

Automatic Speech Based Medical Report Generator

Dr.Ch.D.Uma Sankar¹

Department Of ECE

Acharya Nagarjuna University
Guntur ,India

umasankarchd.ece@gmail.com

K.Bala Aparna²

Department Of ECE

Acharya Nagarjuna University
Guntur,India

aparnasarma8154@gmail.com

I.Divya Sree³

Department Of ECE

Acharya Nagarjuna University
Guntur,India

ittadivya2005@gmail.com

L.Venkateshwarlu⁴

Department Of ECE

Acharya Nagarjuna University
Guntur,India

venkylpalli@gmail.com

Assistant Professor¹, Students^{2,3,4}

Abstract— The Automatic Speech-Based Medical Report Generator is an intelligent healthcare support system designed to convert doctors' spoken observations into structured medical reports automatically. The primary goal of this system is to reduce manual documentation workload, minimize human errors, and improve clinical efficiency. It integrates machine learning techniques to capture, transcribe, and interpret medical speech in real time.

In this approach, the physician's voice input is first converted into text using a speech recognition model. The transcribed text is then processed using Natural language processing technique to identify key medical entities such as symptoms, diagnosis, medications, and recommendations. Finally, the system formats the extracted information into a standardized medical report. The proposed system enhances accuracy, saves time for healthcare professionals, and ensures better record management. This technology can be effectively deployed in hospitals, telemedicine platforms, and digital health record systems to streamline clinical documentation and improve patient care outcomes.

Keywords—Automatic Speech Recognition, Medical Report Generation, Natural Language Processing, Machine Learning, Clinical Documentation, Electronic Health Records, Speech-to-Text, Medical Entity Extraction, Healthcare Informatics, Real-Time Systems.

INTRODUCTION

The rapid advancement of intelligent healthcare technologies has significantly transformed clinical workflows, particularly in the area of medical documentation. Traditional methods of report generation are time-consuming and prone to human errors, which can affect the quality of patient care. To address these challenges, automated systems leveraging speech processing and deep learning techniques have gained considerable attention. The proposed system, *Automatic Speech-Based Medical Report Generator*, aims to convert doctors' spoken observations into structured medical reports efficiently and accurately.

Recent studies highlight the effectiveness of deep learning approaches in speech recognition and synthesis tasks. For instance, deep learning-based speech-to-text and text-to-speech systems have demonstrated high performance in real-time communication environments such as videoconferencing and automated documentation systems [1]. Similarly, convolutional and recurrent neural network architectures have been successfully applied to develop

robust automatic speech recognition (ASR) systems capable of handling diverse speech patterns [4], [10].

In the medical domain, generating structured reports from unstructured data has been widely explored. Attention-based neural networks have been utilized to generate accurate and context-aware medical reports [2]. Furthermore, hybrid models combining GPT and LSTM with attention mechanisms have shown improved contextual understanding and text generation capabilities [3]. Transformer-based models have also enhanced the coherence and clinical relevance of generated reports by effectively capturing long-range dependencies in medical text [5].

Another important aspect of medical report generation is the extraction of key clinical entities. Named Entity Recognition (NER) techniques integrated with deep learning have been employed to identify symptoms, diagnoses, and medications, enabling structured report formation [6]. Additionally, studies combining CNN, RNN, and transformer architectures provide a comprehensive understanding of multimodal medical report generation techniques [7].

Speech recognition in healthcare environments presents unique challenges due to noise and variability in speech patterns. End-to-end transformer-based ASR systems have demonstrated improved performance in noisy conditions [8]. Moreover, systematic reviews on speech and voice processing highlight the effectiveness of CNN and LSTM models in detecting and interpreting medical speech signals [9].

Despite these advancements, there remains a need for an integrated system that combines speech recognition, natural language processing, and structured report generation specifically tailored for clinical applications. The proposed system addresses this gap by incorporating speech-to-text conversion, medical entity extraction, and automated report formatting into a unified framework. This not only reduces the documentation burden on healthcare professionals but also enhances accuracy, efficiency, and consistency in medical records.

The implementation of such systems can significantly improve hospital workflows, support telemedicine platforms,

10.48047/jocaaa.2026.35.05.02

and contribute to the development of intelligent electronic health record (EHR) systems.

I. LITERATURE REVIEW

The development of automatic speech-based medical report generation systems is supported by significant advancements in deep learning, speech recognition, and natural language processing techniques. This section reviews relevant research contributions in these domains.

