Speech Recongition on Online session and Document Generation

Mohammad Furkan Munavalli
Dept. of Computer Science and Engg
KLE Technological University, Belgaum, India
02fe23bcs403@kletech.ac.in

Shreenandan Murari
Dept. of Computer Science and Engg
KLE Technological University, Belgaum, India
02fe23bcs414@kletech.ac.in

Mahammada Saqeed Noorashanavar
Dept. of Computer Science and Engg
KLE Technological University, Belgaum, India
02fe23bcs417@kletech.ac.in

Manuja Shetti
Dept. of Computer Science and Engg
KLE Technological University, Belgaum, India
02fe23bcs422@kletech.ac.in

Savita Bagewadi
Dept. of Computer Science and Eng
KLE Technological University, Belgaum, India
savitabagewadi.mss@kletech.ac.in

***Abstract*—The growing reliance on virtual communication systems has changed how individuals conduct meetings, lectures, and collaborative work. However, these online conferences frequently lack automatic mechanisms for recording and summarizing key talks, resulting in the loss of vital information. To address this issue, we describe an intelligent system capable of automatically detecting when an online session begins, recording real-time audio, converting speech to text, and producing succinct, readable summaries. This system leverages powerful Automatic Speech Recognition (ASR) technology, notably the OpenAI Whisper model, to ensure correct transcription.T5, BART, and GPT are sophisticated transformer-based models used for summarization and natural language synthesis. It also addresses common issues like as real-time processing, speaker recognition, and loud surroundings, resulting in consistent performance across a wide range of session types. The system is designed to run with minimum user interaction, increasing productivity and accessibility in educational and professional environments. Test findings demonstrate that it can generate high-quality summaries from live sessions, making it a useful tool for collecting knowledge while decreasing the load of human note-taking.**

***Index Terms*—Automatic Speech Recognition (ASR),Bidirectional and Auto-Regressive Transformers(BART),Bidirectional Encoder Representations from Transformers(BERT) ,GPT**

I. Introduction

The increasing usage of virtual platforms like Google Meet, Microsoft Teams, and Zoom has revolutionized communication in educational, corporate, and social contexts. However, these platforms frequently lack built-in features for automatically transforming spoken input into organized, searchable text. As a result, critical information presented during meetings or lectures is frequently lost due to insufficient or inconsistent manual note-taking.

This work describes a fully automated system for capturing, transcribing, and summarizing online sessions with little human intervention. The technology starts by identifying the beginning of a virtual meeting and recording the live audio stream. Transcription is done in real time with OpenAI's Whisper model, which is noted for its multilingual support and robustness in noisy conditions. The system uses the BART model for summarization, which is a sophisticated transformer architecture designed to generate high-quality abstractive prose. BART generates succinct, context-aware summaries that retain the core content and flow of the original discussion.

The end result is a well-structured document that may be reviewed, archived, or shared. This solution is notably useful in educational settings for students who miss classes or require easily available resources, as well as in companies where efficient, accurate documentation boosts productivity. With its modular design and platform-agnostic approach, the system provides a scalable and practical solution for turning unstructured speech into meaningful written material in today's virtual communication environment.

II. Literature Survey

Martin Radfar and Rohit Barnwal *et al.* [1] developed a real-time speech recognition system especially for Google Meet. Their solution employs deep learning-based automatic voice recognition models such as Wav2Vec 2.0 and Whisper, which

Author(s)	Objective	Model(s)	Methodology	Advantages	Disadvantages	Results
Martin Radfar *et al.*	Real-time speech recognition for Google Meet	Wav2Vec 2.0, Whisper	Uses deep learning-based ASR with NLP techniques to improve transcription accuracy	High accuracy; supports multiple languages	Struggles with overlapping speech, accents; computationally intensive	Improved transcription accuracy over traditional models
Anmol Gulati *et al.*	Enhance live meeting transcription with context	Transformer-based ASR	Trained on large-scale datasets with fine-tuning for meeting scenarios	Context-aware transcription; better speaker identification	Requires large labeled datasets; high processing time	higher word accuracy in benchmark tests
Ayushi Trivedi *et al.*	explores speech-to-text with deep learning	RNN, HMM,	sequential modeling of phonemes for speech recognition	Efficient for small datasets; interpretable models	Poor performance in noisy environments; outdated models	Moderate performance; outperformed by modern DL models
Xavier Anguera *et al.*	ASR development for low-resource languages	Kaldi ASR, Grapheme-based models	Uses phoneme recognition with dynamic programming for alignment	Works with minimal labeled data; scalable	High WER in noisy conditions; issues with spontaneous speech	Achieved 20% WER; feasible for low-resource setups
Mitushi Kohli *et al.*	AI summarization for YouTube and Podcasts	LLMs (Gemini Pro, GPT)	Speech-to-text + NLP summarization with cosine similarity for evaluation	Fast and flexible summarization for diverse video content	Needs fine-tuning for dialogues; scalability limits for long content	Effective summarization; improves accessibility for users

