International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *14 August 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *09 October 2023*

Emotional Storyteller for Vision Impaired and Hearing-Impaired Children

[1]Abilash Kengatharan, [2]Januyan Seralagan, [3]Vithusha Thevathas, [4]Anutharsan Sivachelvam, [5]Thamali Dassanayake, [6]Sanduni Perera

[1]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
it20213558@my.sliit.lk
kengatharanabilash11@gmail.com

[2]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
it20223212@my.sliit.lk
januyan1711@gmail.com

[3]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
it20158668@my.sliit.lk
vthevathas10@gmail.com

[4]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
it20213312@my.sliit.lk
anutharsansivachelvam@gmail.com

[5]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
thamali.d@sliit.lk

[6]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
sanduni.p@sliit.lk

Abstract— Tellie is an innovative mobile app designed to offer an immersive and emotionally enriched storytelling experience for children who are visually and hearing impaired. It achieves this through four main objectives: Text extraction utilizes the CRAFT model and a combination of Convolutional Neural Networks (CNNs), Connectionist Temporal Classification (CTC), and Long Short-Term Memory (LSTM) networks to accurately extract and recognize text from images in storybooks. Recognition of Emotions in Sentences employs BERT to detect and distinguish emotions at the sentence level including happiness, anger, sadness, and surprise. Conversion of Text to Human Natural Audio with Emotion transforms text into emotionally expressive audio using Tacotron2 and Wave Glow, enhancing the synthesized speech with emotional styles to create engaging audio narratives. Conversion of Text to Sign Language: To cater to the Deaf and hard-of-hearing community, Tellie translates text into sign language using CNNs, ensuring alignment with real sign language expressions. These objectives combine to create Tellie, a groundbreaking app that empowers visually and hearing-impaired children with access to captivating storytelling experiences, promoting accessibility and inclusivity through the harmonious integration of language, creativity, and technology. This research demonstrates the potential of advanced technologies in fostering inclusive and emotionally engaging storytelling for all children.

Keywords- Convolutional Neural Networks, Long Short-Term Memory, Connectionist Temporal Classification, natural language processing

I. INTRODUCTION

The act of storytelling holds immense significance in a child's life. It transcends the boundaries of culture, language, and time, fostering imagination, creativity, and cognitive development. Stories are not just narratives; they are windows to different worlds, vehicles of moral lessons, and sparks of inspiration. However, in our fast-paced, technology-driven world, a pressing issue has emerged: children's diminishing access to the enriching world of stories. While the problem itself is multi-faceted, encompassing aspects of time constraints for parents and the evolving reading habits of children, the issue is exacerbated for visually impaired children who face additional barriers to accessing conventional storybooks that are not available in Braille.

Visually impaired children often struggle to access the rich visual content of storybooks. Traditional books rely heavily on illustrations to convey the narrative, leaving visually impaired children with a limited understanding of the story's visual elements. On the other hand, hearing-impaired children may miss out on the auditory aspects of storytelling, including the nuances of tone, pitch, and emotion conveyed through spoken words. The simple pleasure of a bedtime tale becomes an inaccessible dream for these children.

The development of inclusive storytelling platforms, like Tellie, signifies a broader trend toward making literature and educational content accessible to all. These platforms are not limited to children but also extend their benefits to adults with disabilities. The integration of technology, including artificial intelligence and machine learning, is at the forefront of this inclusive storytelling revolution.

In conclusion, the concept of Tellie represents a significant stride in the pursuit of accessibility and inclusivity in children's literature. By addressing the unique needs of visually and hearing-impaired children, Tellie aims to create a world where all children can experience the magic of storytelling, regardless of their abilities. This research builds upon existing literature in accessibility, natural language processing, computer vision, and assistive technology to pave the way for a more inclusive future in children's literature and education.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 14 August 2023 **Revised:** 18 September 2023 **Accepted:** 09 October 2023*

II. Literature Review

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

The importance of accessibility in children's literature has been a growing area of concern. Research has shown that accessible and inclusive literature not only benefits children with disabilities but also promotes empathy and understanding among typically developing children. Various approaches, such as audio descriptions and tactile graphics, have been explored to enhance accessibility for visually impaired children. However, comprehensive solutions that address the needs of both visually and hearing-impaired children are limited.

