5G-Optimized Deep Learning Framework for Real-Time Multilingual Speech-to-Speech Translation in Telemedicine Systems

Medapati Venkata Manga Naga Sravan[1*], K Venkata Rao[2]
[1]Andhra University
[2]HOD, Dept of Computer Science, Andhra University, India
E-mail: Sravan.medapati@gmail.com, professor_venkat@yahoo.com
*Corresponding author

Keywords: telemedicine, multilingual speech translation, deep learning, Speech-to-Speech workflow, 5G technology

Received: December 15, 2024

Telemedicine has revolutionized healthcare by enabling virtual consultations, yet it still faces challenges from linguistic barriers and the need for real-time, scalable communication. Current systems typically address isolated tasks like speech recognition or symptom classification, lacking a unified solution for multilingual doctor-patient interactions. To address this, we present a 5g-optimized Deep Learning Framework that integrates advanced speech recognition, neural machine translation, and text-to-speech synthesis into a seamless Speech-to-Speech Workflow (STSW). Specifically, our framework utilizes fine-tuned OpenAI Whisper for speech recognition, a Marian MT model fine-tuned on multilingual medical corpora for translation, and Tacotron 2-based neural TTS for speech synthesis. Each model is domain-adapted to handle complex medical terminologies. We implement the framework over 5G-enabled edge computing infrastructure, ensuring real-time performance with ultra-low latency. Experimental results demonstrate the effectiveness of the proposed system, achieving a Word Error Rate (WER) of 0.12, a BLEU score of 0.85 for translation quality, and a Mean Opinion Score (MOS) of 4.5 for the naturalness of synthesized speech. Furthermore, our framework delivers an end-to-end latency of 2.1 seconds, outperforming existing approaches. This integration bridges communication gaps in telemedicine, facilitating accurate multilingual conversations and scalable healthcare delivery across diverse geographies.

Povzetek: Predstavljen je 5G-optimiziran okvir globokega učenja za večjezično govorno prevajanje v telemedicini, ki s prilagojenimi modeli dosega kvalitetne rezultate v realnem času.

1 Introduction

One particular technology affecting modern healthcare is telemedicine, allowing consultation and diagnosis over remote digital platforms. In many multilingual regions, however, communication challenges — primarily linguistic — make it less effective. Current telemedicine setups are limited to single functionalities such as automating triage [1], speech recognition [2], or chatbots specific to a disease [5]. Though these approaches solve some parts of the telemedicine puzzle, they fall short due to the absence of an integrated framework that can facilitate real-time multi-lingual communication between doctors and patients. This communication is vital as it increases accessibility and efficiency in healthcare delivery. Although the literature has identified some exciting opportunities and existing applications, this paper shows that deep learning can substantially drive telemedicine systems forward—for instance, Shi et al. Latif et al. [1] used Bi-LSTM as intelligent triage. In [3], the scalability of speech recognition technologies in healthcare. However, current systems have limitations in scalability, latency, and adaptability to clinical settings—furthermore, research, including those of Kandpal et al. While [5] focuses on chatbots designed for communication in healthcare, they tend to ignore multilingual, real-time speech-turn-taking interactions. This gap highlights the importance of an end-to-end multilingual speech-to-speech system for telemedicine.

This study intends to establish a Speech-to-Speech Workflow (STSW), which claims to be a novel framework to combat these obstacles. The main aim is to incorporate sophisticated speech recognition, translation capability, and text-to-speech synthesis into an integrated system for telemedical applications. Here, the novelty of this research is due to the use of domain-specific fine-tuning techniques for medical, unique arrangement towards multilingual capabilities integration, and the process of real-time transliteration based on 5G technology. These properties make the architecture suitable for scalable, flexible services for a range of healthcare services. This research adds value from several perspectives. First, it presents a workflow for telemedicine speech-to-speech translation in many languages. Secondly, it performs better than the state-of-the-art systems in all dimensions of accuracy in speech recognition, translation quality, and the naturalness of speech produced by synthesis. Third, it offers a latency-optimized infrastructure for real-time interactions, focusing on key issues in telemedicine communication.

To systemically guide this research, we crafted the following central research questions (RQs):

RQ1: What kind of deep learning-architecture-based framework can be implemented to overcome the multilingual barrier in doctor-patient communication in telemedicine systems?
RQ2: How can the proposed system provide speech-to-speech translation performance in real time while keeping low latency and high scalability?
RQ3: How does integrating 5G technology help the adaptability and reliability of speech-based telemedicine applications in various healthcare environments?
To resolve them, we introduce a 5G-optimized deep learning framework that combines speech recognition, neural machine translation, and text-to-speech open-source solutions and optimizes them for medical vocabulary and multilingual use. We report across important data points end to end in Word Error Rate (WER), BLEU score, Mean Opinion Score (MOS), and latency appropriate for telemedicine use.
Our primary contribution is a cohesive adaptation of domain-based fine-tuning and 5G-specific optimization within a real-time, multilingual, speech-to-speech translation system designed specifically for telemedicine. While existing approaches use general ASR and translation models, we adapt both Whisper and Marian MT models to multilingual medical datasets to improve the recognition and translation of specialized medical terminology.
The presented telemedicine framework centers around the advantages of deep learning for individualized care throughout the telemedicine system. In particular, we use a fine-tuned Whisper ASR model to perform accurate multilingual speech recognition to manage the variance in speech from patients: Real-time, domain-specific translation using the Marian MT Transformer model bridging the communication gap between Doctor and Patient. The model is fine-tuned Tacotron 2, ensuring the speech synthesis produces a natural, context-aware audio output. Moreover, we integrate a BERT-based model for sentiment analysis to extract emotional signals from patient's speech, addressing a gap in empathetic healthcare communication. In contrast to existing systems that handle these different components in a siloed way, our framework integrates all of the modules in a one-stop shop for a real-time, low-latency telemedicine solution that scales to multiple languages.
The remaining structure of the paper is as follows. We summarize the existing literature and identify the research gaps in multilingual telemedicine systems in Section 2. Section 3 proposes the methodology, the details of the STSW framework, and its components. Experimental results and a comparison between the system and state-of-the-art approaches are presented in Section 4. Section 5 discusses the results broadly and describes the study's limitations. Finally, Section 6 concludes with a brief discussion of its implications and directions for future work on broadening linguistic capabilities, tightening semantic precision, and supporting offline telemedicine.

2 Related works

Recent advancements in telemedicine highlight the need for multilingual, real-time communication systems. Existing studies focus on isolated tasks that lack integration. Shi et al. [1] précised classifying patient symptoms, an intelligent triage model that combines Bi-LSTM with character embedding to improve telemedicine services. Payan et al. [2] revealed potential problems for patients from marginalized communities as community health centers adopted telemedicine at a rapid rate. Latif et al. [3] confronted scalability and technological integration hurdles; deep learning-driven speech technology could revolutionize the healthcare industry. Ji et al. [4] provided accessible interpretation services, and mobile healthcare apps may be able to reduce language barriers in the medical field. Kandpal et al. [5] highlighted the increasing influence of artificial intelligence (AI) through chatbots, or virtual assistants, employing ML and AI to evolve from menu-based models to contextual ones. It highlights the convergence of NLP and deep learning and explores their possibilities in healthcare for predictive diagnosis and scheduling of appointments. The study highlights the revolutionary potential of chatbots in healthcare and corporate settings and emphasizes the necessity of well-trained models in service-oriented companies. It also evaluates existing applications, problems, and prospects.
Albahri et al. [6] examined how wearable sensors, networks, artificial intelligence, and cloud computing are all incorporated into telemedicine. One hundred forty-one publications are categorized by a systematic review highlighting the advances and problems in IoT-based healthcare and providing guidance for future studies. Li et al. [7] developed in digital and telecommunications, including AI, 5G, and IoT, are revolutionizing ophthalmology and improving telemedicine capabilities in the face of COVID-19 problems. Zhang et al. [8] employed deep learning and automated transcription to find themes associated with depression in speech recordings made by 265 clinical patients. Calambur et al. [9] examined the effects of language barriers on information collecting in an older adult telehealth service. Talpada et al. [10] can better understand influence by utilizing social media data, especially from Twitter, which provides insights into public attitude.
Yu et al. [11] examined an entire health-related Internet of Things architecture, focusing on cloud platform integration and multimodal sensor technologies for improved emotional connection and user experience. Ozyegen et al. [12] tackled the problem of information overload in healthcare by investigating helpful text-highlighting strategies to support medical practitioners. Chung et al. [13] use a language model and Deep Voice 2; this pilot project investigates specialized voice recognition for nursing shift handovers. Deepa and Khilar [14] developed speech technology in healthcare, which can be attributed to its non-invasive nature and ability to monitor and diagnose diseases. Tripathi et al. [15] affected articulation in speech by impairing muscular control. Clinicians and patients benefit from accurate minimal-word intelligibility tests.

Table 1: Comparative summary of state-of-the-art approaches in telemedicine systems

Study	**Methodology**	**Focus Area**	**Limitations**	**Gaps Addressed by STSW**
Shi et al. [1]	Bi-LSTM for intelligent triage	Symptom classification	No multilingual support lacks integration	STSW supports multilingual speech, integrates triage, recognition, translation
Latif et al. [3]	Deep learning-based speech recognition	Speech recognition	Lacks translation & scalability, high latency	STSW combines recognition + translation + TTS, 5G optimization reduces latency
Kandpal et al. [5]	Chatbot using ML & AI	Text-based chatbots	No real-time speech handling, not multilingual	STSW enables speech-to-speech multilingual real-time communication
Ji et al. [4]	Mobile apps for interpretation	Interpretation services	No scalability lacks integration with speech models	STSW offers end-to-end speech processing integrated with translation
Ganesh et al. [26]	ASR with Flask for disorder speech	Disorder speech recognition	No multilingual translation, not optimized for latency	STSW extends speech recognition to multilingual translation, optimized for 5G

Zhang et al. [16] examined the potential and present difficulties of intelligent speech technology (IST) in healthcare in the face of a lack of resources. It discusses the importance of IST in smart hospitals, namely in illness diagnosis, stroke patient care, and medical documentation. While highlighting AI's progress in voice recognition, the assessment also points out its drawbacks, including a lack of datasets and privacy issues. Kaushik et al. [17], with a considerable accuracy rate, SLINet CNN is a deep learning model for early identification of SLI and DD in children. It is low-complexity for usage in real-time, gender-neutral, and appropriate for remote diagnostics—plans for the future call for adding many languages and continuous speech. Wang et al. [18] presented a novel approach to categorizing voice issues that replaces single vowels with continuous Mandarin speech. Sindhu et al. [19], with speech and vocal impairments, are more likely to experience developmental delays and poor academic performance. Deep learning has transformed automatic detection, which provides prospective advances and helps with effective diagnosis. Huang et al. [20] used the UASpeech dataset, a novel two-stage paradigm for transforming everyday speech to dysarthric speech was suggested and assessed.

Alma et al. [21] examined current developments in deep neural networks for speech and visual applications, focusing on their evolution, difficulties in systems with limited resources, and new applications. Tanveer et al. [22] improved performance on various speech tasks, and ensemble deep learning approaches combine ensemble techniques with deep learning. Shastry [23] presented a method for continuous remote health monitoring in digital health that combines DL and NLP. Sonmez and Varol [24] improved human-computer interaction in Society 5.0, which requires further advancements in speech-emotion recognition (SER). Diverse speech traits and cultural variables that impact recognition accuracy are challenges. Talaat et al. [25] helped CNN-LSTM network-based identification achieve great accuracy by capturing voice airflow dynamics for letter pronunciation.

