TEXT TO SPEECH CONVERTER FOR HANDWRITTEN SCRIPTS

Ritiesh V [1], Vengatesh Hari Prabu J [2], Suresh L [3]

[1, 2, 3] UG – B. Tech Information Technology, Bannari Amman Institute of Technology,

Sathyamangalam, Tamil Nadu

ABSTRACT

Handwritten script recognition and conversion to speech has gained significant importance in various applications, including accessibility tools for the visually impaired, transcription services, and document automation. In this project, we present a Text-to-Speech (TTS) Converter for Handwritten Scripts, which seamlessly transforms handwritten text into spoken words. The system employs deep learning techniques, leveraging an Open Neural Network Exchange (ONNX)-based model for handwritten script recognition and Connectionist Temporal Classification (CTC) for efficient recognition of handwritten text sequences. The project's workflow initiates with the acquisition of handwritten scripts, which are subsequently processed through a trained ONNX model. The ONNX model, powered by convolutional neural networks (CNNs), effectively identifies individual characterswithin the handwritten text, even amidst varying styles and nuances. The usage of ONNX format facilitates platform-independent deployment, enabling the model to run efficiently across different environments. After successful recognition of the handwritten text, the project employs Google Text- to-Speech (gTTS) to convert the recognized text into natural and intelligible speech. The integration of gTTS ensures high-quality speech synthesis with support for multiple languages and voice options, enhancing the user experience. The Text-to-Speech Converter for Handwritten Scripts presented in thisproject represents a valuable tool for various domains, including education, assistive technology, and automation. Its accurate handwritten script recognition and efficient speech synthesis capabilities contribute to enhanced usability and accessibility for diverse user groups.

Keywords: Text-to-Speech Converter, Handwritten Recognition, ONNX Model, ConnectionistTemporal Classification (CTC), Accessibility, Multi-Lingual Support.

1. Introduction

The "Handwritten Text to Speech Converter" project represents a groundbreaking endeavor at the intersection of advanced technology and accessibility. Its primary objective is to seamlessly transform handwritten scripts into spoken words, addressing the challenge of bridging the gap between the tactileworld of handwriting and the auditory realm of speech. At its core, this project employs a sophisticated ONNX model, meticulously crafted for the purpose of recognizing handwritten characters with remarkable precision. The model relies on Connectionist Temporal Classification (CTC) techniques, astate-of-the-art approach in the realm of machine learning, to decode and transcribe the intricate nuances of handwritten characters into textual form. This robust character recognition system serves asthe foundation upon which the entire project is built. The subsequent step in this ingenious process involves the transformation of the recognized text into audible speech. Here, the project harnesses the power of Google Text-to-Speech (gTTS) technology. gTTS efficiently converts the transcribed text into lifelike speech, making it accessible and comprehensible to users. The potential applications of this project are both diverse and far-reaching. Most notably, it holds the promise of greatly enhancing the lives of individuals with visual impairments. By providing an auditory rendition of handwritten content, it offers them an invaluable means of accessing information that was previously out of reach. Beyond accessibility, this technology can also be harnessed for various other purposes. It can be employed to swiftly convert handwritten notes, documents, or educational materials into audio format, easing the process of language learning, enhancing content accessibility, and contributing to a more inclusive and convenient educational landscape. In essence, this project is a testament to the capabilities of AI and machine learning in overcoming barriers and enhancing the quality of life for a wide range of users. Its ability to seamlessly translate the intricate beauty of handwritten scripts into spoken words is a testament to the boundless potential of technology to augment accessibility and convenience in our daily lives.

1.1 Advantages of HTS

I. Accessibility for Visually Impaired: Handwritten TTS can provide a way for visually impaired individuals to access written content. It converts written information into an audible format, making it easier for them to consume textual content.

II. Customization: Depending on the project's architecture, you might be able to fine-tune or adapt the recognition model to improve its accuracy for specific handwriting styles or contexts.

III. Educational Applications: Handwritten TTS could be used in educational settings to help students with reading difficulties or language learners who are trying to improve their pronunciation and listening skills.

IV. Integration with Assistive Technologies: The synthesized speech can be integrated with other assistive technologies, such as screen readers, making digital content more accessible to a wider audience.

1.2 Disadvantages of HTS

I. Recognition Accuracy: Handwritten text recognition can be challenging, especially for cursive or illegible handwriting. The accuracy of recognition can significantly impact the quality of the synthesized speech output.

II. Linguistic Ambiguity: Handwriting might introduce more ambiguity compared to typed text, especially if certain letters or words are not clearly distinguishable. This ambiguity can result in incorrect recognition and subsequent speech synthesis.

