A review on techniques for speech to text conversion and text summarization

TO be done:

Result paraphrase

Discussion 4.1 4.2 , tables paraphrase

Paraphase as much as possible

if we need to add more diagram, pls make them (if possible) and send them to me

Im tried adding formula where it was relevant not sure about any other place

This table is temporary for overview

Table of Contents

[1. Abstract 2](#_Toc180533977)

[2. Introduction 2](#_Toc180533978)

[3. Methods 3](#_Toc180533979)

[3.1 Cloud based 3](#_Toc180533980)

[3.2 Neural Network 4](#_Toc180533981)

[Long Short-Term Memory (LSTMs) 4](#_Toc180533982)

[Recurrent Neural Networks (RNN) 4](#_Toc180533983)

[Extractive Summarization 5](#_Toc180533984)

[3.3 Transformer based 5](#_Toc180533985)

[Generative Adversarial Networks (GANs) 5](#_Toc180533986)

[Two-Stage Transformer 6](#_Toc180533987)

[Dual-encoding using transformer 6](#_Toc180533988)

[3.4 PodSumm Architecture 7](#_Toc180533989)

[4. Discussion 7](#_Toc180533990)

[4.1 Speech to Text conversion 7](#_Toc180533991)

[4.2 Text summarization 9](#_Toc180533992)

[5. Results 10](#_Toc180533993)

[References 12](#_Toc180533994)

# 1. Abstract

Speech-to-text and text summarization are two key areas, which are part of automatic natural language processing, which have been very useful in recent times as there has been an upsurge in the amount of digitized information as well as the need for efficient transcription and distilling information into its essence. This review paper evaluates the conventional approaches and structural patterns registrable in the recent innovations across both opuses with a keen interest on their complexities, centricity, and space of application. The first issue addressed is related to the transcription of speech into text. These include deep learning and end to end models like transformer networks for neural network based speech transcription as well as cloud based services. A number of strengths and weaknesses of these methods are assessed relative to precision, time, and flexibility towards different languages and trends within same languages. Then the paper moves on to examine the text summarisation techniques. Two types of text summarization techniques are compared, namely extractive and abstractive text summarization, emphasizing advances by transformer models such as BERT and GPT in generating context sensitive summaries. The last focus of the paper is on the speech recognition and summation technologies in practice as used by virtual assistants, meeting transcription applications, and content generation tools.

keywords

# 2. Introduction

The introduction of artificial intelligence (AI) and natural language processing (NLP) tools has revolutionized the way content is consumed. Speech-to-text transformation and text summarization stand out as linguistics technologies that are pivotal to effective interaction and access to information in the present global village. With the emergence of various voice-based, automatic transcription and text saturation over usage, the need for efficient speech-to-text translation and text summarization has grown greatly. The speech recognition technology such as ASR is of great importance since it encodes spoken words and phonemes into textual representation, hence supporting the ability to search and utilize voice data on various platforms. Before, systems ASR were based on statistical approaches of the analysis of the speech signal, such as Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) – which they used together with the language and acoustic models to derive a reasonable accuracy. Nevertheless, with the development of deep learning and end-to-end neural models, ASR has evolved and now these techniques offer higher accuracy rates and an ability to recognize varying accents, languages, and contexts more effectively. Text summarization uses the information available in long texts to produce an overview that is shorter in terms of content and wording. It can also either be extractive which selects certain key sentences in the original text or be abstractive where a new summary is created that states the main point of the original text. Furthermore, in recent years transformer-based models such as BERT and GPT have successfully performed both extractive and abstractive types of text summarization achieving new heights in coherence and target relevance. This review paper aims to provide a comprehensive overview of current speech-to-text and text summarization methods, examining their underlying technologies, strengths, limitations, and real-world applications. By exploring these interrelated fields, we highlight the latest innovations and identify future directions to improve the integration of these technologies in diverse industries.

