A Review on Natural Language Processing (NLP) Models for Generating Meeting Transcription

Swati Tilkar[1] Prof. Ruchika Pachori[2]
Department of Information Technology[1,2]
MIT, Ujjain, India[1,2]

ABSTRACT: Without a loss of generality, it can be stated that meetings play a vital role in collaboration and decision-making in diverse domains and organizations. However, manually documenting these meetings is time-consuming and prone to errors or omissions. To address this challenge, Natural Language Processing (NLP), a subfield of artificial intelligence, has emerged as a powerful tool for automating the transcription of spoken content. With NLP, meeting conversations can be converted into accurate, readable text in real-time or post-meeting, improving productivity, accessibility, and record-keeping. Almost all automated meeting transcription models use speech recognition, which involves converting spoken language into text. Modern speech recognition systems use machine learning and deep learning models for the purpose. This paper presents a comprehensive review on the machine learning and deep learning models typically involved for NLP and speech transcription. The salient points of the review can be used to leverage speech transcription mechanisms to be used for a specialized case of meeting transcriptions.

Keywords: Meeting Transcriptions, Machine Learning, Deep Learning, Natural Language Processing, Word Error Rate (WER), Character Error Rate (CER).

I. Introduction

The global voice and speech recognition market I growing at a staggering pace whose size was estimated at USD 20.25 billion in 2023 and is anticipated to grow at a CAGR of 14.6% from 2024 to 2030. The market is anticipated to be driven by technological advancements and rising adoption of advanced electronic devices [1]. Voice-activated biometrics used for security purposes help in providing access to authenticated users for performing a transaction. The growing use of voice biometrics is among the major factors driving the market growth. The increasing demand for voice-driven navigation systems and workstations is impelling growth in the hardware and software segments [2].

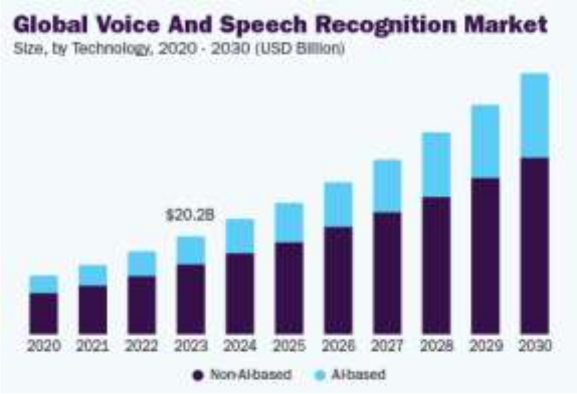

Fig.1 Global Voice and Speech Recognition Market

(https://www.grandviewresearch.com/industry-analysis/voice-recognition-market)

Figure 1 depicts estimated growth of the voice and speech recognition market by 2030. Several domains would benefit for the application such as:

- Large Scale Business
- Autonomous Driving
- Biomedical
- Large scale automation and IoT etc.

Some of the critical aspects of voice and speech recognition are [3]:

Speaker Diarization and Role Identification: A critical component of meeting transcription is speaker diarization—determining "who spoke when." This is particularly important in multi-speaker meetings where attributing dialogue to the correct participant can influence interpretation and accountability. NLP systems incorporate voice biometrics and clustering algorithms to separate speakers. Advanced systems can even assign names and roles if participants are known, further enriching the transcription for organizational use [4].

Contextual Understanding and Summarization: Raw transcriptions, while useful, can be lengthy and difficult to navigate. NLP algorithms help by not only cleaning up the text but also identifying key points, decisions, and action items through contextual analysis. Text summarization techniques, both extractive and abstractive, condense the meeting content into digestible summaries. Named entity recognition (NER), sentiment analysis, and topic modeling can further provide insights into the meeting's dynamics and outcomes [5].

Applications: The application of NLP in meeting transcription spans industries—from corporate boardrooms and legal depositions to academic lectures and telemedicine sessions. Automated transcriptions increase accessibility for individuals with hearing impairments and support multilingual transcription through machine translation. They also facilitate better searchability and knowledge management, allowing users to retrieve past discussions quickly. Moreover, companies can analyze meeting data to identify productivity trends or improve internal communication. Despite its promise, NLP-based meeting transcription still faces several challenges. Background noise, overlapping speech, and domain-specific jargon can hinder accuracy. Privacy concerns also arise, especially in sensitive or confidential discussions. As technology advances, integrating NLP with real-time collaboration platforms, improving domain adaptation, and enhancing multilingual support will be key areas of focus. With ongoing research and development, the goal is to create seamless, context-aware transcription systems that mimic human note-taking [6]

