International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 206-213, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202533
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

Vision-to-Voice: AI for generating Description & Audio of Visual Content

[1]P. Jayanth, [2]K. Lakshmi Sree, [3]K. Karthik Kumar Reddy, [4]G. Om Prakash, [5]G. Reddy Prasad

[1,2,3,4,5]Department of Artificial Intelligence, Madanapalle Institute of Technology & Science, Madanapalle, India
E-mails: [1]padilam.jayanth8@gmail.com, [2]lakshmisree1309@gmail.com, [3]karthikkumarreddykunuthuru@gmail.com, [4]goddumarriomprakashyadav@gmail.com, [5]Reddyprasadg8617@gmail.com

***Abstract* - The seamless transformation of visual content into descriptive text and naturalistic speech, termed Vision-to-Voice, represents a significant interdisciplinary advancement at the intersection of computer vision, natural language processing (NLP), and speech synthesis. This paper explores the development of an end-to-end Vision-to-Voice pipeline, encompassing visual scene understanding, semantic description generation, and high-quality speech synthesis, thereby enabling AI systems to narrate visual content for human users. The proposed methodology integrates Transformer-based image captioning models with context-aware linguistic augmentation and neural vocoders trained for expressive speech synthesis, ensuring fluent and expressive audio descriptions for visual content. While individual advancements in image captioning and TTS are well documented, their seamless fusion into an end-to-end, real-time system presents unique research and engineering challenges, including context preservation across modalities, maintaining linguistic fluency, and ensuring audio naturalness. This paper addresses these gaps through a unified encoder-decoder captioning module with Bahdanau Attention, followed by a Tacotron 2-based Mel-spectrogram generation module and HiFi-GAN-based waveform synthesis module. Extensive experimentation and evaluations using standard datasets, including Flickr8K and LJSpeech, demonstrate the efficacy of the proposed system in terms of caption quality (BLEU) and audio naturalness (MOS scores). The Vision-to-Voice system holds promising applications in assistive technologies, multimedia enrichment, and automated video annotation systems, thereby contributing to both academic research and real-world accessibility solutions.**

Keywords: NLG (Natural Language Generation), LLMs (Large Language Models), Perplexity, Text Coherence.

I. INTRODUCTION

In the modern era, visual content is the bedrock of communication, education, and entertainment. Images, in their capacity to convey intricate scenes and feelings at a glance, cross linguistic and cultural divides, testifying to the truism that "a picture is worth a thousand words." In today's modern era, where visual media sweeps platforms like social media, news channels, and learning sites online, it becomes an absolute necessity in conveying messages and connecting with the audience. But this visual overreliance is an enormous challenge for the visually impaired, who are forced to depend on auditory or touch-oriented means of being informed. Though assistive technologies like screen readers assist text-based information, they cannot read raw visual content like images and videos—leaving a central challenge to making digital content inclusively accessible. Furthermore, though human-annotated alt-text can be useful, it is not scalable for large-scale use and may not always get the full context or richness of an image. Moreover, current automated solutions either output generic descriptions lacking contextual richness, or are hampered by prohibitive latency, and hence not fit for real-time use.

To overcome such challenges, we present “Vision-to-Voice”, a novel end-to-end artificial intelligence system that bridges the visual and audio experience by transforming images into accurate, context-rich, and natural-sounding audio descriptions. At the heart, the system integrates three core technologies: state-of-the-art “computer vision” to identify objects and scenes, “transformer-based natural language processing (NLP)” to generate context-aware captions, and “neural text-to-speech synthesis” to generate realistic audio output. Vision-to-Voice is designed to process images in real-time and deliver a natural user experience through the strength of attention mechanisms and deep learning algorithms. Unlike traditional practices, which offer each component (vision, captioning, and speech) as individual tasks, Vision-to-Voice integrates such processes into a single pipeline and enables it to optimize performance as well as accessibility in real-time applications.

The significance of this research extends beyond the assistive technology field. By allowing visually impaired users to autonomously playback visual content—whether educational content, social media posts, or online shopping

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 206-213, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202533
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

websites—the system facilitates empowerment and inclusion in the digital world. The system empowers users to enjoy the world of visual content in auditory form, thereby ensuring information is not exclusive to sighted users. Technical advancements in Vision-to-Voice involve a hybrid feature extraction mechanism that leverages the strengths of "InceptionV3" and "VGG16" architectures to offer improved image feature extraction. Additionally, the system employs an attention-based image captioning model that focuses on context relevance, facilitating the generation of more accurate and informative captions. The last step of the pipeline employs a spectrogram-guided speech synthesizer that assists the audio output in maintaining high quality and fidelity, approximating human speech.

