Article Received: *12 August 2023* ***Revised:*** *29 September 2023* ***Accepted:*** *16 October 2023*

Audio Transcription and Summarization System using Cloud Computing and Artificial Intelligence

[1]Kaushal Rajendra Khonde, [2]Dr. Jaimeel Shah, [3]Dr. Pratik Patel
[1]Research Scholar, Computer Science & Engineering Department,
Parul Institute of Engineering & Technology, Parul University,
Vadodara, Gujarat, India
2203032010017@paruluniversity.ac.in
[2]Assistant professor, Computer Science & Engineering Department,
Parul Institute of Engineering & Technology, Parul University,
Vadodara, GujaratI India
jaimeel.shah@paruluniversity.ac.in
[3]Assistant professor, Computer Science & Engineering Department,
Parul Institute of Engineering & Technology, Parul University,
Vadodara, GujaratI India
pratik.patel2988@paruluniversity.ac.in

Abstract- In the modern era, organizations increasingly rely on virtual meetings to address customer issues promptly and effectively. However, dealing with recorded customer calls can be arduous. This review abstract introduces an innovative methodology to summarize audio data from customer interactions, which can streamline virtual meetings. Leveraging a speech recognizer, like AssemblyAI's API, the methodology converts audio data into text, and then employs a Graph-theoretic approach to generate concise summaries.

This review abstract delves into the growing prominence of cloud-based AI and ML services in the tech industry. It underscores the unique competitive strategies and focuses of major players, namely Amazon, Microsoft, and Google, in the realm of AI and ML platform development. The analysis explores these companies' internal applications and external ecosystem, dissecting their respective AI and ML development strategies. Finally, it predicts future directions for AI and ML platforms, including potential business models and emerging trends, while considering how Amazon, Microsoft, and Google align their platform development strategies with these future prospects.

Keywords- ASR, Open AI’s Whisper ASR, AWS, Punctuation Prediction, Cloud-Based AI, Machine Learning, Amazon, AI and ML Platforms, Speech Recognition, AWS Transcribe, Virtual Meeting Platforms, ChatGPT, Audio Data Summarization

I. INTRODUCTION

Developing a comprehensive software solution that seamlessly integrates with popular meeting platforms such as Google Meet, Microsoft Teams, Skype, and various others is a visionary undertaking in today's digitally driven world. The purpose of this review paper is to outline a groundbreaking software concept aimed at transforming the way we interact in meetings by harnessing cutting-edge technologies. This software will revolutionize the meeting experience by converting spoken dialogue into written text and subsequently summarizing the discussions using OpenAI's ChatGPT. In this introduction, we will explore the underlying motivations and provide an overview of the components that will be essential in the development of this transformative software.

Meetings, whether they are for business collaborations, educational sessions, or social interactions, are an integral part of our daily lives. However, as our reliance on digital communication platforms has grown, so too has the complexity of managing, organizing, and extracting valuable insights from these meetings. With the surge in remote work and the importance of efficient, accessible, and comprehensive meeting documentation, there is a pressing need for a powerful and automated solution that can facilitate these processes.

This envisioned software solution responds to this imperative by integrating AWS Transcribe, a state-of-the-art Automatic Speech Recognition (ASR) service, with OpenAI's ChatGPT, a revolutionary language model, to transform the way we conduct and analyze meetings.

AWS Transcribe, as a prominent component of this software, will serve as the backbone for transcribing the spoken words of multiple participants into written text. This technology will not only make meetings more accessible for those who may have difficulty in comprehending spoken language, such as individuals with hearing impairments, but also for anyone looking to revisit or review the content of a meeting. AWS Transcribe's remarkable accuracy and adaptability in various languages will ensure that every participant's voice is captured and converted into text effectively.

OpenAI's Whisper Model, on the other hand, offers an exceptional capability to not only convert speech to text but also to intelligently summarize the discussions. As the need for concise and actionable insights from meetings continues to grow, ChatGPT can play a pivotal role in delivering relevant summaries that capture the essence of the discussions. This summarization functionality is particularly crucial in cases where meetings involve complex topics or lengthy discussions.

Our review will also consider the challenges that may arise during the implementation of this software, including technical limitations, privacy concerns, and the need for user-friendly interfaces. Moreover, we will analyze the real-world feasibility and performance of such a system, given the ever-evolving landscape of AI, machine learning, and remote communication technologies.