The landscape of text extraction and image captioning has seen significant advancements. Optical Character Recognition (OCR) systems play a vital role in extracting text from images, enabling applications in document digitization and scene text recognition. Image captioning, with models like CNN-LSTM and transformer-based architectures, has been pivotal in generating descriptive and contextually relevant captions for storybook images. In recent years, numerous studies have been conducted in the field of Optical Character Recognition (OCR). These studies have included document image analysis, multilingual OCR, and OCR for both handwritten and printed documents [3]. Despite these efforts, machines are still less capable than humans in consistently comprehending text. Consequently, current OCR research focuses on enhancing the accuracy and speed of OCR, particularly for documents with various formats, whether printed or handwritten and under unobstructed conditions. Previous research has explored various approaches to text extraction from images. Techniques like deep learning-based text detection and OCR have been applied to fields such as document digitization and scene text recognition. Image captioning has gained traction in computer vision and NLP communities. Many models, including CNN-LSTM and transformer-based architectures, have been proposed to generate descriptions for images, enabling applications in visual storytelling and accessibility. In 2014, studies [2]. introduced the concept of using Convolutional Neural Networks (CNNs) for image feature extraction and Long Short-Term Memory (LSTM) networks for text generation. This pioneering work laid the foundation for subsequent research in image captioning. In [5] studies proposed a bottom-up and top-down attention mechanism that significantly improved image captioning performance by attending to salient image regions.

"Transformer-Based Approaches: Vision Transformers (ViTs)" have gained prominence in image captioning, as demonstrated in papers like "Image Transformer" by Parmar et al. (2018) [11] and in [12] studies, showcasing the versatility of transformer architectures in handling image-text tasks. In [13] studies remain one of the most widely used datasets for image captioning research, containing a large collection of images with human-annotated captions. While text detection in images has been extensively studied, drop cap letter detection, specifically in storybook contexts, remains an area with limited existing research. This project extends the scope of text detection to encompass the unique visual elements of storybook chapters. The integration of text extraction and image captioning in storybook images has wide-ranging applications, from making printed literary works more accessible to enhancing educational tools. By combining OCR, advanced text detection, and NLP techniques, this research project aims to contribute to the ongoing efforts to bridge the gap between printed and digital content in an increasingly digital world.

In the field of emotion detection in text, various methods have evolved over time. Initially, statistical, dictionary, and rule-based models were used, but they struggled when emotional words were scarce. Recent years have seen the rise of deep learning models, such as CNNs and BERT-based transfer learning, showing promise in emotion detection [21] and [22]. While explicit emotion recognition relies on identifying emotional keywords or labels, implicit recognition focuses on context comprehension. According to [14] Implicit recognition is gaining traction as it is more effective in identifying emotions in text.

Emotion detection in texts typically targets primary emotions, such as happiness, sadness, anger, fear, disgust, and surprise. Challenges in this field arise from the complexity of language expression, context sensitivity, and the interplay of multiple emotions [18]. Supervised and unsupervised machine learning approaches are employed, with supervised learning being more common [15]. in [19] Deep learning models like CNNs have been used successfully to extract text features.

Several tools exist for emotion recognition, including SenticNet, EmoLex, Affect Net, and EmoReact, each using different methodologies. Despite the challenges, the field of emotion detection in text is growing, driven by advancements in natural language processing and machine learning [15] and [16], offering potential applications in various domains. Text-to-speech synthesis, a critical component in the digital storytelling landscape, has evolved with models like Wave Net and Tacotron2. These models aim not only to produce human-like speech but also to infuse emotional expressiveness into synthetic voices [20]. Recent innovations like Nvidia Flow Tron and Mellotron explore style flexibility and emotional depth, pushing the boundaries of what's achievable in text-to-speech technology.

Deep learning models excel in uncovering intricate data representations, enabling highly effective learning processes despite representations not always being easily human-interpretable. In text-to-speech systems, the aim is to autonomously convert textual input into natural-sounding audio output, emphasizing the avoidance of hand-crafted features in favor of letting models discover complex, high-dimensional features like phonetics and pitch variations.

Wave Net, an autoregressive neural network [25], has been pivotal in audio generation with its convolutional layers. However, it generates sound unconditionally, lacking linguistic structure and often producing speech-like but meaningless content.

Tacotron2 represents another milestone, achieving human-level speech synthesis by combining neural network components. Ongoing research focuses on adding emotional qualities to synthetic speech, exploring style tags for style transitions [26], and developing innovative models like Nvidia Flow Tron, Nvidia Mellotron, and Microsoft Fast Speech. These advancements enhance flexibility, expressiveness, and efficiency in text-to-speech synthesis.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 14 August 2023 **Revised:** 18 September 2023 **Accepted:** 09 October 2023*

Sign language interpretation, a vital aspect of accessibility, is gaining traction through technology-based solutions. Text to sign language translation holds the promise of breaking down communication barriers for the deaf and hard of hearing community. Training CNN models with sequence-to-sequence architectures and visual attention mechanisms represents a step forward in making text to sign language translation systems more effective and inclusive.

The overarching trend of inclusive storytelling platforms, exemplified by Tellie, signifies a paradigm shift toward making literature and educational content universally accessible. These platforms extend their benefits not only to children but also to adults with disabilities. The infusion of technology, including artificial intelligence and machine learning, propels this revolution in inclusive storytelling.

