**Handicap Assistant – TTS and STT**

Siri Smitha Joginapally – BG98738

Sai Kowshik Mayur Kaluvayi - IG25413

**University of Maryland, Baltimore County (UMBC)**

**DATA 606: Capstone in Data Science**

**Dr. Tony Diana**

**May 13, 2024**

ABSTRACT

Text-to-speech (TTS) and Speech-to-text (STT) have advanced recently. Using a few advanced models available in Transformers we would like to build a model that will be useful for people suffering from disability to communicate with others easily. TTS components help convert text into synthesized speech, whereas the components in STT will help transcribe spoken speech into written text. We can integrate TTS and STT into Handicap Assistant which gives a solution for communication accessibility giving independence to the users.

INTRODUCTION

Research States that communication between a person suffering from a hearing problem and a dumb person can be a major disadvantage. This creates a very little place for them to communicate with others. This can cause them loneliness and depression (Jean, C., & Peter, B. (2004)). This communication gap can be covered using advanced technologies like that of Text-to-Speech and Speech-to-Text. The TTS and STT will help enhance the accessibility and communication for individuals suffering from disabilities. In our project, we built a handicapped assistant integrated with TTS and STT models. For building a Text-to-Speech (TTS) model we used BARK and Google’s Text-to-Speech this will help produce naturally sounding speech using input text. Whereas Speech-to-Text will be using Wav2Vec2ForCTC and Wav2Vec2Tokenizer which will help generate written text from the audio input.

We built a Handicap Assistant using a powerful architecture that is Transformers. We used BART (Bidirectional and Auto-Regressive Transformers) for text summarization. We do text summarization on the large text so that we can just convert the summary of the text to speech which will enhance communication and help users to process information more efficiently. Then the use of pre-trained models like BARK, GTTS, Wav2Vec2ForCTC, and Wav2Vec2Tokenizer helped us build a seamless and good performance TTS and STT.

With the development of the Handicap Assistant, our main aim is to empower individuals with disability and make them more independent. This will help them be more social with the society. This can be used as a communication tool with which we can reduce the communication gap between people and people with disability.

LITERATURE REVIEW

The field of Data Science technologies has witnessed advancement in recent years. In the domain of Text-to-Speech (TTS) and Speech-to-Text (STT), we have seen recent developments like Transformers, BARK, and many more with these models help in developing models like TTS and STT which help enhance communication for individuals with disabilities. Transformers have improved natural language processing tasks, including TTS and STT. This is because of their ability to capture long-range dependencies and contextual information effectively (Vaswani et al., 2017).

Automatic Text Summarization systems were designed to effectively summarize large volumes of text data this is created using deep learning models like Transformers. As there is an increase of the data or text in the modern world text summarization can be very helpful for summarizing the text for easy understanding and communication. So, Transformers have a few pre-trained models like that of BART, T5, and Roberta for text summarization which will provide accurate results. (A. Venkataramana et al., 2022).

The advancements in TTS technologies like that of BARK and Google-TTS are pre-trained models with the help of which we will be able to create API. BARK helps generate natural-sounding speech from textual input and helps in improving communication. GTTs also work similarly this is a cloud-based solution, that generates speech from text input. These will be mainly helpful in building the communication gap with disabled people.

Wav2Vec2ForCTC and Wav2Vac2Tokenizer are the recent advancements for generating STT. These models use CNN (Convolutional Neural Networks) and CTC (Connectionist Temporal Classification). These can accurately recognize the speech or audio and convert it into text very precisely. (Baevski, A., & Auli, M. (2020))

The Transformer-based models like BART, BARK, GTTs, Wav2Vec2ForCTC, and Wav2Vac2Tokenizer have made it easy to build TTS and STT models. With the integration of TTS and STT along with text summarization we will be able to build a Handicap Assistant model that will reduce the communication gap with disabled people.

DATASET DESCRIPTION AND INSIGHTS

The dataset used to train the model is taken from the Hugging Face Repository this dataset consists of articles from the CNN/DailyMail dataset.

Table 1. Dataset Variables

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| Article | String | The article variable consists complete article written by journalists at CNN/DailyMail. |
| Highlights | String | This variable consists of the summary of the article. |
| Id | String | This is the unique ID given for each article. |

The CNN/DailyMail dataset is used mainly for natural language processing tasks like text summarization. This dataset contains news articles with corresponding summaries, making it suitable for training and testing text summarization models.

ALGORITHM

**The BART** model is a sequence-to-sequence transformer architecture. This was introduced by Facebook AI research. BART is combined with auto-regressive generation with denoising autoencoding objectives. This will allow it to use large text for pretraining and then fine-tune downstream tasks. It can be fine-tuned for tasks like text generation, summarization, and language understanding (Lewis et al., 2020). BART will be used for text summarization that will help summarize the text which will make it easy to convert into the speech.