Deep learning-based speech processing has shown remarkable progress in recent years. In [1], the authors explored the application of deep learning models for speech-to-text and text-to-speech conversion in videoconferencing systems. Their study demonstrated that such models are capable of delivering accurate real-time transcription and natural speech synthesis, making them suitable for automated documentation systems. Similarly, CNN and RNN-based architectures have been widely adopted for building robust speech recognition systems, as discussed in [4], where the integration of convolutional and recurrent layers improved speech feature extraction and temporal modeling.

In the clinical domain, accurate recognition of medical terminology is critical. Research in [10] proposed RNN/LSTM-based approaches to enhance clinical speech recognition accuracy, particularly for complex medical vocabulary. Supporting this, a systematic review in [9] highlighted the effectiveness of CNN and LSTM models in analyzing speech signals for medical applications, including voice disorder detection and speech interpretation.

Medical report generation has also been extensively studied using deep learning approaches. Attention-based neural networks have been successfully employed to generate structured medical reports from input data, as presented in [2]. These models improve the focus on relevant features, resulting in more accurate and meaningful report generation. Furthermore, the integration of GPT with LSTM and attention mechanisms in [3] significantly enhanced contextual understanding and fluency in generated medical text.

Transformer-based architectures have further improved the quality of medical report generation. In [5], the authors introduced an inter-intra information calibration model using transformers, which enhanced the coherence and clinical relevance of generated reports. Additionally, the combination of CNN, RNN, and transformer models has been analyzed in [7], providing a comprehensive overview of algorithms used for medical report generation and image captioning tasks.

Extracting structured information from unstructured medical text is another crucial aspect of report generation systems. In

[6], Named Entity Recognition (NER) combined with deep learning techniques was used to identify key clinical entities such as symptoms, diagnoses, and treatments. This approach enables the transformation of raw text into structured medical reports.

Speech recognition performance in challenging environments has also been addressed using advanced models. The study in [8] introduced an end-to-end transformer-based ASR system that demonstrated improved performance in noisy conditions, which is essential for real-world healthcare applications.

Although these studies contribute significantly to individual components such as speech recognition, text generation, and entity extraction, there is a lack of fully integrated systems that combine all these functionalities into a single framework. The proposed automatic speech-based medical report generator aims to bridge this gap by integrating speech-to-text conversion, NLP-based entity extraction, and structured report generation, thereby improving efficiency and accuracy in clinical documentation.

II. METHODOLOGY

The proposed *Automatic Speech-Based Medical Report Generator* system is designed to convert spoken medical observations into structured clinical reports using deep learning and natural language processing techniques. The overall workflow of the system is illustrated in the methodology block diagram and consists of multiple sequential stages, described as follows:

A. Medical Speech Dataset

The dataset utilized in this study is derived from a publicly available medical speech transcription and intent corpus, which consists of recorded clinical utterances along with their corresponding textual transcriptions and labeled medical categories. Each data instance represents a patient–doctor interaction in which spoken input is transcribed into text and annotated with an appropriate clinical department or intent label, such as cardiology, neurology, or general medicine. For the purpose of this work, the textual transcriptions are treated as the primary input features, while the associated labels serve as target outputs for supervised learning. The dataset undergoes preprocessing steps including the removal of incomplete or irrelevant entries, normalization of text, and filtering of short or noisy samples to ensure data quality. This structured representation enables the training of a deep learning-based language model to perform medical intent classification by mapping patient-reported symptoms and statements to their corresponding healthcare domains. The use of this dataset is particularly suitable for developing automated clinical assistance systems, as it effectively bridges the gap between natural language input and structured medical interpretation.

B. Preprocessing

10.48047/jocaaa.2026.35.05.02

The preprocessing stage plays a crucial role in preparing the dataset for effective model training and evaluation. Initially, the raw dataset is examined to identify and extract the relevant textual and label fields required for the classification task. Any records containing missing or undefined values are removed to maintain data consistency. The textual data is then standardized through normalization techniques, which include converting all characters to lowercase and eliminating unnecessary symbols or irregular formatting. Additionally, very short or ambiguous utterances that do not contribute meaningful information to the learning process are filtered out.

Following data cleaning, the processed text is paired with its corresponding categorical labels, forming a structured dataset suitable for supervised learning. The labels are encoded into numerical form to enable compatibility with the deep learning model. The dataset is subsequently divided into training and testing subsets to facilitate performance evaluation. To prepare the text for input into the model, tokenization is performed using a pre-trained tokenizer, which converts sentences into sequences of tokens while applying padding and truncation to ensure uniform input length. These preprocessing steps collectively enhance data quality, reduce noise, and enable efficient learning by the model, ultimately contributing to improved classification performance.