II. LITERATURE REVIEW

Table 1: Analysis of reviewed approaches, Algorithm and dataset availability: -

Reference	***Author***	***Technique /Algorithm***	***Dataset***	***limitation***	***Research Scope***
Voice Recognition Systems in the Cloud Networks: Has It Reached Its Full Potential?	Asian Journal of Applied Science and Engineering, Volume 8, No 1/2019	Transfer learning and recurrent neural networks	Vast and diverse collection of voice recordings, representing various languages, accents, and speech patterns	Accuracy is still challenge in noisy or complex audio environments	Developing more user-friendly interfaces for voice recognition systems to encourage broader adoption
Artificial Intelligence and Machine Learning Capabilities and Application Programming Interfaces at Amazon, Google, and Microsoft	Boyan Liu, MIT Sloan School of Management on May 7, 2022	AWS, Microsoft, and Google have their own AI and ML platforms, covering IaaS, PaaS, and some SaaS in all aspects and providing private, public, hybrid and multi-cloud deployment methods	The cloud platforms provide more than just a development environment and infrastructure for model deployment. Model lifecycle management, observability, end-to-end security, cost management.	The cloud platforms are mostly deployed locally for security. Many platforms support low-code or even no-code operations, such as visual drag-and-drop functionality.	The platform ecosystem is also extraordinarily active, such as technology companies providing Feature stores based on cloud platforms to join more complementors.
On Speech Recognition Algorithms	Shaun V. Ault, Rene J. Perez, Chloe A. Kimble, and Jin Wang	Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), K-Means Clustering Algorithm, Expectation-Maximization (EM) algorithm	A sequence of acoustic vectors, Y = y1, y2, ..., yT. A set of acoustic vectors Y and a set of unobserved latent data or missing values Z	The Gaussian Mixture Model (GMM) prevents an HMM from taking the full advantage of the correlation that exists among the frames of a phonetic segment because it assumes a conditional independence.	Develop new language models that can better predict the next word based on the previous words and the context.
The automated method of Summarization of Audio Data	Jatin Pardhi, Tanmay Kharik Visvesvaraya National Institute of Technology	Text Summarization: Extractive text summarization using weighted frequency calculation and thresholding Graph-theoretic Approach: Cosine similarity	A set of audio recordings of call center conversations	The accuracy of the speech recognition and text summarization models can be affected by the quality of the audio recording and the complexity of the text, respectively.	Develop new text summarization approaches that can better capture the meaning of the original text, especially for complex or nuanced texts.
Development of Multilingual Speech Recognition and Translation Technologies for Communication and Interaction	Ali A. AL-Bakhrani, Gehad Abdullah Amran , Aymen M. Al-Hejri S. R. Chavan, Ramesh Manza, and Sunil Nimbhore	Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are used for speech recognition and translation.	A large dataset of audio recordings and text transcripts is required to train a deep learning model for speech recognition and translation.	Deep learning models can be computationally expensive to train and require a large dataset of training data.	Research is ongoing to improve the accuracy and efficiency of deep learning models for speech recognition and translation.

Article Received: *12 August 2023* ***Revised:*** *29 September 2023* ***Accepted:*** *16 October 2023*