**BARK** is a transformers-based text-to-audio model. This can generate realistic, multilingual speech with background music, noise, and sound effects if required. This is a pre-trained model which can be used to create TTS models and interfaces. This will be used to convert the summarized text into speech. This will be a humanistic speech.

**GTTs** Stands for Google Text to Speech. GTTs is a Python library that provides an interface to Google’s text-to-speech which will give the speech taking the text input. The generated speech will be saved into a file or can be played directly using audio playback (Khete, T., & Bakshi, A. (2022)). This is used to generate the speech because this can generate accurate results when compared to that of BARK.

“**Wav2Vec2ForCTC” and “Wav2Vec2Tokenizer”** are deep learning models that were developed for automatic speech recognition (ASR). These are developed by Facebook AI Research. Wav2Vec2ForCTC is the natural network architecture for ASR whereas Wav2Vec2Tokenizer is useful for preprocessing the input audio data before it is fed to the model. These two can be combined to recognize the speech and then convert it into text (KDNuggets. (2021)).

OUTCOME

As the dataset was too big. A sample data is taken this text is first cleaned once the dataset is cleaned the most frequent words that occurred in the text the plotted in the graph. We can see that the words that occurred most frequently are “the, in, and, to, a, says, they, leifman, of, are, is”.

A graph of a number of words

Description automatically generated

Figure 1: Most Frequent Words in the Text

A pie chart with different colored triangles

Description automatically generated

Figure 2: Sentiment Analysis of the Text

Sentiment analysis is done on the text it can be observed that the text contains 8.6% positive words, 14.6% negative words and 76.8% of the text is neutral. From this, it can be said that the text can be mostly neutral.

“The ninth floor of the miamidade pretrial detention facility is dubbed the forgotten floor . Inmates with the most severe mental illnesses are incarcerated until theyre ready to appear in court . They face drug charges or charges of assaulting an officer charges that judge steven leifman says are usually avoidable felonies . The judge says he has become a strong advocate for changing things in Miami.”

Figure 3: Summarized Text

This is the summarized text the text summarization is done using a version of BART which is a part of Transformers.



Figure 4 & 5: Audio generated BARK Figure 6&7: Audio generated by GTTs

Please click on the icons above to listen to the audio generated by BARK and GTTs. The audio generated by BARK can change every time the code is executed and it will give different audio for the same text every time. It also has a limitation that it is not able to convert large text to audio. In contrast, GTTs will overcome these difficulties and will provide precise outcome every time the code is executed. GTTs are a powerful pre-trained model that can be used to perform TTS successfully.

“BUT OUR HUNDRED PER CENT OFF FROM MAY FOURTH FIRST WE HAVE AN ELECTRIC TOOTHBRUSH WITH EIGHT REPLACEMENTHEADS SCROLLED DOWN”

Figure 4: Text Generated by Wav2Vec2 Models

This is the text generated using Wav2Vac2 models. When these models are interrogated together it dose speech recognition precisely and can be used to generate text from the audio inputs.

CONCLUSION

In Conclusion, the Handicap Assistant gives accurate and efficient results using technologies like Transformers like BART for Text Summarization, BARK and GTTs for building Text-to-Speech (TTS), and Wav2Vec2 models for Speech-to-Text conversion. The integration of these components into the Streamlit interface has enhanced its usability. With this, we can empower users with disabilities to interact with digital content more effectively. This will reduce the communication gap with disabled people in society as well.

ACKNOWLEDGEMENTS

We want to thank Professor Dr. Tony Diana for guiding us in completing our project Handicap Assistance – TTS and STT.

REFERENCES

A.Venkataramana, K. Srividya and R. Cristin, "Abstractive Text Summarization Using BART," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972639.

<https://ieeexplore.ieee.org/document/9972639>

Baevski, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In International Conference on Learning Representations (ICLR).

Jean, C., & Peter, B. (2004). "Recognition of Arm Gestures Using Multiple Orientation Sensors: Gesture Classification." IEEE Intelligent Transportation Systems Conference on Electronics, 13(1), 334-520. <https://www.researchgate.net/publication/336114939_366ICERT-155-pdf>

KDNuggets. (2021, March). How to Convert Speech to Text with Wav2Vec. KDnuggets. Retrieved from <https://www.kdnuggets.com/2021/03/speech-text-wav2vec.html>

Khete, T., & Bakshi, A. (2022). Autonomous Assistance System for Visually Impaired using Tesseract OCR & gTTS. *Journal of Physics: Conference Series*, 2327(1), 012065. <https://doi.org/10.1088/1742-6596/2327/1/012065>

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).