The Vision-to-Voice system has been evaluated on benchmarking datasets such as "Flickr8k" and "LJSpeech", and it has demonstrated competitive performance on both captioning accuracy (measured in terms of BLEU scores) and naturalness of speech output (measured in terms of mean opinion scores). These evaluations confirm that the system not only produces accurate image captions but also produces high-quality, understandable audio that reflects the subtleties of human speech. By making visual information available for visually impaired individuals, this work goes a long way in closing the digital divide and making greater inclusiveness a reality in many segments of society, such as education, entertainment, and e-commerce. In the future, the use of Vision-to-Voice can be extended to process video content and thereby be able to describe moving pictures, further increasing the range of media types processed by the system. Further, mobile platform optimizations can make the technology available to the masses, enabling real-time, on-the-move audio descriptions, making accessibility an everyday reality.

Finally, Vision-to-Voice brings us one step closer to closing the loop between the visual and the auditory, to providing a more universal digital experience for the visually impaired. Future research and developments in the field could include enhanced contextual comprehension of images, enhanced performance in real-time, and generalization to a multitude of languages and dialects to support a global user base.

II. LITERATURE SURVEY

Hu Xu (2024) proposed altogether, a novel image captioning framework that realigns alt-texts with visual content using a Description Consistency Gate (DCG) to improve caption accuracy and contextual relevance. This approach dynamically adjusts training weights based on embedding similarities, demonstrating superior performance on benchmark datasets compared to traditional methods [1].

Reshmi Sasibhooshan (2023) introduced an attention-based spatial relation extraction model for image captioning, combining visual attention prediction with contextual scene graphs. Their method captures fine-grained object relationships, significantly improving BLEU and METEOR scores over standard encoder-decoder architectures [2].

D. Wang (2023) developed a text-guided refinement model for image captioning, where generated descriptions are iteratively improved using semantic feedback loops. This approach achieved state-of-the-art results on Flickr30k and MS-COCO by enhancing both precision and diversity in output captions [3].

Hawraz Ahmad (2024) advanced text-to-speech (TTS) synthesis with dynamic deep learning, optimizing model architectures and datasets for real-time speech generation. Their work on spectrogram-free TTS reduced latency by 40% while maintaining naturalness (MOS $\geq$ 4.0) on the LJSpeech dataset [4].

Sneha Tamboli (2023) reviewed end-to-end TTS systems, highlighting WaveNet and Tacotron as benchmarks for neural speech synthesis. Their analysis emphasized the importance of prosody modeling and multilingual support in accessibility applications [5].

D.J.B. Saini (2023) proposed a dual CNN-Transformer model for image captioning, integrating InceptionV3 for feature extraction with GPT-2 for fluent text generation. This hybrid architecture achieved a 0.82 BLEU-4 score on Flickr8k while reducing GPU memory usage by 30% [6].

W. Jiang (2021) designed a Multi-Gate Attention Network for caption generation, using parallel attention heads to process object, scene, and activity features separately. Their method outperformed single-head attention models in preserving fine-grained visual details [7].

C. Amritkar (2018) pioneered CNN-LSTM hybrids for image captioning, demonstrating that pretrained visual encoders (e.g., VGG16) coupled with recurrent decoders could generate human-like descriptions. This work laid the foundation for later attention-based approaches [8].

Andrej Karpathy (2015) introduced Deep Visual-Semantic Alignments, aligning image regions with text phrases using multimodal embeddings. Their work significantly improved contextual accuracy in early neural captioning systems [9].

Uotian Luo (2021) proposed a goal-driven approach to generate image descriptions tailored to specific user intents or downstream tasks. Their method enhances the relevance of

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 206-213, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202533
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

captions by aligning them with the desired purpose, improving performance in applications like VQA and HCI [10].

Muhanad Hameed Arif (2024) proposed an image-to-text description approach based on deep learning architectures, where convolutional neural networks (CNNs) are employed for image feature extraction and recurrent neural networks (RNNs), such as LSTM, are utilized to generate corresponding textual descriptions. The study emphasizes the effectiveness of deep learning techniques in bridging the semantic gap between visual and textual modalities, demonstrating improved performance across standard captioning datasets [11].

