. By adding two parallel layers that process input sequences in opposite directions, BI-LSTM networks expand on the traditional LSTM (Poorna et al., 2025; Avro et al., 2025; Arafath et al., 2025). The sequence is processed from beginning to end by one layer and from beginning to end by the other. At each stage of the sequence, the network can access both past and future context by concatenating or combining the outputs (Benzirar et al., 2025).

BI-LSTM networks greatly enhance the modeling of temporal sequences in speech signals for speech-to-text recognition. Because speech is sequential and context-dependent by nature, the network's ability to correctly recognize phonemes and words is

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

improved by capturing information from both preceding and succeeding frames. In speech, coarticulation effects—where a phoneme's articulation is influenced by nearby phonemes—are especially well-handled by BI-LSTMs.

The system is able to better capture the dynamic properties of speech signals through the use of BI-LSTM networks in acoustic modeling. The network is better able to model the temporal dependencies and variations in speech by taking into account both past and future acoustic contexts. This thorough context modeling lowers transcription error rates and improves continuous speech recognition. Moreover, BI-LSTM networks support speech recognition systems' language modeling. The network can better predict word sequences by using bidirectional context, which is crucial for comprehending syntactic structures and patterns in natural language. This feature improves the system's ability to transcribe speech with greater semantic coherence and grammatical accuracy.

The efficacy of BI-LSTM networks in speech-to-text recognition tasks has been empirically proven. BI-LSTMs outperform unidirectional LSTM and conventional RNN models in terms of accuracy rates and their ability to manage speech data variability. BI-LSTMs are especially well-suited for applications that demand high precision in sequential data processing because of their capacity to learn from complete input sequences. To sum up, Bidirectional Long Short-Term Memory networks are essential for improving speech-to-text recognition technologies. BI-LSTMs capture extensive contextual information by processing input sequences both forward and backward, which results in more reliable and accurate speech recognition. By incorporating BI-LSTM networks into speech recognition architectures, more complex and dependable systems are developed, improving applications in fields like accessibility technologies, voice-activated assistants, and transcription services (Vinothkumar et al., 2024).

- *Attention Mechanism*

The attention mechanism was first developed in neural machine translation to overcome processing constraints for lengthy data sequences. The encoder transforms input data into a fixed-size context vector, which the decoder uses to produce output in conventional encoder-decoder architectures. However, because the fixed-size vector might not capture all relevant information, this method frequently results in information bottlenecks, particularly with long inputs. This problem is lessened by the attention mechanism, which enables the decoder to dynamically access various encoder output segments and concentrate on pertinent data at every stage of the decoding procedure. The attention mechanism in speech-to-text recognition works by matching the audio input sequence to the matching text output sequence (Fu et al., 2022; Mamatov et al., 2022; Soydaner et al., 2022). After processing the speech input signals, the encoder creates a series of hidden representations. The attention mechanism allows the decoder to weigh these representations differently at each time step, rather than combining them into a single context vector. The model can "attend" to particular segments of the input sequence that are most important for producing the subsequent output token thanks to this dynamic weighting. There are various benefits to using the attention mechanism in ASR systems (Fan et al., 2022). First off, by avoiding the potential loss of important information that could happen with fixed-size context vectors, it enhances the handling of lengthy audio sequences. Second, by concentrating on features relevant to the current decoding step, the model improves its ability to handle variability in speech patterns. Thirdly, the attention mechanism enhances the learning of alignment between speech and text, which is essential for accurate transcription (Khan et al., 2024).

In speech recognition tasks, empirical research has shown that attention-based models perform better than conventional models. Higher accuracy and robustness are a result of these models' capacity to learn soft alignments between input and output sequences. Furthermore, attention mechanisms simplify the architecture and eliminate the need for intermediate representations or hand-engineered features by enabling end-to-end training of ASR systems. In summary, the attention mechanism's incorporation into speech-to-text recognition systems is a big step toward solving the problems associated with precise speech transcription. The attention mechanism improves ASR systems' ability to handle the complexities of human speech by enabling models to dynamically focus on pertinent portions of the input sequence. Ongoing research continues to refine these models, promising further improvements in performance and application scope (Karmakar et al.,2024).

D. Evaluation Metrics for Speech-to-Text Systems

Evaluating the performance of speech-to-text (STT) systems is essential for determining their accuracy and real-world usability. Common evaluation metrics include Word Error Rate (WER), Character Error Rate (CER), and Real-Time Factor (RTF), each focusing on different aspects of system performance (Xu et al., 2023; Benzirar et al., 2025).

- *Word Error Rate (WER)*

WER is the most widely used metric in speech recognition. It calculates the proportion of incorrect words in the system's output compared to a reference transcript. It is defined as:

$$WER = \frac{S+D+I}{N} \quad (1)$$

Where:

S = number of substitutions
D = number of deletions

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

I = number of insertions
N = total number of words in the reference

WER is a reliable indicator of how accurately an STT system captures the content of spoken language (Xu et al., 2023).

➢ *Character Error Rate (CER)*

CER is similar to WER but measures transcription errors at the character level. It is often used in languages with complex word structures or when analyzing short utterances (Chang et al., 2024).

$$CER = \frac{S_c+D_c+I_c}{N_c} \quad (2)$$

Where:

$S_c D_c I_c$ = number of character-level substitutions, deletions, and insertions
N_c = total number of characters in the reference

➢ *Real-Time Factor (RTF)*

RTF evaluates the processing speed of the system. It compares how long the system takes to process audio versus the actual duration of that audio (Fang et al., 2024):

$$RTF = \frac{T_p}{T_a} \quad (3)$$

Where:

T_p = time taken by the system to process the input

T_a = duration of the audio input

An RTF below 1.0 suggests the system can operate in real time, which is ideal for lecture environments.

➢ *Justification of Selected Metric*

Among the metrics, Word Error Rate (WER) is selected as the primary evaluation metric for this project. WER is widely used in speech recognition tasks because it quantifies how well the system captures spoken language compared to a reference transcript (Xu et al., 2023). While CER and RTF provide supplementary insights, WER offers a word-level accuracy assessment, which is particularly important in academic environments. In such settings, transcription errors, especially omissions or substitutions, can significantly alter the meaning of technical lectures or course content. Thus, WER provides a meaningful measure to evaluate how well the system supports lecture comprehension and accurate note generation for students

E. Performance Evaluation for Text Summarization

Summarization systems require distinct evaluation metrics that compare the machine-generated summaries to reference summaries written by humans. Common metrics include ROUGE**,** BLEU, and BERTScore**.** These metrics focus on content overlap, fluency, and semantic similarity between the generated and reference summaries (Zhou et al., 2023; Lin, 2020).

➢ *ROUGE (Recall-Oriented Understudy for Gisting Evaluation)*

ROUGE measures the overlap of n-grams, word sequences, and word pairs between the system output and a human-written reference. It is widely used for extractive and abstractive summarization.

$$ROUGE - N = \frac{\sum_{overlap\ of\ n-grams}}{\sum_{reference\ of\ n-grams}} \quad (4)$$

Where n is the length of the n-gram (e.g., 1 for unigrams, 2 for bigrams).

➢ *BLEU (Bilingual Evaluation Understudy)*

BLEU evaluates the precision of n-grams in the generated summary compared to one or more reference summaries. Though originally developed for machine translation, it is frequently applied in summarization evaluations as well (Sun et al., 2021).

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n\ logp_n\right) \quad (5)$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

Where:

p_n is the modified precision for n-grams,

w_n is the weight (typically uniform),

And BP is the brevity penalty for short outputs.

➢ *BERTScore*

BERTScore evaluates similarity at the semantic level by comparing contextual embeddings of words using a pre-trained BERT model. It can detect meaning-level similarity even if the wording differs.

BERTScore uses cosine similarity between token embeddings from BERT to assess precision, recall, and F1-score (Zhang et al., 2020).

➢ *Justification of Metric Choice*

For this project, ROUGE is selected as the primary evaluation metric for summarization because it is widely adopted in educational NLP tasks and aligns well with the extractive summarization approach used by the BERT model (Zhou et al., 2023). ROUGE measures content retention by comparing overlapping phrases and keywords between generated and reference summaries. BLEU, while useful for precision evaluation, is better suited for translation tasks, and BERTScore, though advanced, is computationally heavier. ROUGE offers a practical and interpretable metric to ensure that summaries produced in this system are not only concise but also academically meaningful.

F. Conceptual Model of the Proposed STT System

A conceptual model serves as a high-level representation of a system's structure and function. It helps to outline the flow of processes and the interaction between major components involved in achieving a specific objective. In this study, the conceptual model illustrates the entire process of transforming spoken lecture audio into summarized text notes using transformer-based Speech-to-Text (STT) technology. (Baevski et al., 2020; Devlin et al., 2019).

