

Enhancing Child Education Through Image Generation from Spoken Descriptions

Mrs. A. Usha Rani
 Department of ECE
 Vignan's Institute of Information
 Technology,
 Visakhapatnam, India
 ushaviit@gmail.com

K. Venkatesh
 Department of ECE
 Vignan's Institute of Information
 Technology,
 Visakhapatnam, India
 karakavenkatesh2003@gmail.com

H. Sathwika
 Department of ECE
 Vignan's Institute of Information
 Technology,
 Visakhapatnam, India
 sathwikahariyala24@gmail.com

M. V. Phanindra
 Department of ECE
 Vignan's Institute of Information
 Technology,
 Visakhapatnam, India
 phanimylaru@gmail.com

G. Karthikeya
 Department of ECE
 Vignan's Institute of Information
 Technology,
 Visakhapatnam, India
 karthikgorle30@gmail.com

Abstract- *This paper presents a pioneering website designed to transform child education by enabling the generation of images from spoken descriptions. Utilizing cutting-edge natural language processing and image generation technologies, our platform empowers young learners to express their ideas verbally and witness them materialize as vibrant illustrations. With this innovative method, we connect the realms of language comprehension and visual interpretation, fostering cognitive growth and enabling more efficient learning processes. With a diverse array of educational content and intuitive tools, our website offers an engaging and immersive experience for children, fostering creativity and communication skills. This research highlights the significance of leveraging technology to enhance educational outcomes and provides a promising avenue for future advancements in the field of child education.*

KEYWORDS: *Image Captioning, Speech to Text, Deep Learning, Image Processing, Speech Recognition, Image Generation*

I. INTRODUCTION

In the realm of child education, where curiosity blossoms and imagination knows no bounds, there exists a world of endless possibilities waiting to be explored. In the contemporary digital era, technology serves as a fundamental force in reshaping the

educational landscape, notably by addressing the varied learning requirements of children. An innovative approach gaining momentum is the generation of images from spoken descriptions, which not only fosters creativity but also enhances comprehension and engagement among young learners. This endeavor harnesses the power of artificial intelligence (AI) and image-processing techniques to translate verbal expressions into vivid visual representations. In this context, we present a pioneering initiative leveraging the UNSPLASH database coupled with Python programming to facilitate this transformative educational experience. Speech recognition technology serves as the cornerstone of this project, enabling the seamless conversion of spoken language into text format. Through the integration of the SpeechRecognition module in Python, the system adeptly captures verbal descriptions articulated by children, irrespective of linguistic nuances or accents. This functionality empowers learners to articulate their ideas freely, fostering a conducive environment for self-expression and communication. Subsequently, the generated textual descriptions serve as the foundation for the image generation process. Leveraging advanced image processing algorithms and the versatile capabilities of the Pillow module in Python, the system translates these textual inputs into visually appealing representations. By analyzing the semantic content of the descriptions, the system retrieves

relevant images from the extensive UNSPLASH database, which encompasses a vast repository of high-quality, royalty-free photographs contributed by a global community of photographers. This integration not only enriches the learning experience but also ensures the availability of diverse and culturally inclusive visual content, catering to the multifaceted interests and backgrounds of young learners. Furthermore, the utilization of Python facilitates the seamless amalgamation of diverse functionalities, thereby streamlining the image generation process. The versatility of Python libraries such as Pillow enables the manipulation of images with precision and efficiency, ensuring the seamless alignment between verbal descriptions and visual depictions. This integration encapsulates the essence of interdisciplinary collaboration, bridging the realms of linguistics, computer science, and education to foster holistic learning experiences. Alongside the core system, a user-friendly web interface was built using HTML and the Bottle framework. This interface acts as a bridge, allowing users to easily interact with the image generation system. This intuitive platform empowers children to engage actively in the learning process, fostering a sense of agency and autonomy in their educational journey. Through a visually appealing and accessible interface, users can articulate their ideas through spoken descriptions and witness the real-time generation of corresponding images, thereby bridging the gap between verbal communication and visual comprehension. Moreover, the educational implications of this innovative approach are profound, transcending conventional pedagogical paradigms. By harnessing the power of image generation from spoken descriptions, educators can cultivate critical thinking skills, creativity, and visual literacy among young learners. This multimodal approach accommodates diverse learning styles, catering to the unique cognitive preferences and strengths of individual students. Additionally, the interactive nature of the platform fosters collaborative learning experiences, as children collaborate to articulate descriptions and interpret visual representations collectively. In conclusion, the integration of speech recognition technology, image generation algorithms, and web development frameworks heralds a new era in child education, characterized by innovation, inclusivity, and engagement. By harnessing the collective potential of these interdisciplinary domains, we aspire to empower the next generation of learners with the tools and resources necessary to thrive in an increasingly complex and dynamic world. Through the synthesis

