TalkTrace: Secure Automated Transcription and Summary Generation

Alfred Joe Devasia
Department of Computer Science and Engineering, Carmel College of Engineering and Technology
Alappuzha, Kerala, India
alfredjoedevaia9497@gmail.com

Nandhana Kunjumon
Department of Computer Science and Engineering, Carmel College of Engineering and Technology
Alappuzha, Kerala, India
nandhanakunjumon2004@gmail.com

Rahul R Krishna
Department of Computer Science and Engineering, Carmel College of Engineering and Technology
Alappuzha, Kerala, India
rahulrkrishnan2004@gmail.com

Safna M S
Department of Computer Science and Engineering, Carmel College of Engineering and Technology
Alappuzha, Kerala, India
safnams2003@gmail.com

Thomas Joseph
Assistant Professor, Department of Computer Science and Engineering, Carmel College of Engineering and Technology
Alappuzha, Kerala, India
thomasjoseph@carmelcet.in

Abstract— Although the importance of accurate documentation of meetings is paramount in modern organizations, the majority of the state-of-the-art transcription services utilize third-party cloud-based AI services, leading to severe concerns over data privacy, security, and user control. This paper proposes a secure and privacy-focused meeting transcription system called TalkTrace, which utilizes automated bot-based audio capture and server-side speech processing by the provider through a locally deployed speech-to-text model. By providing a link to the meeting through the web interface, such as a Zoom or Google Meet link, the automated bot joins the meeting and exits the meeting after the completion of the meeting or manually by the user. The recorded audio is encrypted in real time by employing a hybrid RSA-AES encryption technique to ensure the confidentiality of the recorded data. On a rented server from a hosting provider, speech transcription and speaker diarization take place, with the Whisper model running locally on the server in inference mode, ensuring that audio data is not transmitted to any external AI services or used for training any model. Speaker identification is done using guided verbal introductions that match with the diarized speech using timestamp-based matching methods. To secure the integrity of the transcripts, the final output of the speech-to-text process is hashed using SHA-256 before secure storage. TalkTrace provides an alternative to traditional cloud-based meeting transcription services by using strong encryption, automated meeting integration, and the use of hosted local inference using AI services.

KEYWORDS

Meeting transcription; speaker diarization; RSA; Advanced Encryption Standard (AES); secure hashing

I. INTRODUCTION

The widespread adoption of online and hybrid meeting platforms has significantly changed modern communication and collaboration practices in organizations. Meetings have become a primary medium for discussion, decision-making, and coordination among distributed teams. However, manual documentation of meeting discussions is often inefficient, error-prone, and distracting, leading to incomplete or inaccurate records of important decisions and action items [2]. Automatic meeting transcription and summarization systems have emerged as an effective solution to address these challenges. By leveraging advances in Automatic Speech Recognition (ASR) and Natural Language Processing (NLP), such systems can automatically convert spoken conversations into textual transcripts and generate concise meeting minutes [1]. These technologies reduce human effort and improve productivity, especially in professional and academic environments. Nevertheless, real-world meeting scenarios introduce challenges such as overlapping speech, background noise, and varying speaker accents [3]. Most existing meeting transcription and summarization solutions rely heavily on cloud-based processing architectures [4]. While cloud platforms offer scalability and high computational power, they raise serious concerns related to data privacy, security, and ownership of sensitive meeting content. Organizations conducting confidential meetings often require stronger guarantees such as secure audio handling, encrypted transcript storage, and user-controlled data access [6]. Recent research has explored improved learning techniques and robust models to enhance transcription accuracy and summary quality. Approaches such as semi-supervised learning, SpecAugment-based data augmentation, and transformer-based summarization models have shown significant reductions in Word Error Rate (WER) and improved summary coherence [3,5]. However, these methods are often computationally intensive and unsuitable for real-time or resource-constrained environments. This paper presents TalkTrace, an intelligent and privacy-aware system for automatic meeting transcription and summarization. The proposed system integrates efficient ASR models with NLP-based summarization techniques to generate structured meeting minutes while prioritizing secure audio processing and user-controlled data storage. Unlike conventional cloud dependent systems, TalkTrace emphasizes privacy preservation without compromising transcription accuracy and summary relevance. The main contributions of this work include the design of an end-to-end automated meeting documentation framework,

integration of action-item-driven summarization techniques [7], evaluation of system performance in realistic meeting scenarios, and analysis of privacy and security limitations observed in existing solutions [8]. The remainder of this paper is organized as follows. The next section reviews related work in automatic meeting transcription and summarization. The system architecture and design methodology are then presented, followed by implementation details and experimental setup. System performance and evaluation results are subsequently discussed, and finally, the paper concludes with limitations, future enhancements, and concluding remarks.