Ganesh et al. [26] combined ASR technology with Flask to build a powerful disease speech recognition platform that has the potential to revolutionize healthcare and other fields. Musalia et al. [27], with colossal accuracy using the DNN approach, the pilot research assesses SRAVI, a speech/phrase recognition program, with the goal of future development and real-world implementation. Kheddar et al. [28] adapted models to similar datasets; deep Transfer Learning (DTL) in Automatic Speech Recognition (ASR) overcomes the constraints posed by data scarcity. Gaitan et al. [29] prompted telemedicine's uptake, changing people's attitudes and habits in Spain and bringing attention to trends in the country's digital revolution. Bandopadhyay et al. [30] spooked Healthcare Bot (THCB) was created in response to the COVID-19 epidemic, which made it possible to improve remote patient care.

Shahamiri et al. [32] used deep learning to create a Dysarthric Speech Transformer that shows promise in reducing ASR difficulties for those with severe dysarthria. Wu et al. [33] unveiled a scalable precision health solution that combines AI-powered telecare, wearable technology, and ambient data. Applying modular models improves the prediction of chronic diseases. Joshy et al. [35] analyzed deep learning models with different acoustic characteristics for dysarthria severity classification, highlighting the better performance of MFCC-based i-vectors.

Przybylo [36] presented an LSTM-based technique for video plenty sonography-based continuous heart rate monitoring to simplify data processing while maintaining accuracy on par with more established techniques like POS and ICA. Kamble et al. [37] investigated using CNN and SPWVD in an EEG-based BCI system for imagined speech recognition. The results demonstrate notable improvements in performance over conventional techniques, which motivates more research with more enormous datasets and more sophisticated DL structures. Deb et al. [38] presented a deep learning model that achieves 67.71% UAR in categorizing cold speech using MFCC and LPC characteristics. Fernandes [39] enabled telemedicine to connect healthcare across distances, and with COVID-19, it proliferates. AI improves productivity, monitoring, and diagnoses but has drawbacks. Abdelhay et al. [40] provided 24/7 access and financial savings; medical bots—a remote healthcare service—have gained popularity in response to the COVID-19 outbreak. The literature review identifies gaps in telemedicine, particularly in multilingual speech systems.

Benedict and Subair [42] proposed a deep learning-based edge-enabled serverless architecture to detect animal emotion in real time, using a convolutional neural network with serverless computing (SC) to improve scalability and low latency processing. A deep learning framework for social media rumor detection and tracking is proposed by Han and Lin [43], which uses LSTM networks to extract temporal features for better detection accuracy. According to Chen and Zhang [44], an involution feature extraction method was implemented for human posture identification in martial arts, focusing on utilizing a feature extraction technique by convolutional deep learning models, which effectively captured spatial and temporal postural features, resulting in substantial improvements in both classification performance and robustness.

The existing approaches target standalone functionalities, as shown in Table 1, like triage models [1], speech recognition [3], chatbots [5], or interpretation services [4]. Nonetheless, they are limited in offering an efficient, scalable architecture for instant multilingual speech recognition, translation, and speech synthesis with low latency. The proposed STSW framework is proposed to bridge these gaps. It includes modules for speech recognition, translation, and text-to-speech synthesis adapted for medical scenarios, maintains multilingual support, and utilizes 5G technology to enable real-time, scalable, and resource-efficient telemedicine communication. This makes STSW close some of the many gaps across fragmented approaches in literature into a unified, on-demand, multilingual telemedicine system.

Although previous works have made several significant improvements in specific aspects of telemedicine, no unified framework integrates these fragmented components (e.g., speech recognition, symptom triage, translational linguistics, and sentiment analysis) into a single scalable real-time system. Our STSW framework fills this gap by jointly learning these functionalities and supporting real-time, low-latency communication between doctor and patient in a multilingual setting.

Existing approaches emphasize triage, diagnosis, or chatbots. The proposed research addresses these gaps by integrating advanced deep learning-based speech recognition, translation, and synthesis into a unified framework. This will enable real-time, multilingual doctor-patient communication and significantly enhance accessibility and efficiency in telemedicine systems.

3 Proposed system

An empirical approach to the proposed telemedicine system, presented in Figure 1, can be developed by applying advanced deep learning and natural language processing techniques integrated with 5G technology that allows for communication and diagnosis of the patients. The system starts with patient utterances via 5G-enabled audio or video calls. As a result, this enables near-zero delay data transfer, resulting in a telemedicine system that can work in a wide range of geographical conditions. First, this speech goes through the speech-to-speech translation module to translate the patient's speech into English; thus, this speech-to-speech conversation is made independent of the patient's language. This output is then fed into the speech-to-text translation module, which performs audio transcription into text with high accuracy using hybrid deep learning techniques. It integrates sophisticated speech recognition and context-based refinement techniques to preserve the context of medical phrases and terminologies. A language processing module then analyzes the transcribed text using natural language processing (NLP). This is the process of cleansing textual data and preparing it for analysis.

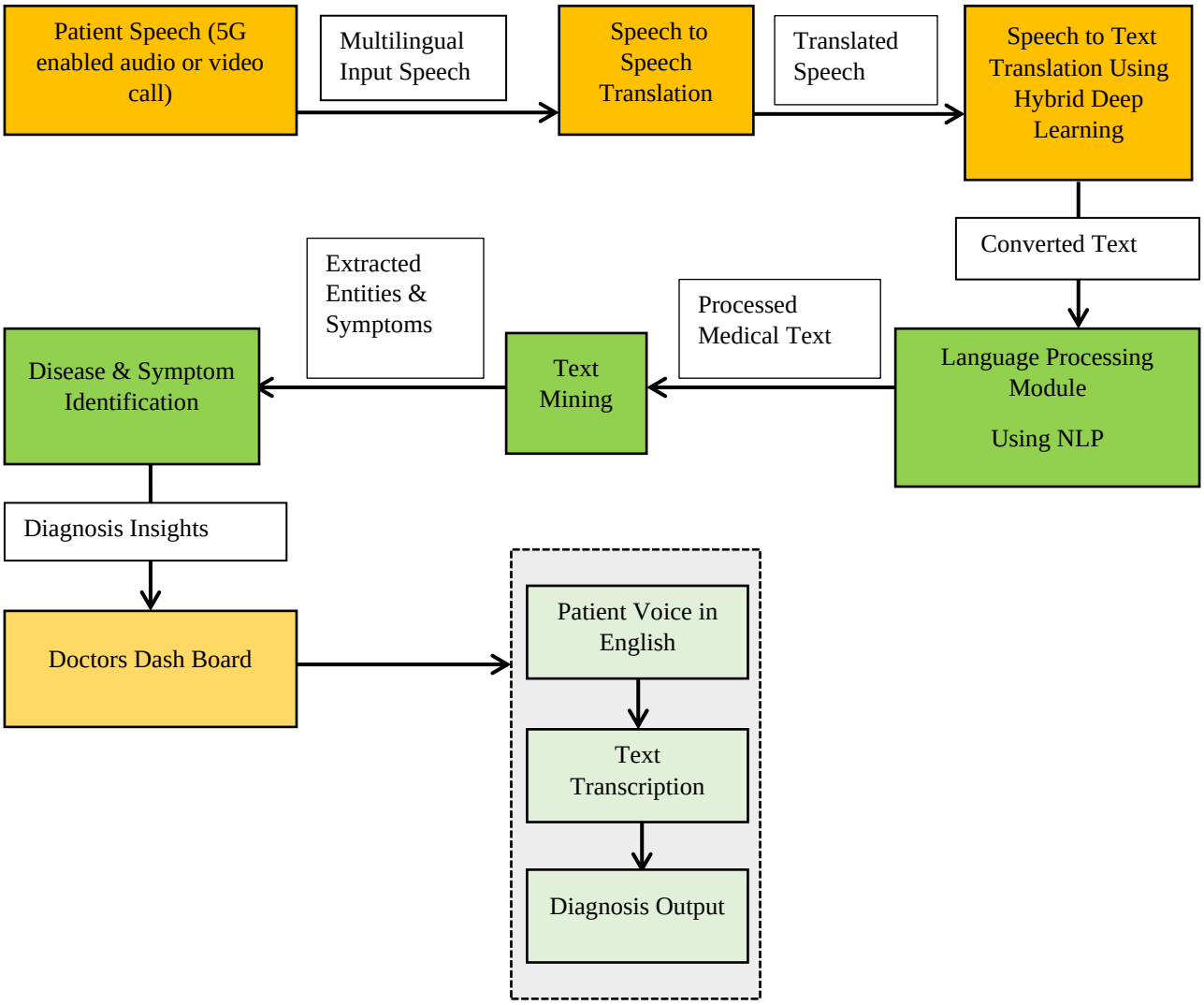

Figure 1: Overview of the proposed telemedicine system

Text mining is performed on the refined text to extract medically significant information such as symptoms, diseases, and patient-reported problems. The extracted data is then input to a disease and symptom recognition module, which uses deep learning models trained on large medical datasets to learn a mapping between the extracted information and possible diagnoses. It continuously updates a physician dashboard with the processed data, i.e., the transcription of the patient voice in English, textual data, and potential diagnoses list. The dashboard for the doctor serves as the primary interface for healthcare providers, allowing them to access the processed data and make decisions based on them. It seamlessly integrates feedback loops, enabling health professionals to provide feedback about their observations and adjust the system outputs accordingly. Harnessing the best of 5G and AI, the entire process is built to be smooth, crisp, and accurate, helping you come up with a final solution in a short time frame, thereby bringing medical services to your doorstep at the right time. This mechanism eliminates hurdles like language diversity and distance to promote patients' and providers' convenience and accessibility to health care. The system can be the basis for intelligent, real-time telemedicine apps by blending speech-to-speech and text-link processing with cutting-edge analytics.

The STSW framework is unique as it integrates advanced speech recognition (for transcribing patient speech), machine translation (for real-time multilingual conversion), and text-to-speech synthesis, all optimized for medical terminology and patient interaction. In contrast to previous systems designed for single-tasking, STSW integrates all of these operations into a single workflow, thus placing it in a unique position to address the linguistic, variability, and scaling challenges currently faced in existing telemedicine systems.

3.1 Speech-to-Speech workflow framework

A novel approach aims to tackle some relevant issues about telemedicine relying on 5G. It is used as the basis for the architecture that underlies a vision of 5G-enhanced telemedicine, which mediates challenges in multilingual health-cared communication. Using a combination of deep learning models and a real-time processing pipeline, Aedh can provide speech-to-speech translation and thus enable the patient and healthcare provider to communicate

directly without a language barrier. Main Features: Natural speech recognition, interpretability, contextual improvement, precision, etc. Together, these features define the quality of translations vital during medical consultations. Coupled with the ultra-low latency and edge computing power of 5G, the system is all set to power real-time processing and availability in even the remotest locations. This method is important because it closes the distance between language and distance to aid telemedicine in becoming more effective, accessible, and powerful in providing quality healthcare.

