SpeakEZ: AI Language Translation Application

Kaali Lovell, Pavan Poliready, Rakesh Tirumala,
Steve St Fleur, Tanvi Prakash Gavali, Venkat Sai Pallapu
Seidenberg School of Computer Science and Information Systems
Pace University, New York, NY, USA

Abstract— SpeakEZ is an artificial intelligent web application designed to create seamless communication between individuals of different speaking languages. Due to the current climate of wars between countries and the open border policies, there have been a migration of people with different linguistic backgrounds, seeking to find a better life from their current volatile locations. SpeakEZ hopes to address this issue by allowing users to input audio and allowing them to transcribe or translate the audio into their preferred language. SpeakEZ uses advanced AI models to accurately transcribe the audio and produce real-time transcriptions. With this, SpeakEZ hopes to break the communication barriers between those of different languages and cultures.

***Keywords*— artificial intelligence, communication, transcription, translation, SMT, NMT**

I. Introduction

With the recent global conflicts around the world in areas such as Isreal, Palestine, Ukraine, Mexico, and Haiti, this has led to a mass migration of individuals relocating to seek a better life. There are currently over 7,000 living languages worldwide and with this migration, individuals of different language backgrounds are constantly coming into contact with one another. Being able to communicate with each other is important for sharing ideas, information, and fostering global understanding amongst each other.

To close the communication gap and provide ease of communication between individuals, this study introduces SpeakEZ, an AI- integrated application for language transcription and translation. SpeakEZ allows users to upload audio files, or record audio using our platform. SpeakEZ also allows the transcription of audio in its original language input and the translation of the transcription into the user's preferred language. The application will also allow users to store their transcribed audio and translations securely for future use, if they are account owners with SpeakEZ.

The security and ease of access with using SpeakEZ, makes it a reliable tool for reducing language barriers between individuals and promoting inclusivity between users.

II. Literature Review

A. Current Solutions

"Google Translate is a free online service that allows global users to translate text into a desired language."[6] Google Translate has been in service since 2006 and has hence implemented different models over time to produce translations for users. As described in [3], initially, Google Translate used statistical machine translations (SMT) to assist with the generations of translations. Later on however, Google Translate implemented neural machine translation (NMT) to help produce translations.

SMT uses Bayes Theorem. Bayes Theorem is a mathematical formula for determining conditional probabilities. "Conditional probability assesses the likelihood of an event occurring based on prior events or data." [2] In the context of Bayes Theorem with respect to translation, the theorem suggests that a sentence in one language corresponds to a sentence in another language. The SMT model breaks the input text into smaller segments, and matches each segment with its equivalent translation in its targeted language. It then reconstructs the sentence back together. In order to decide the correct translated equivalent, the SMT approach uses conditional probability as mentioned before.[3] We will explain this approach using the following example. If within the database containing a large dataset of word, if there has been a large subset of the word "hola" being matched to the word "hello," then the probability of the word "hola" being matched to the word "hello" would be higher than any other Spanish translated word version. This example proves the idea of conditional probability- the likelihood of the event "hola" being chosen based on prior events occurring, "the constant matching of hola being matched to hello in a large dataset".

Advancements in AI have allowed us to introduce neural network models into our applications, such as Google Translate who implemented the model later on. NMT produces more accurate translations and takes up less space, hence less costs. NMT differs from SMT in that NMT does not break down the sentences, rather it tries to translate the sentences as a whole using "deep learning neural networks" to translate the work. NMT models are trained on large datasets of human-translated text and create contextual representations of words or phrases, allowing for the model to return results in a human-like way. Neural networks are trained to think like a human brain to allow the results produced to seem more human-like. The dataset of human-translated text would then in turn produce a translation result that would be more accurate to the way a human would respond in the targeted language.