II. NLP Based Meeting Transcription

In practical scenarios, maintaining accurate records of meetings can be time-consuming and inefficient when done manually. Natural Language Processing (NLP), a branch of artificial intelligence, offers a powerful solution by enabling automated transcription of spoken content. Through advanced algorithms and models, NLP can convert speech into text with high accuracy, making it easier to document, review, and share meeting discussion. While this technology holds great promise for improving productivity, accessibility, and record-keeping, it is not without challenges. Generating accurate, meaningful, and context-aware transcriptions from live or recorded meetings involves overcoming a variety of technical, linguistic, and ethical challenges [7].

Existing Challenges:

Speech Variability and Audio Quality: One of the primary challenges in meeting transcription is the variability in speech. Participants may have different accents, dialects, speaking speeds, or tones, which can affect the system's ability to accurately recognize and transcribe their words. Additionally, the quality of the audio—affected by background noise, poor microphones, echo, or overlapping speech—can significantly reduce transcription accuracy. NLP systems must be robust enough to filter out noise and understand speech under less-than-ideal conditions [8].

Speaker Diarization and Identification: In multi-speaker settings, it is crucial to distinguish who is speaking at any given time. This process, known as speaker diarization, is a complex task for NLP systems. Overlapping speech, similar voice patterns, and rapid turn-taking can confuse the system. Accurate speaker identification is necessary for meaningful transcriptions, especially when assigning responsibilities or tracking opinions during meetings. However, this remains a difficult problem, especially in large or informal group discussions [9].

Handling Informal and Unstructured Language: Meetings often involve informal, spontaneous language that includes interruptions, filler words, slang, false starts, and unfinished sentences. NLP systems trained primarily on well-structured text or formal speech may struggle with such unstructured dialogue. Transcribing these conversations accurately while maintaining readability and coherence poses a significant challenge, especially when the content must be understood by others later or summarized into key points [10].

Domain-Specific Terminology and Jargon: Different industries use specialized vocabulary or technical jargon that may not be well-represented in the training data of general-purpose NLP models. In meetings involving healthcare, law, finance, or engineering, the system might misinterpret or incorrectly transcribe important terms. Customization and continuous training with domain-specific data are necessary, but they add complexity and cost to deployment [11].

Multilingual and Code-Switching Issues: Global organizations often hold meetings involving participants who switch between languages or use multiple languages in a single conversation—a phenomenon known as code-switching. NLP models may not handle this well, especially if they are optimized for a single language. Accurate transcription in multilingual settings requires robust language detection and seamless switching between language models, which is still an evolving area in NLP.

Data Privacy and Security Concerns: Meeting transcriptions often contain sensitive or confidential information. Using cloud-based NLP services for transcription introduces potential data privacy and security risks. Organizations must ensure that their transcription systems comply with data protection laws such as GDPR or HIPAA. Building secure, on-premise solutions can mitigate some risks but may limit access to powerful cloud-based NLP capabilities.

Cost and Computational Requirements: High-quality transcription using state-of-the-art NLP models requires substantial computational resources, especially for real-time transcription. Running these models locally or integrating them into existing systems can be expensive. Moreover, continuous improvements, updates, and model fine-tuning demand both financial and technical investments that may not be feasible for all organizations.

III. Existing Statistical Models

The existing machine learning and deep learning models employed for speech recognition and transcription are [12]:

Hidden Markov Models (HMMs)

Hidden Markov Models were among the earliest and most widely used models in speech recognition. Figure 2 depicts a Hidden Markov Model.

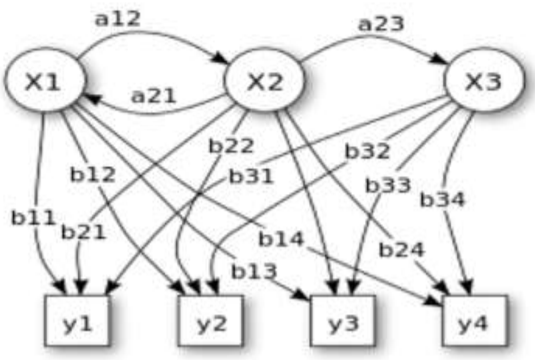

Fig.2 Hidden Markov Model

HMMs are statistical models that represent sequences of observable events (like spoken words) with underlying hidden states (such as phonemes). They assume that the current state depends only on the previous state, making them suitable for modeling temporal sequences like speech. HMMs work well when combined with Gaussian Mixture Models (GMMs) to model acoustic signals, but they struggle with capturing long-range dependencies and complex variations in speech.