Speech to text and text to speech recognition systems - A Review	Ayushi Trivedi,NavyPant, Pinal Shah,Simran Sonik and Supriya Agrawal DepartmentofComputerScience, NMIMS University,Mumbai, India.	Speech recognition: pre-emphasis of signals, feature extraction, and recognition of the signals using Hidden Markov Models (HMMs) Speech to text conversion: Dynamic Time Warping (DTW) and HMMs, Neural Network models, End-to-end ASR	Large datasets of audio recordings and text transcripts are required to train speech recognition and machine translation models.	Speech recognition and machine translation models can be computationally expensive to train and require a large dataset of training data.	Research is ongoing to improve the accuracy and efficiency of deep learning models for speech recognition and translation. For example, researchers are developing new deep learning models that can be trained on smaller datasets and that can operate in real time.
Speech-To-Text Software Design for the High Education Learning	Elena Yashina, Tetiana Rubanik and Andriy Chukhray National Aerospace University Kharkiv Aviation Institute, Chkalova St., 17, 61070 Kharkiv, Ukraine	Hidden Markov Model (HMM), Deep Neural Networks (DNNs)	FLEURS - 102-language dataset for evaluating few-shot learning of speech representations.	Investigating new applications for speech recognition, such as in education, healthcare, and customer service. Speech recognition has the potential to revolutionize the way we interact with computers and devices.	Robust speech recognition: This involves developing speech recognition systems that are robust to noise, accents, and other variations in speech. This is important for making speech recognition systems more practical for real-world use.
Automatic speech recognition: system variability within a sociolinguistically homogeneous group of speakers	Lauren Harrington, Vincent Hughes. University of York. Version Accepted	JiWER (Python package), Amazon Transcribe	DyViS task 2	Small dataset, only considers SSBE speakers	Extend to more speakers and languages, investigate other factors that may affect ASR performance
Speaker Diarization and Identification From Single Channel Classroom Audio Recordings Using Virtual Microphones	ANTONIO GOMEZ, (Senior Member, IEEE), MARIOS S. PATTICHIS, (Senior Member, IEEE), AND SYLVIA CELEDÓN-PATTICHIS	Room Impulse Responses (RIRs) at the virtual microphones are then estimated using acoustic scene simulations. The RIRs are then used to compute a cross-correlation matrix of possible audio sources.	Evaluated on a dataset of classroom audio recordings with 2 to 5 speakers. The recordings were made in a noisy environment with significant crosstalk.	The method requires a rough estimate of the speaker geometry, which can be derived from video recordings. However, the method is not limited by the spoken language or accent of the participants and works well in noisy environments.	The method could be extended to identify speakers in other noisy environments, such as meeting rooms and conference halls. The method could also be used to identify multiple speakers speaking simultaneously to the same microphone.
Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization	Matthew Marge Satanjeev Banerjee Alexander I. Rudnicky School of Computer Science, Carnegie Mellon University Pittsburgh, PA 15213, USA	Amazon Transcribe, A word-level minimum edit distance metric is then used to align the two transcripts and locate disagreements.	The data were sampled from a previously-collected corpus of natural meetings.	The second-pass transcription task was more difficult than the first-pass transcription task. 30% of turkers indicated that the second-pass correction task was difficult, as compared with 15% for the first-pass transcription task.	Proposed corrective strategy could be used to improve the accuracy of transcriptions for other types of speech, such as lecture speech, broadcast news, or telephone conversations.

Speech Recognition in Background Noise by a Convolution Neural Networks Model	Adel Kabirikopaei and Fabio Vitor Department of Mathematics, University of Nebraska at Omaha Omaha, Nebraska, USA	Convolution neural network (CNN) that encodes the audio signal over characters. For each frame, the CNN gives a distribution of different characters. The decoder then chooses the most likely sequence of characters.	The "Voices Obscured in Complex Environmental Settings" (VOiCES) dataset is used to train and evaluate the proposed model. This dataset contains clean speech recorded in different rooms with background noise played concurrently.	The proposed model is only trained and evaluated on the VOiCES dataset, so it is not clear how well it would generalize to other datasets.	The model could be extended to handle other types of noise, such as music or street noise. It could also be evaluated on other datasets to see how well it generalizes.
SummIt: Iterative Text Summarization via ChatGPT	Haopeng Zhang Xiao Liu Jiawei Zhang IFM Lab, Department of Computer Science, University of California, Davis, CA, USA	Large language models like ChatGPT. It enables the model to refine the generated summary iteratively through self-evaluation and feedback, resembling humans' iterative process when drafting and revising summaries.	The experiments are conducted on three benchmark summarization datasets, namely CNN/Daily Mail, XSum, and TACoS	Evaluated on three benchmark summarization datasets, so it is not clear how well it would generalize to other datasets	Could be integrated with other natural language processing tasks, such as question answering and machine translation.
Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization	Yan Li Xinlu Zhang Haifeng Chen University of California, Santa Barbara Microsoft NEC Laboratories America	ChatGPT, a large language model, is evaluated for aspect- and query-based text summarization.	Four publicly available datasets are used in the evaluation: Reddit posts, news articles, dialogue meetings, and stories.	The evaluation is limited to 100 examples selected at random from each test set, due to the lack of an API provided by ChatGPT for processing large amounts of input data.	It could be conducted to evaluate ChatGPT on a larger dataset and to develop methods for improving the quality of LLM-generated summaries.
A Review on Automatic Speech Recognition Architecture and Approaches	Karpagavalli Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore 641 004, India. Chandra Department of Computer Science, Bharathiar University, Coimbatore 641 046, India	A typical speech recognition system is developed with major components that include acoustic front-end, acoustic model, lexicon, language model and decoder.	Training data for speech recognition systems typically consists of recordings of human speech, along with transcripts of the spoken words	Speech recognition systems can be sensitive to noise and background conditions.	It could improve the accuracy and robustness of speech recognition systems, as well as developing new applications for speech recognition technology.