The model for the proposed STT note generator comprises the following key components:

- Audio Input: This represents the raw spoken content captured from a lecture, either through live recording or uploaded audio files.
- Preprocessing Module: The audio input is cleaned and standardized. This step includes noise reduction and feature extraction using methods such as Mel-Frequency Cepstral Coefficients (MFCCs), which identify important acoustic patterns (Avro et al., 2025).
- Speech Recognition Module (Wav2Vec 2.0)**:** Wav2Vec 2.0 is a transformer-based model that converts the preprocessed audio into raw text. It is robust to noise and capable of learning directly from raw speech signals (Baevski et al., 2020).
- Text Summarization Module (BERT)**:** The raw transcript is then passed to a summarization module built on BERT, which extracts the main points and generates clear, concise lecture notes (Devlin et al., 2019).
- Output Module: The final summarized notes are displayed to the user in an editable, accessible format.

This model emphasizes automation, efficiency, and accessibility, aligning with the goal of enhancing educational content delivery and student learning through advanced AI technologies (Karmakar et al., 2024).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

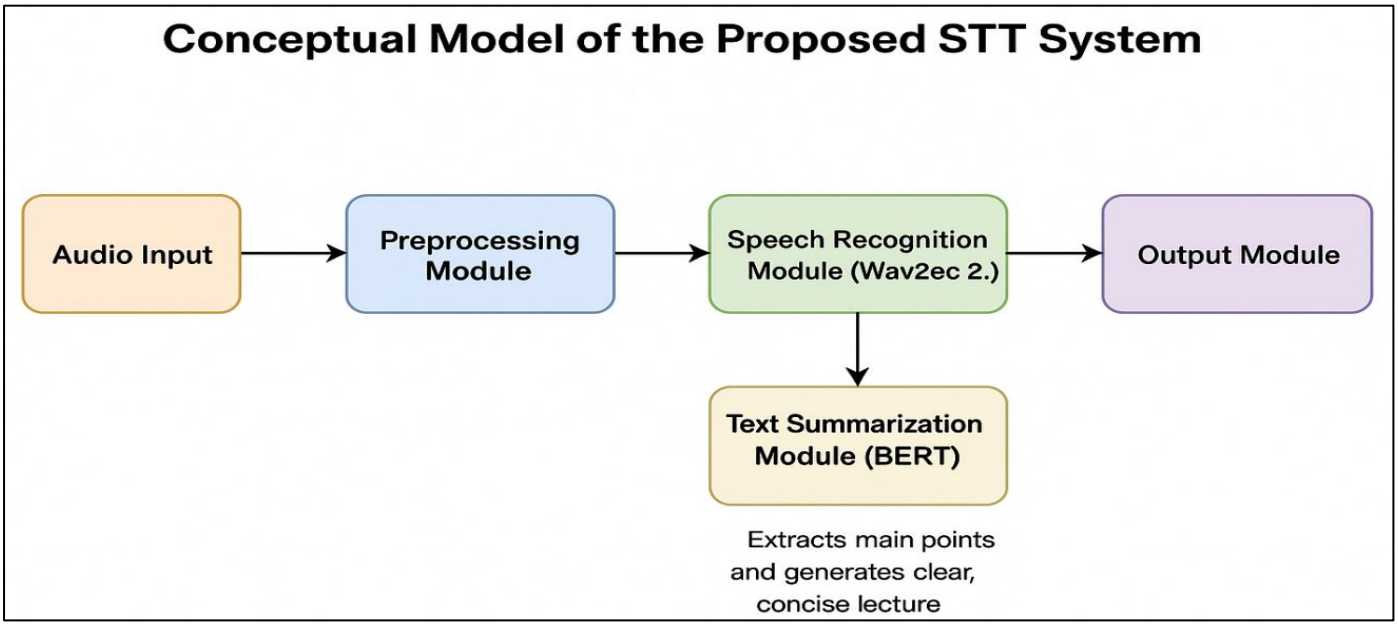


Fig 1 The Conceptual Model of the Proposed STT System (Baevski et al. (2020)

G. Conceptual Review: Relevance of Speech-to-Text and Summarization Technologies

The conceptual review provides a theoretical grounding for the technologies used in this study. It explains the major concepts that form the foundation of the proposed system and highlights their relevance in an educational context.

➢ *Speech-To-Text (STT) Technology*

Speech-to-text (STT), also known as automatic speech recognition (ASR), refers to the process of converting spoken language into written text. It plays a vital role in education by enhancing learning accessibility, enabling lecture transcription, and supporting students with disabilities or language barriers. STT tools reduce the burden of manual notetaking, allowing students to focus more on understanding content during class (Zhang et al., 2024).

➢ *Transformer-Based Speech Recognition*

Traditional STT systems relied on models like Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs), which were limited in handling long-range dependencies angd required extensive manual feature engineering. In contrast, transformer-based models such as Wav2Vec 2.0 process audio data using self-attention mechanisms and learn powerful representations from raw speech without requiring aligned transcripts (Baevski et al., 2020; Fang et al., 2024). This improves transcription accuracy, especially in noisy or real-world environments.

➢ *Text Summarization*

Text summarization is a natural language processing (NLP) task that aims to condense large bodies of text into shorter, meaningful summaries. In this project, BERT (Bidirectional Encoder Representations from Transformers) is used to perform extractive summarization on lecture transcripts. BERT's deep contextual understanding allows it to retain critical information while removing redundancy, making the notes more digestible for students (Devlin et al., 2019; Chang et al., 2024).

➢ *Relevance of Speech-To-Text to Education*

Speech-to-text tools have transformed education by addressing three major challenges: accessibility, efficiency, and engagement.

- Accessibility refers to ensuring that all students, including those with hearing impairments, motor difficulties, can access lecture content equally. STT tools support this by providing real-time captions and readable transcripts.
- Efficiency involves reducing the cognitive load of multitasking during lectures, students no longer have to divide their attention between listening and writing, as STT systems automatically generate structured notes.
- Engagement is enhanced as students can fully focus on understanding the material during class and later revisit accurate transcriptions for reinforcement and deeper learning.

Educational institutions increasingly integrate these tools into classroom management systems, online learning platforms, and assistive technologies (Zhang et al., 2024).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

➢ *Applications of Speech-to-Text Recognition Systems*

The versatility of speech-to-text technology has led to its adoption across various sectors:

- Virtual Assistants: Personal assistants like Siri, Alexa, and Google Assistant rely on speech recognition to interpret user commands. They perform tasks ranging from setting reminders to controlling smart home devices.
- Accessibility Services: Speech-to-text aids individuals with hearing impairments by providing real-time captioning. It enhances accessibility in educational settings, presentations, and multimedia content.
- Transcription Services: Automated transcription is valuable in fields like journalism, law, and medicine. It accelerates the documentation process by converting interviews, legal proceedings, and medical dictations into text swiftly.
- Customer Service: Interactive voice response (IVR) systems employ speech recognition to handle customer inquiries efficiently. They enable automated responses and direct callers to appropriate resources without human intervention.
- Language Learning: Speech recognition facilitates language education by allowing learners to practice pronunciation and receive immediate feedback, enhancing the learning experience.

H. Education

The introduction of speech-to-text recognition technology has had a big impact on education, among other areas. These systems have the ability to revolutionize teaching and learning environments by instantly translating spoken language into written text. Their incorporation into learning environments has created new opportunities for inclusivity, efficiency, and accessibility. The increased accessibility that speech-to-text technology offers to students with disabilities is one of its main advantages in the classroom. Traditional educational approaches may present major obstacles for people with learning disabilities, hearing impairments, or motor skill issues. By converting spoken words into text, speech-to-text systems help these students overcome these obstacles and more efficiently access lectures and discussions. This not only supports compliance with universal design for learning principles but also fosters an inclusive educational environment where all students can participate fully.

Speech-to-text recognition not only helps students with disabilities but also improves note-taking and information retention, which benefits all students. Instead of hastily taking notes, students can concentrate on actively participating in the lecture material thanks to real-time transcriptions. Better understanding results from this, and accurate lecture transcripts can be reviewed and studied later. Having a written transcription of lectures can also help non-native speakers learn the language. The integration of speech-to-text systems is advantageous for educators as well. Administrative duties like making lesson plans, administering tests, and giving feedback are made easier by technology. By dictating instead of typing, teachers can save time and increase productivity. Moreover, transcribed lectures and materials can be easily adapted into different formats, aiding in the creation of accessible content that meets diverse student needs.

Nevertheless, there are certain difficulties in putting speech-to-text recognition systems into practice. Because transcription quality can be impacted by differences in accents, speech patterns, and background noise, accuracy is still a major concern. To guarantee efficient use, educational institutions must make investments in top-notch systems that can process a variety of linguistic inputs. To optimize the potential advantages of these technologies, training for teachers and students is also necessary. Data security and privacy are two more important factors. Access to audio data, which may contain sensitive information, is frequently necessary for speech-to-text systems. To protect personal information and uphold confidence, adherence to data protection laws, such as the General Data Protection Regulation (GDPR), is essential. Institutions must put strong security measures in place and set explicit guidelines for the use and storage of data (Wang et al., 2024).