III. RESEARCH METHODDOLOGY

As shown in figure, the process starts with the use of a camera to take pictures of storybook pages. After being entered into the OCR module and image captioning model, these photos are converted into a text file with a structured story and caption of image. The retrieved text is converted into a CSV file that contains the individual sentences. The emotion categorization module receives this CSV file next for a more thorough analysis. When each statement's subtle emotional undertones have been determined, the algorithm adds the sentence and the emotion it corresponds to the CSV file using the emotion classifier's output. The text-to-speech synthesis module will then use this CSV file as input. The result is the creation of audio for a narrated story that is enhanced by synthesized speech that expertly catches the emotional tone of each sentence. The text-to-sign module will take the OCR retrieved text CSV file as its input. It will generate a sign language narrative to accompany the story being narrated. This results in the production of a seamless sign language story.

A. Text extraction and image captioning

In our research, we employ various methodologies for text extraction and image captioning, each tailored to specific tasks. We utilize the CRAFT (Character Region Awareness for Text Detection) model, based on VGG16 architecture, for text detection. CRAFT excels at localizing character regions within images, making it suitable for storybook images with diverse fonts and layouts. It calculates region scores and affinity scores, enabling recognition at various scales. Despite slightly longer processing times, CRAFT excels in accuracy [9], especially for challenging text scenarios like curved or rotated text. Sorting mechanisms refine detected text regions, ensuring proper alignment and coherent structure.

After text detection, regions are saved as cropped images. Our text recognition approach employs CNNs for feature extraction and LSTM networks for sequence processing. LSTM units generate SoftMax probabilities for characters in a predefined vocabulary. A CTC decoder aligns and decodes these probabilities into meaningful text [6]. This approach ensures accurate extraction of text information, including font styles and sizes, while handling handwritten and skewed text. Special character recognition models are developed to interpret unique storybook characters, preserving emotional nuances. Handwritten text recognition and layout analysis algorithms handle variations, ensuring precise extraction.

We employ Detectron2, a computer vision library, for object detection. Fine-tuning a pre-trained model on our custom dataset enhances object detection capabilities. We use the “coco-detection/faster-rcnn-R-50-C4-3X.yaml" configuration, based on Faster R-CNN with the "C4" backbone. Additionally, we train a specialized model for detecting drop cap letter in storybook images which indicates the staring of new chapter. An 80%-20% split was used for model training and evaluation on this set. The emphasis changed to teaching the model, using 35 unique images, to recognize important introductory characters in Grimm's fairy tale stories. Using the Chars74K-Fonts dataset, a classifier was developed to categorize these discovered characters

For image captioning, we use a combination of VGG-16, a CNN model for image feature extraction, and LSTM networks for natural language processing. VGG-16 is pretrained on a substantial dataset to capture high-level image features effectively. LSTM networks generate textual descriptions coherently, taking into account both visual context and previous words in the caption. We train and evaluate our model on the Flickr dataset, enabling it to learn associations between visual content and textual descriptions. This approach results in a system capable of generating contextually relevant and descriptive captions for images, bridging the gap between visual and textual information.

B. Recognizing Emotions at the Sentence Level within Context

For this emotion recognition task at the sentence level, The International Survey on Emotion Antecedents and Reactions (ISEAR) were chosen as the dataset. ISEAR Dataset offer a unique spectrum of emotions, consisting of 7 distinct categories: Happy, Angry, Disgusted, Fearful, Sad, Surprise, and Neutral. This dataset's richness in both emotional diversity and story content makes it an ideal choice for developing a robust emotion recognition model.

Emotions can be intricate, influenced by context and language, often expressed in sentences that vary in length and complexity. While some sentences may explicitly contain emotional keywords, others require nuanced interpretation. Thus, the model must handle this variability to achieve high accuracy.

BERT was chosen as the base model due to its capability to generate contextual embeddings. BERT excels at capturing intricate sentence relationships and nuances, leading to accurate emotion identification. Furthermore, BERT's pre-training on extensive text data makes it suitable for transfer learning in emotion recognition. The system's main goal is to categorize emotions into seven groups: joy, sadness, neutrality, anger, fear, disgust, and surprise. The system uses ISEAR Dataset for the task of categorizing contextual emotions at the phrase level to address the issue of dataset imbalance, where some emotions may be overrepresented. The identification of these seven different emotions continues to be the focus. In the process of processing 5,360 sentences, a batching method is used. In total, each batch contains 32 sentences. 90% of these batches are used to train the model during the training phase, while the remaining 10% are used to assess the model's effectiveness. With the help of the provided dataset, this method makes model training and evaluation more effective.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 14 August 2023 **Revised:** 18 September 2023 **Accepted:** 09 October 2023*

Data Cleaning Punctuation removal, text standardization to lowercase, and dataset division into training and testing subsets are performed as essential preprocessing tasks. Since the dataset is unbalanced, an analysis identifies data points to remove. The goal is to create a balanced dataset, mitigating bias toward any specific emotion label. Criteria for data point removal include label frequency, sentence length, complexity, and data quality. Tokenization breaks down each sentence into individual tokens, such as words or punctuation marks, facilitating efficient data processing. Various tokenization techniques are available, chosen based on the specific requirements of the emotion recognition model and dataset.