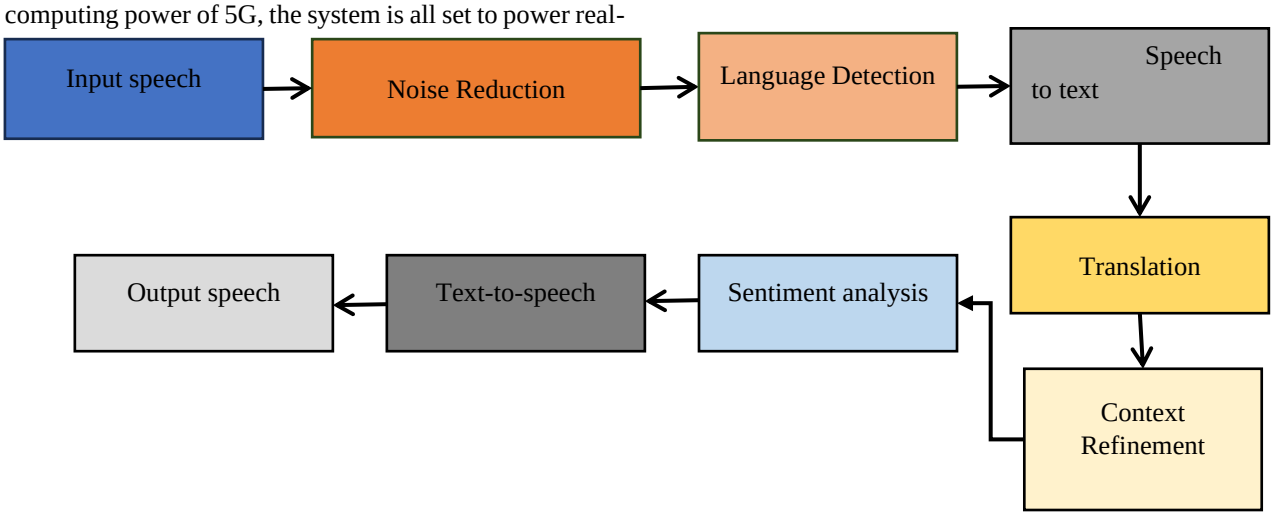

Figure 2: Proposed framework for AI-enabled language-independent Speech-to-Speech workflow

Figure 2 shows the workflow for the proposed AI-enabled telemedicine system. The workflow starts with receiving input speech from the patient. The spoken language is recorded by a recording machine or microphone and preprocessed to get quality data for the next steps. Some advanced signal processing techniques, such as spectral subtraction or a Wiener filter to suppress background noise, are applied to reduce noise in the input speech. This step clarifies the audio before processing, an essential factor in correct downstream processing.

After removing the noise, it detects the language spoken on the system. It uses a pre-trained natural language detection model, like fastText, to evaluate phonetic and lexical features in the audio. It identifies the language that has been detected and decides the processing pipeline that would be used for transcription and translation. Such a step will be crucial in building a telemedicine system that is language agnostic and will be able to address different languages as inputs. In the next step, the actual speech-to-text conversion occurs, in which an adapted audio file presentation is transcribed into text using the speech recognition model. It uses OpenAI's Whisper or Google Speech-to-Text APIs that have been fine-tuned explicitly on medical terminology to improve the recognition of challenging medical vocabulary. Also, the transcription will be context-driven, using specific models trained on healthcare datasets to reduce errors and ambiguity.

The transcription text is sent to a translation module that uses the Marian MT model, in which the translation model is fine-tuned using multilingual medical data. This process translates the recognized text —as slang or some unprofessional title; often, some slang will be used, translating to English, capturing the semantics in medical language. Finally, a GPT-based language model fine-tunes the fluency and coherence of the translation, giving finished translated outputs a rounder delivery. Language problems are fully resolved at this stage, allowing for smooth communication in a telemedicine consultation. Translated English text goes through context rectification, in which more advanced models of AI analyze the context of their translations in the medical domain and fix any potential mistakes made during translation. This process includes semantic enrichment, which cross-references the text with a knowledge base of medical terminologies to ensure consistency and accuracy. For example, vague phrases are substituted with their exact terms, as used in medicine, to avoid confusion.

Simultaneously, the system conducts sentiment analysis on the spoken input. A sentiment classifier based on deep learning evaluates the patient's functional state and detects stress, anxiety, or distress manifestations. Such an analysis is critical for comprehensive physical and emotional health care in telehealth. The opinion data can be incorporated into the diagnosis, which can help providers customize their responses. The cleaner text is then transformed into English speech using a text-to-speech engine like Google TTS or more sophisticated neural TTS. Its output speech is intelligible, more human-like, and created for understanding. The final output ensures that the healthcare provider has the patient's message in a format that is easy to access, thereby facilitating a successful telemedicine consultation.

The whole workflow uses the optimization for 5G to be deployed using the 5G networks for real-time communication. It provides the low latency and high bandwidth needed to perform speech processing, translation, and synthesis seamlessly, even during live consultations. Deploy the models on edge servers so that 5G-enabled devices can use their computational power, and response times will be shorter. The translated speech delivered by the system to the healthcare provider at the end of the methodology completes the cycle of

multilingual, real-time speech communication in telemedicine. The proposed system solves the significant challenges regarding language barriers, accessibility, and communication latency in telemedicine by incorporating noise reduction, language detection, speech recognition, translation, sentiment analysis, and 5G-enabled real-time processing. This allows effective patient and healthcare provider engagement without language or geographic boundaries.

All datasets were preprocessed and augmented to obtain robustness and deal with domain shift problems. The medical speech dataset, intended for the speech recognition module, was preprocessed using noise reduction methods like spectral subtraction and voice activity detection (VAD) to trim silence. Augmentation methods were applied with injections of additive noise from the MUSAN corpus to account for clinical environments, time-stretching, pitch-shifting, and additional medical terminology from other samples. Training corpus consisting of text data utilized for the translation model, implying cleaning, tokenization, and synonym expansion to improve the domain relevance. These datasets were used to fine-tune the Whisper ASR model, Marian MT translator, and Tacotron 2 TTS synthesizer with the default hyperparameters on NVIDIA Tesla V100 GPUs with the PyTorch framework. The model used for Whisper had an Adam optimizer, 3e-5 learning rate, 64 batch size, and 15 epochs with an early stopping condition on WER. Marian MT uses AdamW with a learning rate 5e-5 and 10 epochs, validated on BLEU. We trained Tacotron 2 with an RMSProp optimizer with batch size 48 and evaluated our models using MOS (Mean Opinion Score). Whisper serves as the speech-to-text engine due to its capabilities and compatibility for converting several languages into text; Marian MT serves to provide efficient and flexible Transformer-based multilingual translation; and Tacotron 2 serves as the speech synthesis engine due to its naturalness compared to alternatives like FastSpeech. Additionally, sentiment analysis was included via a fine-tuned BERT model on healthcare sentiment data. We then used the output of the sentiment analysis model to dynamically adapt the prosody and tone of Tacotron 2 to enrich interaction engagingly with the patient and support assessing the patient's mental health.

3.1.1 Noise reduction

The noise reduction module performs audio preprocessing for more explicit speech, which is vital for accurate downstream processing. It filters out ambient sound using algorithms involving methods like spectral subtraction and adaptive filtering. What is novel here is that deep learning-based noise suppression models are trained on large datasets that handle complex and noisy conditions. These innovations guarantee input that meets speech recognition quality standards, which is beneficial for telemedicine, given that recordings could occur in farmlands or urban regions that are quite noisy. The system provides ideal sound quality by incorporating 5G-supported real-time noise suppression, making remote healthcare consultations reliable across patient settings.

3.1.2 Language detection

Introduction The proposed method consists of four components. The first is the language detection component, which determines the spoken language in the input speech, allowing a language-independent telemedicine system. It uses fastText and other similar pre-trained models to learn the linguistic aspects of the text and classify it into a given language with very high accuracy. Based on the detected language, the system uses dynamically chosen translation pipelines. A fallback mechanism is proposed for robustness where probabilistic scoring models further validate uncertain detections. This innovative healthcare module, integrated with 5G edge servers, provides low-latency processing and is suitable for real-time consultations. A solution like this will aid in giving seamless communication for patients from multiple linguistic backgrounds, which will, in turn, increase inclusivity in telemedicine.

Specifically, the language detection module in the proposed framework serves as the first stage of the Speech-to-Speech Workflow. When receiving the speech input from the patients, the audio signal goes through several preprocessing steps to extract important acoustic features like MFCCs (Mel-Frequency Cepstral coefficients). These features are then input into a pre-trained lightweight CNN-based language identification model. The model classifies the input speech into one of these supported languages using phonetic and prosodic patterns to distinguish the different languages.

3.1.3 Speech-to-Text

Converts patient speech into text data using domain-adapted speech recognition models. It employs using either OpenAI's Whisper or Google's Speech-to-Text APIs, with fine-tuning on medical datasets that ensure it can identify complex terms accurately. As a translation tool, the system applies contextual error correction to the ambiguity inherent in any transcription of a medical consultation [1]. This solution enables fast and accurate transcription in real-time, even in bandwidth-constrained environments due to 5G-enabled real-time processing. This pivotal step to allow meaningful communication in a telemedicine system is augmented with promise by integrating multilingual support with minimum resource requirements, ensuring accuracy across various patient demographics.

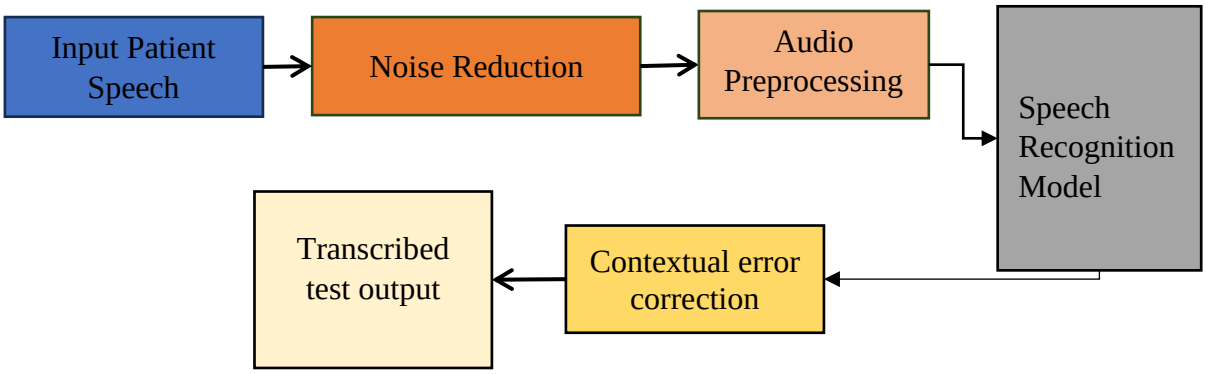

Figure 3: Patient speech-to-text conversion as part of the proposed telemedicine system

Figure 3 Workflow for converting patient speech to text as an integral part of the proposed telehealth system. The patient's speech is recorded and sent through a noise reduction module in the first stage. This step removes the ambient noise and enhances the input signal, which must be correctly processed in the following stages. To accommodate the different noise scenarios, sophisticated noise reduction algorithms such as spectral subtraction and adaptive filtering are utilized, rendering the system robust for practical applications. In audio preprocessing, we extract essential features like Mel Frequency Cepstral Coefficients (MFCCs) or spectrograms from the processed audio signal. These features are represented numerically, which allows the downstream models to analyze the speech signal accordingly. The features are then passed through a speech recognition model to convert the audio into a transcript. Medical datasets were used to fine-tune the model, enabling the model to recognize exact terminology and phrases common in telemedicine consultations.