II. RELATED WORK

Wav2Letter++ is a high-performance, open-source end-to-end Automatic Speech Recognition (ASR) framework developed using modern C++. It supports multiple training criteria such as Connectionist Temporal Classification (CTC), Auto Segmentation Criterion (ASG), and sequence-to-sequence models, enabling flexibility in ASR research. The framework is optimized for scalability and efficient training on large speech datasets using multi-GPU architectures. Due to its speed and modular design, it is widely adopted in academic and industrial research. However, the system primarily focuses on transcription accuracy and computational efficiency, without addressing privacy concerns, secure audio handling, or encrypted storage of meeting transcripts [1].

This work presents a comprehensive survey of recent techniques used in automatic meeting minutes generation. It explores the integration of speech recognition, speaker diarization, keyword extraction, and natural language processing–based summarization to generate structured meeting records. The study highlights significant productivity improvements achieved by automating documentation in organizational environments. It also discusses challenges such as overlapping speech, background noise, and accurate speaker identification. Despite these advancements, most systems reviewed rely heavily on cloud-based or centralized processing, which raises concerns related to data privacy, security, and user ownership of meeting content [2].

Improved Noisy Student Training introduces a semi-supervised learning approach aimed at enhancing speech recognition performance by leveraging large volumes of unlabeled audio data. The method combines self-training with SpecAugment-based data augmentation to improve model robustness. Experimental results demonstrate significant reductions in Word Error Rate (WER) on benchmark datasets such as LibriSpeech, showcasing the effectiveness of the approach in noisy and low-resource environments. However, the training pipeline is computationally intensive and requires high-end hardware resources, making it less suitable for real-time, local, or resource-constrained deployments [3].

The Online Meeting Summary Generator proposes an automated system for generating summaries from online meeting recordings. It integrates speech-to-text conversion with natural language processing techniques such as text summarization and keyword extraction to identify key discussion points. The system assists users in tracking decisions, action items, and meeting outcomes, making it suitable for remote and hybrid work environments. While effective in enhancing meeting documentation, the system places limited emphasis on security aspects such as end-to-end encryption, transcript integrity verification, and user-controlled data storage [4].

This study investigates the impact of domain-specific terminology on the performance of automatic meeting summarization systems. The authors analyze how specialized vocabulary affects summary coherence and accuracy in meeting minutes generation. The findings highlight challenges in handling diverse meeting contexts. However, the work focuses on summarization quality and does not address real-time transcription, secure meeting capture, or privacy-aware system design [5].

This paper presents a machine learning–based framework for automatically generating minutes of meetings from recorded discussions. The system applies speech-to-text conversion followed by text classification and keyword extraction to identify important content. The generated minutes are structured to highlight major topics and decisions. Despite its effectiveness, the framework does not consider data security, encrypted storage, or privacy protection in online meeting environments [6].

This work focuses on generating concise meeting summaries by identifying and prioritizing action items within long meeting transcripts. The authors use transformer-based natural language processing models to recursively summarize topic-specific segments of meetings. The approach improves the relevance and usefulness of generated meeting minutes. However, the system assumes pre-existing transcripts and does not address secure audio

III. SYSTEM ARCHITECTURE

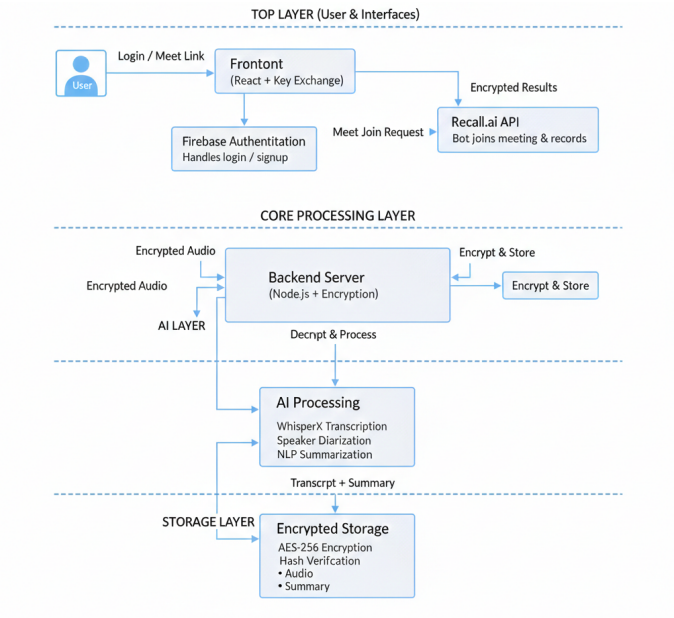

Figure 1: TalkTrace Architecture

IV. ARCHITECTURE DESIGN

A. Architecture Overview

The platform is an end-to-end transcription solution for meetings that is privacy-preserving and reduces cloud deployment. It minimizes the need for cloud processing while offering an automated online meeting approach. It features a privacy-respecting architecture that encrypts portions of private audio data and then sends it for processing in a locally trusted server architecture. Such an architecture removes the risk of privacy-invading transcription that other platforms face when sending data all the way across the globe to transcribe platforms in the cloud.