Although Google Translate has implemented more advanced models over time, their results are still limited and produce less accurate outcomes. Google Translate tends to mistranslate colloquial terms, and rare spoken languages. "Research shows that native speakers often find its output comprehensible but not flawless, highlighting the need for further editing to achieve accuracy." [9]

B. Relevancy

SpeakEZ differs from Google Translate in that it focuses on recognizing audio rather than using a text-based approach. Due to this, SpeakEZ accepts audio input rather than Google Translate which uses text inputs. SpeakEZ's model also focuses on encoding and decoding which ensures accuracy in transcription and translation within the application. Users can upload audio files or record audio using the platform. They are also able to transcribe the audio in the spoken language, as well as translate the transcription into their chosen language. The model implemented within SpeakEZ updates it database daily with over 680,000 hours of supervised multilingual data sourced from the web. This increases accuracy and reliability of the results produced through SpeakEZ. The constant improvements through the database will allow users to communicate with confidence to others of a different language background.

This study is valuable for users of different backgrounds- it will facilitate ease of communication between individuals of different languages. By using audio as inputs, language translation is more efficient and faster compared to traditional typed inputs for translation.

C. Product Requirements

To develop the SpeakEZ, we integrate the following key features:

1) **Responsive Web Design:** Users can access SpeakEZ on any device, making it universally accessible. They can record, transcribe, and translate audio seamlessly from their chosen platform.

2) **Real-Time Accuracy**: Neural networks enable SpeakEZ to deliver accurate and appropriate results.

METHODOLOGY

D. User Frontend

In order to deliver a seamless user experience and ensure the team collaborates effectively, we used a comprehensive set of tools and technology to develop the SpeakEZ application. We used tools such as GitHub, VSCode, and Figma. GitHub allowed for the development and collaboration of code. It allows the team to work remotely and still be connected and up-to-date with the code produced. VSCode is used as our chosen code editor. VSCode is flexible and versatile- it supports multiple programming languages and extensions. For brainstorming the UI features of the application, we used Figma to collaborate on the visual aspects. To keep track and prioritize tasks, we used Atlassian Jira software to keep the team organized and focused.

On the technical side, we implemented HTML5, CSS, and JavaScript as the functional programming language- it is the basis for our application. We also implemented React which supports in building dynamic and interactive user interfaces. To create a more modern interface, we implemented libraries such as MaterialUI, and TailwindCSS. Vite was also used to assist with faster development of our front-end application. Firebase is implemented in the backend as our database for securely storing the users' transcriptions and translations.

By integrating these tools and technologies, SpeakEZ is a reliable and user-friendly platform for enhancing communication between multilingual individuals globally.

E. Machine Learning Model

SpeakEZ implements a whisper model presented by OpenAI. As stated by OpenAI [7], OpenAI's whisper model is a sophisticated speech recognition system capable of handling multilingual speech recognition, translation, and language identification. As mentioned before, it is trained on a dataset of over 600,000 hours of multilingual data. The model is precise in that it recognizes speech regardless of the accent, colloquial terms used, and background noise presented. [1] States the whisper model uses an encoder-decoder transformer which breaks down the audio into 30 second chunks and then converts the audio into a log-Mel spectrogram (a spectrogram where the audio frequencies are converted to a mel scale- a unit for pitch [8]). This transformed data is then fed into an encoder where it is processed, and the decoder produces a corresponding text based on the prior trained data from the web. This approach enables the model to effectively perform tasks such as language identification, multilingual transcription, and speech translation.

F. Firebase Backend API

As addressed in [4], Firebase is built on Google Cloud infrastructure in which developers use the service to store user generated content. In firebase users are able to store images, audio, videos, and other forms of content. For the SpeakEZ application, we will be deploying firebase in the backend to store the user's transcribed and translated work. Firebase is reliable for its security, scalability, and robustness.

III. PRODUCT RESULTS

A widely used metric for evaluating transcription accuracy is the Word Error Rate (WER). The WER calculates the proportion of errors in a transcribed text compared to a reference text. [5] breaks down the WER score of the OpenAI Whisper and Google's Speech-to-Text which we will break down now. OpenAI's Whisper model achieves a median WER of 8.06%, demonstrating a significantly higher accuracy compared to Google Translate's WER, which ranges from 16.51% to 20.63%. This indicates that Whisper produces fewer errors in transcription, making it more reliable for accurate audio processing. As a result, incorporating Whisper into our application enhances its overall accuracy and trustworthiness, ensuring a better user experience.