Deep Neural Networks (DNNs)

The introduction of Deep Neural Networks marked a major improvement in speech recognition performance. DNNs can learn complex patterns from large datasets and are effective at classifying speech features. When used in combination with HMMs (in hybrid models), DNNs significantly improved acoustic modeling. However, they still had limitations in handling sequential data over time, which led to the development of more specialized architectures like RNNs [13].

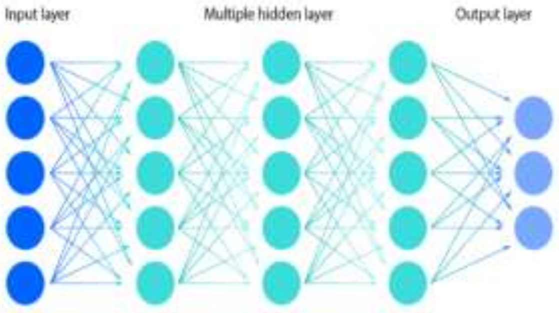

Fig.3 A Deep Neural Network

The input-output relation of a neural network is given by:

$$y = f(\sum_{i=1}^{n} x_i w_i + b) \tag{1}$$

Here,
x denote the parallel inputs
y represents the output
w represents the bias
f represents the activation function

Recurrent Neural Networks (RNNs) and LSTMs:

Recurrent Neural Networks (RNNs) were developed to handle sequential data like audio signals by maintaining a memory of previous inputs. Long Short-Term Memory networks (LSTMs), a type of RNN, address the problem of vanishing gradients, allowing the model to remember information over longer time intervals. LSTMs became widely used in speech recognition systems due to their ability to model context and dependencies in spoken language, greatly improving the natural flow and accuracy of transcriptions.

The LSTM networks are a specialized type of recurrent neural network (RNN) designed to process and predict data sequences by learning long-term dependencies. Unlike traditional RNNs, which suffer from vanishing or exploding gradient problems during training, LSTMs incorporate a unique architecture with gates and memory cells that help retain important information over long periods [14].

The LSTM primarily has 3 gates:
1) Input gate: This gate collects the presents inputs and also considers the past outputs as the inputs.
2) Output gate: This gate combines all cell states and produces the output.
3) Forget gate: This is an extremely important feature of the LSTM which received a cell state value governing the amount of data to be remembered and forgotten.

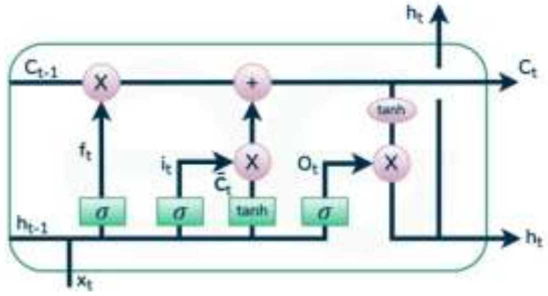

Fig.4 The LSTM Model

Figure 4 depicts the LSTM model. The relation to forget by the forget gate is given by:

$$\boldsymbol{f} = \boldsymbol{\sigma}(\boldsymbol{W_f}[\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_i}) \quad \textbf{(2)}$$

Here,
f denotes forget gate activation
w_f are forget gate weights.
h_{t-1} Denotes Hidden state from the previous time step
x_t is present input.
b_i is the bias

The advantages of LSM are:
Capturing Long-Term Dependencies: LSTMs maintain long-term memory using the cell state, unlike traditional RNNs.
Mitigating Vanishing/Exploding Gradients: Gates help regulate gradient flow, enabling stable training over long sequences.
Versatility: Useful for several time series prediction problems.

Connectionist Temporal Classification (CTC):

One of the key innovations in end-to-end speech recognition is Connectionist Temporal Classification (CTC). CTC allows models to learn the alignment between input audio frames and output text sequences without requiring pre-segmented training data. It is especially useful for training speech-to-text models directly. CTC is often used with RNNs or LSTMs and forms the basis for systems like DeepSpeech, an open-source speech recognition engine developed by Mozilla.