of verbal and visual modalities, we endeavor to cultivate a generation of creative thinkers, equipped with the skills and competencies to shape a brighter future for themselves and society at large.

II. LITERATURE SURVEY

S2IGAN breaks new ground in speech-to-image generation, empowering unwritten languages. It forgoes text, utilizing a Speech Embedding Network (SEN) and a Relation-Supervised Densely-Stacked Generative Model (RDG) to translate spoken descriptions into realistic images. Rigorous testing on CUB and Oxford-102 datasets confirms S2IGAN's ability to create high-quality, speech-driven images, establishing a solid foundation for future speech-to-image advancements [1].

This research tackles direct speech-to-image translation, a game-changer for communication and artistic creation in languages without written forms. Our approach leverages a speech encoder trained in tandem with a pre-existing image encoder through a teacher-student method. This combined knowledge is then fed into a generative adversarial network to synthesize images directly from speech. Experiments confirm the system's ability to translate raw speech into images, bypassing text entirely. Ablation studies offer valuable insights into the model's inner workings [2].

This research investigates deep learning for image captioning, a cornerstone of computer vision and natural language processing. Current methods depend on manually labeled images, a laborious process. We introduce a method that merges real data with synthetic images generated by a GAN-based system and leverages attention-based captioning. This approach not only improves captioning for synthetic images but also enhances the quality of captions for real data, confirmed by qualitative and quantitative analyses [3].

This research introduces S2IGAN, a groundbreaking framework for speech-to-image translation. S2IGAN bypasses text entirely, utilizing a speech embedding network and a generative model to create high-fidelity images directly from spoken descriptions. Evaluations on established datasets demonstrate the system's success in capturing the relationship between speech and images, paving the way for advancements in cross-modal learning [4].

This research delves into unsupervised Recognition of Speech Automatically (RSA), aiming to break free from the limitations of labeled data for every language. It explores techniques like autonomous sub-word and word modeling, speech segmentation, and speech-to-text mapping. By analyzing these approaches, the study seeks to understand the inherent limitations and minimum data needs for ASR, particularly in resource-scarce languages. This knowledge can then be used to optimize resource allocation and development efforts in this field [5].

This research tackles scalable audio generation for image captioning datasets, crucial for learning language grounded in visual cues. A dual encoder pre-trains deep networks to represent both audio and images, aligning their core characteristics. A masked margin softmax loss further enhances model performance, achieving top results on the Flickr8k Audio Captions Corpus. However, human evaluations of retrieved audio highlight discrepancies with automatic metrics, emphasizing the need for robust assessment methods in this field [6].

This research presents S2IGAN, a groundbreaking framework for Speech-to-Image generation. S2IGAN overcomes the limitations of text-based methods by directly translating spoken descriptions into realistic images. It leverages a two-part system: a speech embedding network and a generative model. The first network captures the essence of speech descriptions, informed by corresponding images. The generative model then uses this understanding to create images that closely match the spoken content. Rigorous testing confirms S2IGAN's ability to produce high-quality, semantically accurate images, paving the way for text-free S2IG applications [7].

III. METHODOLOGY

The methodology for the project "Generating Images from Spoken Descriptions for Child Education" entails a multifaceted process aimed at creating a system that can produce images based on verbal descriptions, particularly targeted at enhancing child education. The initial step involves meticulous data gathering from reputable sources such as the UNSPLASH database, ensuring a wide array of high-quality images relevant to various educational topics suitable for children. This curated dataset undergoes rigorous preprocessing to filter out any images that

may be inappropriate or irrelevant to the educational context. The core functionality of the system lies in its ability to accurately transcribe spoken descriptions into text format. To achieve this, the project leverages the powerful speech recognition module available in Python. This module plays a pivotal role in converting the verbal input provided by users or educators into a textual representation that can be comprehended and processed by the innovative system.