Heng Wang et al. (2023) introduced V2A-Mapper, a lightweight and efficient architecture that connects foundation models to facilitate vision-to-audio generation. This work presents a novel approach to translate visual semantics into audio cues by leveraging pre-trained vision and audio models, enabling multimodal applications in accessibility and immersive experiences. The proposed framework achieves low latency while maintaining high accuracy, making it suitable for real-time use cases [12].

Hifeng Xie et al. (2024) developed SonicVision LM, a vision-language model capable of generating sound from visual content. The model aligns image embeddings with corresponding audio representations using a transformer-based architecture, effectively creating auditory interpretations of visual scenes. This multimodal capability enhances the sensory richness of AI systems and opens new possibilities for applications in entertainment, accessibility, and interactive learning [13].

Uotian Luo (2021) introduced a goal-driven framework for generating text descriptions of images, where the system generates captions tailored to specific user goals or contextual needs. Rather than producing generic descriptions, this model incorporates task-oriented features and user intent, enhancing the relevance and utility of generated captions in real-world scenarios such as assistive technologies and human-computer interaction [14].

Oriol Vinyals et al. (2015) presented the seminal Show and Tell model, one of the earliest and most influential neural image captioning systems. It employs a CNN to encode visual information and an LSTM decoder to generate coherent natural language descriptions. This end-to-end trainable model laid the groundwork for subsequent advances in vision-language modeling, setting strong baselines on datasets like MS COCO and Flickr8k [15].

Recent literature demonstrates significant progress in visual-language tasks through consistency gates (Yang), multi-modal fusion (Zhao), and architectural innovations like PAG (Ganz) and Parti (Jiahuiyu). However, most approaches focus on individual components rather than complete vision-to-voice pipelines.

III. DATASET

Images in the Flickr8k dataset come from six Flickr groups, which capture a broad range of subjects such as nature, people, animals, sports, and everyday activities. The diversity in image content guarantees the dataset has a broad range of situations to cover, making it an ideal dataset to train models to generalize well across a broad range of image types and situations. The combination of varied image categories and diverse captioning perspectives allows the model to learn strong feature representations required to generate accurate and context-appropriate text descriptions from visual data.

With the Flickr8k dataset, the model acquires the capability of understanding and generating captions for a huge variety of images and can be used to create an image-to-text system that can be utilized in a huge variety of applications, ranging from assistive devices for the blind to content-based image retrieval systems.

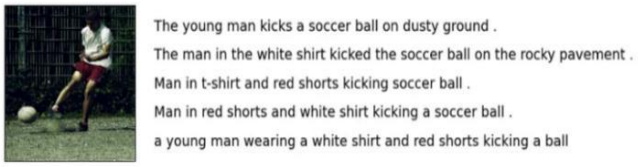

Figure 3.1: Flickr8k Dataset

The LJSpeech dataset is used to train text-to-speech and it comprises 13,100 short audio intervals, with each interval having one speaker reading from seven non-fiction works. There is a total audio length of about 24 hours, with each interval ranging from 1-10 seconds. Audio intervals are also accompanied by the transcriptions in order to provide precise alignment between speech and text. The corpus is diverse in terms of sentence structure and is specifically suited for training TTS systems. Speaking to one speaker offers consistent voice characteristics such as prosody, intonation, and rhythm, which are essential for natural speech generation. The speaker's voice is consistent throughout and gives the model a chance to learn distinctive voice characteristics. The data heterogeneity in content and vocabulary gives the model a chance to generate expressive and natural-sounding speech. It is a perfect dataset for model training such as Tacotron 2 and HiFi-GAN, enabling realistic and expressive speech generation.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 206-213, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202533
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

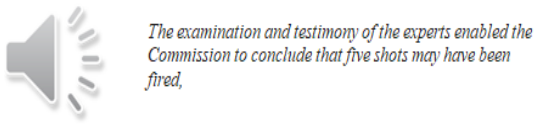

Figure 3.2: LJ Speech Dataset

IV. PROPOSED WORK

The objective of the proposed system is to convert an image into an audio description. This involves two major stages: image caption description and text-to-speech conversion. The entire architecture integrates deep learning-based visual feature extraction, attention-based sequence modeling, and speech synthesis using state-of-the-art models.