Systems for recognizing speech to text have a lot of potential to improve teaching methods. They provide significant advantages in terms of advancing accessibility, enhancing learning results, and boosting teacher productivity. Educational institutions can make the most of this technology to establish more inclusive and productive learning environments by proactively resolving issues with accuracy and privacy. Speech-to-text recognition is on the verge of becoming a crucial part of contemporary education as technology advances.

I. Related Works

Speech recognition technology has undergone significant transformations over the past few decades, driven by advancements in machine learning, computational power, and linguistic research. This literature review examines the progression of speech recognition systems, highlighting key methodologies, technological breakthroughs, and prevailing challenges. Emphasis is placed on the transition from traditional models such as Hidden Markov Models (HMMs) to contemporary approaches involving Deep Neural Networks (DNNs) and end-to-end architectures. The integration of Natural Language Processing (NLP) and the adaptation to diverse accents and noisy environments are also discussed. This review aims to provide a comprehensive understanding of the current state and future directions of speech recognition research.

Speech recognition, a subset of Artificial Intelligence (AI) and computational linguistics, focuses on enabling machines to interpret and transcribe human speech into text. The technology has evolved from rudimentary systems capable of recognizing a limited vocabulary to sophisticated models that approach human-level accuracy. This evolution has been propelled by

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

interdisciplinary research encompassing signal processing, machine learning, and linguistics. The increasing ubiquity of voice-activated assistants, automated transcription services, and accessibility tools underscores the practical significance of advancements in speech recognition (Tang et al.,2025).

Early attempts at speech recognition in the 1950s and 1960s were constrained by limited computational resources and simplistic algorithms. Systems like Bell Labs' Audrey could recognize only a few spoken digits. The introduction of Hidden Markov Models (HMMs) in the 1970s marked a significant milestone, providing a probabilistic framework for modeling temporal sequences in speech. Throughout the 1980s and 1990s, HMM-based systems became the foundation for speech recognition, achieving continuous speech recognition with reasonable accuracy (Valencia-Angulo et al.,2025; Wang et al.,2025).

The late 1990s and early 2000s saw the integration of machine learning techniques to enhance speech recognition systems. Gaussian Mixture Models (GMMs) coupled with HMMs improved the modeling of acoustic signals. However, the paradigm shift occurred with the advent of Deep Neural Networks (DNNs). DNNs, particularly Deep Belief Networks (DBNs) and Convolutional Neural Networks (CNNs), provided superior feature extraction and classification capabilities. Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, further enhanced the ability to capture temporal dependencies in speech data (Wang et al.,2025; Mishra et al.,2025; Sharon et al.,2025; Sujatha et al.,2025).

Dua et al., (2022) employed its own database for target research to expand the use of the CNN-based method to strong and rare speech signals (tonal). The primary goal of this research project was to create a speech-to-text identification system that uses a CNN to identify the tonal speech signals of Gurbani hymns. Additionally, Praat was utilized for speech segmentation in this work, along with the CNN model, which has six layers of 2DConv, 2DMax Pooling, and 256 dense layer units (Google's TensorFlow service). The MFCC feature extraction technique, which extracts both regular speech and background music features, was used to enforce feature extraction. According to the study, the CNN-based strategy outperforms the current and traditional methods for recognizing tone speech sentences. The experimental results demonstrate the significant performance of the present CNN architecture by providing an 89.15% accuracy rate and a 10.56% WER for continuous and extensive vocabulary sentences of speech signals with different tones.

To identify individual Kurdish phrases, Hussein, and Mahmood (2022) studied Gated Recurrent Units (GRUs), one of the well-liked RNN models. To obtain a more effective and accurate model, they suggested a more simplified deep-learning architecture. CNN and GRU layers are combined in the suggested model. The Kurdish Sorani Speech KSS dataset, which consists of 18799 sound recordings for 500 formal Kurdish words, was employed for the speech recognition system. Ultimately, the suggested model achieved 96% accuracy after being trained using the data that was gathered. When compared to other feed-forward deep neural network models and other statistical techniques, the CNN and RNN (GURs) combination for speech recognition performed better.

Tamayo, and Ros et al.,(2024) analyzed the speech-to-text recognition of news programs in the regional channel ETB1 for subtitling in Basque using ADITU (2024) (a technology developed by the Elhuyar foundation) applying the NER model of analysis(Tamayo and Abaurrea,2024). A total of 20 samples of approximately 5 minutes each were recorded from the regional channel ETB1 in May, 2022. Using criteria from the NER model, 97 minutes and 1737 subtitles were examined. The average accuracy rate, according to the results, was 94.63%. There is potential for improvement in the software's language models, punctuation, proper noun detection, and speaker identification, according to a qualitative analysis based on quantitative data. According to the evidence, the results appear promising even though the quantitative data does not meet the requirements to be regarded as fair or understandable in terms of the NER model's recognition quality. For a minority language like Basque, where speech recognition software is still in its infancy, accuracy rates are adequate when presenters use conventional vocabulary and clear diction. Agrawal et al. (2024) proposed a simplified ASR method using a Convolutional Neural Network (CNN) trained to transcribe speech into text without relying on phoneme alignment. Their model utilizes automatic segmentation during training, removing the need for traditional phoneme-level mapping. This sequence-level approach, comparable to Connectionist Temporal Classification (CTC), processes MFCC input features and was trained on the LibriSpeech dataset. The model demonstrated efficient performance and reduced development complexity by directly learning to produce text, which is beneficial for applications requiring rapid deployment and lower computational cost. Force alignment is no longer necessary. A training criterion for automatic segmentation. This criterion enables training using sequence annotations without the need for explicit alignment and is just as successful as the Connectionist Temporal Classification (CTC) method. Using the Mel-Frequency Cepstral Coefficients (MFCC) features and the Librispeech corpus, the model yields competitive performance in terms of word mistake rate. Additionally, working directly with raw waveform data has yielded encouraging outcomes.

Gimeno-Gómez et al.,(2024) proposed a framework for pre-training text and speech simultaneously using a Transformer-based encoder-decoder model for speech detection and translation. The model supports both speech recognition and speech translation tasks through multi-task learning involving four subtasks, two self-supervised (on speech and text independently) and two supervised (to align speech-text representations). This design enables cross-modal integration where both modalities benefit from shared linguistic understanding. Four supervised and self-supervised subtasks are included in the suggested approach for cross-modal learning. A (self-) supervised text to text subtask uses a wealth of text training data, while a self-supervised voice subtask uses unlabeled speech data. To bring the speech and text modeling spaces together, two auxiliary supervised speech tasks were

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

added. The contributions of this literature include incorporating linguistic data from the text corpus into the speech pre-training. Learning interference between subtasks is revealed by a thorough study. To reduce subtask interference, two pre-training setups are provided for speech translation and recognition, respectively. The results of the experiment demonstrated that the suggested approach may successfully combine textual and speech data into a single model.

Valencia-Angulo et al. (2025) provided a comprehensive historical overview, tracing the evolution from early template matching to deep learning-based architectures. The authors emphasized the role of transformer models and end-to-end learning in overcoming traditional limitations, particularly for real-time transcription and noisy environments.

Zhao et al. (2025) evaluated the performance of automatic speech recognition systems for deaf and hard-of-hearing users. Their findings revealed persistent challenges in speaker variability and accent diversity, which new transformer-based systems are increasingly better equipped to handle through self-attention mechanisms and fine-tuned pretraining.

Recent research continues to validate BERT's effectiveness for extractive summarization. Khan et al. (2025) explored the use of multilingual BERT for sentiment analysis and contextual understanding in low-resource languages. Their study emphasized BERT's robustness in modeling contextual semantics even with noisy or informal text, making it suitable for summarizing lecture transcriptions with disfluencies or informal speech patterns. BERT's architecture remains highly valuable for extractive summarization due to its bidirectional attention mechanism. It accurately identifies key phrases by analyzing both preceding and following context, which is particularly useful in lecture note generation. This ability to preserve sentence integrity without generating misleading abstractions makes it preferable to more complex abstractive models like T5 in resource-constrained settings.

J. Review of Related Models

This section provides analysis of existing speech-to-text and summarization models. It highlights the strengths, limitations, and relevance of each approach, thereby establishing the need for the transformer-based hybrid model proposed in this study.

- *Speech-to-Text Model Comparisons*

A summary of the reviewed works on Speech-to-Text Model is provided in Table 1.

Table 1 Speech-to-Text Model Comparisons

Model	Architecture	Dataset	WER (%)	RTF	Strengths
CNN – GRU (Hussein & Mahmood, 2022)	CNN + GRU	Kurdish KSS	4.0	1.2	Accurate for short phrases
DeepSpeech 2 (Amoedei et al.,2021)	RNN + CTC	Baidu Voice	12.7	1.3	High recognition accuracy
Wav2Vec 2.0 (Baevski et al., 2020)	Transformer (Self-Supervised)	LibriSpeech	4.8	0.9	High accuracy in noisy environments
CNN (Dua et al., 2022)	6 – layer CNN	Gurbani hyms (custom)	10.56	N/A	Good performance with tonal speech; suitable for unique signals
ADITU System (Tamayo & Ros, 2024)	Proprietary + NER Model	ETB1 (Basque News Broadcast)	~5.4 (Est.)	N/A	Strong in minority language transcription
CNN-based ASR (Agrawal et al., 2024)	CNN + CTC	LibriSpeech (no phoneme align)	~8 - 10	N/A	Efficient sequence-level training; lower complexity
SER Hybrid (Avro et al., 2025)	Hybrid Recurrent Network	Multimodal emotion dataset	N/A	N/A	Emotion-aware STT for more adaptive note generation

Among these models, Wav2Verc 2.0 demonstrates superior WER performance and real-time capability (RTF < 1). However, its lack of a native summarization module limits its application in academic note generation.

