CS5302 Project Proposal

Speech Recognition and Translation System For Medical Communication

Group 15

Talha Ahmed
24100033

Syed Muqeem Mahmood
24100025

Faizan Ali
24100065

Maryam Shakeel
24100295

Nehal Ahmed Shaikh
24020001

Ubaid Ur Rehman
24020389

26 February 2024

1 Introduction

This proposal outlines a Speech to Machine Translation + Speech (SMTS) system, specifically designed for healthcare. This system can potentially help overcome language barriers in healthcare, facilitating clear communication and thereby improving the quality of patient care and outcomes. Our system is built upon four key components:

1. **Speech Recognition**: This is the first step where the system listens to the spoken words (e.g a patient's symptoms) and *transcribes* them into English for easier processing.
2. **Large Language Model (LLM) + Retrieval Augmented Generation (RAG)**: The English text is then fed into a pre-trained LLM to fetch accurate and contextually appropriate response (e.g the diagnosis of the patient) from a set of curated medical documents.
3. **Machine Translation**: Once the diagnosis has been done, the system then translates the text into the desired language (can be different from the patients language).
4. **Text-to-Speech Synthesis**: The translated text is finally converted back into speech in the translated language.

The ultimate goal is to provide accurate translation in real time, making communication in important areas like healthcare easier.

2 Problem Motivation

Language barriers in healthcare are a significant challenge that can lead to misunderstandings, misdiagnoses, and even incorrect treatment plans. This is particularly problematic in multicultural societies or regions where patients and healthcare providers may not share a common language. Traditional translation methods, such as human interpreters or basic translation software, often fall short. They can be slow, leading to delays in urgent care situations. They can also be inaccurate, especially when dealing with complex medical terminology, potentially leading to serious consequences for patient care.

Our proposed Speech to Machine Translation + Speech (SMTS) system is designed to address these challenges. It makes use of the capabilities of state-of-the-art algorithms in the fields of Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS) synthesis. In addition, it serves as a virtual medical expert, providing reliable healthcare advice in a conversational manner. Therefore, unlike traditional methods, the SMTS system is not only designed to understand and accurately translate medical terminology but also to replace a medical expert with a medical chatbot, making it a more reliable solution for healthcare communication.

By implementing the SMTS system, we aim to enhance communication in healthcare settings, making it more efficient and reliable. This will not only improve the quality of patient care but also make healthcare services more accessible to non-native speakers.

3 Goal of the Project

The objective of our project is to develop a Speech to Machine Translation + Speech (SMTS) system that can interpret, translate, and vocalize spoken language in real-time. The system is designed to tackle the challenge of language barriers in healthcare settings, specifically focusing on the accurate translation of medical terminology and contexts.

Our project is aimed at healthcare professionals and patients who face language barriers during their interactions. It is also beneficial for medical institutions seeking to improve their services in multicultural environments and for researchers interested in the application of language translation technologies in healthcare.

To put our work in context with other notable developments in the field:

- **Translatotron and Translatotron 2 by Google AI**: These systems offer end-to-end speech-to-speech translation and retain the original speaker's voice.
- **Simultaneous Speech-to-Speech Translation System with Neural Incremental ASR, MT, and TTS**: This system consists of three fully-incremental neural processing modules for ASR, MT, and TTS.
- **ESPnet-ST**: ESPnet-ST is a framework that integrates ASR, MT, TTS, and ST into a unified platform to support a wide range of languages and domains, making it a versatile tool for various speech translation tasks.

Therefore, while there are existing solutions that offer speech recognition, machine translation, and text-to-speech synthesis, our project stands out by integrating these components into a single system specifically tailored for healthcare. Furthermore, our system is designed to support a wide range of languages, promoting inclusivity and accessibility in healthcare environments.

4 Project Workflow and Resources

Our project can be divided into several key components, each supported by a set of potential resources. The aim is to use a combination of pre-trained models to achieve our overarching project goal.

4.1 Resources

The resources we will be using for this project are advanced tools and models that specialize in different aspects of language processing.

- **Speech Recognition**: We have access to Wav2Vec 2.0 and Whisper, both known for their excellent speech recognition performance in various languages and conditions. Whisper even transcribes almost any language into English.
- **Language Detection**: In the event we require language detection capabilities then FastText might be be able to help us for learning word embeddings and accurate language identification from text.
- **Machine Translation**: We have access to MarianMT or mT5 due to their extensive language pair support and multilingual capabilites.
- **Text-to-Speech (TTS)**: For converting translated text into natural-sounding speech, we can employ Tacotron 2 or WaveNet.