are augmented with natural language processing approaches for increased accuracy. The technique is highly accurate and supports several languages. However, it problems with overlapping speech, multiple accents, and significant computing costs. Despite these obstacles, the system achieved higher transcribing accuracy than traditional methods.

Anmol Gulati *et al.* [2] aimed to improve live meeting transcription using contextual knowledge. They trained transformer-based automatic speech recognition models on broad datasets before fine-tuning them with meeting-specific data. This technology enables context-aware transcriptions and more accurate identification of specific speakers. The key drawbacks are the need for big labeled datasets and longer processing times. Nonetheless, their approach outperformed benchmarks in terms of word correctness.

Ayushi *et al.* [3] explores speech-to-text application using standard models such as Recurrent Neural Networks (RNN) and Hidden Markov Models (HMMs). Their approach utilized sequential modeling of speech signals for phoneme detection, which is both efficient and interpretable, particularly on small datasets. However, these models perform poorly in noisy environments and are surpassed by more contemporary deep learning algorithms. The overall performance was moderate..

Xavier *et al.* [4]focused on developing automatic voice recognition algorithms for Catalan and Spanish languages with very limited resources . They used Kaldi ASR and grapheme-based models to align audio and text using phoneme identification and dynamic programming. The system performs well with limited labeled data and is scalable. However, it struggles with spontaneous speech and has a high word error rate (20) in noisy situations. Despite these problems, the system demonstrated suitability for low-resource ASR tasks.

Mitushi Kohli and Ansh Choudhary *et al.* [5] developed an AI-powered summarizer for YouTube videos and podcasts using huge language models like Gemini Pro and GPT. Their method combines AI-based speech-to-text conversion with natural language processing to achieve text summarization, which is measured using cosine similarity. This technology delivers quick and effective content summarizing that can be tailored to different video types, making it more accessible to viewers. The hurdles include fine-tuning of conversational transcripts and scalability issues while processing large videos. Nevertheless, it produces useful video summaries.

III. PROPOSED METHODOLOGY

A. Speech Recongition with Whisper

Whisper is a sophisticated speech recognition system that employs deep learning to convert spoken words into written text. It works effectively in a variety of languages and can recognize diverse accents and noisy backgrounds, making it ideal for real-world online meetings.Whisper is constructed with a particular type of AI called a transformer, which allows the model to understand the context of what is being said rather than simply individual words. This allows it to make more accurate transcriptions.The model was trained using a large volume of speech from various persons, languages, and locations. It will accept audio input and then divide/separate it into 30 second batches, with each batch going through encoding (using the attention model) and decoding (for position). This process continues until all of the audio is transcribed. This was trained, like GPT, for 680,000 hours. This model was trained using a variety of audio formats such as clips, podcasts, TedTalks, interviews, and so on. This model was developed by Openai in 2022.