As for the user experience, SpeakEZ provides a login page for users to sign up for our application. The signup will allow users to securely store their transcriptions, and translations for future use. The user will be able to upload a .mp3 file and or record an audio clip using our application. The user will be able to transcribe their audio, and then if desires, translate their transcription into their desired language.

These features will provide the user with an efficient and seamless user experience. Figures 1 to 6 demonstrate the features mentioned.

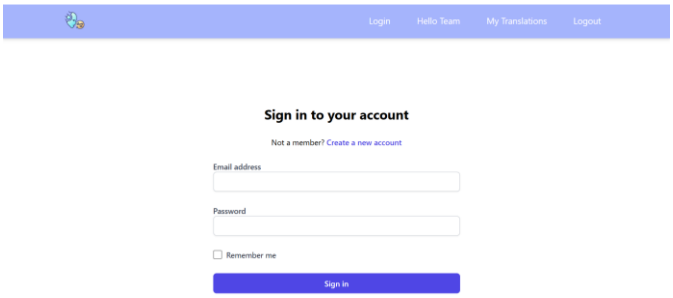


Fig 1. Home Login Page for User

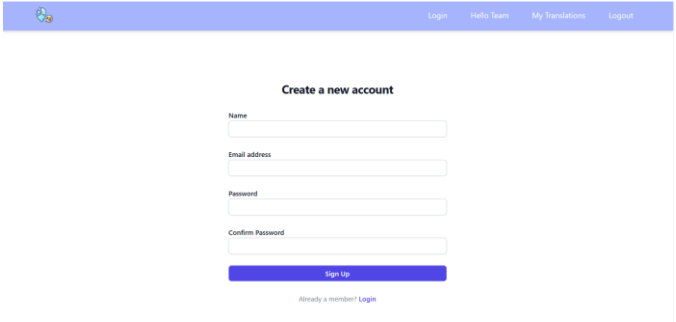


Fig 2. Sign-up Page for User

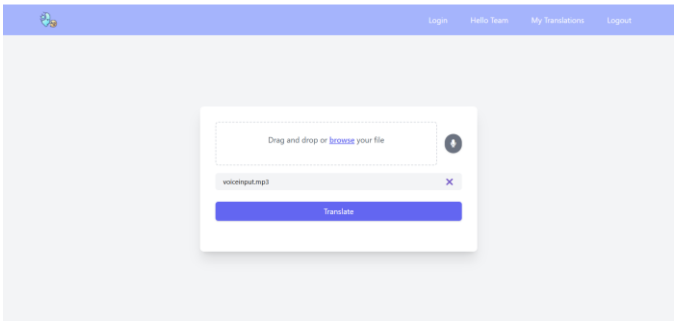
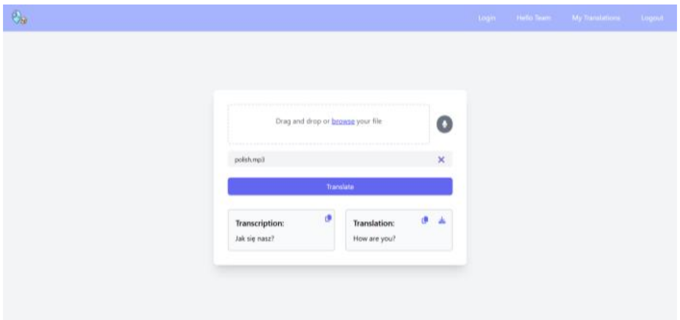


Fig 3. Audio Upload/ Record Page



Transcription:
Jak się nasz?

Translation:
How are you?