Cutting-edge machine learning techniques like Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs) power the core image generation process. The sophisticated models are trained on preprocessed image datasets paired with their corresponding speech descriptions. This training allows the system to grasp the complex relationships between spoken language and visual elements. This understanding empowers it to synthesize high-quality images that accurately reflect the spoken input and its context. Additionally, the system leverages the Python Pillow library for advanced image processing, ensuring the generated images maintain clarity, quality, and strong alignment with the spoken descriptions.

This integration is vital for seamless image manipulation and enhancement, ultimately contributing to the overall effectiveness of the system. The user interaction aspect is addressed through the development of a user-friendly website using HTML and the Bottle framework. This website serves as the interface through which users can input their spoken descriptions and subsequently receive the generated images in response. The website's design and functionality are carefully crafted to provide a smooth and intuitive user experience, catering specifically to educators and individuals involved in child education. A crucial phase in the methodology involves rigorous testing and validation of the system. This includes evaluating the generated images against the original spoken descriptions to assess accuracy and coherence. User feedback and input are also solicited during this phase to identify any areas of improvement or refinement needed to enhance the system's performance and usability. Overall, the methodology adopts a systematic and iterative approach, encompassing data collection, preprocessing, speech recognition, image generation, website development, testing, and user feedback. These comprehensive steps are geared towards developing a robust and effective system that leverages the power of

technology to enhance child education through visually engaging learning aids.

IV. PROPOSED MODEL

A) Block Diagram of our work

Captures spoken words and converts them to a digital format. The system works by first capturing the speech signal, which is then preprocessed to remove noise and other unwanted sounds. The system analyzes the speech signal to extract key characteristics, which are features of the signal that can be used to identify the spoken words. These features are then compared to a database of known words or phrases, and the best match is identified as the output text.

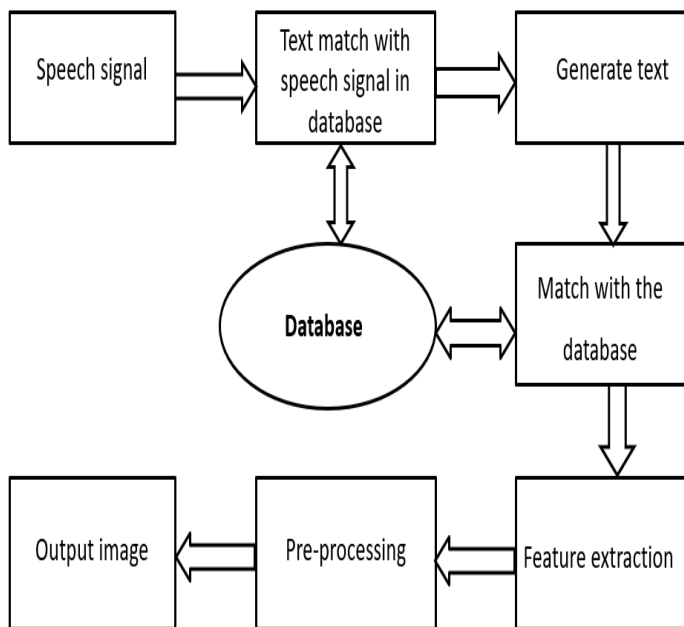


Figure 1: Block Diagram of Speech-to-Image Conversion

Speech signal: This is the raw audio input that the speech recognition system receives.

Processing: The given input speech signal is filtered to remove noise and other unwanted sounds.

Feature extraction: The system analyzes the speech signal to extract key characteristics. These characteristics might include the speaker's pitch, the

overall signal strength, and Mel-frequency cepstral coefficients (MFCCs).

Match with database: The features extracted from the pre-processed speech signal are then forwarded to this stage, where a comparison occurs. Here, a vast database containing a collection of known words or phrases comes into play. This database likely stores reference features for each word or phrase it encompasses.

Text match with speech signal in the database: The system finds the best match between the features of the speech signal and the words or phrases in the database.

Generate text: After meticulously comparing the features and identifying the closest match within the database, this stage translates the outcome into text. The system essentially retrieves the word or phrase from the database that aligns most closely with the features extracted from the user's speech input. This retrieved word or phrase represents the recognized text corresponding to the spoken language.