An Iterative Dual Pathway Structure for Speech-to-Text Transcription	Beatrice Liem Haoqi Zhang Yiling Chen School of Engineering and Applied Sciences Harvard University Cambridge, MA 02138 USA	An iterative dual pathway structure for speech-to-text transcription. In this structure, participants in either path can iteratively refine the transcriptions of others in their path while being rewarded based on transcriptions.	Uses audio recordings of Harvard undergraduate students as the dataset	Does not evaluate the proposed method on a large and diverse dataset.	Potential to improve the accuracy of speech-to-text transcription by using more sophisticated algorithms to divide audio files into segments.
Is automatic speech-to-text transcription ready for use in psychological experiments?	Kirsten Ziman, Andrew C. Heusser, Paxton C. Fitzpatrick, Campbell E. Field, Jeremy R. Manning	A state-of-the-art speech recognition algorithm (Google Cloud Speech API) to automatically transcribe audio data.	The dataset consisted of 240 audio recordings of participants' verbal responses in a list-learning experiment.	Research is needed to evaluate the performance of speech recognition algorithms on a larger and more diverse dataset.	Evaluate the performance of speech recognition algorithms on a larger and more diverse dataset, and investigate the potential to improve the accuracy of speech-to-text transcription by using more sophisticated algorithms to divide audio files into segments.

III. METHODOLOGY

1. Data Collection and Preprocessing:
Virtual meetings like skype, google meet, microsoft teams, jio meet etc for real time data collection. Collect a diverse dataset of audio recordings representing different languages, accents, and speaking styles. Ensure data quality by removing noise, enhancing audio quality, and preparing the data for ASR model training.

2. Selection of ASR Services:
Choose ASR services from Amazon Transcribe, Google Cloud Speech-to-Text, Microsoft Bing Speech to Text, Deepgram, Azure Cognitive Speech Recognition, and IBM Watson Speech to Text based on your project requirements, cost considerations, and service-specific features.

3. Data Annotation:
Annotate the collected audio data by transcribing them into text. These transcriptions will serve as ground truth for training and evaluation.

4. Model Training and Fine-Tuning:
For each selected ASR service, train models using the annotated data. Fine-tune the models to adapt them to the specific domain, if required.

5. Evaluation and Benchmarking:
Evaluate the performance of each ASR service using established metrics such as Word Error Rate (WER) and Character Error Rate (CER). Benchmark the services in terms of accuracy, speed, and cost.

6. Integration with Cloud Computing:
Leverage cloud computing infrastructure to deploy the selected ASR services. Set up auto-scaling and load balancing to handle varying workloads efficiently.

7. API Integration:
Utilize the APIs provided by the chosen ASR services to integrate them into your applications. Ensure compatibility with different programming languages and platforms.

8. Real-Time Processing:
Implement real-time speech recognition by using the streaming capabilities provided by the cloud services. Optimize the processing pipeline to reduce latency.

9. Language Support and Multilingual Processing:
Explore the language support of the ASR services and ensure they can handle multiple languages and dialects. Implement language detection for multilingual applications.

10. Security and Compliance:
Address security concerns by implementing encryption, authentication, and authorization mechanisms. Ensure compliance with data protection regulations, especially when dealing with sensitive data.

11. Cost Management:
Monitor and manage the costs associated with the usage of cloud-based ASR services. Optimize resource allocation and usage to minimize expenses.

12. Error Handling and Feedback Loop:
Implement robust error handling to manage cases where ASR services may provide inaccurate transcriptions. Create a feedback loop to continuously improve the ASR system.