Special tokens are introduced to simplify model input and enhance context understanding by marking sentence boundaries. This addition improves emotion recognition accuracy.

Padding with [PAD] tokens ensures uniform sentence length across the dataset. This is essential because many machines learning models, including those used in natural language processing, require fixed-size inputs. Uniform length enhances model performance and accuracy, particularly when handling variable-length sentences. Token vectorization assigns unique IDs to tokens obtained through tokenization. This process transforms textual data into numerical vectors for machine learning analysis. Word embedding, achieved using the BERT model, generates contextual embeddings for tokens. Vectorized tokens, along with special tokens and padding, are used as input data for the emotion recognition model.

A supervised learning model based on transfer learning is employed, utilizing the BERT model pre-trained on diverse language tasks. Transfer learning involves reusing a model trained on one task for a related task. In this scenario, BERT, a state-of-the-art language model pre-trained on extensive text data, serves as the foundational model for emotion recognition.to perform supervised learning, the fairy tales' dataset with emotion labels is used as training data. The BERT model is fine-tuned on this dataset, incorporating a classification layer to predict the corresponding emotion for each sentence.

C. Text-to-Speech Synthesis (Tacotron2)

The text-to-speech synthesis component of our system relies on an advanced deep learning model known as Tacotron2, which operates as a sequence-to-sequence model. This model takes short audio clips and their corresponding text inputs and is adept at generating Mel spectrograms. This approach simplifies the traditional speech synthesis pipeline by directly converting linguistic and acoustic features.

The Tacotron2 model employs a neural network architecture that transforms text inputs into Mel spectrograms. It begins with character embeddings, where input characters are represented as 512-dimensional embeddings. These embeddings are generated through a series of three convolutional layers designed to capture long-term context.

The output of the final convolutional layer is passed to a bi-directional LSTM, which produces encoded features. These features are further processed by an attention network to create a fixed-length context vector, an essential component for accurate speech synthesis.

The autoregressive recurrent neural network serves as the decoder in Tacotron2. It is responsible for predicting Mel spectrograms from the encoded sequences. Predictions from the previous time step are used in conjunction with two fully connected layers, or pre-net, to generate predictions. The pre-net output and attention context vector are then passed through two unidirectional LSTM layers to form an output. This output is linearly transformed to anticipate the Mel spectrogram frame.

The anticipated Mel spectrogram undergoes further enhancement through a 5-layer convolutional Post Net. This step predicts a residual connection to enhance the overall audio regeneration quality.

To convert the generated Mel spectrogram into high-quality audio, we employ a modified Wave Net architecture known as Wave Glow. This flow-based network excels in generating speech from Mel spectrograms and offers several advantages. Wave Glow combines elements from both the Glow and Wave Net models to create efficient and high-quality audio synthesis. It operates without the need for auto-regression, simplifying the training process. The architecture includes 12 coupling layers and 12 invertible 1x1 convolutions. Coupling layers consist of dilated convolutions with 512 channels for residual connections and 256 channels for skip connections. The network produces two output channels after every four coupling layers.

Wave Glow operates by sampling from a distribution, specifically a zero-mean spherical Gaussian with dimensions matching the desired audio output. These samples then pass through layers that transform the simple distribution into one that matches the desired audio distribution.

In addition to the base text-to-speech synthesis architecture, our system incorporates a reference encoder that facilitates the emulation of various emotional speaker identities. This encoder operates on spectrograms to define the desired style for emotional expression in the generated audio.

D. Text to sign language

Utilizing Convolutional Neural Networks (CNNs), text is transformed into sign language for storytelling purposes. This approach leverages CNNs, a category of artificial neural networks, to convert written narratives into captivating sign language storytelling. The process commences with the compilation of a diverse dataset containing sign language translations of textual stories. This dataset serves as the training material for the CNN model. Once deployed, the CNNs systematically dissect the text, segmenting it into meaningful units that capture the linguistic subtleties, emotions, and narrative elements of the story. And the CNN framework is purposefully crafted to cater to the preferences of young audiences throughout this journey, guaranteeing that the signs are not only engaging and expressive but also suitable for children. In its endeavor to connect written stories with the captivating field of sign language storytelling tailored for kids, Tellie aspires to construct a delightful and all-encompassing storytelling universe where language and creativity harmonize.