The system consists of five components at a high-level: a front-end web-based interface, an automated meeting bot, a secure back-end interface, a hybrid encryption component, a hybrid local transcription and analysis engine. The process begins with a user providing an authenticated meeting link for either Zoom or Google Meet. The back-end subsequently triggers an automated meeting bot that joins the meeting and records audio. The recorded audio is subsequently encrypted by the Hybrid Encryption Algorithm (combining RSA and AES).

The difference comes in the third component. Typically, it's on an external server for both conversion and transcription, however TalkTrace utilizes a locally trusted (still a virtual component within TalkTrace) Whisper algorithm from OpenAI for the speech to text conversion. The same happens with the transcription as original audio is not sent to any outside audio to text processors. The transcript is subsequently speaker diarized and compared with speaker identity matchers and then a hash is created from the transcript to ensure integrity.

B. Component Architecture

TalkTrace consists of the following key components:

Web Client: The primary user interface is the web client. Authentication, getting the meeting link, requesting encrypted audio, decrypting the audio when necessary for safe handling, and displaying the transcript are some of its duties. Speech recognition is not done by the web client; in order to protect privacy and security, transcription is done within the system's locally controlled server environment.

Automated Meeting Bot: The bot joins the meeting link and simply sits in a meeting. It listens (or, records, with blended audio) without video, or participant details and leaves the meeting either when the meeting time is over or the user ends the session.

Backend Service Layer: Authentication, encrypted audio storage, encryption API endpoints, bot lifecycle operations (joining, leaving, and status tracking), and local transcription execution are all handled by the backend. Crucially, no external services are able to access unencrypted audio or transcripts from the backend.

Encryption Module: A hybrid cryptography layer is utilized for safe transfer and storage of audio. AES is implemented for audio fast encryption, while RSA is reserved for transmission of AES session keys between the system elements.

Local Transcription Engine: This module carries out speech recognition using the Whisper model, which has been implemented locally within the controlled environment of the server. It processes speech-to-text, speaker identification, identity inference, as well as transcript integrity analysis without relying on cloud services.

C. Bot-Based Meeting Capture Architecture

Meeting automation through a bot is a critical component of TalkTrace. Once a valid meeting URL is received, the backend launches a bot instance that joins the meeting as a regular participant. The bot captures only the mixed audio stream and does not access video feeds, chat messages, or other participant data.

The bot operates in the background, continuously monitoring meeting status. Recording stops automatically when the meeting ends or when a manual stop request is issued by the user. The captured audio is temporarily buffered in memory and then forwarded to the encryption module. Raw audio data is never stored or processed in unencrypted form within the backend.

D. Security and Encryption Architecture

TalkTrace implements a hybrid RSA-AES encryption model to balance security and efficiency. Since RSA encryption has higher computational overhead and is unsuitable for large data streams, AES-256 is used to encrypt the meeting audio using a randomly generated session key for each recording.

To securely transmit the AES session key, it is encrypted using the user's RSA public key. This ensures that the backend can encrypt the audio and session key but cannot decrypt them. Only the frontend, which holds the RSA private key, can decrypt the AES key and subsequently the audio data.

This design establishes a strong security boundary: the backend can lock data but cannot unlock it, while the frontend can unlock data but does not generate encryption keys. The architecture follows zero-trust principles and privacy-by-design standards.

E. Data Flow and Communication Sequence

TalkTrace operates through the following sequence:

- The user verifies and enters the meeting link into the web interface.
- The backend validates the link and activates the meeting bot.
- The bot joins the meeting and records the audio stream.
- Audio is encrypted in real time using AES, and the AES key is encrypted using RSA.
- The encrypted audio, initialization vector (IV), and encrypted AES key are securely stored.
- The frontend retrieves the encrypted payload via secure API calls.
- The AES key is decrypted using the RSA private key.
- The audio is decrypted and processed locally for transcription and analysis.
- The final transcript is hashed and stored to ensure integrity verification.