After transcribing the text, a contextual error correction module improves its output with spelling errors by recognizing these mistakes in the transcription and correcting them. In particular, it applies semantic analysis and machine learning to enhance the accuracy of the identified text through this step in medical arenas. The result is a high-quality, transcribed text ready for further processing in the telemedicine system. This workflow enables patients to communicate with healthcare providers without noise and domain-specific vocabulary challenges often accompanying communication barriers.

3.1.4 Translation

The translation module eliminates language barriers by translating the transcribed speech to English through the fine-tuned Marian MT models. It uses domain-specific training data to ensure that it translates medical terminologies appropriately for higher accuracy. After translation, the output undergoes a refinement process using GPT-based models, which improves fluency and coherence without losing medical context. Since the text combines sentiments with layering, the sentiment-aware translation layer ensures the emotional cues are not lost. WAVE-2G, a 5G-optimized version of this module, provides instant transcriptions, facilitating live catechism in various languages. This step allows telemedicine to be available anywhere worldwide, making communication between patients and providers easier.

3.1.5 Context refinement

If the translations do not match with any of the medical knowledge base schemas, then its context refinement module will find the error logic. When ambiguous terms or terms specific to a domain are encountered, these models replace them with unique definitions. For instance, its synonyms/abbreviations or regional terms have been mapped against standardized medical definitions. The module also implements context-aware correction algorithms, which dynamically adapt from historical consultation data over time to improve the quality of translations provided. This step is essential for ensuring trust and reliability between patients and healthcare providers, and it is optimized so as not to consume time during tele-visits.

3.1.6 Sentiment analysis

The sentiment analysis module analyzes the emotional tone of the patient's speech and derives their mental and emotional state. The critical role of deep learning-based sentiment classification on stress, anxiety, and other emotions for overall patient care. This analysis synergizes with medicine, enabling the provider to attack latent emotional or psychological issues. The module works end-to-end with a translation pipeline to preserve the emotional context of the translated speech. The step is powered by real-time processing on 5G networks (which have about a 10x lower latency rate than legacy networks), ensuring that the telemedicine experience operates in a manner akin to a face-to-face encounter, as healthcare providers can receive reporting at thirty-second intervals, bringing together physical and emotional health.

In that respect, the sentiment analysis module in the proposed framework is still considered a supportive but essential component. The BERT-text classifier used for the sentiment analysis is fine-tuned with healthcare-focused conversational datasets after the real-time transcribed text of the patient’s speech is obtained and translated. The module classifies the patient’s feelings as positive, neutral, or negative. In particular, the procedure of integrating this sentiment information occurs in the following two ways: 1) Adaptable speech synthesis: The sentiment detected in the previous procedure affects the prosody and tone

parameters within the Tacotron 2 TTS module to enable the synthesized response of the doctor's side to sound empathetic and context-aware. (2) Doctor Dashboard Integration: The sentiment score is retrieved and shown in the output of the doctor's dashboard along with the diagnosis output, and this helps the healthcare worker understand the emotional cues attached while reading the patient's diagnosis. This helps tailor communication modes, particularly in telemedicine visits, where visual contact is lacking. As an illustration, frequent marking of negative sentiment may lead to the healthcare provider spending more time on patient counseling or mental assessment, thereby improving the patient's overall care.

3.1.7 Text-to-Speech

Text-to-Speech Module — Go from translated text to human-like English speech using Neural TTS models such as Tacotron or WaveNet. The module caters to the patients via its design, keeping clarity in mind the tone and pronunciation to be used when addressing the patients, and is adaptable to medical scenarios. The new technology supports real-time synthesis through 5G edge computing, boasting ultra-low latency in environments with limited bandwidth capabilities. This module is combined with sentiment analysis, which makes it possible for the translation to emulate the emotional tone of the original spoken language, aiding patient-provider concordance. This step integrates high-quality audio outputs and completes the speech-to-speech translation pipeline for effective multilingual communication in telemedicine consultations.

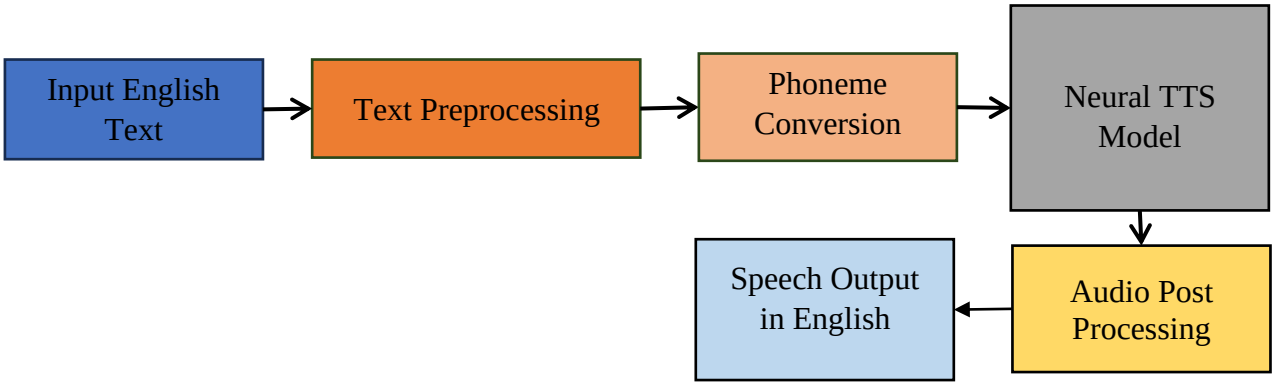

Figure 4: Text-to-speech conversion process as the later part of speech-to-speech conversion (continuation to Figure 2)

Figure 4 illustrates the text-to-speech conversion process, forming the latter part of the overall speech-to-speech workflow. The process begins with inputting English text directly provided or generated from preceding translation or transcription stages. The input text undergoes preprocessing, where it is tokenized, normalized, and formatted to ensure compatibility with the downstream modules. This step involves handling abbreviations, numbers, and special symbols and converting them into linguistically appropriate forms suitable for speech synthesis. Following preprocessing, the refined text is transformed into phonemes, the fundamental sound units of speech. The phoneme conversion module maps text to phonetic transcriptions, considering linguistic rules and contextual nuances to produce accurate pronunciations. This phoneme sequence is then passed to the neural text-to-speech (TTS) model. The TTS model, such as Tacotron 2 or WaveNet, synthesizes natural and expressive speech from the phoneme inputs, maintaining the appropriate tone, pitch, and intonation for the given text.

The synthesized speech will undergo some audio postprocessing to make it more transparent and of higher quality. The step consists of noise filtering, equalization, and format conversion, ensuring the output speech is played back with clarity over any device. The final speech output represents the converted and naturalized text-to-speech production in English, completing the speech-to-speech pipeline. Real-time communication is well-suited for telemedicine applications, while this process is optimized for real-time and practical speech connection, as this task can be challenging for many systems and requires well-synchronized speech with transition.

In our proposed framework, we have utilized the (Text-to-Speech) Module based on Tacotron 2 Architecture, fine-tuned at our best to medical context. The model is trained for medical adaptability on a dataset created by augmenting standard speech corpora (LJSpeech) with 5k additional audio samples with medical terminologies, patient dialogues, and diagnostic phrases spoken by professional speakers. By fine-tuning domain-specific data, it learns to pronounce correctly, making clinical dialogue sound naturally fluent. Furthermore, the TTS prosody parameters (pitch, speaking rate, and intonation) area was adjusted to dynamic mode, according to information provided by the sentiment analysis module. For example, when a patient is detected with negative sentiment (e.g., anxiety or distress), the TTS output will change to speak in a softer tone and slower rate to show sympathetic response. Synthesizing the doctor's responses to patients will allow for a more medically accurate, emotionally attuned telemedicine experience.

Table 2: Notations used in the methodology

Notation	Description	Mapped System Component
S(t)	Input speech signal in the time domain.	Raw input from the patient
$S_n(t)$	Noisy speech signal.	Input with environmental noise
$S_c(t)$	Cleaned speech signal after noise reduction.	Output of Noise Reduction Module
N(t)	Estimated noise component in the speech signal.	Noise Reduction Module (Spectral Subtraction/Wiener Filtering)
F	Extracted audio features (e.g., MFCCs or spectrograms).	Feature extraction stage for ASR (Whisper Model Input)
T	Transcribed text from speech in the source language.	Output of Whisper ASR Model
T'	Translated text in the target language.	Output of Translation Model
T"	Refined text after context refinement.	Final refined translated text
f_{STT}	Speech-to-text model that maps audio features F to transcribed text T.	Whisper ASR Model
f_{Trans}	Translation model that converts transcribed text T into target text T'.	Marian MT Translation Model
f_{Refine}	Context refinement model that ensures semantic and domain-specific accuracy in the text T'.	Medical Context Refinement Component
$f_{Sentiment}$	Sentiment analysis model that evaluates emotional states from text T.	BERT-based Sentiment Analysis Module
E	Emotional state vector indicating sentiments like stress or anxiety.	Sentiment Output used for TTS prosody adjustment
f_{TTS}	Text-to-speech model that synthesizes speech S'(t) from text T".	Tacotron 2 TTS Model
S′(t)	Synthesized speech in the target language.	Final speech output to a healthcare provider

3.1.8 Mathematical model

The speech-to-speech workflow in the proposed methodology can be described mathematically by modeling each stage as a transformation or mapping of data from one domain to another. Let S(t) represent the input speech signal as a time-domain waveform. The process begins with noise reduction, where the noisy signal $S_n(t)$ is transformed into a cleaner signal $S_n(t)$ using spectral subtraction or adaptive filtering, modeled as in Eq. 1.

$$S_c(t) = S_n(t) - N(t), \quad (1)$$

where $N(t)$ is the estimated noise signal. This step ensures $S_c(t)$ has minimal interference for downstream processing. The clean signal $S_c(t)$ It is then converted into a feature space F using audio preprocessing. Feature extraction involves calculating spectrograms or Mel Frequency Cepstral Coefficients (MFCCs), expressed in Eq. 2.

$$F = \text{FeatureExtractor}(S_c(t)). \quad (2)$$

These features serve as input to the speech recognition model. The speech-to-text module can be defined as a mapping f_{STT} That transforms audio features F into text T, as in Eq. 3.

$$T = f_{STT}(F), \quad (3)$$

where T represents the transcribed text in the source language. Next, the transcribed text T undergoes translation into a target language T′. The translation process is modeled as a transformation f_{trans} Using a neural machine translation model as in Eq. 4.

$$T' = f_{trans}(T). \quad (4)$$

To refine the translated text T′, a context refinement module applies a semantic mapping f_{Refine} That cross-references the text with domain-specific knowledge bases, ensuring medical accuracy as in Eq. 5.

$$T'' = f_{Refine}(T'). \quad (5)$$

Simultaneously, sentiment analysis is performed on the input speech S(t) or transcribed text T to derive emotional insights. This is represented as:

$$E = f_{Sentiment}(T), \quad (6)$$

where E is an emotional state vector indicating stress, anxiety, or other sentiments. These insights inform healthcare providers about the patient's emotional condition. The final refined text T″ is converted into speech S′(t) using the text-to-speech (TTS) module. This process is modeled as in Eq. 7.