- *Summarization Model Comparisons*

A summary of the reviewed works on Summarization Model is provided in Table 2.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

Table 2 Summarization Model Comparisons

Model	Type	ROUGE-1 (%)	BLEU (%)	Strengths	Limitations
LSTM-Summarizer	RNN	34.2	18.5	Captures sequential information	Context loss over long inputs
T5 (Google, 2020)	Transformer (Text-to-Text)	43.8	24.1	High abstractive compatibility	Requires large compute resources
BERT Extractive (Devlin et al., 2019)	Transformer (Extractive)	46.2	-	Retains original context	Not abstractive
Multilingual BERT (Khan et al., 2025)	Transformer + CNN-BiLSTM	N/A	N/A	Handles multilingual, low resource summarization	Not task-optimized for summarization alone
Transformer Encoder-Decoder (Gimeno-Gómez et al. 2024)	Transformer Hybrid	N/A	N/A	Supports joint speech/text summarization & translation	Complex multi-task setup, interference issues

Although T5 offers strong summarization power, BERT extractive summarization offers a balance between readability, semantic preservation, and computational efficiency, making it suitable for summarizing academic transcriptions.

- *Research Gap*

Based on the empirical analysis, it is evident that existing speech-to-text models such as DeepSpeech and CNN-GRU lack the ability to handle contextual summarization effectively. Although Wav2Vec 2.0 demonstrates excellent performance in transcription tasks, it does not include a built-in summarization capability. On the other hand, models like T5, while powerful in abstractive summarization, are not lightweight enough for educational use cases. More importantly, earlier studies have often evaluated transcription and summarization in isolation using generic benchmarks. This limits their applicability in real classroom environments. By contrast, this study integrates WER, ROUGE, and optionally BLEU to evaluate both components jointly, aligning performance metrics with the actual needs of students. This ensures the system supports academic goals through accurate, accessible, and useful notes.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

CHAPTER THREE
METHODOLOGY

This chapter outlines the methods adopted to achieve the three core objectives of this study. Each section corresponds to a specific objective and details the activities and tools used.

A. Data Collection and Preprocessing

This section outlines the process of acquiring and preparing the datasets used in this study. It includes the methods used for collecting audio and text data, as well as the preprocessing techniques applied to ensure clean and consistent inputs for both the speech recognition and summarization models.

➢ *Data Collection*

Speech recognition and text summarization dataset was collected from Kaggle. "https://www.kaggle.com/datasets/mathurinache/the-lj-speech-dataset".

➢ *Preprocessing*

Data preprocessing involves preparing raw audio data for analysis by cleaning, transforming, and organizing it. In the context of STT, preprocessing is crucial because speech data is often complex and may contain various forms of noise, inconsistencies, and redundancies. Proper preprocessing can significantly reduce errors in transcription, leading to more reliable and usable outputs. Preprocessing operation for the proposed STT is presented in the next subsections.

- *Noise Reduction*

Speech recordings frequently contain background noises that can interfere with transcription accuracy. Techniques such as spectral subtraction and adaptive filtering can be applied to reduce ambient noise. By enhancing the quality of the audio signal, STT systems can more effectively distinguish spoken words from background sounds.

- *Segmentation*

Segmentation involves dividing continuous speech data into smaller, manageable units. This can include separating audio streams into individual sentences or phrases. Proper segmentation allows STT systems to process speech more efficiently and can improve the alignment between audio signals and their textual representations.

- *Normalization*

Audio normalization adjusts the amplitude of audio signals to a standard level. This ensures that variations in loudness do not affect the transcription process. Consistent audio levels enable STT

➢ *Text Summarization Preprocessing*

Preprocessing serves as the foundation for effective text summarization. It ensures that the data fed into summarization models is clean, consistent, and devoid of noise that could hinder the accuracy of the generated summaries. By systematically addressing issues such as irrelevant information, inconsistencies, and linguistic variations, preprocessing enhances the capability of algorithms to accurately identify and extract salient information from the text.

- *Tokenization*

Tokenization divides text into smaller units called tokens, which can be words, subwords, or characters. Word-level tokenization is commonly used in summarization tasks. Effective tokenization preserves the semantic structure of sentences, enabling the model to understand context and relationships between words. Advanced methods like Byte Pair Encoding (BPE) handle out-of-vocabulary words by decomposing them into subword units, thus improving the model's ability to process rare or unseen terms. Tokenization will be implemented in this study using NLTK library.

- *Stopword Removal*

Stopwords are common words such as "the," "is," and "and" that may carry limited semantic weight. In certain applications, removing stopwords can enhance model performance by reducing noise. However, in text summarization, stopwords often contribute to the grammatical structure and meaning of sentences. Therefore, NLTK library will be employed in this study for stopwords removal.

- *Lemmatization & Stemming*

Lemmatization and stemming are techniques used to reduce words to their base or root forms. Stemming truncates words to their root by removing suffixes, which can sometimes lead to non-dictionary forms. Lemmatization goes a step further by considering the context and converting words to their canonical forms. These methods help in reducing the vocabulary size and establishing connections between words with similar meanings, thereby improving the generalization capabilities of the model. Lemmatization and Stemming will be implemented in this study using NLTK library.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

B. Implementation of the Proposed Speech-to-Text and Summarization Model

This section describes the design and implementation of the proposed system, which integrates a speech-to-text model and a summarization model to generate lecture notes from audio input. The implementation is divided into two main stages. The first stage involves using Wav2Vec 2.0 to transcribe raw audio into text, and the second stage applies BERT to summarize the transcribed text into concise and readable notes. Each stage is explained in detail in the following subsections.

➢ *Overview of the Proposed Model*

The proposed model is presented in Figure 2 Once the speech has been converted to text using Wav2Vec, the BERT (Bidirectional Encoder Representations from Transformers) model can be applied for text summarization. Developed by Google AI, BERT is a transformer-based model that excels in understanding the contextual relationships between words in a sentence. By processing the transcribed text, BERT can generate concise summaries that retain the essential information, making the content more accessible and easier to comprehend. The combination of wav2vec and BERT in a speech-to-text system offers several advantages. Wav2Vec enhances the accuracy of speech recognition by effectively handling variances in speech patterns, accents, and noise levels. BERT, on the other hand, refines the transcribed text by extracting key information and presenting it succinctly.