This pipeline ensures end-to-end security of confidential data while maintaining a seamless and privacy-focused user experience.

V. IMPLEMENTATION

A. Frontend Implementation

TalkTrace develops its frontend as a web application using modern JavaScript frameworks to build an interactive user interface. User authentication is implemented through stateless session management using JSON Web Tokens (JWT). Users can enter a meeting URL, monitor recording progress, initiate transcription, and review the completed transcript through the interface. The frontend makes use of browser-based crypto libraries to carry out the RSA key decryption and AES audio decryption operations. The encrypted audio data is transmitted securely to the backend for the decryption and transcription process to take place within the controlled server environment.

B. Automated Bot Implementation

The meeting automation bot operates as a backend service that enables users to control Zoom and Google Meet through browser automation and media capture functionalities. The bot joins meetings with participant-level permissions and records system audio streams in real time. Bot lifecycle management is handled through secure backend APIs, allowing users to start and stop bots while monitoring their current status. The recorded audio streams into a buffer where encryption begins immediately, ensuring that no plaintext audio files are stored on the backend server.

C. Encryption and Secure Storage Implementation

The encryption process begins when the backend retrieves the meeting audio stream, either through download or direct receipt. Each meeting session generates a unique 256-bit AES key along with a one-time initialization vector (IV). The audio buffer is encrypted using AES-256-CBC to ensure both strong security and efficient performance for large media files. The AES key is then encrypted using the user's RSA public key, creating an encrypted key payload. The backend stores the encrypted audio along with the IV and encrypted AES key. During download, these components are sent to the frontend, where decryption is performed using the user's private key. This design ensures that unauthorized entities cannot access any meeting data.

D. Local Transcription Engine

TalkTrace uses Whisper locally for speech recognition. Audio input is processed into 30-second segments to produce Log-Mel spectrograms. Timestamped text segments are produced by feeding these into Whisper's encoder-decoder architecture. Since all computations are done locally, neither the audio data nor the transcription results are sent to any external AI cloud for training.

E. Speaker Diarization and Identification

Speaker diarization is used to determine "who spoke when" during the meeting. The process begins with voice activity detection (VAD) to isolate speech segments. Voice embeddings are then generated to create speaker fingerprints, which are clustered to distinguish different speakers. TalkTrace adopts a guided introduction method for speaker identification. During the initial phase of the meeting, statements such as "My name is John" are detected using JavaScript regular expressions applied to Whisper's timestamped transcription. Identified names are mapped to diarized speaker clusters based on timestamp alignment.

F. Transcript Integrity and Trust Management

After transcription and speaker labeling, post-processing and normalization are performed. A SHA-256 cryptographic hash of the final transcript is generated and stored alongside the document. Any modification to the transcript produces a new hash, ensuring tamper detection. This integrity mechanism enhances trust and supports auditability of meeting transcripts. Future improvements may include integrating blockchain-based notarization to create a permanently immutable transcript record for high-compliance applications.

VI. EXPERIMENTAL RESULTS

A. Speech Recognition Accuracy

The transcription component of TalkTrace uses OpenAI Whisper as its base model which achieves more than 90% accuracy when transforming meeting audio into text. The model exhibits strong performance across different accents and speech rates while handling typical online meeting background noise. The system shows high transcription accuracy which enables reliable results for tasks that follow such as speaker identification and summarization.

B. Speaker Diarization Performance

The WhisperX diarization system enables speaker identification to achieve 90% precision when it matches spoken audio to corresponding text sections. The system employs this method to successfully identify different speakers and match their spoken words to the correct audio parts. The process of using guided verbal introductions and matching timestamps automatically produces accurate meeting records which enhance the security of the presentation and make it easier to read.

C. Local NLP Processing and Privacy Preservation

The organization ensures complete data privacy by executing all natural language processing operations through local systems which utilize offline language models including Ollama and Mistral for their language processing needs. TalkTrace maintains total user privacy by not using cloud-based Natural Language Processing services which keep confidential meeting information within the user's workspace. The design prevents data leakage while it enables organizations to maintain compliance with their rigorous data security standards.

D. Encryption and Data Security

TalkTrace implements a hybrid encryption system that uses AES 256 for bulk audio data protection and RSA 2048 for secure key distribution. The military-grade cryptographic system protects all stored and sent meeting information from unauthorized access by making it confidential and impossible to read. The system uses cryptographic hashing procedures to protect transcript data from

unauthorized access which ensures data integrity and prevents any changes to the information.