4.2 Proposed Pipelines for Speech to Machine Translation + Speech System

This section outlines the proposed approaches for developing a comprehensive SMTS system. The approaches are divided into two pipelines and a hybrid approach.

4.2.1 Pipeline 1: Using Pre-existing Models

This pipeline involves the use of established models for each stage of the process:

- **Speech Recognition:** We'll use the Whisper tool, known for its robust speech-to-text capabilities across multiple languages. You can find it on OpenAI's GitHub repository.
- **Response Generation:** We'll employ the LLM RAG model available at Hugging Face, using a curated set of medical documents to generate an appropriate response in English.
- **Machine Translation:** We'll use either the Google Translate API or a machine translation model like MarianMT for translating the generated response.
- **Text-to-Speech:** We'll convert the translated text into speech using models like Tacotron 2 or WaveNet. These models can be fine-tuned for numerous languages.

4.2.2 Pipeline 2: Fine-tuning on Medical Documents

This pipeline is similar to Pipeline 1, with a key difference in the generation of responses. Here, we'll fine-tune the RAG model on a specific set of medical documents for more accurate and contextually relevant responses.

4.2.3 Hybrid Approach

This approach combines the strengths of both pipelines for optimal accuracy and efficiency. It includes Whisper for speech recognition, a fine-tuned RAG model for response generation, reliable translation APIs/models, and advanced text-to-speech technology. We can report results from both pipelines (in terms of metrics like BLEU or ROUGE) and compare performance.

4.2.4 Additional Notes

- **For RAG and Querying:** In all pipelines, we will utilize Chroma DB for transforming our medical documents into appropriate vector databases for efficient querying and retrieval of contextual information in response to the user's prompt.
- **For Deployment:** Our current strategy involves the deployment of Language Model (LLM) models using Gradio. We are optimistic about its capabilities for easy deployment. However, should it not meet our requirements, we remain open to exploring alternative solutions in the future.
- **For LLM:** We are currently exploring the possibility of utilizing Mistral, particularly the version with 7 billion parameters. However, the financial implications of accessing this model remain unclear at this stage. We are actively seeking cost-effective solutions to this challenge.

4.3 Potential Datasets for Medical Document Processing

- Shaip's Healthcare Datasets
  Shaip offers diverse healthcare datasets, including physician dictation audio data, transcribed medical records, and EHRs. These datasets cover 31 specialties, offering de-identified, HIPAA-compliant data suitable for AI model training in speech recognition and medical documentation. This could prove invaluable for developing models that understand and generate medical documentation or for speech recognition tasks in the medical domain.
- McGill-NLP's MeDAL Dataset
  The MEDAL dataset is curated for NLP pre-training in the medical domain, focusing on abbreviation disambiguation. It's designed for enhancing models' understanding of medical texts, making it suitable for projects aiming to generate or process medical documentation accurately.
- MS²: Multi-Document Summarization of Medical Studies
  MS² contains over 470k documents and 20K summaries from scientific literature for automating the literature review process in medical research. It supports the development of systems for assessing and aggregating evidence across studies, available online. This could be particularly useful for developing models that need to understand and synthesize medical research findings.

4.4 Project Timeline

In this section, we present a tentative timeline for our project. This timeline provides an overview of the various stages of the project, from model selection to final adjustments. However, the timeline is flexible and may be adjusted as the project progresses to accommodate any unforeseen challenges or changes in scope.

Time	Task
1-2 Weeks	Literature review and setup of speech recognition and language detection.
3-4 Weeks	Integration of translation and TTS.
5-6 Weeks	Component fine-tuning along with system integration and testing.
7-8 Weeks	UI development and final adjustments.

Table 1: Project Timeline

5 Conclusion

In this project, we will be using a variety of models that specialize in different aspects of language processing. These include both proprietary models like Whisper and open-source models like Tacotron.

The techniques we aim to use include using pre-trained models like RAG, fine-tuning these models on specific datasets, and potentially training models from scratch if necessary. The justification for these techniques is to make use of pre-existing models while tailoring them to our specific needs.

The models we aim to use, such as the 7 billion parameter version of Mistral, are large because the complexity of the tasks they are performing requires a high level of sophistication. However, we remain mindful of the potential financial implications and are actively seeking cost-effective solutions.

In terms of practicality and usability for the end-user, our strategy involves deploying our models using tools like Gradio. We also plan to fine-tune our models on specific medical documents to ensure the generated responses are accurate and contextually relevant. This user-centered approach will help ensure that our system is practical and easy to use in a real-world healthcare setting.