# 3. Methods

The methods for speech to text conversion and text summarization can be divided into categories such as cloud based, Neural Network based and Transformer based. Using the combination of these approaches multiple architectures are created. Some of these architectures are PodSumm architecture [11] and BertSumm (transformer) architecture

## 3.1 Cloud based

Cloud-based speech-to-text methods leverage advanced machine learning models hosted on cloud platforms to convert spoken language into text. Services like **Google Cloud Speech-to-Text**, **Amazon Transcribe**, and **Microsoft Azure Speech Service** provide scalable solutions that support real-time and batch transcription, automatic punctuation, speaker identification, and multiple language support. These platforms are equipped with state-of-the-art deep learning models, ensuring high accuracy and performance. Additionally, cloud-based services eliminate the need for local infrastructure, offering accessibility from anywhere, easy integration into applications, and the ability to handle large volumes of audio data efficiently.

Though the cloud based methods are effective over the time for the production use the cost of the services may add up and cost heavily with increase in the number of the consumers. Cloud based methods are best used for a small scale project. This methods are easy to setup and implement due to them being heavily abstracted.

## 3.2 Neural Network

### Long Short-Term Memory (LSTMs)

The paper [2] dives into the design and usefulness of LSTM systems, emphasizing their capacity to capture successive conditions in content information Through an examination of context-based text generation techniques, the author showcases the efficacy of LSTM networks in generating coherent and contextually relevant textual content. text generation models, highlighting their significance in tasks such as machine translation, dialogue systems, and summarization. Furthermore, the author presents experimental results and performance evaluations to demonstrate the effectiveness of LSTM networks in context-based text generation, paving the way for advancements in natural language processing and artificial intelligence research." Through a detailed examination of context-based text generation techniques, the author demonstrates the effectiveness of LSTM networks in generating coherent and contextually relevant textual output. The article discusses strategies for integrating contextual cues into LSTM-based models, showcasing their applicability in diverse domains such as machine translation, dialogue generation, and summarization.

### Recurrent Neural Networks (RNN)

The paper [1] and [3] present us with structure of RNN network and its various implementations such as RNN Transducer. RNNs are ideal in tasks such as speech to text and text summarization RNNs have a unique perspective on speech because they continuously strive to convert audio features into text sequences. This enables them to transform audio features into textual forms. Such an ability is advantageous as it allows a network to understand context in the spoken language. RNNs are therefore able to compensate for varying patterns of speech and pronunciation accents that may exist across different individuals. These transcribers are more effective for RNNs’ greater transcription success increases their effectiveness in a broad range of applications including voice-command mechanics and automated transcription systems.

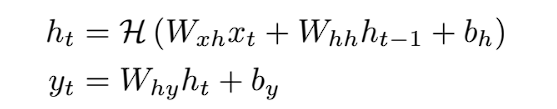
Deeps RNNs also consists of various layers such as Connectionist Temporal Classification, RNN transducer which are decoded by a decoding function and its output is regularised using Regularization process. The paper experiment on various combinations of the layer to find the most optimal and low error producing layer.

Figure 1 Hidden layer function in RNN

### Extractive Summarization

The authors of paper [4] utilizes Extractive summarization [8] to generate a text summary of the input text. The input text is pre-processed to clean and tokenize the data, followed by the extraction of features such as word embeddings, which represent the semantic meaning of words in a numerical format. Next, a neural network model, such as a Recurrent Neural Network (RNN) or a Long Short-Term Memory (LSTM) network, is trained on labelled data to learn the importance of sentences based on their context and content. During inference, the model evaluates each sentence, assigning a score that reflects its significance in the overall text. Sentences with the highest scores are then selected and combined to form a coherent summary, ensuring that the key information from the original text is retained while reducing redundancy. Finally, the selected sentences are organized and formatted to produce the final summary.

## 3.3 Transformer based

### Generative Adversarial Networks (GANs)

The overview [2] comprehensively explores the application of GANs in creating literary substance, highlighting their potential over different spaces. The GANs are built over the top of transformers [6]. By analyzing the basic standards and designs of GANs in content era, the overview illustrates the points of interest and challenges related with this approach. Assessment strategies particular to GAN-based content era, along with relevant evaluation measurements, are examined to evaluate the quality and coherence of created content. Moreover, the study addresses eminent progressions, rising patterns, and future investigate bearings in leveraging GANs for text generation, underscoring their importance in progressing the field of common language preparing. The overview fastidiously analyzes the utilization of GANs over a range of text generation assignments, counting language modeling, dialogue generation, and story generation. By dissecting the engineering and preparing methods of GANs in text generation. The study explains the challenges and openings characteristic in this approach.