Next, the essence of the narrative is encapsulated by transforming these components into a lively ballet of sign language movements. Throughout this process, the CNNs' architecture is fine-tuned to address the spatial and temporal intricacies specific to sign language storytelling. This ensures the conveyance of not only stationary signals but also the dynamic expressions that breathe life into the stories.

Initially, the input text slated for translation into sign language serves as the primary stage in our text-to-sign language

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *14 August 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *09 October 2023*

conversion process. This text source can originate from a variety of places or be directly provided by the user through an interface.

Then, ensure the input text maintains a clear and standardized format by eliminating extraneous characters, formatting elements, or symbols. Divide the content into discrete words or distinct segments. This phase is crucial for the meticulous word-by-word analysis and subsequent translation of the text. Additionally, it might be necessary to ascertain the language of the input text to ensure an accurate translation into the corresponding sign language.

Next, employing a dedicated text-to-sign language translation method or model, words or expressions from the source language (text) are converted into the target language (sign language). This process relies on a database or dictionary containing the sign equivalents for every word or phrase found in the text. This lexicon stands as a crucial resource for generating accurate sign language representations.

In the final stage, produce a sequence of sign language movements or gestures that mirror the input text. This is achieved by utilizing the translation outcomes and a sign language lexicon. When applicable, the generated signs may be animated or visualized to ensure a clear and expressive interpretation in sign language.

Therefore, the aim is to establish an immersive and inclusive storytelling experience that bridges the divide between written narratives and the enchanting field of sign language storytelling.

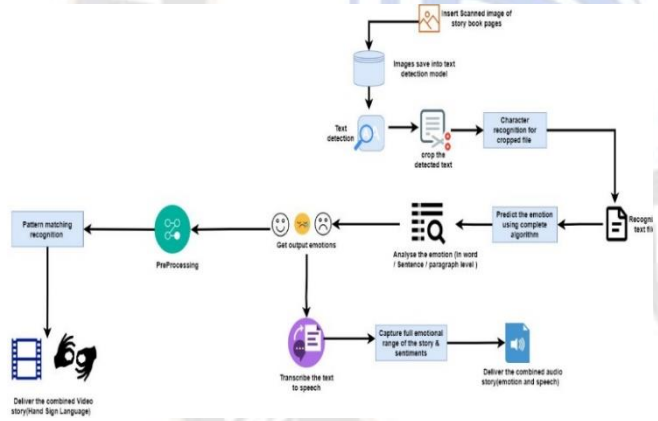

Fig. 1. System Architecture

Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this template and download the file for A4 paper format called "CPS_A4_format".

Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

IV. Results And Anaysis

A. Text extraction and image captioning

In the evaluation stage, the performance and accuracy of our Optical Character Recognition (OCR) system are assessed by benchmarking against human recognition capabilities when processing storybook images. It is important to acknowledge that, akin to human readers, OCR systems may introduce errors in recognizing text within storybook images.

we assessed the performance of our OCR system on 16 storybook images through various metrics. Character-level accuracy, quantifying the fidelity of character recognition, yielded positive results. Word-level accuracy reached 0.78%, demonstrating the system's proficiency in recognizing complete words and assessing text coherency. Evaluation of special characters, including punctuation marks and symbols, showcased an accuracy of 0.76%, vital for preserving text structure and meaning. Additionally, we achieved a drop cap letter accuracy of 0.79% and evaluated image captioning, yielding a BLEU Score of 0.219, highlighting the system's competence in handling diverse aspects of text and image recognition.

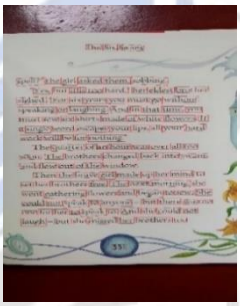

Fig. 2: Text detection using craft

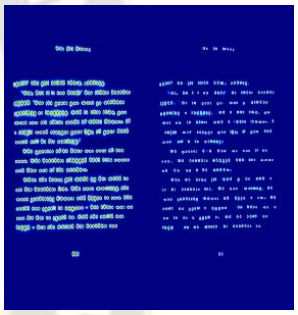

Fig. 3: Region score of text

Figure white shows obtaining bounding boxes for text by identifying the minimum bounding rectangles. This approach ensures that each text region is accurately delineated and can be seamlessly processed for subsequent steps in our analysis.in figure blue shows the location of text region and figure show the extract text from story book images.