Process Description: Preprocess the text description. Remove punctuation and capitalization (for simplicity in child education). Split the description into keywords or short phrases. Identify key concepts in the description (e.g., animals, objects, colors).

Image Search: Use the UNSPLASH API to search for images based on the identified concepts. Filter results for age-appropriateness and educational value (consider using UNSPLASH SFW filters).

Image Selection: Find an image (or a few images) that best illustrates the description. You can implement a scoring system based on keyword matching or relevance ranking provided by the UNSPLASH API.

Image Processing: Resize or crop the image(s) for better display. Apply basic image editing for clarity or educational purposes (e.g., highlighting key elements). Use the Pillow module for image manipulation tasks.

Image Refinement: Use the Bottle framework to create a dynamic webpage. This webpage displays the user's spoken description and also displays the generated image(s).

Output image: In essence, the speech recognition system takes an audio speech signal as input, pre-processes it to eliminate noise, extracts characteristic features, compares these features against a database of known words or phrases, and finally generates the recognized text corresponding to the closest match.

B) SOFTWARE DESIGN

Flowchart of the working procedure of the project:

The process starts with the user interacting with the webpage. The user provides a spoken description through the microphone, which is captured and converted into text by the speech recognition module. A Python script then takes over to process the text description. It extracts key elements from the description to identify the most relevant content.

Next, the script leverages an API to search the UNSPLASH database that visually relevant images based on the extracted keywords. The retrieved images are then filtered to ensure they meet the set criteria, such as being copyright-free and appropriate for children.

Noisy environments can make it challenging to extract clean features from the speech signal. Speech characteristics can differ between individuals due to factors like age, gender, and accent. The system's performance might be impacted if the speaker's voice isn't well-represented within the training data used to build the recognition model. The system's ability to recognize words or phrases is inherently limited by the database it references.

Children with learning differences or those who benefit from visual aids might find this tool particularly helpful.

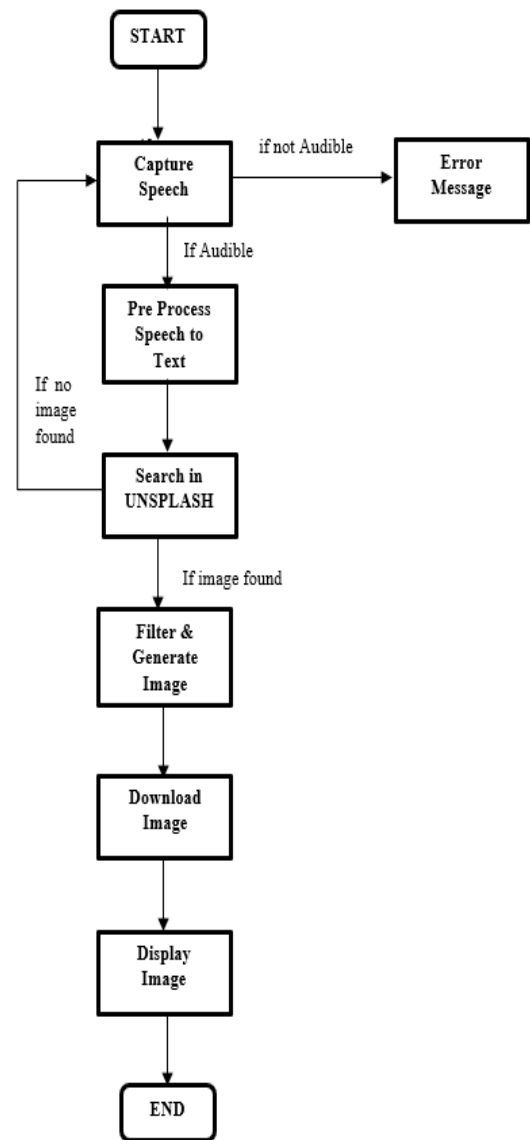





Figure 2: Flowchart of generating images from spoken descriptions

From the filtered options, the system selects an image and downloads it. The Pillow library is used to prepare the image for display, which might involve resizing or formatting adjustments. Finally, the Bottle framework comes into play to generate a webpage that displays the chosen image alongside the original spoken description. The user can then see the webpage with the image corresponding to their description.