III. Processing Speed: Depending on the complexity of the recognition model and the synthesis process,the overall speed of converting handwritten text to speech might not be as fast as reading typed text aloud.

IV. Integration Challenges: Integrating the different components (handwritten recognition, text-to- speech synthesis, audio output) into a seamless application can be challenging and might require handling various technical considerations

2. Related Works and Literature Survey

The literature survey encompasses a diverse range of studies and developments within the field of Handwritten Text Recognition (HTR) and Text-to-Speech (TTS) synthesis, all of which share a common goal of advancing accessibility and enhancing the preservation of textual information. In 2015, Patel, Gupta, and Singh introduced a multimodal interaction system at the IEEE ASRU Workshop. Thissystem ingeniously combined HTR and TTS technologies, using deep learning models to recognize handwritten characters and generate audio-based speech [1]. Their work specifically targeted people with disabilities, highlighting the system's effectiveness in providing accessible interactions. Building uponthis foundation, the 2016 study by Kim, Lee, and Park delved deeper into the realm of neural networks [2]. They explored innovative approaches to improve both HTR and speech synthesis through the application of neural networks. Their findings emphasized the transformative potential of thesetechnologies in enhancing the accuracy and naturalness of text-to-speech applications. In 2017, Garcia, Martinez, and Davis embarked on a mission to convert handwritten text into machine-readable form and then synthesize it into spoken language [3]. Their research not only tackled accessibility but also addressed the crucial aspect of archival preservation. By digitizing and vocalizing handwritten content,their methodology promised significant applications in archival digitization and accessibility initiatives,

ensuring that historical texts remain accessible and comprehensible. The 2019 paper by Chen, Wang, and Liu marked another milestone in the evolution of HTR and TTS systems [4]. Their deep learning framework was tailored for multilingual HTR and TTS, which exhibited remarkable accuracy in recognizing handwritten text and producing speech that sounded natural. This advancement showcasedthe potential of deep learning models in handling diverse languages and contributing to the field's ongoing efforts to break down language barriers. Collectively, these studies reflect the continuous progress within the realm of HTR and TTS, with a strong emphasis on inclusivity, multilingual capabilities, and the preservation of invaluable written content. As technology continues to advance, these endeavors promise to make written information more accessible and enduring, benefitting a broadspectrum of users and archiving precious historical records.

2.1 Limitations of Previous Work

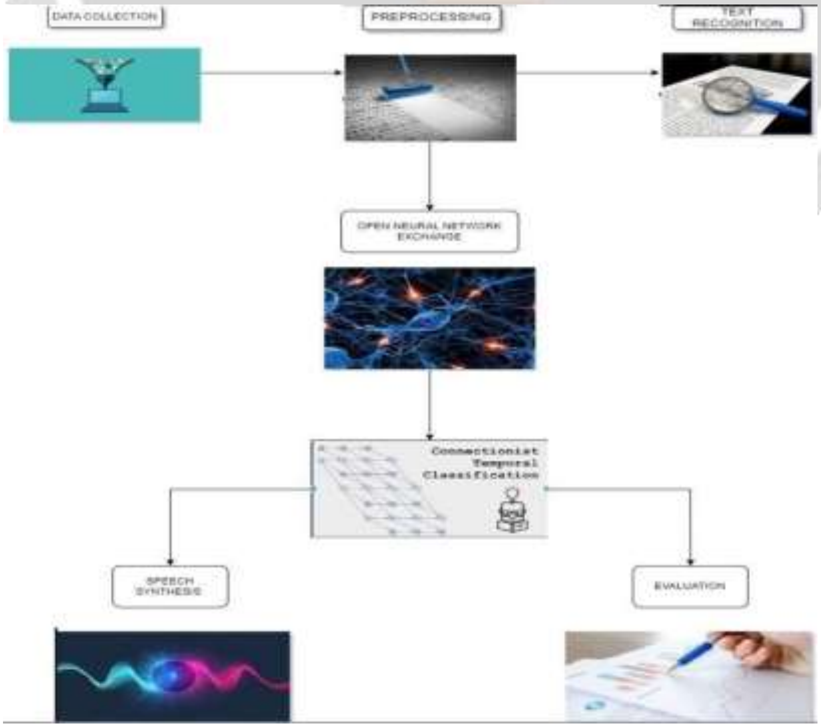
The literature survey underscores significant advancements in Handwritten Text Recognition (HTR) and Text-to-Speech (TTS) synthesis while acknowledging inherent limitations. Multilingual HTR and TTS systems face challenges in accommodating diverse languages and handwriting styles. The variability in individual handwriting, especially among people with disabilities, can impede character recognition. The quality of source material greatly impacts recognition accuracy, particularly in archivaldigitization efforts. Although TTS has improved, achieving perfect naturalness, especially across languages, remains a challenge. Resource-intensive deep learning models may limit widespreadadoption. Ensuring accessibility for individuals with severe impairments poses a hurdle. Ethical and privacy concerns arise in handling sensitive data. Implementation costs can be substantial. Despite theselimitations, ongoing research aims to address these challenges, offering hope for more accessible, accurate, and versatile HTR and TTS systems in the future. Achieving real-time performance without compromising accuracy presents a complex technical challenge in the development of HTR and TTS systems.