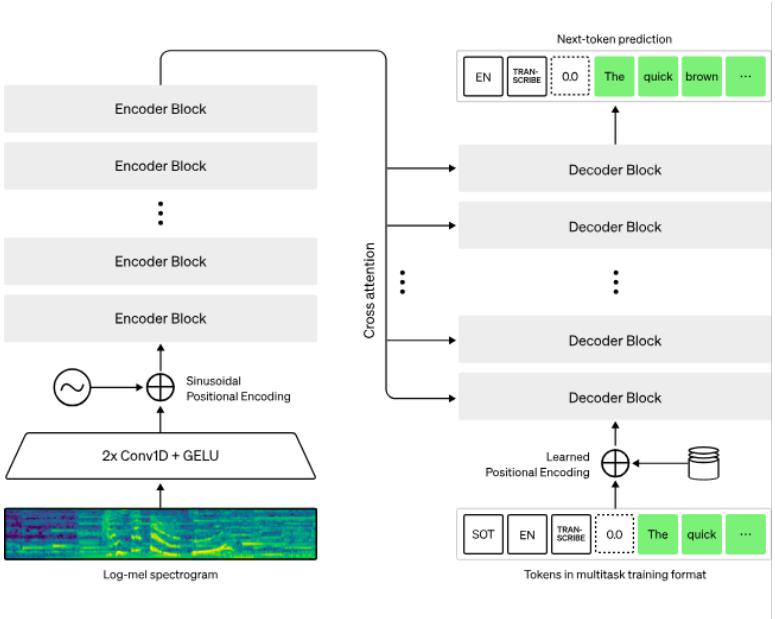

Fig. 1. Pipeline diagram of whisper

B. Text Refinement and Summarization using BART

Even after the audio is translated into text using Whisper, the result may contain errors such as repeated words, filler phrases, and unstructured sentences. To address this, we apply BART (Bidirectional and Auto-Regressive Transformers), an advanced AI model capable of cleaning and summarizing text.

BART rewrites the rough text, making it cleaner and more orderly. It removes extraneous portions, improves phrase flow, and highlights key topics from the discourse. This aids in the conversion of lengthy, disorganized transcripts into well-structured summaries.

In our project, we are creating this system as a browser extension (or platform add-on), so users can utilize it during or after an online session. The extension captures the transcription and processes it through BART, and automatically generates a neat summary document. This makes it easier for users to evaluate and discuss the essential points from their meetings. The BART model is a combination of BERT (generation) and GPT (understanding), utilizing BERT encoding and GPT decoding.It randomly masks the word and then attempts to recreate the original word.

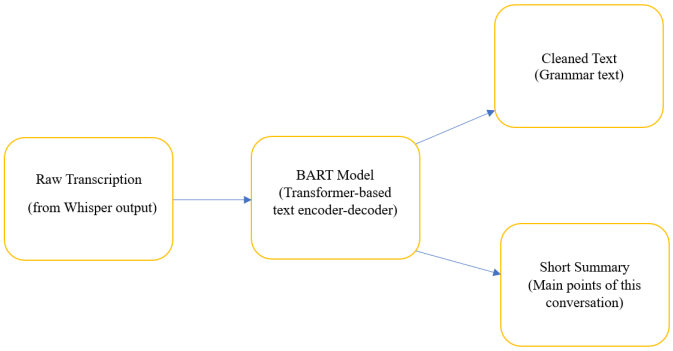

Fig. 2. Pipeline diagram of BART

C. Working of our model

Audio input: The technology starts by recording live audio from online sessions like meetings, webinars, and lectures. This is accomplished in real time via an integrated extension or utility that detects and records the audio stream without user intervention.

Speech-to-Text Conversion (Whisper): Audio is recorded and converted to text using the Whisper paradigm. Whisper is a deep learning model that works well in noisy circumstances and can recognize a variety of accents and languages. It generates a thorough transcription of the session, although the text may still contain filler words or incorrect formatting.

Text Cleaning and Structuring (BART): Whisper transcriptions are cleaned and organized using the BART methodology. BART removes filler words (such as ”um” or ”you know”), corrects grammar errors, and organizes the content into logical paragraphs. This step makes the transcript considerably more readable and understandable.

Summarization: The system uses BART to summarize the cleaned text. It focuses on capturing the major points, critical decisions, and important issues covered during the session. The end result is a concise and clear summary that provides users with a rapid overview without the need to read the entire transcript.

Document Generation: The detailed transcript and summary are combined into a final document. This document is well-organized and ready to share, store, or use for future reference. It can be saved as a PDF, Word document, or displayed within the extension interface for convenient access.