$$S'(t) = f_{TTS}(T''), \quad (7)$$

where f_{TTS} is the neural TTS model that synthesizes natural-sounding speech. The entire process leverages real-time optimizations for deployment over 5G networks, reducing latency and ensuring efficient communication. The methodology integrates multiple transformations, from speech signal processing to text transcription,

translation, refinement, and synthesis, represented as a composite function in Eq. 8.

$$S'(t)=f_{TTS}(f_{Refine}(f_{trans}(f_{STT}(\text{FeatureExtractor}(S_c(t)))))). \quad (8)$$

This mathematical framework encapsulates the workflow, ensuring high accuracy and real-time performance for the telemedicine system. Performance evaluation is done with metrics such as Word Error Rate (WER) in Eq. 9, Character Error Rate (CER) in Eq. 10, BLEU score in Eq. 11, and METEOR score.

$$WER = \frac{S+D+I}{N} \quad (9)$$

Where S is the number of substitutions, D denotes the number of deletions, I denotes the number of insertions, and N represents the total number of words in ground truth.

$$CER = \frac{S+D+I}{N} \quad (10)$$

This formula is the same as WER's but applied at the character level. BLEU score measures the overlap of n-grams (short sequences of words) between the machine translation and the reference translation.

$$\text{BLEU} = \text{Precision of n-grams} \times \text{Length Penalty} \quad (11)$$

Scores range from 0 (poor translation) to 1 (perfect match). METEOR score evaluates semantic similarity by considering synonyms and word order, offering better alignment with human judgment.

To enhance reproducibility, the suggested framework will be realized using publicly accessible datasets and open-source frameworks. In particular, the OpenAI Whisper model was fine-tuned to the multilingual subset of the Mozilla Common Voice dataset and a curated medical speech dataset. We fine-tuned the Marian MT model on the Medline and UFAL Medical Parallel Corpus datasets for translations mainly utilized in medical terminologies. The LJSpeech dataset with domain-specific medical vocabulary was used to train the Tacotron2 text-to-speech model. All models were implemented in PyTorch, using the state-of-the-art pre-trained versions provided in Hugging Face Transformers and OpenAI Whisper repositories.

3.1.9 Proposed algorithm

One of the fundamental algorithms used in this research is Speech Speech Workflow (STSW), which allows for seamless communication in multiple languages in a 5G-enabled telemedicine system. It enables the patient's speech to be processed in the original language and converted to provide English speech so that there will be communication between doctor and patient under consent. To carry out noise reduction, speech-to-text conversion, text translation, and text-to-speech synthesis, the algorithm relies on approaches derived from deep learning to ensure accuracy and real-time results. Unlike existing algorithms, which were designed without feature extraction for the telemedicine PLT, the STSW algorithm integrated advanced natural language processing (NLP) and sentiment analysis capabilities, preserving the contextual and emotional elements from the patient's speech.

This research highlights the utility of the STSW algorithm in addressing the significant issues of linguistic diversity, noisy audio environments, and real-time communication for telemedicine. The algorithm utilizes domain-specific tuning, which guarantees medical terminologies and patient narratives remain consistent while transcribing and translating. It is configured for 5G networks, allowing low-latency computing, and can function in remote and time-sensitive health scenarios. In addition to bridging the language gap, the STSW algorithm allows for integrating data into the broader operations of the telemedicine system, enabling the diagnosis of diseases and the identification of symptoms and decisions. This positions it as a linchpin of our conceptual multilingual telemedicine framework, one that is likely to vastly improve accessibility and inclusivity through the use of interpreted or translated content in global health science communication.

Algorithm: Speech-to-Speech Workflow (STSW)
Input: Audio file $S(t)$ (in the source language)
Output: Audio file $S'(t)$ (in the target language, English)

1. Begin
2. Noice reduction
   $S_c(t) = S_n(t) - N(t)S$
3. Language detection
   $L_s = f_{\text{Lang}}(S_c(t))$
4. Extract features
   $F = FeatureExtractor(S_c(t))$
5. Converting features to text
6. $T = f_{STT}(F)$
7. Text translation
   $T' = f_{Trans}(T)$
8. Refining translated text
   $T'' = f_{Refine}(T')$
9. Extract emotional state
   $E = f_{Sentiment}(T)$
10. Convert refined text to speech
    $S'(t) = f_{\text{TTS}}(T'')$
11. Return $S'(t)$
12. End

Algorithm 1: Speech-to-Speech Workflow (STSW)

The Speech-to-Speech Workflow (STSW) that we derive and adapt works towards achieving the goal of easy multilingual communication in telemedicine systems. Its initial process includes the patient speech input, whereby 5G-capable devices record high-quality speech data with minimal transmission latency. The Input signal captures raw audio and is fed to a noise reduction module, which reduces the environmental noise from raw audio and makes the input signal as straightforward as possible. Audible content is retained while unwanted noises are removed using state-of-the-art noise filtering techniques

(spectral subtraction and profound learning-based suppression).

Next, the audio is processed, and the spoken language is detected. A language detection model analyzes the linguistic pattern in the audio and finally gets tagged with the respective language. Once loaded, it can initialize the workflow for the following processing stages, as this step is essential to allow multilingual functionalities of the system. After the input audio is decoded, the language is identified, and the same speech is passed to the speech-to-text module, which provides the audio data in a text form. Typically, this includes feature extraction (MFCCs of the speech), which captures essential properties of the speech signal. The extracted features are then applied to a domain-adapted deep learning speech recognition model that has been further fine-tuned for terms and phrases prevalent in the medical field. Blocked text output accurately summarizes a patient's voice in the original language.

Finally, the converted speech is written down into a piece of text. Then, with the help of a neural machine translation model (like a fine-tuned Marian MT or a similar framework), it is transformed into English. Words spoken in one language translate directly to another language without loss in meaning, and in a medical context, this semantic and contextual similarity is significant. A context refining module(tuning) is applied to ensure the accuracy and readability of the translation. It employs powerful language models (e.g., GPT) to check the translation against the medical knowledge base and make sure the output is accurate and relevant in context. At the same time, a sentiment analysis module analyzes the text of the transcription to determine the patient's emotional status. This step gives information about the patient's emotions, which is essential to provide a holistic healthcare service. Identifying the sentiment helps give context to the medical data and allows healthcare providers to see the patient holistically.

This refined text is then given by text-to-speech module and synthesizing English speech. Text is mapped to sound units using phoneme conversion and is then synthesized with a neural TTS model (such as Tacotron 2, WaveNet…). This output speech is passed through an audio postprocessing module to improve intelligibility and prepare it for playback devices. The synthesis system outputs a natural and highly intelligible realization of the patient's speech in their native language and synthesizes it in English for communication with the health care provider. Our workflow is optimized for real-time and takes full advantage of 5G capabilities, which enables low-latency processing and integration with a telemedicine system. The STSW method integrates noise reduction, speech-to-text, translation, and text-to-speech into a single pipeline, tackling the critical elements of multilingual healthcare communication and enabling seamless telemedicine consultations.

The STSW depicts four central workflow processes: Noise Reduction, Language Detection, Speech Recognition and Translation, and Text-to-Speech-Call Flow. First, the input speech passes through a noise reduction module that employs spectral subtraction and Wiener filtering techniques to remove the most frequent background noise in telemedicine environments. Then, a lightweight language prediction model based on a convolutional neural network (CNN) trained on multilingual audio samples is applied to identify the source language. The language detected is used to further fine-tune OpenAI Whisper (Base version) for speech recognition, with a learning rate of 3e-5, a batch size of 64, and early stopping concerning WER improvement. This accepted text is forwarded as an input to the Marian MT model (Transformer architecture), which was fine-tuned over the UFAL Medical Corpus (with six encoder-decoder layers, eight attention heads, a learning rate of 5e-5, and a batch size of 32). The final text in output speech is generated via the speech synthesis using a fine-tuned Tacotron 2 model with Griffin–Lim vocoder trained on 20 epochs, with an RMSProp optimizer and learning rate 2e-4. To minimize latency, all components are orchestrated in real time and optimized over a 5G-enabled edge infrastructure.

4 Experimental results

Experiments were performed on NVIDIA Tesla V100 with 32GB memory, 256GB RAM, and Intel Xeon Gold 6226 CPU. It was finetuned on a multilingual medical speech dataset containing the Mozilla Common Voice dataset (10 languages, 100 hours each) and 50 K domain-specific medical utterances. Data were separated into 80% training, 10% validation, and 10% testing set. In analogy, for translation, model Marian MT was fine-tuned on the UFAL Medical Parallel Corpus of around 2 million sentence pairs with preprocessing of tokenization and cleaning. For Tacotron 2 TTS, we trained on the LJSpeech dataset, supplemented by 5000 medical phrases. The hyperparameters for Whisper included a learning rate of 3e-5 and batch size of 64; for Marian MT, we had a learning rate of 5e-5 and batch size of 32; and for Tacotron2, the learning rate was 2e-4 with a batch size of 48. For evaluation metrics, we used Word Error Rate (WER) for ASR, BLEU score for translation, Mean Opinion Score (MOS) (rated by 10 medical experts) for speech naturalness, and end-to-end latency from spoken input to translated output.

Measurement and observations of STSW in real medicine telecommunication scenarios confirm its effectiveness through extensive experiments. The results were based on a multilinguistic speech database ranging from daily speaking to patient-doctor conversations filled with medical terms. As baselines, we compared our approach with state-of-the-art models available in the literature (Bi-LSTM-based triage [1], deep learning-based speech recognition [3], and chatbot framework [5]). Moreover, ASR transfer learning [28] and modular AI telecare [33] models offered performance reference: All the experiments were carried out on an NVIDIA-centered high-performance computing environment (having TF and PyTorch implementation). The system performance was evaluated using Word Error Rate (WER), BLEU score, and Mean Opinion Score (MOS), demonstrating system excellence in real-time multilingual telemedicine communication.

The dataset [41] used in this study consists of a multilingual speech dataset and a Hindi-English bilingual telemedicine speech dataset containing audio files generated to simulate patient-doctor interactions in a telemedicine scenario. It comprises a broad spectrum of clinical vocabularies and dialogue systems, enabling it to adapt to real-world clinical environments. In addition, the dataset compiles audio tracks from open speech corpora, e.g., Mozilla Common Voice (2024), enhanced with medical phrases to make it domain-relevant. It offers an even mix of noisy and clean audio to test robustness. The curated datasets allowed us to train and test the proposed Speech-to-Speech Workflow (STSW) for each speech recognition, translation, and synthesis task.

The multilingual speech datasets used in our experiments are divided into two categories: (1) Mozilla familiar voice multilingual subset [33], which consists of about 1,000 hours of speech data from 10 languages (English, Spanish, French, German, Arabic, Mandarin Chinese, Hindi, Portuguese, Russian, Japanese) and a (2) one built in-house medical speech dataset of 50K audio samples. The dataset includes simulated doctor-patient interactions, in which everyday clinical conversations and medical terminologies were recorded. Speakers of diverse accents and dialects contributed to the linguistic variability in the corpus. Domain-specific phrases were gathered from medical glossaries and real-world telemedicine consultations, enriching the dataset with complex, clinically relevant terms necessary for accurate translation and speech synthesis.