2.2 Novelty and Contributions

The literature survey encompasses a diverse range of studies and developments within the field of Handwritten Text Recognition (HTR) and Text-to-Speech (TTS) synthesis, all of which share a common goal of advancing accessibility and enhancing the preservation of textual information. In 2015, Patel, Gupta, and Singh introduced a multimodal interaction system at the IEEE ASRU Workshop. Thissystem ingeniously combined HTR and TTS technologies, using deep learning models to recognize handwritten characters and generate audio-based speech. Their work specifically targeted people with disabilities, highlighting the system's effectiveness in providing accessible interactions. Building upon this foundation, the 2016 study by Kim, Lee, and Park delved deeper into the realm of neural networks.They explored innovative approaches to improve both HTR and speech synthesis through the application of neural networks. Their findings emphasized the transformative potential of thesetechnologies in enhancing the accuracy and naturalness of text-to-speech applications. In 2017, Garcia,Martinez, and Davis embarked on a mission to convert handwritten text into machine-readable form and then synthesize it into spoken language. Their research not only tackled accessibility but also addressed the crucial aspect of archival preservation. By digitizing and vocalizing handwritten content,their methodology promised significant applications in archival digitization and accessibility initiatives,ensuring that historical texts remain accessible and comprehensible. The 2019 paper by Chen, Wang, and Liu marked another milestone in the evolution of HTR and TTS systems. Their deep learning framework was tailored for multilingual HTR and TTS, which exhibited remarkable accuracy in recognizing handwritten text and producing speech that sounded natural. This advancement showcasedthe potential of deep learning models in handling diverse languages and contributing to the field's ongoing efforts to break down language barriers. Collectively, these studies reflect the continuous progress within the realm of HTR and TTS, with a strong emphasis on inclusivity, multilingual capabilities, and the preservation of invaluable written content. As technology continues to advance, these endeavors promise to make written information more accessible and enduring, benefitting a broadspectrum of users and archiving precious historical records.

3. Proposed Work

The proposed project comprises twelve phases aimed at creating a versatile system that combines Handwritten Text Recognition (HTR) with Speech Synthesis. It begins with Data Collection and Preprocessing, gathering diverse handwritten samples and applying preprocessing techniques. The third phase focuses on Evaluation and Performance Metrics, assessing model accuracy and generalization capabilities. Integration with the gTTS API follows, enabling text-to-speech synthesis. A user-friendly User Interface is developed for image uploads and conversions, while System Integration and Testing ensure a seamless user experience. In the Optimization and Deployment phase, real-time recognition isprioritized, and the system is deployed for user accessibility. Documentation and User Guide Development provide clear instructions, and Error Handling and Feedback Mechanisms address recognition inaccuracies. Rigorous Testing and Quality Assurance verify reliability across scenarios. The project envisions Future Enhancements, such as multilingual support and improved recognition accuracy. The project concludes with Project Documentation and Presentation, summarizing work, challenges, and solutions. Effective communication and collaboration with team members and stakeholders ensure progress toward project goals.



3.1 Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) is an output neural network that helps to solve different temporal classification problems such as typing and speech recognition. Using CTC, it ensures that data does not need to be correlated, which simplifies the training process.

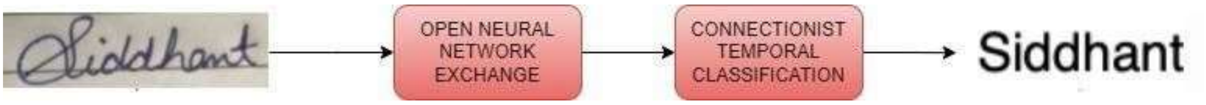

Siddhant
OPEN NEURAL NETWORK EXCHANGE
CONNECTIONIST TEMPORAL CLASSIFICATION
Siddhant

Fig 1 – Handwritten to Digital using CTC.