13. Continuous Monitoring and Maintenance:
Monitor the performance of the ASR services in real-world scenarios. Perform regular updates and maintenance to keep the system up-to-date.

14. Scalability and Redundancy:
Ensure the system is designed for scalability and accommodates increased demand. Implement redundancy and failover mechanisms to guarantee high availability.

15. ChatGPT Summarization:
Implement an additional module for chat-based summarization using GPT-3 or a similar language model. This module should take the transcribed meeting content and generate concise summaries of the discussion. Utilize GPT-3's natural language understanding to create coherent and context-aware meeting summaries.

16. Integration with Transcription:
Ensure seamless integration between the ASR system and the ChatGPT summarization module. Transcribed content should be passed to the summarization module for generating meeting summaries.

17. Summarization Parameters:
Define parameters for summarization, such as desired length or level of detail for the meeting summary. Users may want brief summaries or more detailed ones based on their preferences.

18. Summarization Delivery:
Provide users with options for receiving meeting summaries, including email delivery. Users should be able to choose when and how they want to receive the summaries.

19. Summarization Accuracy Monitoring:
Continuously monitor the accuracy and coherence of the generated summaries. Implement mechanisms to address and learn from any summarization errors or inconsistencies.

20. User Feedback Loop:
Establish a feedback loop where users can provide input on the quality and usefulness of the meeting summaries. Use this feedback to improve the summarization module over time.

21. Documentation and Reporting:
Maintain comprehensive documentation of the ASR integration process, including setup, configurations, and troubleshooting steps. Generate regular reports on system performance and usage.

IV. PROPOSED ALGORITHM

1. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is an algorithm that converts spoken language into text. ASR algorithms are used in a variety of applications, such as voice transcription, voice assistants, and speech-to-text translation.

ASR algorithms typically work by first extracting features from the audio signal. These features may include the pitch, intensity, and duration of the sound. The ASR algorithm then uses these features to identify the words that were spoken.

There are two main types of ASR algorithms: statistical and machine learning algorithms. Statistical ASR algorithms use statistical models to predict the most likely sequence of words that were spoken. Machine learning ASR algorithms use machine learning models to predict the most likely sequence of words that were spoken.

Machine learning ASR algorithms are generally more accurate than statistical ASR algorithms. However, they require more data to train and can be more computationally expensive to run.

ASR is used in a variety of applications, including:

Voice transcription: ASR can be used to transcribe audio recordings into text. This can be useful for creating transcripts of meetings, lectures, and other events.

Voice assistants: ASR is used in voice assistants such as Siri, Alexa, and Google Assistant. Voice assistants allow users to control devices and access information using spoken language.

Speech-to-text translation: ASR can be used to translate spoken language into text. This can be useful for translating conversations between people who speak different languages.

2. Amazon Transcribe

Amazon Transcribe is a cloud-based automatic speech recognition (ASR) service that converts spoken audio into text. It uses a deep learning model to transcribe audio in real time or in batch mode. Amazon Transcribe supports over 87 languages and dialects, and it can be used to transcribe a variety of audio sources, including meetings, lectures, podcasts, and customer service calls.

Amazon Transcribe works by first extracting features from the audio signal, such as the pitch, intensity, and duration of the sound. The ASR model then uses these features to predict the most likely sequence of words that were spoken.

Amazon Transcribe is trained on a massive dataset of transcribed audio, which allows it to achieve high accuracy transcription. Amazon Transcribe also uses a variety of techniques to improve the accuracy of transcription, such as custom vocabulary support and speaker identification.

Amazon Transcribe can be used for a variety of purposes, including:

Voice transcription: Amazon Transcribe can be used to transcribe audio recordings into text. This can be useful for creating transcripts of meetings, lectures, podcasts, and customer service calls.

Real-time transcription: Amazon Transcribe can be used to transcribe audio in real time. This can be useful for providing real-time transcripts to viewers of a live stream or to participants in a meeting.

Translation: Amazon Transcribe can be used to translate transcribed audio into another language. This can be useful for translating audio recordings of meetings, lectures, and podcasts into a language that is more accessible to a wider audience.

To use Amazon Transcribe in ASR application, you will need to create an AWS account and obtain an AWS access key ID and secret access key. You will also need to create a Transcribe service role.