### Two-Stage Transformer

In the paper [5], the authors proposed and created a two-stage approach using the transformers [6] for variable-length summary has been proposed. Initially, the text segmentation module makes use of a pre-trained BERT and a bidirectional LSTM network to segment the input text. Next, BERTSUM or the BERT based extractive summarization model is constructed in order to perform the essential task of extracting the most salient sentence from each segment. With respect to training the two-stage summation model, the first step would involve employing the extracted sentences to train or language build the document summarization model depicted in the second stage. Thereafter, the extracted segments are used to train the specific summarisation module depicted in the first stage with the consideration of the outputs of this module and the documents summarisation module from the second stage. The parameters defining the segment summarization module are perturbed based upon the cumulative loss scores of the document summarization and segment summarisation modules. After this stage of training, an alternate training method termed as collaborative training is performed in which the parameters defining the segment summarisation and documents summarisation modules are trained repeatedly until the parameters converge. To test the framework the outputs from the segment summarization module are integrated, to produce variable-length, one-shot, abstractive summary. For evaluation, the BERT-biLSTM-based text segmentation model is evaluated using ChWiki\_181k database and obtains a good effect in capturing the relationship between sentences. Finally, the proposed variable-length abstractive summarization system achieved a maximum of 70.0% accuracy in human subjective evaluation on the LCSTS dataset.

### Dual-encoding using transformer

Recurrent neural network-based sequence-to- sequence attentional models have proven effective in abstractive text summarization [8]. In the paper [7], the author has proposed a model for abstractive text summarization using a dual encoding. The proposed method employs a dual encoder including the primary and the secondary encoders. Specifically, the primary encoder conducts coarse encoding in a regular way, while the secondary encoder models the importance of words and generates more fine encoding based on the input raw text and the previously generated output text summarization. The two level encodings are combined and fed into the decoder to generate more diverse summary that can decrease repetition phenomenon for long sequence generation. The experimental results on two challenging datasets (i.e., CNN/DailyMail and DUC 2004) demonstrate that our dual encoding model performs against existing methods.

## 3.4 PodSumm Architecture

PodSumm first generates a transcript of the podcast audio using an ASR module and parses the text transcript into individual sentences [10]. It then uses a text summarization model to select relevant sentences, along with their time offsets in the audio, and generates the final audio summary associated with the text summary. Each stage is discussed in detail below. 1) Automatic Speech Recognition: ASR methods perform the task of automatic speech-to-text transcription. As the purpose of this work is not to develop a new ASR system or improve on an existing one, we choose to use a well- known and publicly-available solution for this task, namely AWS Transcribe. 2) Text Processing: The transcripts obtained from the ASR module contain the text for the individual words and punctuation marks, their start and end times in the audio, and their confidence scores regarding the prediction. The author choose to use an open-source library for NLP, namely spaCy3, to parse the text into individual sentences with their corresponding start and end times. Additionally, they force a sentence break when a pause of over two seconds between words occurs. 3) Text Summarization: We generate text summaries by selecting relevant sentences from the transcripts, using automatic extractive summarization. They used the recently-proposed Pre- Summ4 model [10], which builds upon BERT [10] to obtain a sentence level encoding, and stacks inter-sentence Transformer layers to capture document-level features for summarization. The final trained model showed the best performance (ROUGE-L F-score of 0.64) in comparison with the baseline method (ROUGE-L F-score of 0.52)