Fig 4. Transcription & Translation Page

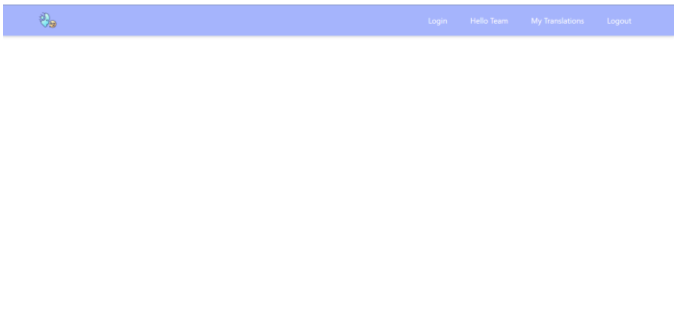


Fig 5. My Translation Page (will be updated)

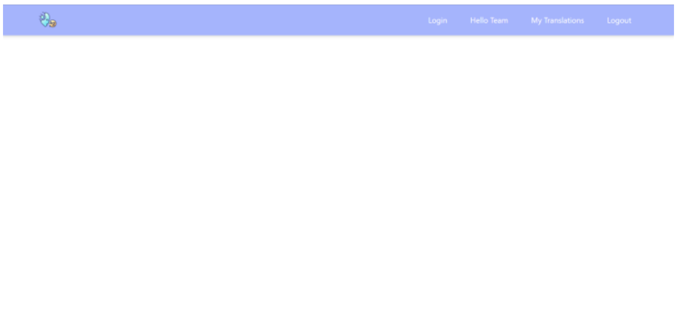


Fig 6. My Team Page (will be updated)

IV. Conclusion

In conclusion, SpeakEZ is a translator web application that processes audio input from a file or recorder, converts it into a transcribed text, and translates it into a specified language. By enhancing communication and bridging language barriers, it aims to bring the world closer through AI.

Future development of the project can involve seeking advice from professors and incorporating customer feedback to create more user stories, making the application more seamless and intuitive. Additional work could include adding functionalities like supporting larger file sizes for uploads, enabling live translation while receiving audio input, and allowing users to upload reports, technical conference papers, or documents for direct translation into multiple languages. Another feature to consider is adding timestamp text while transcribing for better organization. Furthermore, the implementation of voice volume controls would also make it easier for those who rely on auditory outputs for clarity or have visual impairments. The dual-option output would enhance accessibility, catering to those of diverse user needs.

There is significant potential to optimize the performance of the application, making it more efficient. SpeakEZ has the potential to develop into a marketable product. With recent advancements in AI models and frameworks, we can leverage this technology to make a positive impact on the world.

References

[1] A. Radford, J. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Dec. 2022. Available: https://cdn.openai.com/papers/whisper.pdf

[2] Chenna, “Bayes Theorem - Proof, Formula and Solved Examples,” BYJUS, 2019. https://byjus.com/maths/bayes-theorem/

[3] F. Chessa and G. Brelstaff, “Going beyond Google Translate?,” P-arch (Science and Technology Park of Sardinia), vol. 1, pp. 108–113, Sep. 2011, doi: https://doi.org/10.1145/2037296.2037324.

[4] Firebase, “Cloud Storage | Firebase,” *Firebase*, 2019. https://firebase.google.com/docs/storage

[5] “Gladia - OpenAI Whisper vs Google Speech-to-Text vs Amazon Transcribe: The ASR Rundown,” *Gladia.io*, 2024. https://www.gladia.io/blog/openai-whisper-vs-google-speech-to-text-vs-amazon-transcribe

[6] Google Translate — Remotely Nomad, “Remotely Nomad,” Remotely Nomad, 2014. https://remotelynomad.com/shop/p/google-translate (accessed Nov. 19, 2024).

[7] “Introducing Whisper,” Openai.com, 2022. https://openai.com/index/whisper/

[8] L. Roberts, “Understanding the Mel Spectrogram,” Medium, Mar. 14, 2020. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[9] P. Stapleton and B. Leung Ka Kin, “Assessing the accuracy and teachers’ impressions of Google Translate: A study of primary L2 writers in Hong Kong,” English for Specific Purposes, vol. 56, pp. 18–34, Oct. 2019, doi: https://doi.org/10.1016/j.esp.2019.07.001.