Once you have created an AWS account and obtained an AWS access key ID and secret access key, you can start using Amazon Transcribe in your software. To do this, you will need to install the AWS SDK for your programming language.

The AWS SDK provides a variety of APIs for accessing Amazon Transcribe services. You can use these APIs to start and stop transcription jobs, get the results of transcription jobs, and translate transcribed audio into another language.

Some additional considerations for using Amazon Transcribe in your software:

Custom vocabulary support: If your audio recordings involve specialized topics or jargon, you can improve the accuracy of transcription by using a custom vocabulary. To create a custom vocabulary, you will need to create a list of words and phrases that Amazon Transcribe should be aware of.

Speaker identification: Amazon Transcribe can be used to identify the speaker of each utterance in the transcribed audio. This can be useful for displaying the speaker's name in the transcript or for addressing the speaker directly.

Pricing: Amazon Transcribe is a pay-as-you-go service. You are charged based on the amount of audio that you transcribe.

3. OpenAI's Whisper Model

OpenAI's Whisper model is a large language model (LLM) that is trained on a massive dataset of text and audio. Whisper can perform a variety of tasks, including speech recognition, translation, and summarization.

Whisper is trained using a technique called supervised learning. This means that Whisper is given a set of training data that includes both the audio and the corresponding text transcript. Whisper learns to predict the text transcript from the audio by analyzing the features of the audio signal.

Whisper uses a variety of techniques to improve the accuracy of speech recognition, including:

Attention: Whisper uses attention to focus on the most important parts of the audio signal. This helps Whisper to distinguish between different words and to recognize words in noisy environments.

Context: Whisper uses context to predict the next word in a sentence. This helps Whisper to recognize words that are difficult to hear or that are not clearly pronounced.

Language model: Whisper uses a language model to predict the next word in a sentence based on the words that have already been spoken. This helps Whisper to recognize words that are ambiguous or that have multiple meanings.

OpenAI's Whisper model can be used for a variety of purposes, including:

Speech recognition: Whisper can be used to transcribe audio recordings into text. This can be useful for creating transcripts of meetings, lectures, podcasts, and customer service calls.

Translation: Whisper can be used to translate transcribed audio into another language. This can be useful for translating audio recordings of meetings, lectures, and podcasts into a language that is more accessible to a wider audience.

Summarization: Whisper can be used to summarize transcribed audio. This can be useful for creating summaries of meetings, lectures, and podcasts.

To use OpenAI's Whisper model in your software, you will need to create an OpenAI account and obtain an API key. You can then use the OpenAI API to submit audio recordings to Whisper for transcription.

The OpenAI API provides a variety of methods for transcribing audio with Whisper. You can use these methods to transcribe audio in real time, to transcribe audio in batch mode, and to translate transcribed audio into another language.

Some additional considerations for using OpenAI's Whisper model in your software:

Custom vocabulary support: Whisper can be used with a custom vocabulary to improve the accuracy of transcription for specialized topics or jargon. To create a custom vocabulary, you will need to create a list of words and phrases that Whisper should be aware of.

Speaker identification: Whisper can be used to identify the speaker of each utterance in the transcribed audio. This can be useful for displaying the speaker's name in the transcript or for addressing the speaker directly.

Pricing: OpenAI's Whisper model is a paid service. You are charged based on the amount of audio that you transcribe.

V. ANALYSIS AND PERFORMANCE OF THE CLOUD COMPUTING ASR SERVICES

Amazon Transcribe is a cloud-based automatic speech recognition (ASR) service that converts spoken audio into text. It is highly accurate and supports over 87 languages and dialects. Amazon Transcribe is also scalable, so it can handle large volumes of audio data.

Google Cloud Speech-to-Text is another cloud-based ASR service that is highly accurate and supports over 122 languages and dialects. It is also scalable and can handle large volumes of audio data. Google Cloud Speech-to-Text also offers a number of features that make it well-suited for businesses, such as custom vocabulary support and speaker identification.

Microsoft Bing Speech to Text is a cloud-based ASR service that is highly accurate and supports over 90 languages and dialects. It is also scalable and can handle large volumes of audio data. Microsoft Bing Speech to Text also offers a number

***Article Received:** 12 August 2023 **Revised:** 29 September 2023 **Accepted:** 16 October 2023*

of features that make it well-suited for businesses, such as custom vocabulary support and speaker identification.