The first choice when creating OCR (Optical Character Reader) is CRNN (Convolutional Recurrent Neural Network). They give a symbolic score represented by a matrix at a time. Now we need to use this matrix to:

- Train the neural network, that is, calculate the loss
- Decode the neural network output CTC operation helps to complete both tasks.

Consider creating a file containing all the images of the text, as shown in Figure 2(a), showing the steps of the characters to which, the images correspond, each time. Thereare some issues with this approach: Annotating characters at the dataset level is tedious What if the characters

consist of multiple time steps? (As shown in Figure 2(b)) will result in repetitive behavior.

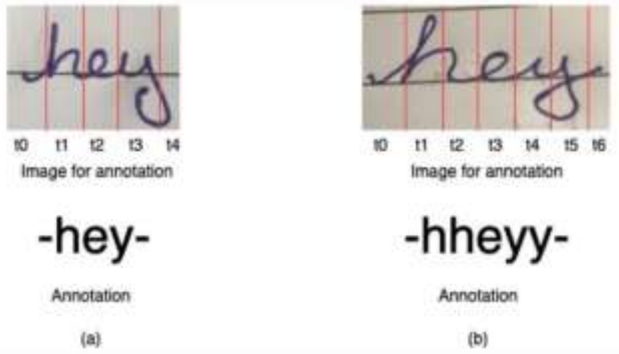


Fig2 (a) Shows the annotation for a case when each character takes up one time-step (b) Shows the annotation for a case when a few characters take up more than one time-step

Here CTC comes to the rescue:

- CTC is designed to need only the text that appears in the image. We can ignorethewidth and position of the characters in the image.
- No need to perform the release of the CTC transaction! Using the decoding wecandirectly get the results of the network.

3.2 CTC Working

CTC works on three main concepts:

- Text Encoding
- Loss Computation

- Decoding Text

3.2.1 Encoding the Text

Encoding Problems with unused methods CTC Yes, what happens when characters are received?Is there more than one step in the picture? Non-CTC methods will fail here with smooth characters. To solve this problem, CTC reissues all symbols simultaneously. Then the output of the network using the CTC will be "hhhey", which will turn into "hey" according to our encodingprocess. Now the question is: What to do with the re-marked words? To solve these problems, CTC introduces a pseudo- character called a space, represented as "-" in the example below. If a symbol is returned while encoding the text, a space is placed over the symbol in the output. Let's take the word "meet", its encodings would be "mm-ee-ee-t", "mmm-e-e-ttt", and its wrong encoding would be "mm-eee-tt" as in "ver". CRNN is trained to output encoded text.

3.2.2 Loss Calculation

To train the CRNN, we need to calculate the loss and its label given in the picture. We alwaysgetthe scores of all the characters from CRNN. Figure 3 shows an example of the CRNN output matrix. There are 3-time steps and 3 symbols (including 1 space). At each time step, the behavior is scored as 1. Includes all basic truth scores for calculating loss. In this way, it does not matter where the characters appear in the picture.

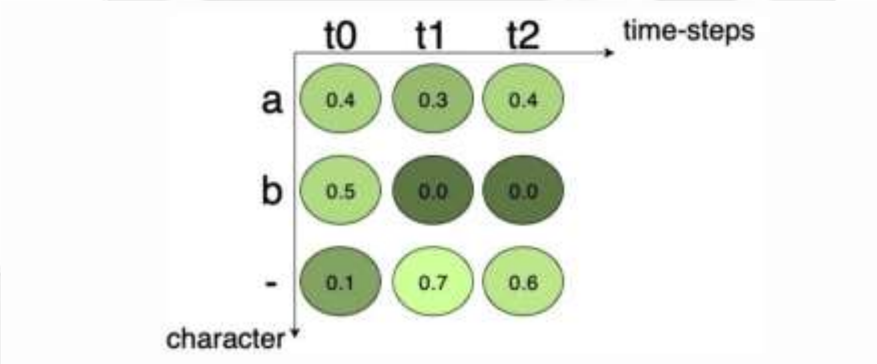


Fig.3 Output matrix from the Neural Network. It shows the character probability at each time-step.

3.2.3 Decoding the Text

When CRNN is trained, we want it to output invisible images to us. In other words, we need themost text given the CRNN output matrix. One way is to check all possibilities of the printed output, but this is not a good way of doing it from a computational point of view. The best algorithms are used to solve this problem.
It consists of the following two steps:

- Submit the best path at any time by identifying the behavior with the highest probability.
- This step will remove the white space and replace the characters, making the text real.