V. RESULTS AND DISCUSSION

TABLE 1. PRESENTS THE RESULTS OF SPEECH-TO-IMAGE CONVERSION, DISPLAYING THE INPUT SPOKEN PHRASES AND THEIR CORRESPONDING OUTPUT IMAGES. THIS CONCISE SUMMARY ENCAPSULATES THE TRANSFORMATION PROCESS, SHOWCASING THE EFFECTIVENESS OF CONVERTING SPEECH INTO VISUAL REPRESENTATIONS.

Input Speech	Output Image
<p>“Elephant ”</p>	
<p>“ TIGER”</p>	
<p>“ ZEBRA”</p>	

VI. CONCLUSION

In conclusion, the creation of a web-based platform for speech-to-image conversion in child education represents a significant advancement in leveraging technology to enhance learning experiences. Through the integration of speech recognition and image retrieval technologies, this study has demonstrated the feasibility of creating an interactive learning environment tailored to the diverse needs and preferences of young learners. The experimental results indicate promising outcomes in accurately transcribing spoken words into visual representations, fostering engagement, and promoting knowledge retention among children aged 5 to 10 years old.

This research contributes significantly to the field of educational technology by introducing a novel approach to content delivery that aligns with the multimedia-rich preferences of digital-native learners. By amalgamating speech and image technologies, this study has opened new avenues for creating immersive and personalized learning experiences that stimulate curiosity and creativity. The findings highlight the potential of speech-to-image conversion as a valuable tool for educators to enhance pedagogical practices and facilitate active learning.

Looking ahead, further research and development efforts are warranted to refine the system's accuracy, scalability, and accessibility. Exploring additional features such as real-time feedback mechanisms and adaptive learning algorithms could enhance the platform's effectiveness in addressing individual learning needs. The merging of speech and image technologies in classrooms is poised to transform how children learn as technology advances.

In summary, this study underscores the transformative potential of speech-to-image conversion for child education and lays the groundwork for continued innovation in the field of educational technology.

VII. ACKNOWLEDGMENT

The authors would like to express our gratitude to Mrs. A. Usha Rani, Assistant Professor, Department of Electronics and Communications Engineering for her valuable guidance, advice, and support in successfully doing this project. Our sincere thanks to Dr. L. Rathaiyah, Chairman, of Vignan's group of institutions, for his co-operation and providing facilities for doing this project. We would like to extend our gratitude to Dr. V. Madhusudhan Rao, Rector, Mr. B. Srikanth, CEO, Vignan Vizag Group, Dr. J. Sudhakar, Principal, Vignan's Institute of Information Technology (VIIT), Visakhapatnam, Dr. Ch. Ramesh Babu, Head of the Department, Electronics and Communication Engineering for their valuable suggestions and encouragement for completion of this project. Lastly, we extend our gratitude to all our Teaching and Non-Teaching staff and all my friends, who directly or indirectly helped us in this Endeavour.

REFERENCES

- [1] D. Harwath et al., "Jointly discovering visual objects and spoken words from raw sensory input," in 2018.
- [2] X. Wang et al., "S2IGAN: Speech-to-image generation via adversarial learning," in 2020.
- [3] J. Li et al., "Direct speech-to-image translation," in 2020.
- [4] MOHAMMED BENNAMOUN, (Senior Member, IEEE) et al., "Text to Image Synthesis for Improved Image Captioning" in 2020.
- [5] A. Haque et al., "Audio-linguistic embeddings for spoken sentences," in 2019.
- [6] D. Harwath et al., "Unsupervised learning of spoken language with visual context," in 2016.
- [7] Xinsheng Wang et al., "Generating Images From Spoken Descriptions," in 2021.
- [8] D. Merx et al., "Language learning using speech to image retrieval," 2019.
- [9] G. Ilharco et al., "Large-scale representation learning from visually grounded untranscribed speech," in 2019.
- [10] O. Scharenborg et al., "Speech technology for unwritten languages," in 2020.
- [11] S. Reed et al., "Learning deep representations of fine-grained visual descriptions," in 2016.
- [12] R. Prenger et al., "Waveglow: A flow-based generative network for speech synthesis," in 2019.
- [13] A. Karpathy et al., "Deep visual-semantic alignments for generating image descriptions," in 2015.
- [14] S. Palaskar et al., "Learned in speech recognition: Contextual acoustic word embeddings," in 2019.
- [15] W. Li et al., "Object-driven text-to-image synthesis via adversarial training," in 2019.