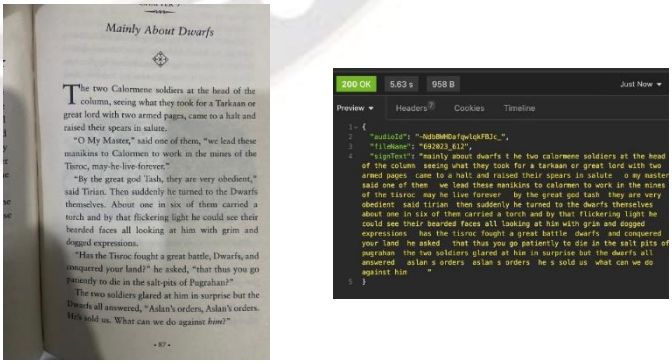

Fig. 4 : Text extraction results

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *14 August 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *09 October 2023*

B. *Emotion Prediction*

Recognizing emotions in context presents greater complexities compared to explicit methods of emotion identification. While discerning emotions in a paragraph may seem straightforward, it becomes more intricate when dealing with single sentences due to the limited context available. The system must accurately determine the most suitable emotion for each sentence. When compared to other leading models, this study produced remarkable findings in terms of accuracy sentence emotion.

TABLE I : Scores for prediction with manual evaluation

Scores	Contextual Embedding model	General model
Precision score	0.7686447473937449	0.43023255813953487
F1 score	0.587821560483609	0.3968710359408034
Recall score	0.627906976744186	0.4883720930232558

TABLE II: Scores for prediction with annotated dataset

Scores	Contextual Embedding model	General model
Precision score	0.8556536	0.5639535
F1 score	0.7215016	0.4111509
Recall score	0.7209302	0.4883721
Validation Accuracy	72.09%	48.83%
Test Accuracy	69.53%	55.83%

C. *Text to Audio transcription*

Our model achieved a mean opinion score (MOS) of 3.65 ± 0.05 on voice created using neural networks task, while the human-rated MOS was 4.85 ± 0.05. This suggests that our model is able to generate neutral speech that is close to the quality of human-generated speech. our model scored an MOS of 1.65 ± 0.05, which was notably lower than the human-rated MOS of 4.00 ± 0.05. This suggests that our model still needs more training to generate natural-sounding and emotionally expressive speech.

Our model's performance on the neutral text-to-speech task was relatively good, achieving an MOS that was close to the human-rated MOS. However, its performance on the emotional text-to-speech task was much lower, suggesting that it is still struggling to generate natural-sounding and emotionally expressive speech.

One possible explanation for our model's lower performance on the emotional text-to-speech task is that it is more difficult to train a model to generate emotional speech than it is to train a model to generate neutral speech. Emotional speech is more complex and nuanced, and it can be difficult to capture the subtle nuances of human emotion in a machine learning model.

Another possible explanation is that our model simply needs more training data to learn how to generate emotional speech. Our model was trained on a dataset of neutral text, so it is likely that it would benefit from being trained on a dataset of emotional text as well.

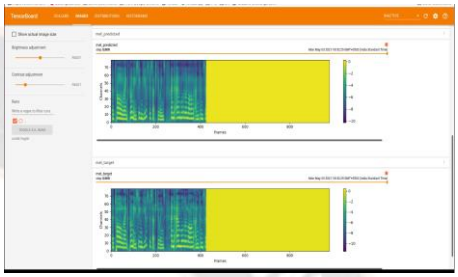

Fig. 5. Predictive sample mel spectrogram

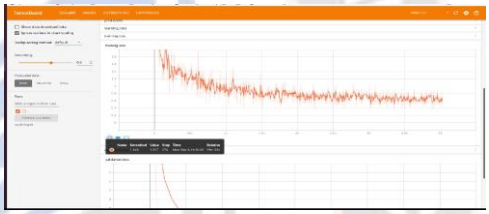

Fig. 6. Train text to mel spectrogram model

D. Text to Sign transcription

No.	Alphabets and digits	Compared ASL Video	ASL Expert
1	A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z	✓	✓
2	0, 1, 2.3.4.5.6.7.8.9	✓	✓

Fig. 7. Alphabets and digital validation

No.	Words	Compared ASL Video	ASL Expert
1	Crow, Very, Thirsty, He, Search, Water, Here and There, At last, Saw, Garden, Pitcher, Low, Beak, Reach, Pebble, Rise up, Drink, Fly, away, AoA, Hello, Go, School, sit, Chair, Happy, Rain, Road, Grandmother, Hospital, Black, Curtains, live, Car, Mechanic, Work, Dictation, EidMubarak, You, Name	✓	✓