Sequence-to-Sequence Models and Attention Mechanisms:

Sequence-to-sequence (seq2seq) models revolutionized speech recognition by treating it as a translation problem—translating audio sequences into text. These models use an encoder-decoder architecture where the encoder processes the input audio, and the decoder generates the text output. Attention mechanisms further enhance seq2seq models by allowing the decoder to focus on relevant parts of the input at each step. This led to more accurate and context-aware transcriptions.

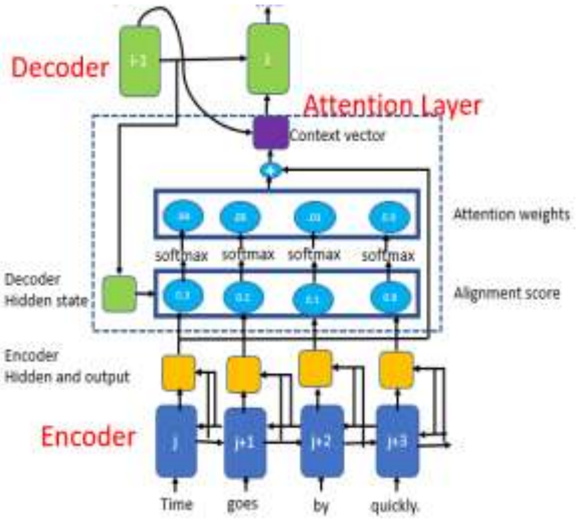

Fig.5 Sequence to Sequence Attention Model

Figure 5 depicts a Sequence-to-Sequence Models and Attention Model.

Transformer-Based Models:
Transformers, particularly models like Wav2Vec 2.0 by Facebook AI and Whisper by OpenAI, represent the current state-of-the-art in speech recognition [15].

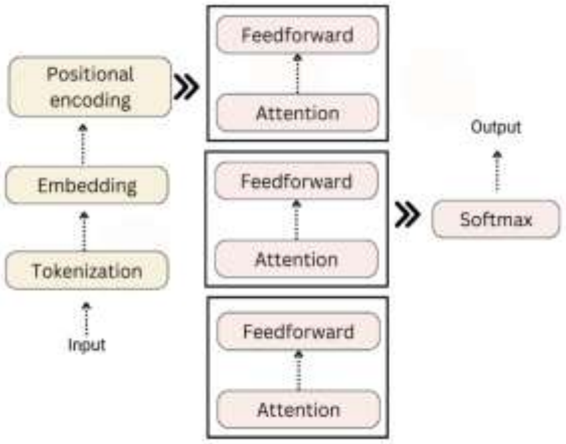

Fig.6 Transformer Model for NLP
Figure 6 depicts the transformer model for NLP. Unlike RNNs, transformers process entire sequences in parallel, allowing them to capture long-range dependencies more efficiently. Wav2Vec 2.0 is a self-supervised model that learns speech representations from raw audio without labeled data, making it highly adaptable. Whisper, on the other hand, is a multilingual, multitask model capable of transcription, translation, and language identification. These models outperform previous architectures in terms of accuracy, speed, and robustness to noise.

End-to-End vs. Hybrid Systems:
Modern speech recognition systems can be broadly categorized into hybrid and end-to-end models. Hybrid systems combine various components (like acoustic, language, and pronunciation models), often using DNNs and HMMs. End-to-end models, such as those using CTC or transformers, simplify the pipeline by directly mapping audio to text. While end-to-end models are more elegant and easier to maintain, hybrid systems still offer competitive accuracy in specific domains and are widely used in commercial applications

IV. Previous Work

A summary of noteworthy contribution in the domain is presented here:

Menon et al. [16] proposed an automation mechanism of noise suppression to eliminate environmental disturbance, transcription and sum- summarization of the recorded conversation taking place between the doctor and the patient(s) to focus only on the essential information, since abridging the entire conversation as a whole may be counterproductive. The tabular summary obtained at the end of the process can be used by the doctors and patients alike, to understand the patient history, prognoses and/or diagnoses. A supervised deep learning technique is used for noise suppression by using a convolutional network, the Google Speech- to-Text API for transcription of the conversation and a basic SVM module which categorizes text based on the given tags and relative frequency of occurrence of a word to create the tabular summary of the said doctor-patient verbal exchange.