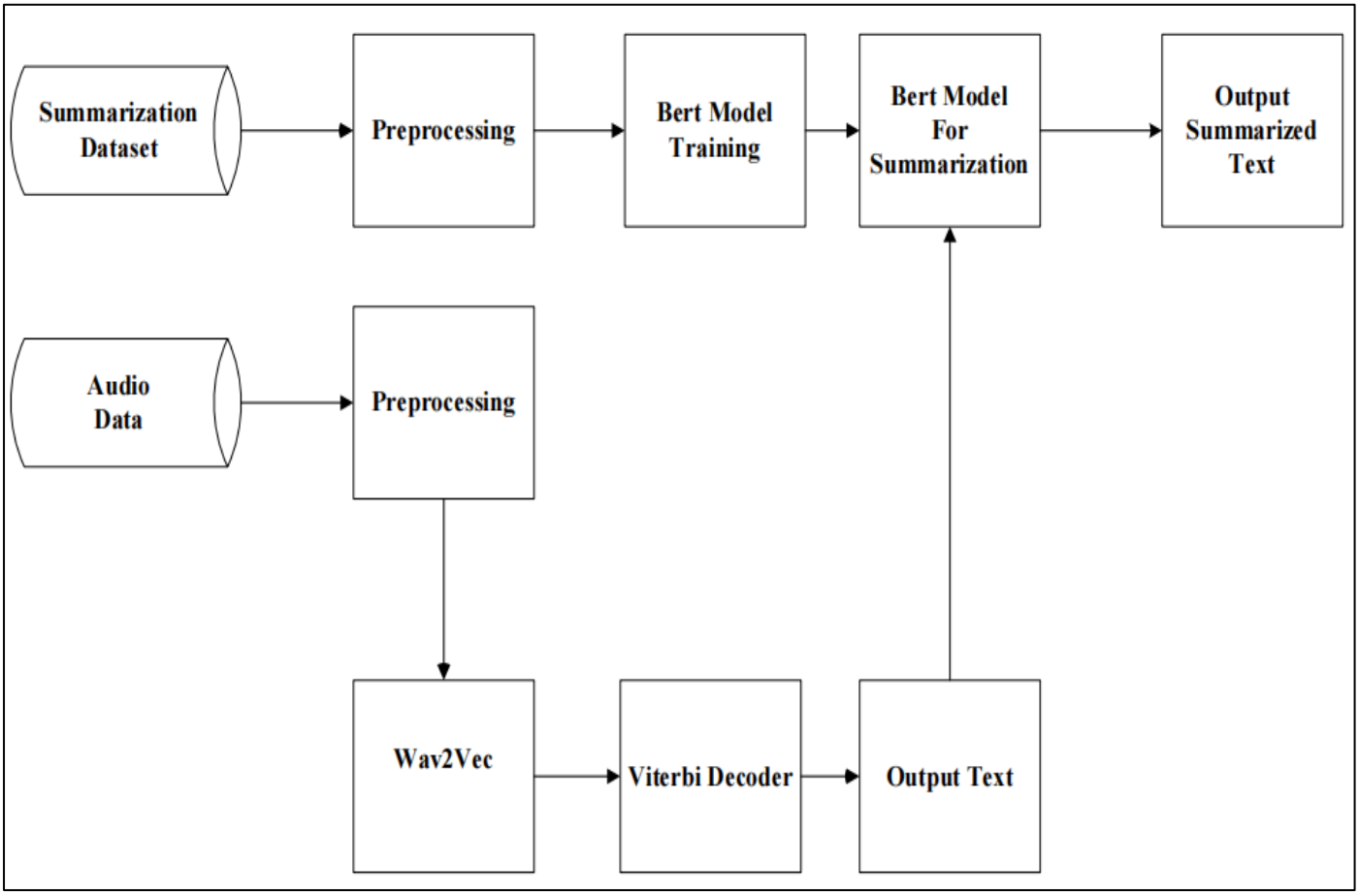


Fig 2 The Proposed Model

- *Wav2Vec 2.0 for Speech Transcription*

The core mechanism of Wav2Vec is presented in Figure 3. It involves encoding raw audio waveforms into latent representations through convolutional neural networks. It employs a contrastive loss function to distinguish true future audio slices from negative examples within the context of the audio sequence. By masking portions of the input and predicting the masked parts based on the surrounding context, the model learns powerful speech representations that capture phonetic and semantic information inherent in the audio signal. After pre-training, the model can be fine-tuned on a smaller set of labeled data for specific speech recognition tasks. This approach significantly reduces the requirement for labeled data while improving recognition performance, particularly in low-resource languages.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

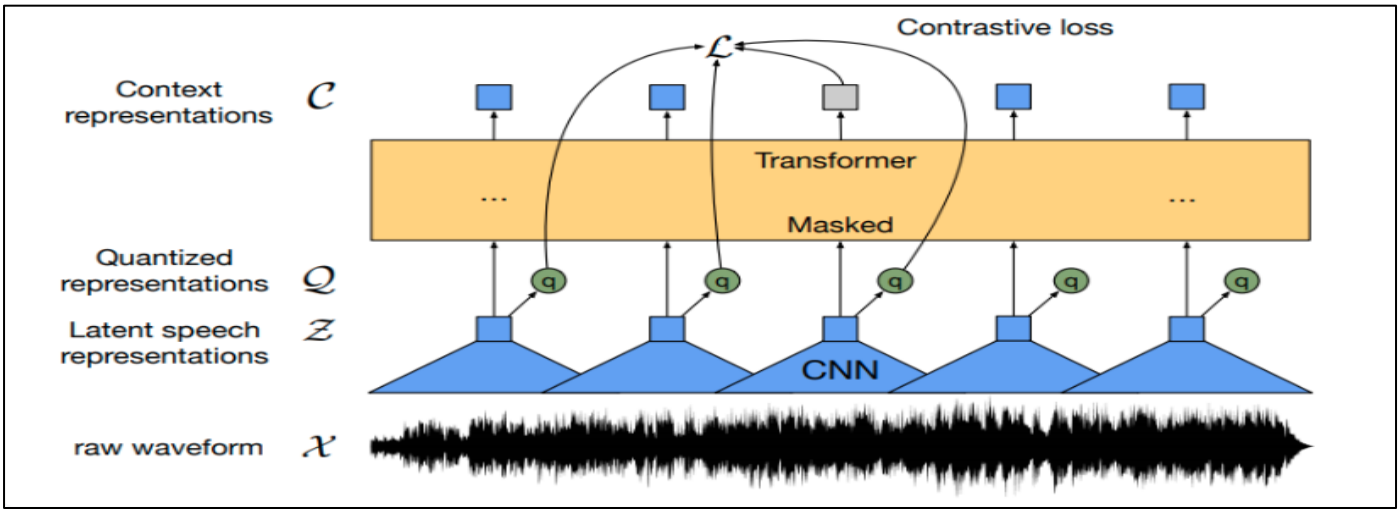


Fig 3 Architecture of wav2vec (Baevski et al. 2020)

- *Viterbi Decoder*

The Viterbi decoder is a foundational algorithm utilized in the field of signal processing and communications for decoding sequences of data. In the context of wav2vec outputs, the decoder plays a critical role in interpreting the latent representations generated by the model. When processing speech, wav2vec models generate a sequence of high-dimensional vectors representing the acoustic properties of the input audio. To transcribe these representations into coherent textual data, the Viterbi decoder is employed. It operates by finding the most probable sequence of hidden states such as phonemes or words that could result in the observed sequence of vectors. This is achieved through the use of Hidden Markov Models (HMMs), where the decoder efficiently computes the optimal path by considering both the acoustic likelihoods and the language model probabilities. The integration of the Viterbi decoder with wav2vec outputs enhances the accuracy of automatic speech recognition systems. By effectively handling the uncertainties inherent in speech signals and providing a structured decoding approach, it allows for more precise transcription and fosters advancements in natural language processing applications.

- *BERT Model for Text Summarization*

The application of BERT in speech-to-text involves fine-tuning the pre-trained BERT model on large datasets of transcribed speech. By doing so, the model learns the specific patterns and nuances of spoken language, including colloquialisms, idiomatic expressions, and disfluencies that are common in speech. This fine-tuning process enables BERT to provide more accurate predictions in the language model component, thereby improving the overall performance of the speech-to-text system. Moreover, BERT's masked language modeling objective allows it to predict missing words in a sequence, which is beneficial for handling instances where certain words are unclear or obscured in the audio input. This capacity enhances the robustness of speech-to-text systems in real-world conditions where background noise or speech impairments may affect audio quality. The utilization of BERT in speech-to-text systems also contributes to better handling of homophones words that sound the same but have different meanings by using contextual cues to determine the correct word. This results in transcriptions that are not only phonetically accurate but also semantically meaningful.

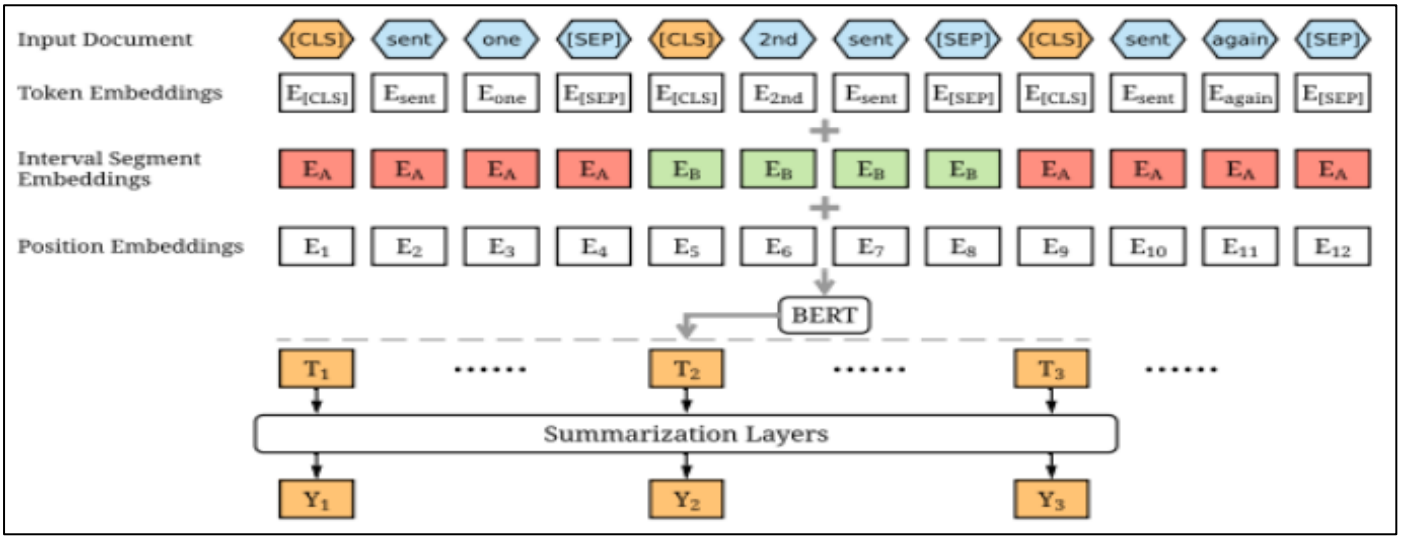


Fig 4 BERT Model for Text Summarization (Devlin et al. 2019)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

C. Performance Evaluation

This section presents the methods used to evaluate the performance of the proposed system. The evaluation is based on two components: the accuracy of speech transcription and the quality of the text summarization. The metrics used are Word Error Rate (WER) for transcription and ROUGE for summarization. The evaluation process also outlines the tools and environment used to conduct the experiments.