E. Comparison table

Feature	Otter.ai	Fireflies.ai	Zoom AI Companion	TalkTrace
Automatic Meeting Recording	Yes	Yes	Yes	Yes
Speech-to-Text Transcription	Yes	Yes	Yes	Yes
Speaker Diarization	Limited	Yes	Limited	Yes
Local / Controlled Processing	No	No	No	Yes
Encrypted Audio Storage	Partial	Partial	Partial	Yes
Transcript Integrity Verification	No	No	No	Yes
Audit Trails / Versioning	No	No	No	Yes
Privacy-Focused Architecture	Limited	Limited	Limited	Yes

Figure 2: Comparison with existing system

The table presents a comparative evaluation of four meeting transcription platforms: Otter.ai, Fireflies.ai, Zoom AI Companion, and TalkTrace, based on several important functional and security-related features. All the tools support automatic meeting recording and speech-to-text transcription, which are essential for capturing and documenting discussions. However, some systems provide only limited support for speaker diarization, making it difficult to accurately identify different speakers. Most existing platforms rely on cloud-based processing, which limits user control over data handling. Security features such as encrypted audio storage are only partially supported, while transcript integrity verification and audit trails are generally absent. In contrast, TalkTrace provides full support for these privacy-focused capabilities.

System	Model Size	WER	Accuracy	Speed / Latency	Key Focus
SpecAugment	Depends on base model	6.8%	92%	Not focused on inference speed	Accuracy improvement technique
wav2letter++	100M	5.0%	~95%	10–140 ms / sample	Extremely fast decoding
Whisper Base (our)	74M	5%	95%	Near real-time GPU	Robust real-world transcription (Best Balanced Option)
Preech: A System for Privacy-Preserving Speech Transcription.	No single model size	4–10%	90–96%	Not latency-focused	Privacy-preserving transcription with improved WER over offline ASR
Configurable Privacy-Preserving Automatic Speech Recognition	No fixed model size	8–44%	56–92%	Near real-time possible	Configurable privacy using separation + discretization modules

Figure 3: Comparison with reference paper

The table compares different automatic speech recognition (ASR) systems based on model size, word error rate (WER), accuracy, processing speed, and their primary research focus. SpecAugment is mainly an accuracy improvement technique that enhances model performance but does not prioritize inference speed. wav2letter++ uses a large model and is designed for extremely fast decoding with high accuracy. Whisper Base, used in our system, provides a balanced approach with a smaller model size, low WER, and near real-time GPU processing, making it suitable for practical transcription tasks. Other systems such as Preech and Configurable Privacy-Preserving ASR focus mainly on privacy protection rather than latency optimization.

VII. CONCLUSION

Because TalkTrace prioritizes user data protection through its automated transcription system that ensures transcription accuracy, it addresses significant issues that current meeting transcription systems encounter. The system's bot-based meeting capture system, which encrypts user data end-to-end and uses Whisper-based local transcription without a cloud connection, allows for secure data processing. When paired with timestamped introduction-based identity verification and cryptographic security techniques, the system uses speaker identification to authenticate transcript content. TalkTrace's system offers businesses a safe and efficient transcription solution that doesn't require cloud services like other systems do.

References

[1] Pratap, V., Hannun, A., Xu, Q., Cai, J., Kahn, J., Synnaeve, G., Liptchinsky, V., Collobert, R. Wav2Letter++: The fastest open-source speech recognition system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6465–6469.

[2] Zhang, A., Li, Y., Wang, H. Advances in automatic meeting minutes generation. *IEEE Access*, vol. 8, pp. 154321–154335, 2020.

[3] Park, D. S., Zhang, Y., Jia, Y., Han, W., Chiu, C.-C., Li, B., Wu, Y., Le, Q. V. Improved noisy student training for automatic speech recognition. In *Proceedings of Interspeech*, 2020, pp. 2817–2821.

[4] Li, B., Zhao, J., Liu, K. Online meeting summary generation using speech and natural language processing. *International Journal of Speech Technology*, vol. 24, no. 3, pp. 635–646, 2021.

[5] Koay, J., Roustai, A., Dai, X., Burns, D., Kerrigan, A., Liu, F. How domain terminology affects meeting summarization performance. arXiv preprint arXiv:2011.00692, 2020.

[6] Pandya, A., Gawande, N. Automatic generation of minutes of meetings using machine learning and NLP. *International Journal of Scientific Research in Science, Engineering and Technology*, 2022.

[7] Golia, L., Kalita, J. Action-item-driven summarization of long meeting transcripts. arXiv preprint arXiv:2312.17581, 2023.

[8] Vaddi, A., Swapna, K. Automatic meeting minutes generation using NLP techniques. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 2025.