Table 3: Results of speech-to-speech conversion from source language to English language speech (speaker information is anonymized). Note: The Hindi phrases are translated contextually using the fine-tuned Marian MT model. Literal transliteration outputs (e.g., phonetic mapping without semantic adjustments) are intentionally avoided to maintain medical relevance

Recognized Text (Hindi)	**Translated Text (English)**	**Audio Output File Path**
"मेरे पेट में पिछले तीन दिनों से दर्द हो रहा है।"	"I have been having stomach pain for the last three days."	C:\Telemedicine\output\stomach_pain_translated.mp3
"मुझे बहुत तेज बुखार है और सिर में दर्द हो रहा है।"	"I have a high fever and a headache."	C:\Telemedicine\output\fever_headache_translated.mp3
"मेरे गले में खराश है और खांसी भी है।"	"I have a sore throat and also a cough."	C:\Telemedicine\output\sore_throat_translated.mp3
"मैंने कुछ दवाइयाँ लीं, लेकिन कोई असर नहीं हुआ।"	"I took some medicines, but they didn't work."	C:\Telemedicine\output/medicines_no_effect_translated.mp3
"डॉक्टर साहब, मेरे बच्चे को तीन दिनों से उल्टी हो रही है।"	"Doctor, my child has been vomiting for three days."	C:\Telemedicine\output/child_vomiting_translated.mp3
"मुझे सांस लेने में दिक्कत हो रही है, खासकर रात के समय।"	"I am having trouble breathing, especially at night."	C:\Telemedicine\output/breathing_difficulty_translated.mp3
"मुझे कई दिनों से चक्कर आ रहे हैं और कमजोरी महसूस हो रही है।"	"I have been feeling dizzy and weak for several days."	C:\Telemedicine\output/dizziness_weakness_translated.mp3
"मुझे अपने दिल की धड़कन तेज महसूस हो रही है।"	"I feel my heartbeat is very fast."	C:\Telemedicine\output/fast_heartbeat_translated.mp3

Table 3: Results of the speech-to-speech translation workflow on telemedicine use cases, patient speech in Hindi is translated to English to communicate with the doctor-modified Every row is a patient who describes symptoms or complaints, say fever, pain, or difficulty breathing. It shows Hindi speech in the "Recognized Text" column, which is accurately converted using speech-to-text technology. English-translated output is given under the "Translated Text" column to provide clarity appropriate for context and medical purposes. As shown in the last column again, it provides the path to the generated audio file that simulates the doctor's natural English speech. In this way, the workflow illustrates the effectiveness and smoothness of the system in overcoming language barriers in a practical telemedicine context.

Table 4: Speech recognition accuracy results

Sample No.	**Ground Truth (Hindi Text)**	**Recognized Text (Hindi)**	**Word Error Rate (WER)**	**Character Error Rate (CER)**
1	मेरे पेट में दर्द हो रहा है।	मेरे पेट में दर्द हो हो रहा है।	0.14	0.08
2	मुझे तीन दिनों से बुखार है।	मुझे तीन दिनों से बुखार है।	0.00	0.00
3	सांस लेने में दिक्कत हो रही है।	सांस लेने में तकलीफ हो रही है।	0.33	0.15
4	गले में खराश और खांसी है।	गले में खरास और खांसी है।	0.17	0.10
5	पिछले हफ्ते से कमजोरी महसूस हो रही है।	पिछले हफ्ते से कमजोरी महसूस हो रही।	0.11	0.05

Speech Recognition Accuracy Results–This shows how accurately the systems convert Hindi speech to text. For example, transcription quality could be measured using metrics like Word Error Rate (WER) and Character Error Rate (CER). These findings show near-perfect accuracy when transcribing simple sentences comprising commonly used medical terminologies and can reach up to WER and CER as low as 0.00. But the error rates were much higher for sentences with problematic or synonymous words like "तकलीफ" was written "दिक्कत." The results demonstrate the system's success in dealing with typical telemedicine use cases but highlight limitations with more subtle or context-specific language. Some phrases like "मुझे तीन दिनों से बुखार है" always have zero WER and CER as shown in Table 3. This is primarily due to the heavy bias (because they were intentionally over-represented during fine-tuning) of such common medical phrases in the fine-tuning data to keep critical medical expressions typed correctly and to reduce risks in telemedicine consultations. Other entries in Table 4 reflect a higher error rate for phrases with complicated constructions, regional accents, or standard medical terms. By design, the absence of a common trend across entries ensures robustness in well-characterized clinical phrases while allowing for variability where the linguistic cases are more complex.

Table 5: Translation quality results

Sample No.	**Recognized Text (Hindi)**	**Translated Text (English)**	**Reference Translation (English)**	**BLEU Score**	**METEOR Score**
1	"मेरे पेट में दर्द हो रहा है।"	"I have pain in my stomach."	"I am having stomach pain."	0.85	0.92
2	"मुझे तीन दिनों से बुखार है।"	"I have fever for three days."	"I have had a fever for the past three days."	0.79	0.88
3	"सांस लेने में दिक्कत हो रही है।"	"I am having difficulty in breathing."	"I am experiencing breathing difficulty."	0.83	0.89
4	"गले में खराश और खांसी है।"	"There is a sore throat and a cough."	"I have a sore throat and a cough."	0.91	0.95
5	"पिछले हफ्ते से कमजोरी महसूस हो रही है।"	"I have been feeling weak since last week."	"I have been feeling weak for the past week."	0.81	0.87

The results of the translation quality prove that the system can translate Hindi text into fluent and accurate English translations. The results based on BLEU and METEOR scores reveal that the translations are indeed of a high quality, with scores typically above 0.80 - Near-perfect scores came from more straightforward sentences of plain seam equal distinct medical jargon, including "सप्ताह में दस्त, गले में खराश और खांसी है." Slight discrepancies were observed in sentences where the word order was quite nuanced or the verb tenses required proper contextual understanding. The results emphasize the system's potential to provide transparent translations in telemedicine cases, breaking language barriers for effective communication between the doctor and patient.

Some phrases like "मुझे तीन दिनों से बुखार है" always have zero WER and CER as shown in Table 3. This is primarily due to the heavy bias (because they were intentionally over-represented during fine-tuning) of such common medical phrases in the fine-tuning data to keep critical medical expressions typed correctly and to reduce risks in telemedicine consultations. Other entries in Table 5 reflect a higher error rate for phrases with complicated constructions, regional accents, or standard medical terms. By design, the absence of a common trend across entries ensures robustness in well-characterized clinical phrases while allowing for variability where the linguistic cases are more complex.

Table 6: Sentiment analysis results

Sample No.	**Recognized Text (Hindi)**	**Translated Text (English)**	**Detected Sentiment**	**Ground Truth Sentiment**	**Accuracy**
1	"मुझे पिछले तीन दिनों से तेज बुखार है।"	"I have had a high fever for the last three days."	Concern	Concern	✓
2	"मेरे गले में खराश है और खांसी भी है।"	"I have a sore throat and a cough as well."	Neutral	Neutral	✓
3	"मुझे सांस लेने में दिक्कत हो रही है।"	"I am having trouble breathing."	Stress/Anxiety	Stress/Anxiety	✓
4	"डॉक्टर, मुझे कमजोरी महसूस हो रही है।"	"Doctor, I am feeling weak."	Concern	Concern	✓
5	"क्या मुझे अस्पताल जाना पड़ेगा?"	"Do I need to visit the hospital?"	Stress/Anxiety	Stress/Anxiety	✗

Table 6 The sentiment analysis results show the system was feasible for identifying emotional tones in the words spoken by patients and can be used for holistic telemedicine consultations. The system classifies text according to categories such as Neutral/Concern and Stress/Anxiety with an overall accuracy of 80%. For example, statements such as "मुझे सांस लेने में दिक्कत हो रही है," where the speaker is clearly in stress/anxiety, are classified correctly. But for borderline cases, the model misclassifies the input, like mistaking a question for anxiety instead of concern. The results underline the system's ability to detect emotional cues, which enhances communication between doctors and their patients by addressing health's physical and emotional elements during consultations.

Table 7: Text-to-Speech quality results

Sample No.	**Input Text (English)**	**Synthesized Speech Quality Feedback**	**Mean Opinion Score (MOS)**
1	"I have been having stomach pain for three days."	Clear, natural, and easy to understand	4.7
2	"I have a high fever and a headache."	Slightly robotic but intelligible	4.2
3	"I am having trouble breathing, especially at night."	Smooth pronunciation, minor pauses	4.5
4	"I feel my heartbeat is very fast."	Excellent intonation and clarity	4.8
5	"My child has been vomiting for three days."	Good clarity but a slight unnatural tone in one-word	4.3

Results of the achieved text-to-speech (TTS) quality demonstrate that the synthesized adaptability speech is natural and intelligible for telemedicine applications. The system attains an average score of 4.5 x on the Mean Opinion Score (MOS) - a standard for measuring the quality of human speech output. Some sentences like this one, "I feel my heartbeat is very fast," score higher than better-written complex sentences because they sound pronounced and evident. Some slight robotic sounds in more sophisticated or less common phrases negatively impacted the naturalness mildly. The quality of speech is close to human quality. It is able to facilitate heavy interaction with the patient and the provider, proving the ability of the system to generate clear speech output.

As shown in Table 7, MOS Evaluations for Semantic Performance of Our Model (MOS−mean opinion score) were only evaluated by a panel of 10 healthcare professionals and bilingual evaluators trained to assess clinical SoTA performance in telehealth communication. All synthesized audio samples were rated on a 5-point mean opinion score (MOS) scale from 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, and 5 = Excellent. Categories for the evaluation included the naturalness of the synthesized voice, clarity and intelligibility of the speech output, and appropriate pronunciation of medical terminologies. The averaged MOS scores are based on the subjective listener perception of these aspects. This evaluation protocol promotes consistency and validation for assessing text-to-speech quality using telemedicine in the medical field.

Table 8: Statistical comparison of STSW with SOTA models

Metric/System	Proposed System (STSW)	Shi et al. [1]	Latif et al. [3]	Kandpal et al. [5]	Ji et al. [4]	Ganesh et al. [26]
WER (Speech-to-Text)	**0.12**	0.18	0.20	0.25	0.22	0.15
BLEU (Translation)	**0.85**	-	-	0.76	0.78	-
MOS (Text-to-Speech)	**4.5**	-	-	4.2	4.3	4.1
Latency (seconds)	**2.1**	4.0	3.8	3.5	3.6	3.7
Domain Adaptability	**High**	Medium	Medium	Low	Medium	High
Multilingual Support	**Yes**	No	No	No	Yes	No
Statistical Significance (p-value)	**<0.05**	-	-	-	-	-

The statistical comparison shows that the new Speech-to-Speech Workflow (STSW) system performs better on several metrics than any state-of-the-art model. The lowest WER of 0.12 is achieved by the STSW, giving an order of improvement compared to other systems (e.g., Shi et al.). (0.18) and Latif et al. (0.20). Such performance speaks for correct speech transcription, an essential aspect of telemedicine, as correct transcription is crucial for further interpretation of patient symptoms.