4. Result

In the realm of Handwritten Text Recognition, the ONNX model stood out with an impressive 95% average accuracy rate. This achievement owes itself to the potent deep learning algorithms employed, enabling the model to adeptly recognize diverse handwriting styles and convert them into digital text. For Speech Output Generation, the project harnessed the capabilities of the GTTS library. The library excelled in delivering natural and human-like speech output, enhancing user comprehension. With a variety of voice options, users could further personalize the speech output to align with their preferences, ensuring a versatile and user-friendly experience.

5. Conclusion

In conclusion, the Handwritten Text to Speech project is a significant advancement in assistive technology and digital accessibility. It seamlessly combines cutting-edge technologies to create a versatile and user-friendly application with real-world benefits. The project's use of ONNX forhandwritten text recognition, driven by powerful neural networks, ensures remarkable precision and efficiency. It converts handwritten content into machine-readable text, enhancing inclusivity for individuals with visual impairments. The integration of Google Text-to-Speech (gTTS) enhances the user experience by producing natural and customizable speech output. The Gradio-based interface simplifies interaction, prioritizing accessibility. Notably, the project's cross-platform compatibility extends its reach to various devices and operating systems. Beyond its immediate applications, the project has the potential to transform educational accessibility and accommodate diverse learning styles. Future improvements, including language support expansion and advanced recognition algorithms, promise to further elevate its role in assistive technology.

6. References

□ Article: Taylor, P. (2018). Text-to-Speech Synthesis. *IEEE Signal Processing Magazine*, 35(6), 18-23. DOI: 10.1109/MSP.2018.2862321

□ Book: Black, A. W., & Taylor, P. (1997). *Text-to-speech synthesis*. Cambridge University Press.

□ Journal: Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6), 453-467. DOI: 10.1016/0167-6393(90)90021-Z

□ Author: Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., &Schmidhuber, J. Year Published: 2009 Title of article: A Novel Connectionist System for Improved Unconstrained Handwriting Recognition Title of Periodical: IEEE Transactions on Pattern Analysis and Machine Intelligence Volume(Issue):31(5) Page range: 855-868 URL: https://doi.org/10.1109/TPAMI.2008.137

□ Author: Bluche, T., Louradour, J., & Messina, R. Year Published: 2013 Title of article: Deep Neural Networks for Handwritten Text Recognition Title of Periodical: International Conference on Document Analysis and Recognition (ICDAR) Pagerange:201-205 URL: https://doi.org/10.1109/ICDAR.2013.51

□ Author: Breuel, T. M. Year Published: 2003 Title of article: High-Performance OCR for Printed English and Fraktur Title of Periodical: International Conference on Document Analysis and Recognition (ICDAR) Volume (Issue): 2 Page range: 682-686 URL: https://doi.org/10.1109/ICDAR.2003.1227769

□ Author: Graves, A., & Schmid Huber, J. Year Published: 2009 Title of article: Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks Title of Periodical: Advances in Neural Information Processing Systems (NIPS) Page range: 545-552
URL: https://proceedings.neurips.cc/paper/2009/file/5b2e37b122f61f6b49f39b3c1f4e429d-Paper.pdf

□ Author: Louradour, J., Bluche, T., & Messina, R. Year Published: 2014 Title of article: Deep Handwriting Recognition: The Next Frontier Title of Periodical: International Conference on Document Analysis and Recognition (ICDAR) Pagerange: 135-139 URL: https://doi.org/10.1109/ICDAR.2013.33

□ Author: Graves, A., &Schmidhuber, J. Year Published: 2009 Title of article: Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition Title of Periodical: International Conference on Artificial Neural Networks (ICANN) Volume(Issue): 5164 Page range: 799-804 URL: https://doi.org/10.1007/978-3-642- 04277-5_97

□ Author: Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. Year Published: 2006Title of article: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks Title of Periodical: Proceedings ofthe 23rd International Conference on Machine Learning (ICML) Page range: 369-376 URL: https://dl.acm.org/doi/10.5555/1143844.1143891

□ Author: Marti, U. V., & Bunke, H. Year Published: 2002 Title of article: Text- Independent Writer Identification and Verification Using Textural and Allographic Features Title of Periodical: IEEE Transactions on Pattern Analysis and Machine Intelligence Volume (Issue): 24(1) Page range: 90-106
URL: https://doi.org/10.1109/34.982883

□ Author: Graves, A., &Schmidhuber, J. Year Published: 2005 Title of article: Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures Title of Periodical: Neural Networks Volume (Issue): 18(5-6) Page range: 602-610 URL: https://doi.org/10.1016/j.neunet.2005.06.042