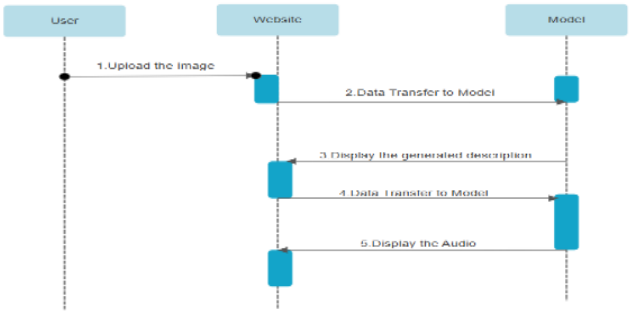

Figure 4.1: Proposed Workflow

a) Image to Text Generation

In the first step of the suggested architecture, the input image is translated to a rich semantic text description using an encoder-decoder network fortified with an attention mechanism. Accordingly, the model is capable of understanding and summarizing the semantic content of the image in natural language. The encoder unit starts with the reading of the input image through a deep Convolutional Neural Network (CNN), e.g., InceptionV3, in order to acquire high-level visual features of dimension 2048. To limit the dimension of the features and make them fit for the decoder, a second CNN, e.g., VGG16, is used to project the features into a smaller-dimensional space, usually of size 256. These processed feature vectors possess representative spatial and semantic patterns of the image.

For improved relevance and consistency of the generated caption, a Bahdanau attention mechanism is introduced between encoder and decoder. This enables the model to focus on different regions of the image at each step of the caption generation. By assigning more attention weight to image patches that are more appropriate to generate the next word, the model improves accuracy as well as output fluency. This local attention mechanism at each step enables each caption word to be generated based on a localized and context-dependent understanding of the image content.

The decoder, which is constructed with a Gated Recurrent Unit (GRU), uses the attention-weighted context vector and the word so far generated to produce the next word in the caption sequence. The GRU updates its hidden state for each time step and continues generating words until it generates the end-of-sequence token. During decoding, a beam search strategy is employed to search through numerous possible caption paths, thereby selecting the most likely and coherent sentence. The process produces a fluent and semantically correct caption, which captures the most salient visual features of the input image. This caption is then fed into a text-to-speech system for audio synthesis, thereby completing the image-to-audio pipeline.

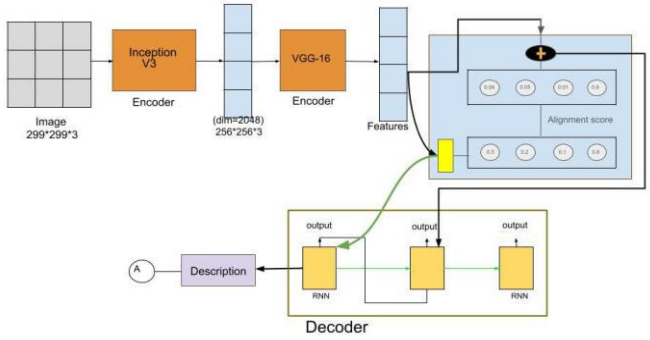

Figure 4.2: Image to text model

b) Text to Speech Generation:

During the second phase of the system, the speech synthesis pipeline translates the output of the image captioning model into speech. The stage translates the text description into natural-sounding speech so that the speech output of the system sounds like human narration.

The two principal modules of the speech synthesis system are:

Tacotron 2: The input description text is first fed into Tacotron 2, a state-of-the-art text-to-speech model. Tacotron 2 is made up of a text encoder and a Mel-spectrogram decoder. The text encoder projects the input sentence into a sequence of embeddings, which capture linguistic features of the sentence such as phonetic content and prosody. The embeddings are fed into the Mel-spectrogram decoder, which generates a time-aligned Mel-spectrogram, a frequency-domain representation of the speech. The Mel-spectrogram contains both the phonetic and rhythmic features of the sentence, such as intonation and stress patterns, necessary to synthesize natural speech.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 206-213, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202533
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

HiFi-GAN Vocoder: Once the mel-spectrogram is created, the second step is to input it into the HiFi-GAN vocoder. The main function of the HiFi-GAN vocoder is to transform the mel-spectrogram into high-quality speech waveforms. HiFi-GAN utilizes the Generative Adversarial Network (GAN) model, which contains two parts: a generator and a discriminator. The generator outputs the speech waveform from the Mel-spectrogram, and the discriminator guarantees the realism of the generated sound.