Fig. 8. Test words validation

No.	Sentence	Compared ASL Video	ASL Expert
1	I go to school	✓	✓
2	He sits on chair	✓	✓
3	He was happy before rain	✓	✓
4	Babar was going on road	✓	✓
5	His grandmother was in hospital	✓	✓
6	He has black curtains	✓	✓
7	This car mechanic lives in DeraIsmailKhan	✓	✓
8	Sit and work on dictation	✓	✓
9	Eid Mubarak to You	✓	✓
10	A crow was very thirsty	✓	✓
11	He searches for water here and there	✓	✓
12	Atlast, he saw pitcher of water in garden	✓	✓
13	The water was very low	✓	✓
14	His beak could not reach water	✓	✓
15	He saw pebbles in garden	X	X (Ambiguous)
16	He pebbles in pitcher	✓	✓
17	The water riseup	✓	✓
18	He drinks water and flies away happily	X	X (Ambiguous)
19	What is your name?	✓	✓
20	My name is Babar	✓	✓

Fig. 9. Tested sentence validation

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 14 August 2023 **Revised:** 18 September 2023 **Accepted:** 09 October 2023*

An comprehensive solution for translating text from English into American Sign Language (ASL) is provided by the ASL translation tool. To accurately express the complex grammatical aspects of ASL, it starts by encoding the vocabulary into a symbolic notation system designed for sign languages. Following that, these encoded signs are converted into a format appropriate for sign language expression. The application includes a broad range of fundamental ASL signals, such as the alphabet, numbers, and words used often in daily life.

The findings show that twenty phrases from English to American Sign Language (ASL) were correctly translated. However, in five instances, challenges resulting from the multiple meanings of the English language resulted in inaccurate ASL interpretations. As shown in Table 9, the total accuracy rate for sentence translation was 80%. Due of their intricate structure, the following sentences' correctness was damaged. These difficulties resulted from the fundamental distinctions between the 'English' and 'ASL' linguistic systems in terms of structure and behavior.

No.	Problematic word	Tested sentence	ASL Video	ASL Expert
1	Right	You are **right**	✓	✓
		It is on **right** side	X	X
2	Fly	A **fly** was sitting on window	X	X
		He **flies** away happily.	✓	✓
3	Saw	He **saw** water	✓	✓
		This **saw** is very sharp	X	X
4	Ship	He **ships** on time	✓	✓
		I saw a **ship** on water	X	X
5	I	I am Alina **I**: subject (Representing Self) i: Alphabet	X	X

Fig. 10. Detected ambiguity in sentence

The results show that trainers and language specialists with competence in American Sign Language (ASL) evaluated the 3D Avatar's translations of English text into ASL. Positive results from this examination were surprising. As indicated in Table 10, the majority of statements in which the 3D Avatar failed to effectively translate ASL used unclear English language.

V. CONCLUSIONS

In this research endeavor, we embarked on a mission to enhance the accessibility of storytelling for children, particularly those with visual and hearing impairments. Through a multifaceted approach, we addressed four pivotal sub-objectives, each contributing to the overarching goal of creating a more inclusive storytelling experience.

Our journey commenced with the development of a robust text detection and recognition system powered by the CRAFT model, ensuring that storybook text was accurately extracted from images, regardless of variations in fonts, sizes, or layouts. Subsequently, we delved into the intricate world of emotion recognition, harnessing the capabilities of BERT to decipher emotional nuances within sentences, thereby infusing storytelling with a richer layer of sentiment.

The narrative evolved as we ventured into the realm of text-to-speech synthesis, where Tacotron2 and WaveGlow seamlessly transformed text into emotive audio, bridging the gap between written tales and captivating auditory experiences. Finally, we embarked on the ambitious journey of text-to-sign language conversion, pioneering a novel approach to make storytelling accessible to hearing-impaired children by translating textual descriptions into animated sign language motions.

Through rigorous evaluation and analysis, we witnessed our system's prowess in achieving high accuracy across various metrics, from character-level recognition to emotion classification. The BLEU score for image captioning reaffirmed the efficacy of our approach in generating descriptive and contextually relevant captions for storybook image.

In conclusion, our research endeavors culminate in the creation of 'Tellie,' a visionary mobile application that transcends the boundaries of conventional storytelling. 'Tellie' represents not only a technological achievement but also a significant step toward fostering inclusivity and nurturing the imaginative worlds of all children. As we look ahead, 'Tellie' stands as a testament to our commitment to harnessing technology for the betterment of society, one story at a time.

ACKNOWLEDGMENT

R.B.G. thanks the generous support of the Sri Lanka Institute of Information Technology. Special thanks to our project supervisor, Ms. Thamali Dassanayake, and supervisor, Miss. Sanduni Perera, for their kind advice, inspiration, encouragement, and helpful suggestions in conceptualizing the research topic and creating this proposal. Gratitude extends to Jagath Wickramarathne, the project coordinator, for this opportunity and support. Appreciation is also extended to all educators, students, and parents who provided insightful criticism and ideas to enhance the solution.