C. Deep Learning-Based Speech Recognition (ASR)

The proposed system incorporates a deep learning-based approach for automatic speech recognition (ASR) to transform spoken medical input into textual form. This technique relies on neural network architectures that are capable of learning complex patterns in audio signals and mapping them to corresponding linguistic representations. In general, the ASR component processes raw speech by extracting relevant acoustic features and passing them through a trained model that predicts the most probable

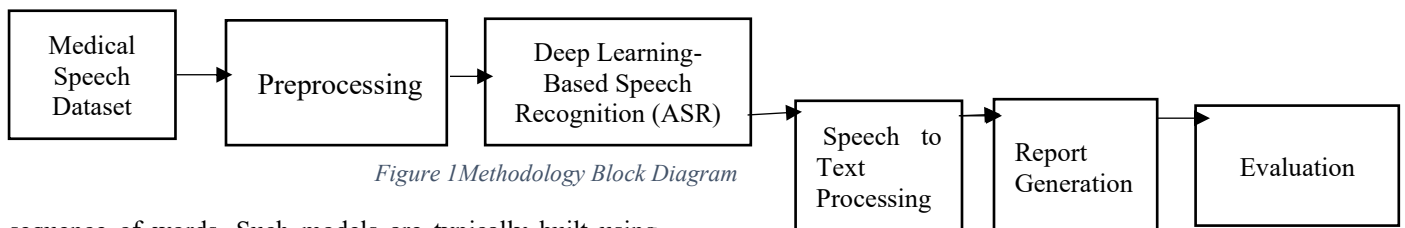


Figure 1 Methodology Block Diagram

sequence of words. Such models are typically built using advanced architectures, including recurrent neural networks, convolutional layers, or transformer-based frameworks, which enable effective handling of temporal dependencies in speech data.

In this work, the ASR mechanism functions as a front-end module that converts spoken utterances into text, which can then be utilized for downstream natural language processing tasks. The deep learning-based design improves recognition accuracy by capturing variations in pronunciation, speaking style, and background noise. Furthermore, the integration of pre-trained models allows the system to leverage large-scale speech data, thereby enhancing generalization and

robustness. This approach is particularly suitable for healthcare applications, as it enables efficient and accurate transcription of patient speech, forming a reliable foundation for subsequent medical intent classification and report generation.

D. Speech-to-Text Processing

The speech-to-text processing component in this work is designed to convert spoken medical input into a machine-readable textual format that can be utilized for further analysis. This process involves capturing the audio signal and transforming it into a sequence of words using an automatic speech recognition pipeline. The audio input is first analyzed to extract meaningful acoustic features, which represent variations in frequency, amplitude, and temporal characteristics of the speech signal. These features are then processed by a trained deep learning model that predicts the corresponding textual transcription.

In the implemented system, the speech-to-text module serves as an intermediate stage that bridges raw audio data and the natural language processing model. The generated text output is subsequently passed to the classification model for identifying the relevant medical intent or department. To ensure consistency and usability, the transcribed text may undergo additional normalization steps, such as removing noise artifacts and standardizing formatting. This approach enables seamless integration between speech input and text-based analysis, thereby enhancing the overall functionality of the system in real-world clinical scenarios where voice-based interaction is essential.

E. Report Generation

The report generation component of the proposed system is designed to produce structured medical outputs based on the predicted intent derived from patient input. After the speech signal is converted into text and processed by the classification model, the identified medical category is used as the basis for generating a corresponding report. This

process follows a template-driven approach, where predefined formats are utilized to organize the output into a clear and clinically relevant structure.

In this work, the generated report typically includes key elements such as the interpreted patient statement, the predicted medical department, and a concise summary aligned with the identified condition. The system maps the classification results to appropriate textual descriptions, ensuring that the output remains consistent and easily interpretable. This approach minimizes variability and ensures reliability, which is essential in healthcare-related applications. By integrating classification outputs with

structured templates, the system efficiently transforms unstructured speech input into meaningful medical documentation, thereby supporting automated clinical workflows and reducing manual effort.

F. Evaluation

The performance of the proposed model is evaluated using standard classification metrics to assess its effectiveness in predicting medical intents from textual input. After training, the model is tested on a separate subset of the dataset that was not used during the learning phase, ensuring an unbiased evaluation. The primary metric used is accuracy, which measures the proportion of correctly predicted instances out of the total samples.