Deepgram is a cloud-based ASR service that is highly accurate and supports over 72 languages and dialects. It is also scalable and can handle large volumes of audio data. Deepgram also offers a number of features that make it well-suited for businesses, such as custom vocabulary support, speaker identification, and real-time transcription.

Azure Cognitive Speech Recognition is a cloud-based ASR service that is highly accurate and supports over 70 languages and dialects. It is also scalable and can handle large volumes of audio data.

IBM Watson Speech to Text is a cloud-based ASR service that is highly accurate and supports over 42 languages and dialects. It is also scalable and can handle large volumes of audio data. IBM Watson Speech to Text also offers a number of features that make it well-suited for businesses, such as custom vocabulary support, speaker identification, and real-time transcription.

Table 2: Analysis and performance summary of cloud computing ASR services: -

Technique	**Task**	**Dataset**	**Results**
Amazon Transcribe	Automatic speech recognition	Various datasets, including meetings, lectures, and podcasts	High accuracy transcription of audio files into text, with support for custom vocabulary and real-time transcription.
Google Cloud Speech-to-Text	Automatic speech recognition	Various datasets, including meetings, lectures, and podcasts	High accuracy transcription of audio files into text, with support for custom vocabulary and real-time transcription.
Microsoft Bing Speech to Text	Automatic speech recognition	Various datasets, including meetings, lectures, and podcasts	High accuracy transcription of audio files into text, with support for custom vocabulary and real-time transcription.
Deepgram	Automatic speech recognition	Various datasets, including meetings, lectures, and podcasts	High accuracy transcription of audio files into text, with support for custom vocabulary, real-time transcription, speaker identification, and other features.
Azure Cognitive Speech Recognition	Automatic speech recognition	Various datasets, including meetings, lectures, and podcasts	High accuracy transcription of audio files into text, with additional features such as speaker identification and real-time transcription.
IBM Watson Speech to	Automatic speech	Various datasets, including	High accuracy transcription of audio files into text, with support for custom vocabulary, real-

Text	recognition	meetings, lectures, and podcasts	time transcription, speaker identification, and other features.
OpenAI Whisper	Automatic speech recognition	Various datasets, including meetings, lectures, and podcasts	High accuracy transcription of audio files into text, with support for multiple languages and dialects, as well as speaker identification and translation.

VI. CONCLUSION

This paper has presented a novel software solution that integrates AWS Transcribe with OpenAI's ChatGPT to generate meeting summaries. The proposed solution addresses the limitations of existing meeting summary tools by providing a comprehensive and accurate summary that is tailored to the specific needs of the users.

The proposed solution works by first using AWS Transcribe to generate a transcript of the meeting audio. The transcript is then passed to ChatGPT, which uses its natural language processing capabilities to generate a summary of the meeting. The summary can be customized to include specific information, such as the meeting agenda, key takeaways, and action items.

The proposed solution has the potential to revolutionize the way we conduct and document meetings. By providing users with a concise and informative summary of the meeting, the solution can help to save time, improve communication, and boost productivity.

The proposed solution also offers a number of other advantages, including:

Scalability: The solution can be scaled to meet the needs of organizations of all sizes.

Flexibility: The solution can be customized to meet the specific needs of each organization

Security: The solution uses secure cloud-based services to protect user data.

The proposed solution is still under development, but it has the potential to become a valuable tool for businesses and organizations of all sizes.

Potential impact of the proposed solution:

Improved meeting productivity: The proposed solution can help to improve meeting productivity by providing users with a concise and informative summary of the meeting. This can help users to quickly identify the key takeaways from the meeting and to focus on the most important action items.

Enhanced communication: The proposed solution can help to enhance communication between meeting participants by providing a common reference point for the meeting discussion. This can help to reduce misunderstandings and to ensure that everyone is on the same page.

Increased transparency: The proposed solution can help to increase transparency in meetings by providing a record of the meeting discussion. This can help to build trust between meeting participants and to ensure that everyone is accountable for their actions.