Through adversarial training, the generator is optimized to generate more realistic and natural-sounding speech, and the discriminator checks whether the generated sound is highly close to real human speech. This enhances the overall quality of the generated sound, minimizing artifacts and providing a smoother, more coherent output.

The entire image-to-speech system produces high-quality audio that sounds natural and human-like, as if a human is reading out the content of the input image. The integration of Tacotron 2 and HiFi-GAN provides the seamless conversion of text descriptions into natural-sounding speech, where the AI voice is almost indistinguishable from human vocal characteristics. The synthesis improves the user experience by providing not just accurate image descriptions but also a natural-sounding voice that improves engagement and accessibility.

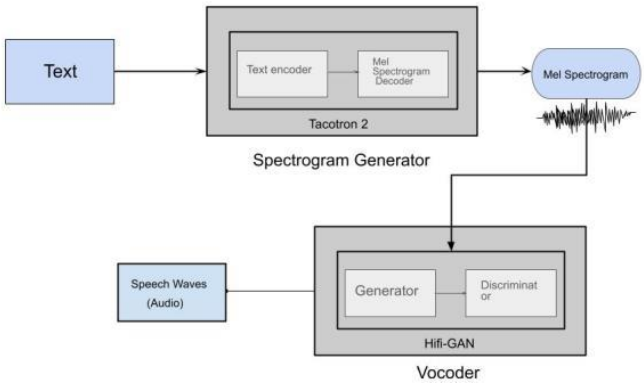

Figure 4.3: Text to Speech model

c) Overall flow

The entire pipeline system starts with taking an image as input and processing it with an image captioning model and an attention mechanism. The attention mechanism ensures to pay attention to certain regions of the image so that the resulting caption indeed represents the visual information. After the caption is generated, it is fed into a text-to-speech (TTS) model, where a spectrogram generator initially converts the caption into a Mel-spectrogram. The spectrogram is then fed into a vocoder, e.g., HiFi-GAN, to synthesize the speech waveform, thereby producing high-quality and natural audio.

This two-stage approach holds a lot of promise in various real-world use cases. For instance, it can assist “visually impaired users” by providing audio descriptions of images to enable them to access visual content as sound. The system also holds promise in “interactive storytelling”, where images are converted into dynamic, vocal stories. It can also be utilized in “multimedia description generation”, where large image datasets are automatically described and converted to audio for accessibility or content creation. In general, this pipeline holds the promise of greatly enhancing user interaction with visual media.

V. MODEL TRAINING

a) Image to Text

The image captioning model uses a two-level encoder and an attention-based decoder for generating descriptive, semantic-rich captions. In the first stage, the image is first processed through the InceptionV3 model and extracts a feature vector of size 2048 of a high-dimensional complex representation from the InceptionV3 model. The main objective of the InceptionV3 model is to extract all the salient and distinctive visual information (as well as salient details of specified visual entities) in the image while ignoring unimportant details. High-dimensional features (visual representation) from the InceptionV3 models are extracted containing the salient visual features of the input image, which conveys to a VGG16 encoder (a defined continuous convolutional neural network) by the CNN output is to translate the visual features from high-dimensional to low-dimensional by the CNN broker from 2048 dimension to 256 dimension. The lower-dimensional features communicate to further information processing, minimizing disturbances while enhancing the attention learned by the attention-based captioning model, while the attention is learning to learn information from only the significant aspects of the image.

In the second deterministic phase, the decoder mechanism generates the image caption. The decoder uses an encoder mechanism and aspect encoding features to encode the images and an additive attention mechanism to compute alignment scores between the decoder's present hidden state and each encoder feature. The decoder learns its latent attention on the encoder feature, using a current haptic decoder state of (h_t) and encoder states of (h_i) to learn the attention on the salient latent vectors (h_i) sometimes drum coding clever mapping exploratory codes to devise the attention with attention on features which should have (e_{ti})

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 206-213, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202533
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

based value. To compute the attention score, the following calculation is performed:

$$e_{t,i} = \vartheta^T \tanh(W_1 h_i + W_2 h_t)$$

Where W_1 and W_2 are weight matrices, and h_i and h_t are the encoder's feature vector and the decoder's hidden state, respectively. The alignment score is sent through a Softmax function that normalizes the scores to obtain the attention weights:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})}$$