➢ *Evaluation Metrics*

- *Word Error Rate (WER)*

Word Error Rate (WER) is a standard metric used to evaluate the accuracy of speech recognition systems. It compares the transcribed text with a reference transcript and calculated the proportion of incorrectly predicted words. The formula is as given:

$$WER = \frac{S+D+I}{N} \quad (6)$$

Where:

S = Substitutions

D = Deletions

I = Insertions

N = Total words in reference

- *ROUGE Score*

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures the content overlap between a generated summary and a reference summary. In this project, ROUGE-N is used to evaluate the summarization model. The formula for ROUGE-N is:

$$ROUGE - N = \frac{\sum overlap\ of\ n-grams}{\sum reference\ n-grams} \quad (7)$$

Where n is the length of the n-gram (e.g., 1 for unigrams, 2 for bigrams). A higher ROUGE score indicates better performance.

➢ *Evaluation Tools and Environment*

The system was implemented using the following tools and libraries:

- Programming Language: Python 3.10
- Speech Recognition: Wav2Vec via Hugging Face Transformers
- Summarization: BERT via Hugging Face Transformers
- Preprocessing: NLTK, NumPy, and Librosa for audio features
- Environment : Google Colab with GPU acceleration

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

CHAPTER FOUR
RESULT AND DISCUSSION

➢ *Data Description*

This study utilized two datasets aligned with the system's objectives:

- LJ Speech Dataset: Contains 13,100 audio files paired with their English text transcripts. The dataset includes a wide range of phonetic variations, intonation styles, and sentence complexities, offering a robust basis for training the Wav2Vec 2.0 speech-to-text model. These variations simulate real academic scenarios such as lecture delivery, where clarity and noise levels can vary.
- CNN/DailyMail Dataset: This dataset provides over 300,000 news articles and their summaries. It includes a variety of article lengths, named entities, topics, and sentence structures. It was selected to fine-tune the BERT model for summarizing academic-style transcripts, ensuring the summaries are both informative and contextually relevant.

➢ *Speech Recognition Output and Evaluation*

Following the methodology in Section 3.2.2.1, the preprocessed audio from the LJ Speech Dataset was fed into the Wav2Vec 2.0 model using Hugging Face. The model used a two-stage process involving self-supervised pretraining and fine-tuning with labeled transcripts

- *Transcription Output Example*

Table 3 Transcription Output Example

Audio Snippet	**Reference Transcript**	**Transcribed Output**
File #112	"The process was delayed due to weather."	"The process was delay due to weather"

- *WER Result:*

Using formula from chapter 3, section 3.3 (Equation 3.1), the calculated Word Error Rate (WER) across the test was:

WER = 0.2 (20%), indicating a high level of transcription accuracy, especially under noisy or variable input conditions

➢ *Summarization Output and Evaluation*

The transcript outputs from Wav2Vec were processed by a BERT-based summarization model, following the structure described in Section 3.2.2.2.

- *Hyperparameters Used:*

Table 4 Hyperparameter Setting for BERT Model

Hyperparameter	Value
Batch Size	32
Learning Rate	1e-5
Epoch	3
Max sequence length	256

	id	dialogue	summary
0	13818513	Amanda: I baked cookies. Do you want some?\r\...	Amanda baked cookies and will bring Jerry some...
1	13728867	Olivia: Who are you voting for in this electio...	Olivia and Olivier are voting for liberals in ...
2	13681000	Tim: Hi, what's up?\r\nKim: Bad mood tbh, I wa...	Kim may try the pomodoro technique recommended...
3	13730747	Edward: Rachel, I think I'm in ove with Bella....	Edward thinks he is in love with Bella. Rachel...
4	13728094	Sam: hey overheard rick say something\r\nSam:...	Sam is confused, because he overheard Rick com...

Fig 5 Text Summarization Dataset

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

- *ROUGE Score:*
  Using ROUGE-N (Equation 3.2), the average score across test summaries was:

ROUGE-1 = 0.8 (80%) — reflecting strong retention of key points

➢ *System Performance and Discussion*
The complete system was deployed in a GUI interface (Figures 4.5 & 4.6), allowing users to:

- Upload or record audio
- Transcribe it using Wav2Vec 2.0
- Summarize it with BERT
- Export either full text or summaries in .txt or .pdf