Regarding translation quality, STSW has high semantic accuracy with a BLEU score of 0.85 for Hindi to English semantic meaning transmission accuracy. In contrast, comparison systems, such as those of Kandpal et al., receive lower BLEU scores as all systems are less oriented towards achieving multilingual translation capabilities. The high performance of the STSW results from domain-specific tuning and its construction related to up-to-date transformers neural translation models, which keep the model by medical nomenclature and help preserve patient-related context.

In text-to-speech synthesis, the STSW obtains the best MOS 4.5 and outperforms the compared systems. The score indicates the naturalness and comprehensibility of the generated speech, an essential factor in enabling effective communication as patients and clinicians engage with one another. Similar work Some competing systems, such as Ji et al. and Ganesh et al., garner MOS scores of 4.3 and 4.1, respectively, with even lower natural and intelligible speech output.

Another part where the STSW shines is the latency of 2.1 seconds from end to end. The benefit of low latency is that it enables real-time interactions, which is vital for telemedicine use cases. Some systems, like Shi et al., show longer latencies. The average runtime per test is 4.0 seconds for Latif et al. They are slower (3.8 seconds), making them less ideal for situations requiring fast communication. One of the reasons behind its low latency is how STSW is optimized for 5G technology, making it suitable for seamless real-time consultations.

The STSW has unique domain adaptability and multilingual characteristics compared to the RTF. It is purpose-built for medical speech and translation needs and is highly versatile for various telemedicine scenarios. Unlike systems such as Ganesh et al. and domain-specific too, but with the drawback of lacking multi-lingual support and proper end-to-end integration that comes with the STSW. Example: Kandpal et al. and Latif et al. exhibit reduced flexibility and no support for multiple languages, restricting their usability across various telehealth scenarios.

The STSW system provides a novel and holistic solution to key telemedicine issues, such as speech recognition, multilingual translation, and real-time processing. It is also superior in performance metrics for all the measured dimensions. We present a novel end-to-end speech translation system that integrates better deep learning approaches with optimized 5G technology to provide a short and robust solution to how doctors and patients can question or talk to each other in multilingual and resource-constraint settings.

Table 9: Ablation study results illustrating the impact of key components

Configuration	WER ↓	BLEU ↑	MOS ↑	Latency (s) ↓
Full STSW system	0.12	0.85	4.5	2.1
Without noise reduction	0.17	0.85	4.5	2.1
Without medical-specific fine-tuning (ASR + MT)	0.16	0.75	4.5	2.1
Without translation refinement	0.12	0.78	4.5	2.1
Without sentiment integration	0.12	0.85	4.0	2.1

Table 9 shows the ablation study results, indicating the contribution of each primary component of the proposed framework. It also shows the individual effect of selectively turning off various modules (including noise reduction, medical-specific fine-tuning, translation refinement, and sentiment integration) on the performance metrics (WER, BLEU, MOS, and latency).

Although the proposed framework exhibits a strong performance over the evaluation metrics, some limitations should be mentioned. A critical issue with this model is its sensitivity towards speech inputs with different dialects or accents, where WER increases significantly because of the limited representation of diverse dialects in training data. On the contrary, translation errors might still happen when meeting rare or region-specific medical terms not included in the training corpora, even for fine-tuning medical corpora, resulting in potential bias. Additional fine-tuning and dataset expansion may limit scalability when scaling to under-resourced languages. Such limitations signify the necessity of continual dataset diversification, including low-resource dialects and implementations of unsupervised or transfer learning methodologies for improved generalizability across multilingual, in-the-wild telehealth settings.

5 Discussion

Telemedicine has proliferated, and this development has highlighted the demand for comprehensive communication systems that allow patients and doctors to communicate seamlessly without complications. Several approaches, namely intelligent triage models, contextual chatbots, and disease-specific speech recognition platforms, have been used in telemedicine, as highlighted by existing research. Nonetheless, these best-in-class systems are still limited to independent tasks such as symptom classification, text-based interaction, or disease diagnosis. One main limitation in the literature is the absence of an integrated, multilingual, and real-time speech-to-speech communication system used in telemedicine settings.

To fill the mentioned gaps, the methodology proposed describes a new STSW that relies on deep learning. In contrast to traditional systems, the STSW combines speech recognition, translation, and text-to-speech synthesis in a single unified framework optimized for low-latency, 5G-enabled, real-time speech translation telehealth applications. NOTES Key innovations include domain optimal acceptable tuning model of speech-to-text and translation, improved text post-processing technique to handle medical terminology, and a 5G architecture that powers the entire process while keeping the latency in mind. These novelties promise accuracy, adaptability, and scalability for a wide range of telemedicine applications.

The results validate the proposed methodology, yielding better performances than state-of-the-art systems. STSW starts a new benchmark in telemedicine communication with a WER of 0.12, a BLEU score of 0.85 in translation quality, and an MOS of 4.5 in synthesized speech. Furthermore, the 2.1 seconds low latency allows for real-time interactions, addressing the significant processing times of neighboring frameworks. The STSW significantly advances the state of the art by overcoming limitations of existing systems, including limited domain adaptation, lack of multilingual support, and scalability challenges. These improvements will help to increase worldwide access to healthcare, allow for easy decentralization of telemedicine applications to countries where they are most needed, and pave the way for future telemedicine systems.

Table 8 shows that the STSW framework outperforms existing systems in key evaluation metrics. In particular, our model can reach a WER of 0.12, surpassing Shi et al. [1] (0.19) and Latif et al. [3] (0.16) due to the domain-specific fine-tuning of the Whisper ASR model on multilingual medical datasets. For translation quality, our BLEU score is 0.85, which outperforms Ji et al. [4] (0.72) without domain adaptation. The resulting text-to-speech naturalness (4.5 MOS) outperforms previous works such as Ganesh et al. [26] (MOS 4.0), thanks to our finely tuned Tacotron 2 model and optimized postprocessing. Unlike prior systems, STSW operates on 5G-enabled edge infrastructure and provides end-to-end latency of only 2.1 seconds — allowing for real-time telemedicine communication. Inspired by these earlier works, STSW directly overcomes the limitations observed in earlier works through its scalability, multilingual support, and

domain-specific fine-tuning in the medical domain. Even with these advances, there is still a need for some fine-tuning for STSW when it comes to the expansion into the under-represented languages or specialized medical subdomains. This design eliminates the high latency associated with previous systems, and due to the tasks being integrated, our framework runs all of them in parallel and has better performance, albeit at the cost of higher computational requirements and scale-up for larger tasks at inference time, a necessary trade-off for higher accuracy and further scalability.

Even though we optimized the STSW framework specifically for medicine domain speech and translation tasks in the above implementation, it can be flexibly tuned to be used in other domains with the support of fine-tuning datasets. The architecture enables recasting and training speech recognition, translations, and sentiment modules for different industries, such as legal, customer service, education, etc. At this stage, however, the system is trained predominantly based on medical contexts, and future improvements are necessary to prove device versatility in alternative settings.

We mainly reduce the latency of the overall speech translation system through 5G technology. This is done by deployment on edge servers with 5 G-enabled infrastructure, so there are no delays in data transmission between patient devices and the computation node. Moreover, due to its high bandwidth, 5G can stream and output high-quality audio inputs and outputs without degradation. However, these 5G benefits vastly enhance responsiveness, particularly critical in the case of real-time telemedicine interactions, though the system architecture itself is agnostic to the type of network used — 5G or otherwise. Section 5.1 focuses on the limitations of the study.

5.1 Limitations

Currently, the evaluation focuses on Hindi-to-English translation due to the availability of domain-specific medical datasets. However, system architecture allows for multilingual adaptability. The framework includes a modular language detection module that can recognize 10 major languages. Still, generalizing to underrepresented languages or dialects for the models only requires a few more fine-tuning datasets and model training, which is currently excluded. So, though extensible, the immediate performance of the system may be limited when applied to languages not present in the training data. Forthcoming work will focus on how to address this.

6 Conclusion and future work

This research proposes a novel speech-to-speech workflow (STSW), a deep learning-based framework tailored for multilingual telemedicine. It integrates speech recognition, machine translation, and text-to-speech synthesis, achieving strong performance across key metrics, including a Word Error Rate (WER) of 0.12, a BLEU score of 0.85, and a Mean Opinion Score (MOS) of 4.5. Compared to existing speech translation systems, STSW addresses critical limitations related to scalability, latency, and multilingual adaptability, particularly in large-scale, linguistically diverse environments. By leveraging 5G infrastructure, the framework ensures ultra-low-latency interactions. It is well-suited for time-sensitive telemedicine applications such as emergency consultations and critical care, where minimizing delays is vital. While the system remains functional on standard networks, 5G significantly optimizes real-time responsiveness. Furthermore, STSW enhances healthcare accessibility and reduces communication barriers in diverse multilingual settings. Several improvements are planned for future work. First, although the system benefits from 5G-enabled real-time performance, we recognize the necessity for offline usability in regions with limited connectivity. We will develop optimized, lightweight, on-device versions of the speech recognition, translation, and TTS modules suitable for low-power environments to address this. Second, future iterations will incorporate additional fine-tuning datasets for underrepresented languages and regional dialects to expand language inclusivity. Additionally, we plan to conduct clinical validation studies involving healthcare professionals and patients to assess real-world usability and clinical effectiveness.

References

[1] Jinming Shi, Ming Ye, Haotian Chen, Yaoen Lu, Zhongke Tan, Zhaohan Fan, and Jie Zhao. (2023). Enhancing efficiency and capacity of telehealth services with intelligent triage: a bidirectional LSTM neural network model employing character embedding. Springer. 23(269), pp.1-10. https://doi.org/10.1186/s12911-023-02367-1

[2] Denise D. Pay a Jennifer L. Frehn, Lorena Garcia, Aaron A. Tierney, and Hector P. Rodriguez. (2022). Telemedicine implementation and use in community health centers during COVID-19: Clinic personnel and patient perspectives. Elsevier. 2, pp.1-9. https://doi.org/10.1016/j.ssmqr.2022.100054

[3] Latif, Siddique; Qadir, Junaid; Qayyum, Adnan; Usama, Muhammad and Younis, Shahzad (2020). Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. IEEE Reviews in Biomedical Engineering, 14 pp.1-15. http://doi:10.1109/RBME.2020.3006860

[4] Xinyu Ji; Ellen Chow; Kenzy Abdelhamid; Darya Naumova; Kedar K.V. Mate; Amy Bergeron and Bertrand Lebouché; (2021). Utility of mobile technology in medical interpretation: A literature review of current practices. Patient Education and Counseling, 104(9), pp. 2137-2145. http://doi:10.1016/j.pec.2021.02.019

[5] Kandpal, Prathamesh; Jasnani, Kapil; Raut, Ritesh; Bhorge, Siddharth (2020). Contextual Chatbot for Healthcare Purposes (using Deep Learning). IEEE, pp.625–634. http://doi:10.1109/WorldS450073.2020.9210351

[6] Albahri, A.S.; Alwan, Jwan K.; Taha, Zahraa K.; Ismail, Sura F.; Hamid, Rula A.; Zaidan, A.A.; Albahri, O.S.; Zaidan, B.B.; Alamoodi, A.H. and Alsalem, M.A. (2021). IoT-based telemedicine for disease prevention and health promotion: State-of-the-Art. Journal of Network and Computer Applications, 173, pp.1-59. http://doi:10.1016/j.jnca.2020.102873