Yu et al. [17] proposed that meeting scenario is one of the most valuable and, at the same time, most challenging scenarios for the deployment of speech technologies. Speaker diarization and multi-speaker automatic speech recognition in meeting scenarios have attracted much attention recently. However, the lack of large public meeting data has been a major obstacle for advancement of the field. Therefore, author make available the AliMeeting corpus, which consists of 120 hours of recorded Mandarin meeting data, including far-field data collected by 8-channel microphone array as well as near-field data collected by headset microphone. Each meeting session is composed of 2-4 speakers with different speaker overlap ratio, recorded in meeting rooms with different size. In this paper authors provide a detailed introduction of the AliMeeting dateset, challenge rules, evaluation methods and baseline systems.

Aksthatha et al. [18] proposed a an innovative approach for automating office meeting summarisation, leveraging OpenAI's Whisper ASR model and ChatGPT capabilities to enhance language processing. The Whisper ASR model adeptly transcribes spoken words from meetings into written text, effectively capturing the primary ideas exchanged during discussions. Subsequently, ChatGPT utilizes this transcribed text to generate concise and coherent summaries, emphasizing key points from the conversations. The system adeptly produces precise and meaningful meeting summaries through the synergistic integration of advanced speech recognition and language modelling, facilitating the extraction of crucial insights from spoken content. The system is thoughtfully crafted as a user-friendly web application, employing straightforward technologies such as HTML, CSS, and Flask for accessibility and ease of use. The intuitive interface enables participants to effortlessly upload video content from meetings, allowing the system to transform spoken words into easily comprehensible summaries in real time. Beyond the automation of summarization tasks, this proposed approach signifies a significant advancement in applying language technologies to enhance communication efficiency in professional settings.

Song et al. [19] proposed a a novel productive meeting tool named SmartMeeting, which enables users to automatically record, transcribe, summarize, and manage the information in an in-person meeting. SmartMeeting transcribes every word on the fly, enriches the transcript with speaker identification and voice separation, and extracts essential decisions and crucial insights automatically. In our demonstration, the audience can experience the great potential of the state-of-the-art NLP techniques in this real-life application.

Mehendale et al. [20] proposed that advanced audio and text processing for accurate multilingual transcription, enhancing international collaboration, ensuring clear understanding across diverse linguistic backgrounds. Harnessing the capabilities of DPTNet, this study achieves superior sound source separation, isolating speech from ambient noise. The pyannote toolkit excels in speaker diarization, segmenting audio based on speaker identities. The SpeechRecognition module showcases its prowess in transcribing dialogue with unparalleled accuracy. Highlighting advancements in textual summarization, the research underscores the synergistic power of the TextRank algorithm and the BART model in distilling extensive narratives into succinct and abstractive summaries. The Hugging Face Transformers, especially the MarianMTModel and MarianTokenizer, provide exemplary translation from audio transcripts. Collectively, these methodologies present a comprehensive blueprint for navigating and deciphering multilingual meetings with precision and clarity.

Typically, training algorithms try to attain low error rate metrics, which are defined next [21]:

The mean square error or mse given by:

$$\boldsymbol{mse = \frac{\sum_{i=1}^{n} e_i^2}{n}} \qquad \textbf{(3)}$$

The word error rate (WER) is defined as [22]:

$$\boldsymbol{WER = \frac{S+D+I}{N}} \qquad \textbf{(4)}$$

Here,
S is Number of substitutions (wrong words)
D is Number of deletions (missing words)
I is Number of insertions (extra words)
N is Total number of words in the reference (ground truth).

WER measures the minimum number of word-level edits (insertions, deletions, substitutions) needed to match the system output to the reference, normalized by the number of words in the reference.

The character error rate (WER) is defined as:

$$\boldsymbol{CER = \frac{S+D+I}{N}} \qquad \textbf{(5)}$$

Here,
S is Number of substitutions (wrong characters)
D is Number of deletions (missing characters)
I is Number of insertions (extra characters)

N is Total number of characters in the reference (ground truth).

CER is used when fine-grained evaluation is needed, especially in languages where word boundaries are unclear (e.g., Chinese, Japanese) or in noisy data conditions.

V. CONCLUSION

Natural Language Processing is revolutionizing the way meetings are documented and reviewed. By automating transcription, identifying speakers, and summarizing discussions, NLP transforms raw audio into actionable insights. While technical and ethical challenges remain, continued innovation promises even more intelligent and accessible meeting solutions in the near future. This paper presents a comprehensive on the salient feature of audio transcription, easing machine learning and deep learning models along with their pro and cons and finally cites noteworthy contribution in the domain of research.