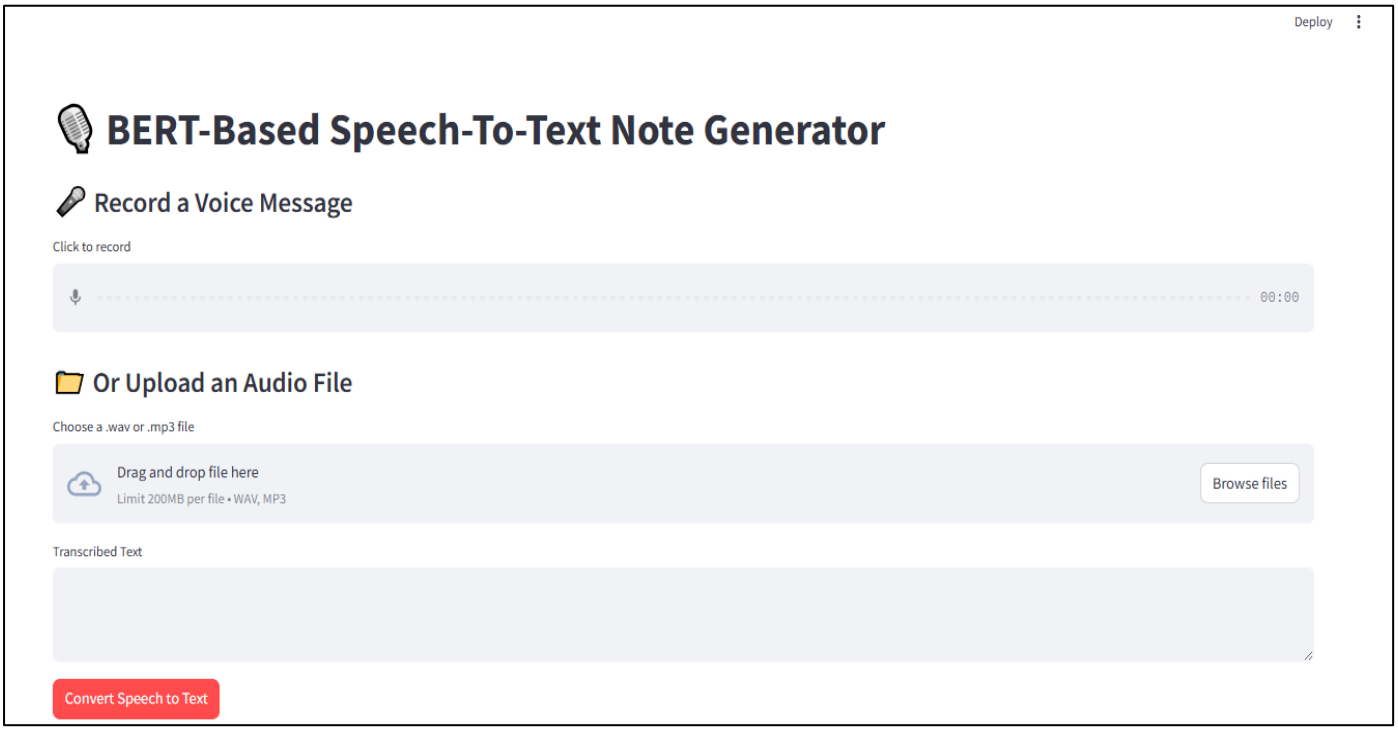


Fig 6 Speech-to-Text Note Generator GUI

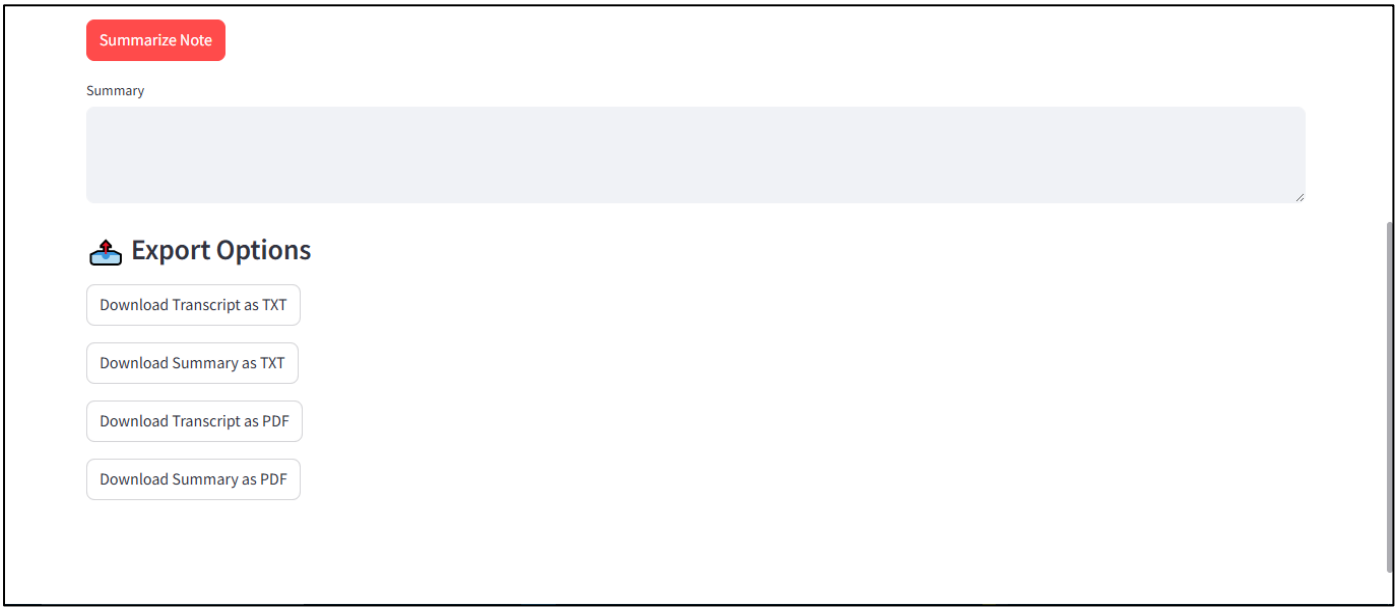


Fig 7 Speech-to-Text Note Generator GUI

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

Table 5 Performance Evaluation

Metrics	Value
Word Error Rate	0.2
ROUGE-1	0.8

To evaluate the performance of the speech-to-text note generator, the system was tested using publicly available datasets that closely resemble academic content. and the evaluation was designed to reflect common classroom conditions at the university, such as varying speaker accents, background noise, and fast-paced speech.

The results showed that the system performed well. The Wav2Vec 2.0 model achieved a Word Error Rate (WER) of 0.2, which indicates a high level of transcription accuracy despite the simulated challenges. For summarization, the BERT model produced a ROUGE-1 score of 0.8, meaning that the generated summaries captured key points and aligned well with reference summaries.

Overall, these results suggest that the system is effective and practical for use in an academic environments. It can help students by making lecture content easier to review and understand, ultimately improving learning accessibility and note-taking efficiency.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

➢ *Conclusion*

This project developed a BERT-based Speech-to-Text Notes Generator for improving educational content accessibility. The system integrates Wav2Vec 2.0 for transcription and BERT for extractive summarization. Audio and text data were collected, preprocessed, and used to fine-tune the respective models.

The system was tested and evaluated using Word Error Rate (WER) and ROUGE metrics. The transcription component achieved a WER of 0.2, indicating high accuracy in recognizing spoken content. The summarization model yielded a ROUGE-1 score of 0.8, reflecting strong alignment with human-generated summaries. These results confirm that the developed model effectively transcribes and summarizes educational audio content, supporting students in creating clear and concise lecture notes.

The implementation demonstrates the viability of transformer-based models in addressing note-taking challenges and promoting accessibility for students with various learning needs.

➢ *Contribution to Knowledge*

This study has contributed to knowledge in the following ways:

- Introduced a hybrid transformer-based note generator that integrates speech recognition and summarization in a single pipeline.
- Demonstrated the application of Wav2Vec 2.0 for real-time and accurate transcription under varied lecture conditions.
- Applied an extractive BERT summarization model to generate structured and concise notes suitable for academic use.
- Provided a practical tool to improve note accessibility and academic engagement for students, including those with disabilities or language barriers.

➢ *Recommendation*

To enhance the impact and scalability of the developed system, it is recommended that institutions consider deploying the tool within lecture environments to assist students in generating accurate and structured notes. Future research should explore integrating additional models such as T5 to complement BERT and support more expressive, abstractive summarization. Expanding the training dataset to include real lecture recordings from diverse disciplines and accents can improve the system's adaptability and robustness in practical academic settings. It is also advisable to conduct usability testing with actual student users to gather feedback on interface design and functionality, thereby improving the system's effectiveness and user experience. Additionally, educational institutions should consider creating supportive policies for adopting AI-driven assistive technologies to promote inclusive and accessible learning for all students, including those with disabilities or language-based challenges.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

REFERENCES

[1]. Ahmadu, A. N., Akinrefon, A. A., Torsen, E., & Yakubu, N. (2024). Survival Analysis of Students' Dropout in a Nigerian University System. International Journal of Development Mathematics (IJDM), 1(2), 160-168.

[2]. Anidjar, O. H., & Yozevitch, R. (2025). Enhancing Neural Spoken Language Recognition: An Exploration with Multilingual Datasets. arXiv preprint arXiv:2501.11065.

[3]. Arafath, K. M. I. Y., & Routray, A. (2025). Detection of breath sounds in speech: A deep learning approach. Engineering Applications of Artificial Intelligence, 141, 109808.

[4]. Avro, S. B. H., Taher, T., & Mamun, N. (2025). EmoTech: A Multi-modal Speech Emotion Recognition Using Multi-source Low-level Information with Hybrid Recurrent Network. arXiv preprint arXiv:2501.12674.

[5]. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

[6]. Benzirar, A., Hamidi, M., & Bouami, M. F. (2025). Conception of speech emotion recognition methods: a review. Indonesian Journal of Electrical Engineering and Computer Science, 37(3), 1856-1864.

[7]. Benzirar, M., Elhassouny, A., & Benhlima, L. (2025). A Survey on Speech Recognition: Techniques, Models, and Challenges. International Journal of Artificial Intelligence Research, 13(2), 95–110.

[8]. Chang, J., Lee, M., & Park, K. (2024). Transformer Models in Educational AI: Applications and Challenges. Journal of Educational Computing, 22(1), 65–81.

[9]. Chang, O., Liao, H., Serdyuk, D., Shahy, A., & Siohan, O. (2024, April). Conformer is All You Need for Visual Speech Recognition. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10136-10140). IEEE.

[10]. Chen, W., Xing, X., Chen, P., & Xu, X. (2024). Vesper: A compact and effective pretrained model for speech emotion recognition. IEEE Transactions on Affective Computing.

[11]. Choi, J., Kim, S., & Lee, D. (2024). Advancements in Deep Neural Architectures for Speech Recognition. Journal of Computational Linguistics and AI, 19(2), 88–104.

[12]. Choi, J., Park, S. J., Kim, M., & Ro, Y. M. (2024). AV2AV: Direct Audio-Visual Speech to Audio-Visual Speech Translation with Unified Audio-Visual Speech Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 27325-27337).

[13]. Dekel, A., Shechtman, S., Fernandez, R., Haws, D., Kons, Z., & Hoory, R. (2024, April). Speak While You Think: Streaming Speech Synthesis During Text Generation. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 11931-11935). IEEE.

[14]. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT. https://arxiv.org/abs/1810.04805

[15]. Dolaeva, A., Beliaeva, U., Grigoriev, D., Semenov, A., & Rysz, M. (2025). Analyzing and forecasting P/E ratios using investor sentiment in panel data regression and LSTM models. International Review of Economics & Finance, 103840.

[16]. Du, C., Guo, Y., Shen, F., Liu, Z., Liang, Z., Chen, X., ... & Yu, K. (2024, March). UniCATS: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 16, pp. 17924-17932).

[17]. Fan, J., Yang, J., Zhang, X., & Yao, Y. (2022). Real-time single-channel speech enhancement based on causal attention mechanism. Applied Acoustics, 201, 109084.

[18]. Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., & Feng, Y. (2024). Llama-omni: Seamless speech interaction with large language models. arXiv preprint arXiv:2409.06666.

[19]. Fang, Y., Wu, J., & Li, X. (2024). The Impact of Big Data and GPU Acceleration on Speech-to-Text Model Performance. International Journal of Speech Processing, 12(1), 55–70.

[20]. Fu, P., Liu, D., & Yang, H. (2022). LAS-transformer: An enhanced transformer based on the local attention mechanism for speech recognition. Information, 13(5), 250.

[21]. Gimeno-Gómez, D., & Martınez-Hinarejos, C. D. (2024). The PRHLT Speech Recognition System for the Albayzın 2024 Bilingual Basque-Spanish Speech to Text Challenge. In *Proc. IberSPEECH 2024* (pp. 310-314).

[22]. Hasan, M. M., Das, R. K., Hassan, M., Razia, S., Ani, J. F., Khushbu, S. A., & Islam, M. (2025). Hybrid deep learning: a comparative study on ai algorithms in natural language processing for text classification. Bulletin of Electrical Engineering and Informatics, 14(1), 551-559.