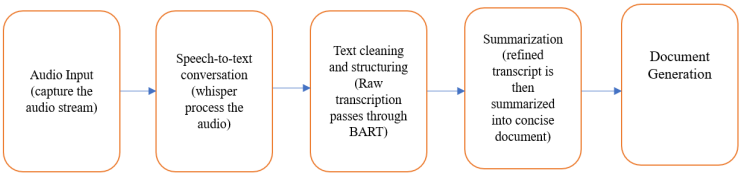

Fig. 3. Pipeline diagram of whisper

IV. Results and Discussion

We tried the Chrome extension on a variety of online platforms, including Google Meet, Zoom, and Microsoft Teams. The application was capable of capturing live audio during online conversations and converting it to text using the Whisper paradigm. Following transcription, the BART model was used to clean up the text and generate an easy-to-read summary. The finished paper includes the entire chat as well as a condensed version with important highlights that could be downloaded or shared.

The BART model improved the raw text by deleting extraneous words, improving syntax, and structuring the content, resulting in concise and effective summaries that highlighted the essential points of the conversation. The transcription results were good, especially in calm areas with clear speech, and Whisper handled varied accents with few problems.

Users found the application handy in real-world scenarios such as classrooms and online meetings. It saved time by decreasing the need for manual note-taking and ensuring that essential facts were not overlooked. Even lengthy meetings were handled efficiently, albeit the time required to prepare the final paper increased somewhat as the session length increased.

There were a few challenges. The extension struggled to handle scenarios in which multiple persons spoke at once or the audio quality was poor. Also, the present version lacks speaker identification, making it difficult to determine who said what. These concerns are being considered in future versions.

In conclusion, the extension met its primary objectives of turning speech into text in real time, cleaning and summarizing the content, and producing a clear document. This study demonstrates how merging powerful AI models with web technologies can provide practical and user-friendly solutions for boosting communication and documentation during online meetings.

TABLE I
COMPARISON OF OUR PROJECT WITH MAJOR MEETING PLATFORMS

Platform	**Generation of Live Captions**	**Summarization**	**Document Generation (like Our project)**
Google Meet	Yes, but only for live or recorded sessions	No	No
Zoom	Yes, via captions and cloud transcripts	No	No (requires external tools)
Microsoft Teams	Yes, supports transcription during/after meetings	Basic summary only (via Copilot)	No
Our Project	Yes, Whisper	Yes, detailed summary via BART	Yes, automatically generated document

V. CONCLUSION AND FUTURE WORK

In this project, we created a Chrome extension that automatically turns audio from live online sessions into a well-structured paper. The system is written in HTML, CSS, JavaScript, and Python Flask, using AI models such as Whisper for speech-to-text conversion and BART for text cleaning and summarization. The plugin operates in real time, allowing users to focus on the meeting rather than manually taking notes. It provides both a full transcript and a brief, easy-to-read summary, allowing users to rapidly grasp the main ideas covered.

This application makes online sessions more accessible and beneficial, particularly for students, distant workers, and professionals who frequently attend virtual meetings. It simplifies paperwork, lowers the likelihood of omitting critical information, and boosts productivity.

In the future, we hope to add capabilities such as speaker identification, which would allow users to see who said what during conversations. We also intend to offer additional languages to make the tool more useful to folks from various places. Our roadmap includes enhancements such as real-time summarization, improved speed for long sessions, and connection with cloud systems such as Google Drive or Microsoft Teams. These enhancements will make the extension even more powerful, adaptable, and appropriate for a broader range of users and scenarios.

REFERENCES

[1] M. Radfar, R. Barnwal, R. V. Swaminathan, F.-J. Chang, G. P. Strimel, N. Susanj, and A. Mouchtaris, "Convrnn-t: Convolutional augmented recurrent neural network transducers for streaming speech recognition," *arXiv preprint arXiv:2209.14868*, 2022.

[2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[3] A. Trivedi, N. Pant, P. Shah, S. Sonik, and S. Agrawal, "Speech to text and text to speech recognition systems-areview," *IOSR J. Comput. Eng*, vol. 20, no. 2, pp. 36–43, 2018.

[4] X. Anguera, J. Luque, and C. Gracia, "Audio-to-text alignment for speech recognition with very limited resources." in *Interspeech*, 2014, pp. 1405–1409.

[5] M. Kohli, A. Choudhary, and D. Gupta, "Youtube, podcast summarizer using ai," 2024.