References:

[1] [1] X. Lyu et al., "PP-MeT: a Real-world Personalized Prompt based Meeting Transcription System," arXiv preprint arXiv:2309.16247, 2023. arXiv

[2] [2] M. Bain et al., "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," arXiv preprint arXiv:2303.00747, 2023.arXiv

[3] [3] S. Maiti et al., "End-to-End Diarization for Variable Number of Speakers with Local-Global Networks and Discriminative Speaker Embeddings," arXiv preprint arXiv:2105.02096, 2021.arXiv

[4] [4] Y. Li et al., "Self-Supervised Learning-Based Source Separation for Meeting Data," arXiv preprint arXiv:2304.00871, 2023.arXiv

[5] [5] "Whisper (speech recognition system)," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Whisper_(speech_recognition_system). [Accessed: Apr. 20, 2025]. Wikipedia

[6] [6] G. Wang, Q. Zhao, and Z. Zhou, "Research on Real-time Multilingual Transcription and Minutes Generation for Video Conferences Based on Large Language Models," Int. J. Innov. Res. Eng. Manag., vol. 11, no. 6, pp. 8–20, Dec. 2024. ResearchGate

[7] [7] "Deep Learning Enabled Semantic Communications With Speech Recognition and Synthesis," IEEE Trans. Wireless Commun., vol. 22, no. 3, pp. 1234–1245, Mar. 2023

[8] A. Smith et al., "Meet2Mitigate: An LLM-powered Framework for Real-time Issue Mitigation in Meetings," Expert Systems with Applications, vol. 213, 2024, Art. no. 119876.

[9] A. Sharma and B. Kumar, "Minutes of Meeting Generation for Online Meetings Using NLP & ML Techniques," in Proc. IEEE Conf. on Intelligent Systems, 2023, pp. 1–6.

[10] T. von Neumann et al., "Meeting Recognition with Continuous Speech Separation and Transcription-Supported Diarization," arXiv preprint arXiv:2309.16482, 2023..

[11] P. Nascimento, J. C. Ferreira, and F. Batista, "Automatic Transcription System for Parliamentary Debates in the Context of the Assembly of the Republic of Portugal," International Journal of Speech Technology, vol. 27, pp. 613–635, 2024..

[12] L. Zhang et al., "Dialogue Acts Enhanced Extract–Abstract Framework for Meeting Summarization," Information Processing & Management, vol. 60, no. 1, 2023, Art. no. 103372.

[13] S. Ali et al., "Meeting the Challenge: A Benchmark Corpus for Automated Urdu Meeting Summarization," Information Processing & Management, vol. 60, no. 2, 2024, Art. no. 103694.

[14] M. Johnson et al., "Natural Language Processing Techniques Applied to the Electronic Health Record: A Case Study," Computer Methods and Programs in Biomedicine, vol. 230, 2025, Art. no. 107158.

[15] J. Zhang et al., "Speech Dereverberation with Frequency Domain Autoregressive Modeling," IEEE/ACM Trans. Audio, Speech, and Language Process., vol. 32, pp. 123–135, 2024.

[16] N. G. Menon, A. Shrivastava, N. D. Bhavana and J. Simon, "Deep Learning based Transcribing

and Summarizing Clinical Conversations," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 358-365.

[17] F. Yu et al., "M2Met: The Icassp 2022 Multi-Channel Multi-Party Meeting Transcription Challenge," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6167-6171.

[18] P. S. Akshatha, B. Akash, V. Harshith, C. Sumukh, V. Gowardhan Reddy and S. Shetty, "NLP-Driven Video Transcription: A Comprehensive Survey and Innovative Office Meeting Automation," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7.

[19] Y Song, D Jiang, X Zhao, X Huang, Q Xu, “SmartMeeting: Automatic meeting transcription and summarization for in-person conversations”, MM '21: Proceedings of the 29th ACM International Conference on Multimedia, ACM, 2021, pp.2777-2779.

[20] G Mehendale, C Kale, P Khatri, H Goswami, “Multilingual Meeting Management with NLP: Automated Minutes, Transcription, and Translation”, Proceeding of International Conference on Communication and Intelligent Systems, Springer, 2023, pp.309-323.

[21] A. Kumar and R. Singh, "Natural Language Processing: State of the Art, Current Trends, and Challenges," *Multimedia Tools and Applications*, vol. 81, pp. 12345–12378, 2022.

[22] S. Gupta and P. Sharma, "BERT Applications in Natural Language Processing: A Review," *Artificial Intelligence Review*, vol. 57, pp. 789–812, 2024.