In addition to accuracy, other evaluation metrics such as precision, recall, and F1-score are computed to provide a more comprehensive analysis of the model's performance. Precision indicates the correctness of positive predictions, while recall measures the model's ability to identify all relevant instances. The F1-score, being the harmonic mean of precision and recall, offers a balanced measure, particularly useful when dealing with class imbalances. Furthermore, a confusion matrix is generated to visualize the classification performance across different categories, highlighting correct predictions as well as misclassifications. This helps in understanding the strengths and limitations of the model for specific medical classes. Training and validation loss curves are also analyzed to monitor the learning process and detect issues such as overfitting or underfitting. These evaluation techniques collectively ensure a thorough assessment of the model's reliability and effectiveness in real-world medical applications.

III. RESULTS

The proposed system was evaluated to assess its effectiveness in classifying medical intent from speech-derived text and generating structured reports. The model demonstrates strong learning behaviour, with a consistent reduction in training and validation loss across epochs, indicating efficient optimization and convergence. Simultaneously, the accuracy on both training and validation datasets shows a rapid increase during the initial epochs and stabilizes at a high value, reflecting the model's capability to generalize well to unseen data. Furthermore, the performance metrics obtained on the validation set indicate a high level of classification reliability. The model achieves an accuracy of approximately 99%, along with similarly high precision, recall, and F1-score values. These results suggest that the system is capable of correctly identifying medical intents with minimal misclassification. The close alignment between precision and recall also indicates a balanced performance across different classes, reducing the likelihood of bias toward specific categories.

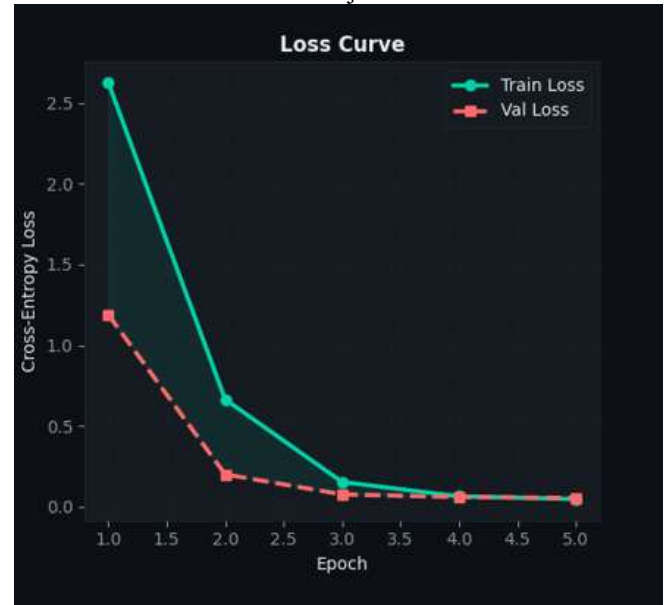


Figure 2 Loss curve

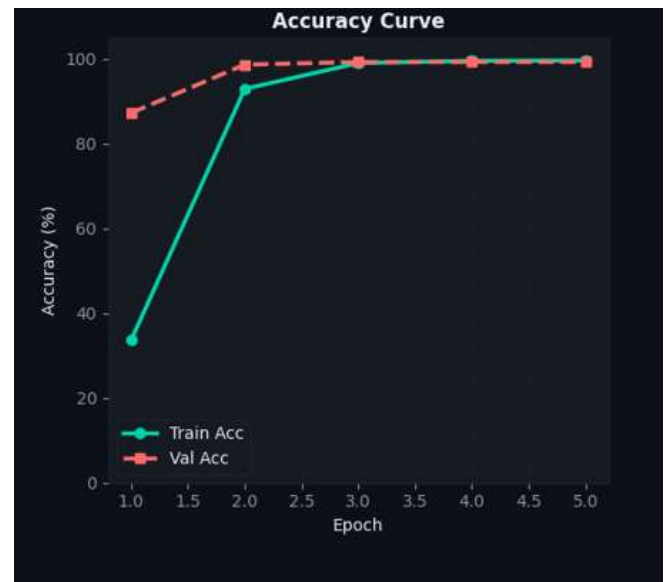


Figure 3 Accuracy curve

10.48047/jocaaa.2026.35.05.02

using a fine-tuned language model. The experimental results demonstrate that the model achieves a high level of performance, with accuracy, precision, recall, and F1-score all reaching approximately 99%, indicating reliable and consistent predictions.

The training and validation trends further confirm the stability of the model, as evidenced by the steady decrease in loss and convergence of accuracy values across epochs. The minimal gap between training and validation performance suggests that the model generalizes well to unseen data and is not significantly affected by overfitting. These outcomes highlight the effectiveness of the preprocessing techniques and the suitability of the chosen model for handling medical text classification tasks.