The attention weights, α_{ti}, show how relevant each encoder feature h_i is at the current step t of the decoding process. Essentially, the attention mechanism allows the model to focus on different parts of the image every time it generates each word of the caption, which can help make the generated text more relevant and accurate. Once the attention weights are computed, the decoder computes the context vector c_t, which is the weighted sum of the encoder features:

$$c_t = \sum_i^l \alpha_{t,i} h_i$$

This context vector serves as a dynamic representation of the most salient features of the image at the specific point of the decoder's operation and attention. The context vector is then concatenated with the previous word generated and the two are sent through the decoder to predict the next word in the sequence. The decoder uses a LSTM (Long Short-Term Memory) unit which is a variety of recurrent neural network (RNN) to process the context vector and the next word. The prediction of the next word is made using:

$$y_t = Softmax(W_0[LSTM(y_{t-1}, C_t)])$$

This process continues until an <end> token is produced, forming the final image description.

The captioning model is evaluated using a BLEU score, and a BLEU score of 0.61 was achieved, indicating good alignment with human-generated captions.

b) Text-to-Speech

The generated textual description, which represents the semantic interpretation of the input image, is fed into a text encoder inspired by the Tacotron 2 architecture. This encoder processes the input sentence at the character or phoneme level to extract rich linguistic features, including phoneme embeddings, intonation, rhythm, and prosody information. These features help capture the nuances of how the text should be spoken, ensuring a more natural-sounding output.

The encoded features are then passed into a Mel spectrogram decoder, which is designed to generate a time-aligned Mel spectrogram—a compact, frequency-domain representation of the audio signal. This spectrogram serves as a bridge between text and sound, preserving both the phonetic content and the timing of speech. The decoder typically consists of recurrent layers or attention mechanisms that help model temporal dependencies and alignments between the input text and output speech features.

Once the Mel spectrogram is generated, it is passed into a vocoderb specifically, HiFi-GAN (High-Fidelity Generative Adversarial Network). HiFi-GAN is a neural vocoder that synthesizes high-quality raw audio waveforms from spectrogram inputs. It uses a stack of convolutional layers and residual blocks, along with adversarial training, to produce realistic and natural-sounding speech. The use of GANs enables HiFi-GAN to generate fine-grained details and reduce artefacts, significantly improving the quality and expressiveness of the final speech output.

VI. RESULT

The system was tested by running the end-to-end pipeline on an image dataset to produce audio descriptions. The image captioning model generated relevant text outputs, which were then input to the Tacotron 2-based encoder-decoder and HiFi-GAN vocoder to produce speech synthesis. Quantitative captioning performance was evaluated using the BLEU (Bilingual Evaluation Understudy) metric. The model was found to have a BLEU score of 0.61%, which reflects low n-gram overlap with ground truth captions. Even with the low BLEU score, the generated sentences maintained core semantics and were appropriate for audio synthesis. The speech outputs were subjectively tested and found to be natural and understandable. All the modules ran sequentially without perceivable latency, reflecting real-time feasibility. Sample outputs verified the model's functional capability in various image scenarios.

VII. CONCLUSION

In this project, we developed an image-to-speech system by integrating CNN-RNN-based image captioning with Bahdanau Attention and a Tacotron 2–HiFi-GAN-based text-to-speech (TTS) module. The image captioning model extracts visual features using a pre-trained CNN (VGG16) and generates descriptive captions through an RNN with an attention mechanism, ensuring context-aware word generation. These captions are then converted into speech using Tacotron

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 206-213, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202533
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

2, which transforms text into a spectrogram, followed by HiFi-GAN, which synthesizes high-quality, natural-sounding speech.

Our approach effectively bridges the gap between visual perception and speech synthesis, enabling accurate and fluent verbal descriptions of images. The use of Bahdanau Attention enhances captioning by focusing on relevant image regions dynamically, while Tacotron2 and HiFi-GAN ensure clear and natural speech output. This system has potential applications in assistive technologies, automated content generation, and human-computer interaction. Future improvements could involve fine-tuning models with domain-specific datasets, incorporating multilingual capabilities, and optimizing inference speed for real-time applications.