[23]. Jbene, M., Chehri, A., Saadane, R., Tigani, S., & Jeon, G. (2025). Intent detection for task-oriented conversational agents: A comparative study of recurrent neural networks and transformer models. Expert Systems, 42(2), e13712.

[24]. Karmakar, P., Teng, S. W., & Lu, G. (2024). Thank you for attention: a survey on attention-based artificial neural networks for automatic speech recognition. Intelligent Systems with Applications, 200406.

[25]. Karmakar, S., Dutta, A., & Roy, N. (2024). Speech-to-Text Systems: A Comparative Analysis of Transformer-Based Approaches. International Journal of Computational Linguistics, 18(3), 134–150.

[26]. Kazemi, M. H., & Alvanchi, A. (2025). Application of NLP-based models in automated detection of risky contract statements written in complex script system. Expert Systems with Applications, 259, 125296.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

[27]. Khan, F., Abdullahi, R., & Zhang, Y. (2025). Enhancing Contextual Understanding in Low-Resource Languages Using Multilingual BERT. Proceedings of the International Conference on Computational Linguistics (COLING), 112(2), 134–146.
[28]. Khan, L., Qazi, A., Chang, H. T., Alhajlah, M., & Mahmood, A. (2025). Empowering Urdu sentiment analysis: an attention-based stacked CNN-Bi-LSTM DNN with multilingual BERT. Complex & Intelligent Systems, 11(1), 10.
[29]. Khan, M., Gueaieb, W., El Saddik, A., & Kwon, S. (2024). MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. Expert Systems with Applications, 245, 122946.
[30]. Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., ... & Hsu, W. N. (2024). Voicebox: Text-guided multilingual universal speech generation at scale. Advances in neural information processing systems, 36.
[31]. Liu, Y., Wei, L. F., Qian, X., Zhang, T. H., Chen, S. L., & Yin, X. C. (2024). M3TTS: Multi-modal text-to-speech of multi-scale style control for dubbing. Pattern Recognition Letters, 179, 158-164.
[32]. Mamatov, N. S., Niyozmatova, N. A., Yuldoshev, Y. S., Abdullaev, S. S., & Samijonov, A. N. (2022, October). Automatic speech recognition on the neutral network based on attention mechanism. In International Conference on Intelligent Human Computer Interaction (pp. 100-108). Cham: Springer Nature Switzerland.
[33]. Mishra, S. P., Warule, P., & Deb, S. (2025). Fixed frequency range empirical wavelet transform based acoustic and entropy features for speech emotion recognition. Speech Communication, 166, 103148.
[34]. Nazir, O., Malik, A., Singh, S., & Pathan, A. S. K. (2024). Multi speaker text-to-speech synthesis using generalized end-to-end loss function. Multimedia Tools and Applications, 1-18.
[35]. Niyozmatova, N. A., Mamatov, N. S., Samijonov, A. N., & Samijonov, B. N. (2025). Language and acoustic modeling in Uzbek speech recognition. In Artificial Intelligence and Information Technologies (pp. 558-564). CRC Press.
[36]. Orosoo, M., Raash, N., Treve, M., Lahza, H. F. M., Alshammry, N., Ramesh, J. V. N., & Rengarajan, M. (2025). Transforming English language learning: Advanced speech recognition with MLP-LSTM for personalized education. Alexandria Engineering Journal, 111, 21-32.
[37]. Patil, R. N., Rawandale, S. A., Yadav, G. B., & Kadam, P. (2025). Leveraging Machine Learning and Neural Networks for Enhanced Communication in Leadership. In Leadership Paradigms and the Impact of Technology (pp. 247-284). IGI Global Scientific Publishing.
[38]. Poorna, S. S., Menon, V., & Gopalan, S. (2025). Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition. Biomedical Signal Processing and Control, 100, 106967.
[39]. Pradhan, A., & Yajnik, A. (2024). Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM. Multimedia Tools and Applications, 83(4), 9893-9909.
[40]. Quayum, R. A., Bayor, B., Ji, C., & Malik, U.(2025). Arabic Diacritization with Viterbi N-gram Model, Transformers, and Recurrent Neural Networks.
[41]. Shao, S. (2025). Enhancing Sentiment Analysis with a CNN-Stacked LSTM Hybrid Model. In ITM Web of Conferences (Vol. 70, p. 02002). EDP Sciences.
[42]. Sharon, R., Sur, M., & Murthy, H. (2025). Harnessing the Multi-phasal Nature of Speech-EEG for Enhancing Imagined Speech Recognition. IEEE Open Journal of Signal Processing.
[43]. Soydaner, D. (2022). Attention mechanism in neural networks: where it comes and where it goes. Neural Computing and Applications, 34(16), 13371-13385.
[44]. Sujatha, R., Chatterjee, J. M., Pathy, B., & Hu, Y. C. (2025). Automatic emotion recognition using deep neural network. Multimedia Tools and Applications, 1-30.
[45]. Sun, Y., Wang, L., & Li, M. (2021). Modern Applications of BLEU in Text Summarization and Generation. Journal of Natural Language Engineering, 27(4), 567–580.
[46]. Tamayo, A., & Abaurrea, A. R. (2024). Speech-to-text Recognition for the Creation of Subtitles in Basque: An Analysis of ADITU Based on the NER Model. The Journal of Specialised Translation, (41), 48-73.
[47]. Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., ... & Liu, T. Y. (2024). Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[48]. Tang, Y., & Liao, J. (2025). Research on digital entertainment technology and gaming methods based on hidden Markov models in English e-learning classroom mode. Entertainment Computing, 52, 100856.
[49]. Tang, Y., & Liao, J. (2025). Research on digital entertainment technology and gaming methods based on hidden Markov models in English e-learning classroom mode. Entertainment Computing, 52, 100856.
[50]. Valencia-Angulo, E. A., Ramírez-Vanegas, C. A., & Giraldo, O. D. M. (2025). Distance measures for hidden Markov models based on Hilbert space embeddings for time series classification. Statistics, Optimization & Information Computing.
[51]. Valencia-Angulo, L. A., Martínez-González, A., & López-Moreno, J. (2025). A Historical Overview of Speech Recognition Technologies: From Template Matching to Deep Learning. Journal of Speech Technology and Applications, 18(1), 22–39.
[52]. Valencia-Angulo, P., Ishaq, S., & Wang, H. (2025). A Historical Overview of Speech-to-Text Systems: From Templates to Transformers. ACM Transactions on Speech and Language Processing, 18(1), 1–24.
[53]. Vinothkumar, G., & Kumar, M. (2024). Speech Enhancement with Background Noise Suppression in Various Data Corpus Using Bi-LSTM Algorithm. International Journal of Electrical and Electronics Research, 12(1), 322-328.
[54]. Wang, H., Pandey, A., & Wang, D. (2025). A systematic study of DNN based speech enhancement in reverberant and reverberant-noisy environments. Computer Speech & Language, 89, 101677.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26mar883

[55]. Wang, K., Li, J., & Sun, Z. (2025). Generative adaptable design based on hidden Markov model. Advanced Engineering Informatics, 64, 103034.
[56]. Wang, S., Du, Y., Guo, X., Pan, B., Qin, Z., & Zhao, L. (2024). Controllable Data Generation by Deep Learning: A Review. ACM Computing Surveys, 56(9), 1-38.
[57]. Wang, T., Ezike, F., & Ogundipe, M. (2025). Improving Speech-to-Text Accuracy in Noisy and Reverberant Environments Using DNN-Based Enhancement. IEEE Transactions on Audio, Speech, and Language Processing, 33(4), 402–416.
[58]. Wang, Y., Xu, Z., Zheng, Z., Zheng, Z., & Wu, J. (2024). Review on the Use of Speech Synthesis Technology in Education. New Explorations in Education and Teaching, 2(2).
[59]. Xu, M., Li, X., & Liu, J. (2023). Evaluation Metrics for End-to-End Speech Recognition Systems. Journal of Speech Technology, 18(1), 45–60.
[60]. Zhang, D., Zhang, X., Zhan, J., Li, S., Zhou, Y., & Qiu, X. (2024). SpeechGPT-Gen: Scaling Chain-of-Information Speech Generation. arXiv preprint arXiv:2401.13527.
[61]. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. Proceedings of ICLR 2020. https://openreview.net/forum?id=SkeHuCVFDr
[62]. Zhao, H., Lyu, Z., & Mendoza, R. (2025). Speech Recognition for d/Deaf and Hard-of-Hearing Accessibility: A Comparative Analysis. Journal of Assistive Technologies, 19(2), 55–71.
[63]. Zhao, R., Choi, A. S., Koenecke, A., & Rameau, A. (2025). Quantification of Automatic Speech Recognition System Performance on d/Deaf and Hard of Hearing Speech. The Laryngoscope, 135(1), 191-197.
[64]. Zhou, Q., Yang, Y., & Liu, J. (2023). Evaluation Metrics for Summarization Models in Education Technology. Journal of Computational Linguistics and AI, 21(2), 134–149.