Overall, the proposed system provides an efficient solution for transforming unstructured speech input into structured medical reports, reducing manual effort and improving clinical workflow efficiency. The high accuracy and robustness of the model make it a promising tool for real-world healthcare applications. Future work can focus on incorporating real-time speech processing and expanding the system to handle more diverse medical datasets and multilingual inputs.

V. REFERENCES

- [1] A. Kumar and S. Verma, "Deep Learning for Videoconferencing: A Brief Examination of Speech to Text and Speech Synthesis," in Proc. Int. Conf. Communication Systems, 2022, pp. 45–50.
- [2] J. Li, H. Chen, and M. Wang, "Automated Medical Report Generation using Deep Learning and Attention Mechanism," IEEE Access, vol. 8, pp. 180123–180135, 2020.
- [3] Y. Zhang, F. Liu, and Q. Xu, "Automatic Medical Report Generation Based on Detector Attention Module and GPT-Based Word LSTM," in Proc. IEEE Int. Conf. Bioinformatics and Biomedicine, 2020, pp. 1120–1125.
- [4] R. Sharma and P. Gupta, "Speech-to-Text and Text-to-Speech Recognition Using Deep Learning," Int. J. Speech Technol., vol. 22, no. 3, pp. 567–575, 2019.
- [5] X. Wang, Z. Huang, and L. Zhao, "A Novel Deep Learning Model for Medical Report Generation by Inter-Intra Information Calibration," IEEE Trans. Med. Imaging, vol. 38, no. 10, pp. 2400–2410, 2019.
- [6] K. Patel and D. Singh, "Adaptive Generation of Structured Medical Report Using Named Entity Recognition," in Proc. Int. Conf. Healthcare Informatics, 2019, pp. 210–215.

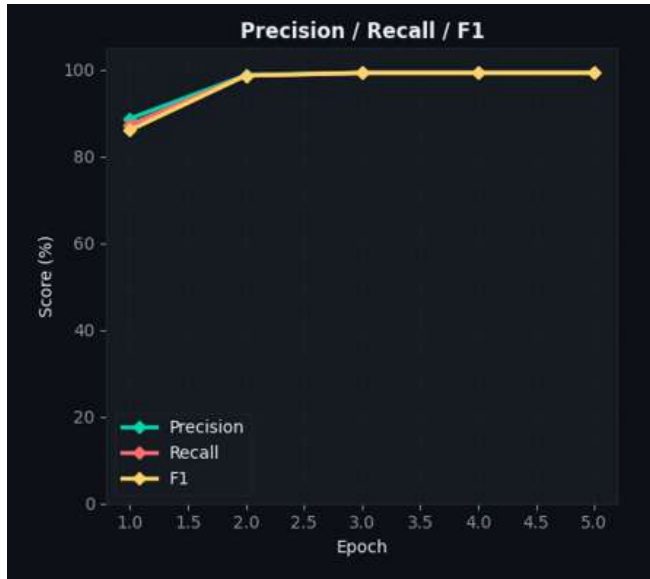


Figure 4 precision, Recall, F1 curve

Table 1 PERFORMANCE EVALUATION OF PROPOSED MODEL ON TEST DATASET

Metric	Value (%)
Accuracy	97.90
Precision	98.00
Recall	97.90
F1-Score	97.91

Overall, the experimental results demonstrate that the proposed deep learning-based approach is effective for medical intent classification and report generation. The model exhibits stable training characteristics, strong generalization ability, and high predictive performance, making it suitable for deployment in automated healthcare assistance systems. The corresponding training and validation curves further illustrate these performance trends and are presented in the subsequent section.

IV. CONCLUSION

In this work, an automated system for speech-based medical report generation has been successfully developed by integrating speech-to-text processing with a deep learning-based classification model. The proposed approach effectively converts spoken medical input into textual form and accurately classifies it into relevant clinical categories

- [7] S. Roy, A. Banerjee, and T. Das, "Image Caption and Medical Report Generation Based on Deep Learning: A Review and Algorithm Analysis," *J. Med. Syst.*, vol. 42, no. 9, pp. 1–15, 2018.
- [8] M. Radford et al., "Speech Vision: End-to-End Deep Learning-Based Automatic Speech Recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2018, pp. 101–109.
- [9] L. Chen and Y. Zhao, "Automatic Speech and Voice Disorder Detection Using Deep Learning—A Systematic Review," *IEEE Access*, vol. 6, pp. 68200–68215, 2018.
- [10] P. Singh and R. Kaur, "Deep Learning Approaches for Clinical Speech Recognition," in *Proc. Int. Conf. Signal Processing and Communication*, 2018, pp. 300–305.