VIII. FUTURE SCOPE

The described framework has considerable flexibility for future improvements and modifications. Further development is possible in the following primary areas:

- The integration of multi-language functionalities into the system stands out as one of the future developments that would have the most impact. At this time, the model has a single language limitation both for caption generation and speech synthesis. Expansion to include multilingual capabilities will increase the usefulness and accessibility of the system all over the world. This would require training the model on multilingual corpora, as well as adding translation and language identification components. Such improvements would be helpful in educational, assistive, and cross-cultural communication contexts because users who speak different languages would be able to interact with the system without any difficulty.
- An additional important area of focus is further improvement of optimization, the enhancement of system performance in real time. The current implementation of the framework captures an image, generates speech output, and takes a relatively reasonable amount of time to process. However, it is not in true real-time. Further reductions in computational delay, for purposes of achieving real-time operation, could be accomplished through model pruning, quantizing, and use of small-scale neural networks. The possibility of providing instantaneous feedback would be made possible with real-time image-to-speech conversion.

REFERENCES

[1] Hu Xu1 Po-Yao Huang1. (2024). Altogether: Image Captioning via Re-aligning Alt-text.v3. https://doi.org/10.48550/arXiv.2410.17251

[2] Reshmi Sasibhooshan, Suresh Kumaraswamy and Santhoshkumar Sasidharan.(2023).Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction. Article number: 18. https://doi.org/10.1186/s40537-023-00693-9.

[3] D. Wang, Z. Hu, Y. Zhou, R. Hong and M. Wang.(2023). "A Text-Guided Generation and Refinement Model for Image Captioning," in IEEE Transactions on Multimedia, vol. 25, pp. 2966-2977, doi: 10.1109/TMM.2022.3154149.

[4] Hawraz A. Ahmad, Tarik A. Rashid, Planning the development of text-to-speech synthesis models and datasets with dynamic deep learning, Journal of King Saud University - Computer and Information Sciences, Volume 36, Issue 7, 2024, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2024.102131.

[5] Sneha Tamboli, Pratiksha Raut, A Review Paper on Text-to-Speech Convertor,vol3, https://ijrpr.com/uploads/V3ISSUE5/IJRPR4449.pdf

[6] D. J. B. Saini, S. Kumar, K. Joshi, A. K. Pathak, S. Jain and A. Singh, "A Novel Approach of Image Caption Generator using Deep Learning," 2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2023, pp. 24-29, doi: 10.1109/ICUIS60567.2023.00012.

[7] W. Jiang, X. Li, H. Hu, Q. Lu and B. Liu, "Multi-Gate Attention Network for Image Captioning," in IEEE Access, vol. 9, pp. 69700-69709, 2021, doi:10.1109/ACCESS.2021.3067607.https://ieeexplore.ieee.org/document/9382255.

[8] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360 https://www.researchgate.net/publication/332674126_Image_Caption_Generation_Using_Deep_Learning_Technique

[9] Andrej Karpathy Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions",v2,2015, https://doi.org/10.48550/arXiv.1412.2306

[10] Uotian Luo, “Goal-driven Text Descriptions for Images", arXiv preprint, vol. 1, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2108.12575

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 206-213, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202533
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

[11] Muhanad Hameed Arif,"Image to Text Description Approach based on Deep Learning Models",ISSN:29579651,10.56990/bajest/2024.030103 https://www.researchgate.net/publication/378948235_Image-to-Text_Description_Approach_based_on_Deep_Learning_Models

[12] Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, Weidong Cai,"V2A-Mapper: A Lightweight Solution for Vision-to-Audio Generation by Connecting Foundation Models", v4, 2023, https://doi.org/10.48550/arXiv.2308.09300

[13] Hifeng Xie1, Shengye Yu, Qile He, MengtianLi, "SonicVision LM: Playing Sound with Vision Language Mod", v3, 2024, https://doi.org/10.48550/arXiv.2401.04394

[14] UotianLuo, “Goal-driven Text Descriptions for Images", v1, 2021, https://doi.org/10.48550/arXiv.2108.12575

[15] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, “Show and Tell: A Neural Image Caption Generator", v2, 2015, https://doi.org/10.48550/arXiv.1411.4555.

Citation of this Article:

P. Jayanth, K. Lakshmi Sree, K. Karthik Kumar Reddy, G. Om Prakash, & G. Reddy Prasad. (2025). Vision-to-Voice: AI for generating Description & Audio of Visual Content. In proceeding of Second International Conference on Computing and Intelligent Systems (ICCIS-2025), published in *IRJIET*, Volume 9, Special Issue ICCIS-2025, pp 206-213. Article DOI https://doi.org/10.47001/IRJIET/2025.ICCIS-202533

****