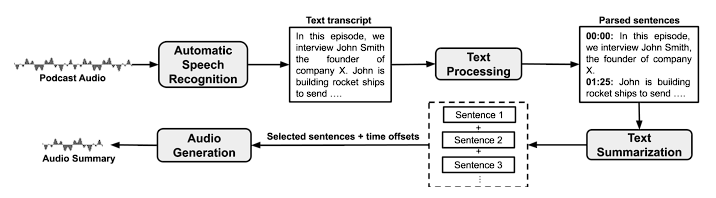


Figure 2 PodSumm Architecture

# 4. Discussion

## 4.1 Speech to Text conversion

In the following table, a comparative discussion is given about various approaches in speech to text and we have discussed major landmark achieved in this field as well as significant advantages and disadvantages to each approach. For example Hidden Markov Models ( HMMs) and Guassian Mixture Models have history with earlier days of speech recognition, where automatic speech recognition was borned outta these statistical models which lead to good useful work done on low resourced data sets etc but NOT in real time The need for more precise and flexible models has led to the increased use of deep learning-based models in projects such as Recurrent Neural Networks(RNNs) and Long Short Memory(LSTM), which are used since they can handle sequences without losing sight of dependencies.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Technique** | **Description** | **Advantages** | **Limitations** | **Examples/Applications** |
| Hidden Markov Model (HMM) | Statistical model that uses probabilities to predict sequences of sounds to words. | Well-established, simple, and interpretable. | Struggles with noisy environments and diverse accents. | Early ASR systems, dictation software. |
| Gaussian Mixture Model (GMM) | Used with HMMs to model different sound distributions. | Improved acoustic modeling over pure HMMs. | Sensitive to variations in accents, noise, and requires large datasets. | Used in early versions of Siri, Google ASR. |
| Recurrent Neural Networks (RNN) | Neural network that processes sequences, making it suitable for speech. | Better at capturing sequential dependencies than statistical models. | Struggles with long-range dependencies and complex speech patterns. | Real-time speech recognition, language translation. |
| Long Short-Term Memory (LSTM) | A type of RNN designed to retain information over long sequences. | Handles long-range dependencies well, robust in noisy environments. | Computationally intensive, difficult to scale to large datasets. | Used in real-time transcription, virtual assistants. |
| End-to-End Models | Models like transformers that directly convert speech to text without intermediate steps. | Simplifies pipeline, high accuracy, adaptable to multiple languages. | Requires large datasets for training, may struggle with low-resource languages. | Wav2Vec, DeepSpeech, automated transcription services. |
| Transformer-Based Models | Deep learning models that process entire speech sequences in parallel. | High accuracy, faster processing, adaptable to various accents and dialects. | Complex to train, resource-intensive, large data requirements. | Wav2Vec 2.0, Google’s Speech-to-Text API. |
| Cloud based Models | Cloud based speech to text model provided by various | Achieves high accuracy in transcribing audio into text, making it suitable for various applications. | Limited customization options are available; also relies on a stable internet connection for optimal performance. | Google Cloud Speech-to-Text, Amazon Transcribe, Microsoft Azure Speech Service, IBM Watson Speech to Text. |

## 4.2 Text summarization

The comparative analysis of text summarization technologies presented in the table reveals a diverse array of approaches, each with its unique strengths and limitations. The table below list the most common used algorithms for text summarization

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Algorithm** | **Key Developments** | **Pros** | **Cons** |
| 2016 | Rule-based Systems | Utilization of linguistic rules and heuristics for text generation. | Offers transparency and interpretability in the generation process. | Scalability is limited; crafting rules manually can be labor-intensive. |
| 2017 | Hidden Markov Models | Application in both speech recognition and generative tasks. | Established approach with a wide range of applications in text modeling. | Struggles to capture complex dependencies between words or phrases. |
| 2018 | Recurrent Neural Networks (RNN) | Enhanced mechanisms for improved context modeling, specifically for sequential data. | Effectively manages and processes sequential information, making it suitable for text generation. | Prone to issues like gradient vanishing; slow convergence during training can hinder performance. |
| 2020 | Generative Adversarial Networks (GANs) | Introduction of improved training techniques to enhance model stability. | Capable of producing diverse and realistic text samples that mimic human-like generation. | Often faces challenges related to training instability and mode collapse, which can affect quality. |
| 2021 | Long Short-Term Memory Networks (LSTM) | Advanced architectures focused on capturing sequential dependencies in text data. | Effectively retains context over longer sequences, improving performance in tasks requiring understanding of continuity. | May still struggle with very long-term dependencies, leading to loss of information over time. |
| 2022 | Facebook’s BART Model | Utilization of a transformer-based architecture specifically designed for summarization tasks. | Produces coherent and concise summaries, enhancing readability and comprehension. | Requires significant computational resources and can be complex to train effectively. |

# 5. Results

This review underscores the rapid progression and growing sophistication of both speech-to-text and text summarization techniques. In the domain of speech-to-text, traditional methods like Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) were foundational, yet they encountered limitations in noisy environments and with diverse accents or dialects. The shift towards deep learning, particularly with Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and more recently, end-to-end architectures such as Convolutional Neural Networks (CNNs) and transformer-based models, has significantly improved transcription accuracy, especially in real-time, multilingual, and noisy settings. Transformer models (e.g., Wav2Vec, DeepSpeech) are particularly promising, enabling faster and more context-aware transcriptions by processing entire speech sequences as a whole rather than in fragments.