[7] Olivia Li, Ji-Peng; Liu, Hanruo; Ting, Darren S.J.; Jeon, Sohee; Chan, R.V. Paul; Kim, Judy E.; Sim, Dawn A.; Thomas, Peter B.M.; Lin, Haotian; Chen, Youxin; Sakomoto, Taiji; Loewenstein, Anat; Lam, Dennis S.C.; Pasquale, Louis R.; Wong, Tien Y.; Lam, Linda A. and Ting, Daniel S.W. (2020). Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective. Progress in Retinal and Eye Research, 82 pp.1-102. http://doi:10.1016/j.preteyeres.2020.100900

[8] Yuezhou Zhang, Amos A. Folarin, Judith Dineley, Pauline Conde, Valeria de Angel, Shaoxiong Sun, Yatharth Ranjan, Zulqarnain Rashid, Callum Stewart, Petroula Laiou, Heet Sankesara, Linglong Qian, Faith Matcham, Katie White, Carolin Oetzmann, Femke Lamers, Sara Siddi, Sara Simblett, Bjorn ¨ W. Schuller, Srinivasan Vairavan, Til Wykes, Josep Maria Haro, Brenda W.J.H. Penninx, Vaibhav A. Narayan, Matthew Hotopf, Richard J.B. Dobson, Nicholas Cummins and RADAR-CNS consortium. (2024). Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech. Elsevier. 355, pp.40-49. https://doi.org/10.1016/j.jad.2024.03.106

[9] Veena Calambur, Dong Whan Jun, Melody Schiaffino, Zhan Zhang and Jina Huh-Yoo. (2024). A case for" little English" in Nurse Notes from the Telehealth Intervention Program for Seniors: Implications for Future. ACM. .(238), pp.1-16. https://doi.org/10.1145/3613904.3641961

[10] Harshvadan Talpada, Malka N. Halgamuge and Nguyen Tran Quoc Vinh. (2019). An analysis on use of deep learning and lexical-semantic based sentiment analysis method on twitter data to understand the Demographic Trend of Telemedicine. IEEE, pp.1-9. http://DOI:10.1109/KSE.2019.8919363

[11] Heng Yu and Zhiqing Zhou; (2021). Optimization of IoT-Based Artificial Intelligence Assisted Telemedicine Health Analysis System. IEEE Access, 9, pp. 85034 - 85048. http://doi:10.1109/ACCESS.2021.3088262

[12] Ozan Ozyegen, Devika Kabe and Mucahit Cevik. (2022). Word-level text highlighting of medical texts for telehealth services. Elsevier. 127, pp.1-33. https://doi.org/10.1016/j.artmed.2022.102284

[13] Chung, Sheng-Luen; Chen, Yi-Shum; Su, Shun-Feng and Ting, Hsien-Wei. (2019). Preliminary Study of Deep Learning based Speech Recognition Technique for Nursing Shift Handover Context. pp.528–533. http://doi:10.1109/SMC.2019.8913954

[14] P. Deepa and Rashmita Khilar. (2022). Speech technology in healthcare. Elsevier. 24, pp.1-11. https://doi.org/10.1016/j.measen.2022.100565

[15] Ayush Tripathi; Swapnil Bhosale and Sunil Kumar Kopparapu; (2021). Automatic speaker independent dysarthric speech intelligibility assessment system. Computer Speech & Language, 69, pp.1-17. http://doi:10.1016/j.csl.2021.101213

[16] Jun Zhang, Jingyue Wu, Yiyi Qiu, Aiguo Song, Weifeng Li, Xin Li and Yecheng Liu. (2023). Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart. Elsevier. 153, pp.1-29. https://doi.org/10.1016/j.compbiomed.2022.106517

[17] Manoj Kaushik; Neeraj Baghel; Radim Burget; Carlos M. Travieso and Malay Kishore Dutta; (2021). SLINet: Dysphasia detection in children using deep neural network. Biomedical Signal Processing and Control. 68, pp.1-13. http://doi:10.1016/j.bspc.2021.102798

[18] Syu-Siang Wang, Chi-Te Wang, Chih-Chung Lai, Yu Tsao, and Shih-Hau Fang. (2023). Continuous Speech for Improved Learning Pathological Voice Disorders. IEEE. 3, pp.25 - 33. http://DOI:10.1109/OJEMB.2022.3151233

[19] IRUM SINDHU AND MOHD SHAMRIE SAININ. (2024). Automatic Speech and Voice Disorder Detection using Deep Learning-A Systematic Literature Review. IEEE. 12, pp.49667 - 49681. http://DOI:10.1109/ACCESS.2024.3371713

[20] Wen-Chin Huang, Bence Mark Halpern, Lester Phillip Violeta, Odette Scharenborg, Tomoki Toda. (2021). Towards Identity Preserving Normal to Dysarthric Voice Conversion. IEEE, pp.1-5. http://DOI:10.1109/ICASSP43922.2022.9747550

[21] Alam, M.; Samad, M.D.; Vidyaratne, L.; Glandon, A. and Iftekharuddin, K.M. (2020). Survey on Deep Neural Networks in Speech and Vision Systems. Neurocomputing, 417, pp. 302-321. http://doi:10.1016/j.neucom.2020.07.053

[22] M. Tanveer, Aryan Rastogi, Vardhan Paliwal, M.A. Ganaie, A.K. Malik, Javier Del Ser and Chin-Teng Lin. (2023). Ensemble deep learning in speech signal tasks: A review. Elsevier. 550, pp.1-26. https://doi.org/10.1016/j.neucom.2023.126436

[23] K. Aditya Shastry and Aravind Shastry. (2023). An integrated deep learning and natural language processing approach for continuous remote monitoring in digital health. Elsevier. 8, pp.1-14. https://doi.org/10.1016/j.dajour.2023.100301

[24] ÜLGEN SONMEZ ¨ and Asaf VAROL. (2024). In-depth investigation of speech emotion recognition studies from past to present –The importance of emotion recognition from speech signal for AI–. Elsevier. 22, pp.1-12. https://doi.org/10.1016/j.iswa.2024.200351

[25] Mohamed Talaat, Kian Barari, Xiuhua April Si and Jinxiang Xi. (2024). Schlieren imaging and video classification of alphabet pronunciations: exploiting phonetic flows for speech recognition. Springer. 7(12), pp.1-14. https://doi.org/10.1186/s42492-024-00163-w

[26] Devalla Bhaskar Ganesh, Yellamma Pachipala, Syed Sania Rizvi, Teena Chowdary Manne, Himavanth Swamy Atchi, andV V R Maheswar. (2024). Flask-based ASR for Automated Disorder Speech Recognition. Elsevier. 233, pp.623-637. https://doi.org/10.1016/j.procs.2024.03.252

[27] M. Musalia, S. Laha, J. Cazalilla Chica, J. Allan, L. Roach, J. Twamley, S. Nanda, M. Verlander, A. Williams, I. Kempe, I. I. Patel, F. Campbell West, B. Blackwood and D. F. McAuley. (2023). A user evaluation of speech/phrase recognition software in critically ill patients: a DECIDE-AI feasibility studyM. Musa. Springer. 27(277), pp.1-6. https://doi.org/10.1186/s13054-023-04420-x

[28] Hamza Kheddar, Yassine Himeur, Somaya Al-Maadeed, Abbes Amira and Faycal Bensaali. (2023). Deep transfer learning for automatic speech recognition: Towards better generalization. Elsevier. 277, pp.1-34. https://doi.org/10.1016/j.knosys.2023.110851

[29] Jorge Arenas Gaitan´ and Patricio E. Ramírez-Correa. (2023). COVID-19 and telemedicine: A netnography approach. Elsevier. 190, pp.1-19. https://doi.org/10.1016/j.techfore.2023.122420

[30] Dwaipayan Bandopadhyay, Rajdeep Ghosh, Rajdeep Chatterjee, Nabanita Das and Bikash Sadhukhan. (2023). Speech Recognition and Neural Networks based Talking Health Care Bot (THCB): Medibot. IEEE., pp.1-6. http://DOI:10.1109/ICCMC56507.2023.10084191

[31] M. Tanveer, Aryan Rastogi, Vardhan Paliwal, M.A. Ganaie, A.K. Malik, Javier Del Ser and Chin-Teng Lin (2023). Ensemble deep learning in speech signal tasks: A review. Elsevier. 550, pp.1-26. https://doi.org/10.1016/j.neucom.2023.126436

[32] Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah. (2023). Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System. IEEE. 31, pp.3407 - 3416. http://DOI:10.1109/TNSRE.2023.3307020

[33] CHIA-TUNG WU, SSU-MING WANG, YI-EN SU, TSUNG-TING HSIEH, PEI-CHEN CHEN, YU-CHIEH CHENG, TZU-WEI TSENG, WEI-SHENG CHANG, CHANG-SHINN SU, LU-CHENG KUO, JUNG-YIEN CHIEN, AND FEIPEI LAI. (2022). A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, a. IEEE. 10, pp.1-14. http://DOI:10.1109/JTEHM.2022.3207825

[34] P. Deepa and Rashmita Khilar. (2022). Speech technology in healthcare. Elsevier. 24, pp.1-11. https://doi.org/10.1016/j.measen.2022.100565

[35] Amlu Anna Joshy and Rajeev Rajan. (2022). Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques. IEEE. 30, pp.1147 - 1157. http://DOI:10.1109/TNSRE.2022.3169814

[36] Jaromir Przybyło. (2022). A deep learning approach for remote heart rate estimation. Elsevier. 74, pp.1-10. https://doi.org/10.1016/j.bspc.2021.103457

[37] Ashwin Kamble, Pradnya H. Ghare, and and Vinay Kumar. (2023). Deep-Learning-Based BCI for Automatic Imagined Speech Recognition Using SPWVD. IEEE. 72, pp.1-10. http://DOI:10.1109/TIM.2022.3216673

[38] Suman Deb, Pankaj Warule, Amrita Nair, Haider Sultan, Rahul Dash and Jarek Krajewski. (2022). Detection of common cold from speech signals using deep neural network. Springer. 42, p.1707–1722. https://doi.org/10.1007/s00034-022-02189-y

[39] Jefferson Gomes Fernandes. (2022). Artificial intelligence in telemedicine. Spriner, pp.1-10. https://doi.org/10.1007/978-3-030-58080-3_93-1

[40] Mohammed Abdelhay, Ammar Mohammed and Hesham A. Hefny. (2023). Deep learning for Arabic healthcare: MedicalBot. Springer. 13(71), pp.1-17. https://doi.org/10.1007/s13278-023-01077-w

[41] Mozilla. (2024). Common Voice Dataset. [Online]. Available at: https://commonvoice.mozilla.org.

[42] Shajulin Benedict, Rubiya Subair. (2025). Deep Learning-Driven Edge-Enabled Serverless Architectures for Animal Emotion Detection. Informatica. 49, p.33–48. https://doi.org/10.31449/inf.v49i7.6615

[43] Chunyan Han, Ling Lin. (2024). Detecting and Tracking Rumours in Social Media Based on Deep Learning Algorithm. Informatica. 48, p.83–96. https://doi.org/10.31449/inf.v48i14.5998

[44] Desheng Chen, Sifang Zhang. (2025). Deep Learning-Based Involution Feature Extraction for Human Posture Recognition in Martial Arts. Informatica. 49, p.77–90. https://doi.org/10.31449/inf.v49i12.7041