Overall, the proposed solution has the potential to significantly improve the way we conduct and document meetings. I am excited to see how this solution is developed and implemented in the future

References

[1] Anusha Bodepudi, Manjunath Reddy, Sai Srujan Gutlapalli, Mounika Mandapuram ”Voice Recognition Systems in the Cloud Networks: Has It Reached Its Full Potential?”in Asian Journal of Applied Science and Engineering, Volume 8, No 1/2019

[2] Artificial Intelligence and Machine Learning Capabilities and Application Programming Interfaces at Amazon, Google, and Microsoft by Boyan Liu Submitted to MIT Sloan School of Management on May 7, 2022 in Partial Fulfillment of the requirements for the Degree of Master of Science in Management Studies.

[3] On Speech Recognition Algorithms by Shaun V. Ault, Rene J. Perez, Chloe A. Kimble, and Jin Wang in International Journal of Machine Learning and Computing, Vol. 8, No. 6, December 2018

[4] The automated method of Summarization of Audio

Data by Jatin Pardhi ofVisvesvaraya National Institute of Technology and Tanmay Kharik of Visvesvaraya National Institute of Technology in research square April 22nd, 2022

[5] Development of Multilingual Speech Recognition and Translation Technologies for Communication and Interaction

by Ali A. AL-Bakhrani1 , Gehad Abdullah Amran , Aymen M. Al-Hejri, S. R. Chavan, Ramesh Manza, and Sunil Nimbhore

[6] Speech to text and text to speech recognition systems-A Review by Ayushi Trivedi,Navya Pant, Pinal Shah, Simran Sonik and Supriya Agrawal Department of Computer Science, NMIMS University, Mumbai,India. In Corresponding Author:Navya Pant in IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 2, Ver. I (Mar.- Apr. 2018), PP 36-43 www.iosrjournals.org

[7] Speech-To-Text Software Design for the High Education Learning by Elena Yashina, Tetiana Rubanik and Andriy Chukhray In National Aerospace University Kharkiv Aviation Institute, Chkalova St., 17, 61070 Kharkiv, Ukraine

[8] Automatic speech recognition: system variability within a sociolinguistically homogenous group of speakers. by Harrington, Lauren and Hughes, Vincent orcid.org/0000-0002-4660-979X (2023) Automatic speech recognition : system variability within a sociolinguistically homogeneous group of speakers. In: Proceedings of the International Congress of Phonetic Sciences (ICPhS) International Congress of Phonetic Sciences, 07-11 Aug 2023 , CZE .White Rose University of York

[9] Speaker Diarization and Identification From Single Channel Classroom Audio Recordings Using Virtual Microphones ANTONIO GOMEZ 1, (Senior Member, IEEE), MARIOS S. PATTICHIS, (Senior Member, IEEE), AND SYLVIA CELEDÓN-PATTICHIS in IEEE Access

[10] Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization by Matthew Marge Satanjeev Banerjee Alexander I. Rudnicky, School of Computer Science, Carnegie Mellon University Pittsburgh, PA 15213, USA

[11] Speech Recognition in Background Noise by a Convolution Neural Networks Model by Adel Kabirikopaei and Fabio Vitor Department of Mathematics, University of Nebraska at Omaha, Nebraska, USA Proceedings of the 2021 IISE Annual Conference A. Ghate, K. Krishnaiyer, K. Paynabar, eds.

[12] SummIt: Iterative Text Summarization via ChatGPT by Haopeng Zhang, Xiao Liu, Jiawei Zhang IFM Lab, Department of Computer Science, University of California, Davis, CA, USA, arXiv:2305.14835v2 [cs.CL] 9 Oct 2023

[13] Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization by Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, University of California, Santa Barbara Microsoft NEC Laboratories America. arXiv:2302.08081v1 [cs.CL] 16 Feb 2023

[14] A Review on Automatic Speech Recognition Architecture and Approaches by Karpagavalli Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore 641 004, India and Chandra, Department of Computer Science, Bharathiar University, Coimbatore 641 046, India In International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.9, No.4, (2016), pp.393-404

[15] An Iterative Dual Pathway Structure for Speech-to-Text Transcription by Beatrice Liem ,Haoqi Zhang, and Yiling Chen, School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA. bliem@post.harvard.edu{hq,yiling}@eecs.harvard.ed

[16] Is automatic speech-to-text transcription ready for use in psychological experiments? by Kirsten Ziman, Andrew C. Heusser, Paxton C. Fitzpatrick, Campbell E. Field, Jeremy R. Manning, © Psychonomic Society, Inc. 2018