In text summarization, extractive approaches, which identify and select important sentences from the original text, have traditionally been effective for simpler summarization tasks. However, they often fail to capture the deeper meaning and contextual nuances of the content. The rise of abstractive summarization methods, powered by advanced neural architectures such as BERT, T5, and GPT, has transformed the field by allowing models to generate entirely new summaries that not only compress the content but also rephrase it in a coherent and human-like manner. These models have shown remarkable improvements in generating concise, accurate, and contextually aware summaries, particularly when trained on large datasets. The use of pre-trained models with fine-tuning has further boosted performance, making them applicable in areas like news summarization, legal document condensation, and academic paper summaries.

While the integration of these techniques into practical applications such as transcription services, meeting summarizers, and personal digital assistants is already evident, challenges persist. Speech-to-text systems still struggle with low-resource languages, speaker diarization, and domain-specific jargon, whereas text summarization systems grapple with issues of factual consistency, fluency, and avoiding repetition. Additionally, the performance of these systems in real-time environments and their ability to handle long-form content efficiently are active areas of research, with recent advances in multi-modal models offering promising future directions.

# References

[1] M. Mehta, K. Gupta, S. Tiwari and Anamika, "A Review on Sentiment Analysis of Text, Image and Audio Data," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1660-1667, doi: 10.1109/ICCMC51019.2021.9418360.

[2] V. Maheshwar Reddy, K. Deepika, K. Adithya Surya Prakash, and M. Sanathan, "A survey on audio analysis: Text characterization and summarization," *World Journal of Advanced Research and Reviews*, vol. 21, no. 03, pp. 1596–1601, 2024, doi: 10.30574/wjarr.2024.21.3.0789.

[3] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," arXiv:1303.5778 [cs.NE], 2013, doi: <https://doi.org/10.48550/arXiv.1303.5778>

[4] Ghadekar, Premanand & Anand, Divsehaj & Gupta, Aryan & Oswal, Preeti & Sharma, Dheeraj & Khare, Shreyas. (2023). Audio Based Text Summarization Using Natural Language Processing. 10.1007/978-981-99-3656-4\_17.

[5] M. -H. Su, C. -H. Wu and H. -T. Cheng, "A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2061-2072, 2020, doi: 10.1109/TASLP.2020.3006731.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N. (2017). Attention is all you need [J]. *Advances in neural information processing systems*, *30*(1), 261-272.

[7] K. Yao, L. Zhang, D. Du, T. Luo, L. Tao and Y. Wu, "Dual Encoding for Abstractive Text Summarization," in *IEEE Transactions on Cybernetics*, vol. 50, no. 3, pp. 985-996, March 2020, doi: 10.1109/TCYB.2018.2876317.

[8] Dhumal, Priyanka & Sutar, Sudarshan & Surve, Indraneel & Munawwar, Mirza & Nanaware, Vishal. (2024). Text Summarization Using NLP. International Journal of Advanced Research in Science, Communication and Technology. 319-324. 10.48175/IJARSCT-18650.

[9] Monteiro, R., Pernes, D. (2023). Towards End-to-End Speech-to-Text Summarization. In: Ekštein, K., Pártl, F., Konopík, M. (eds) Text, Speech, and Dialogue. TSD 2023. Lecture Notes in Computer Science(), vol 14102. Springer, Cham. https://doi.org/10.1007/978-3-031-40498-6\_27

[10] A. Vartakavi, A. Garg and Z. Rafii, "Audio Summarization for Podcasts," 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 2021, pp. 431-435, doi: 10.23919/EUSIPCO54536.2021.9615948.

[11] Vinnarasu, A., and Deepa V. Jose. "Speech to text conversion and summarization for effective understanding and documentation." International Journal of Electrical and Computer Engineering (IJECE) 9, no. 5 (2019): 3642-3648.