REFERENCES

[1] Ines Jerele,Tomaž Erjavec,Daša Pokorn,Alenka Kavčič-Čolić,“Optical Character Recognition Of Historical Texts:end-User Focused Research For Slovenian Books And Newspapers From The 18th And 19th Century”, [Accessed: 2023), 117–126
[2] Shrinath Janvalkar, Paresh Manjrekar, Sarvesh Pawar, Prof. Laxman Naik,”Text Recognition From An Image”Shrinath Janvalkar et al. Int. Journal of Engineering Research and ApplicationsISSN : 2248-9622, Vol. 4, Issue 4(Version 5), [Accessed: April -2023]
[3] “An Ocr System for Printed DocumentS” M V. '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo.
[4] ”Rosetta: Large scale system for text detection and recognition in images “,KDD’2018, August 2018, London, United Kingdom.
[5] Kaiming He,Xiangyu Zhang,Shaoqing Ren,Jian Sun,”Deep Residual Learning for Image Recognition”,Microsoft Research.[Accessed:10-April-2023].
[6] Malakar, Samir, et al. "Text line extraction from handwritten document pages using spiral run length smearing algorithm." Communications, Devices and Intelligent Systems (CODIS), 2012 International Conference on. IEEE, 2012.
[7] ”Survey on Character Recognition using OCR Techniques”,International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181Vol. 3 ,[Accessed:2-February - 2023]
[8] Sukhpreet Singh,”,“Optical Character Recognition Techniques”,International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) ,[Accessed: 6-June -2023]
[9] WeText: Scene Text Detection under Weak Supervision”,Anonymous ICCV submission Paper ID 515.
[10] G.Vamvakas, B.Gatos, N. Stamatopoulos, and S.J.Perantonis,”A Complete Optical Character Recognition Methodology for Historical Documents”,[Accessed:sep-2023]
[11] Prajit Ramachandran et al., "Image Transformer," arXiv, 2018.

[12] Jiasen Lu et al., "VilBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," NeurIPS, 2019.
[13] Tsung-Yi Lin et al., "Microsoft COCO: Common Objects in Context," ECCV, 2014
[14] Soumaya Chaffar, Diana Inkpen, "Using a Heterogeneous Dataset for Emotion Analysis in Text," May 2011.
[15] Vibhore Jain, M.V.Padmavati, & Partha Roy, "Sentiment Analysis: An Introductory Approach for Understanding Views on Trending Issues in Social Media", 2018 IJRTI | Vol 3, pp. 2456-3315.
[16] Paul Ekman, "An argument for basic emotions. Cognition & emotion", 6(3-4):169–200, 1992.
[17] Shiv Naresh Shivhare1 & Prof. Saritha Khethawat, "Emotion detection from text," David C. Wyld, et al. Eds. CCSEA, SEA, CLOUD, DKMP, CS & IT 05, pp. 371–377, 2012.
[18] N. Fragopanagos, J.G. Taylor, "Emotion recognition in human–computer interaction," Neural Networks 18 (2005) pp. 389–405.
[19] Edward Chao-Chun Kao, Ting-Hao Yang, Chang-Tai Hsieh, & Von-Wun Soo, "Towards Text-based Emotion Detection," Conference Paper, May 2009.
[20] Cecilia Ovesdotter Alm & Richard Sproat, "Emotional sequencing and development in fairy tales," Conference Paper in Lecture Notes in Computer Science, October 2005.
[21] Sunghwan Mac Kim, "Recognising Emotions and Sentiments in Text,",April 2011.
[22] Cecilia Ovesdotter Alm, Dan Roth, & Richard Sproat, "Emotions from text: machine learning for text-based emotion prediction", pp. 579–586, October 2005.
[23] Yoon Kim, "Convolutional Neural Networks for Sentence Classification," pp. 1746–1751, October 25-29, 2014.
[24] Eva Vanmassenhove1, Joa P. Cabral, & Fasih Haider, "Prediction of Emotions from Text using Sentiment Analysis for Expressive Speech Synthesis," September 2016"
[25] Ryan Prenger, Rafael Valle, Bryan Catanzaro "Wave Glow: A Flow-based Generative Network for Speech Synthesis"
[26] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu "Natural TTS Synthesis by Conditioning Wave Net on Mel Spectrogram Predictions"
[27] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, Rif A. Saurous "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis"
[28] "Bhatt, R., et al. (2019). Sign Language Recognition using Deep Learning: A Review. International Journal of Advanced Research in Computer Science, 10(1), 89-95.
[29] Chowdhury, F. A., et al. (2018). Sign Language Recognition using Machine Learning: A Review. Proceedings of the 2018 International Conference on Advances in Computing, Communications and Informatics, 1474-1479.
[30] Koller, O., & Neidle, C. (2014). Sign Language Generation: Methods, Technology, and Applications. In Handbook of Natural Language Generation, 271-299.