National Conference on Sustainable Engineering And Management (NCSEM24)
ISSN 2349-6002

# Enhancing Multi-Modal Learning Aid for Dyslexic Education

Anjelo Wincy Bobby. S[1], Benishma. B[2], Blessy. D[3], Doana. S.V[4], Mrs. Sowmiya. R[5]
[1,2,3,4]*Department of Computer Science and Engineering, PJCE.*
[5]*Assistant Professor, Ponjesly College of Engineering.*

***Abstract*— Dyslexia in children can significantly affect their reading, writing, and spelling abilities, presenting unique challenges in their educational journey. To address these challenges, this study proposes a multi-modal approach that aims to enhance the learning process for children with dyslexia.It aims to capitalize on the strengths of individuals with dyslexia,such as their visual processing skills, while addressing their difficulties with reading and decoding text. The methodology employed in this study involves designing a tool to convertspeech into visual representations and vice versa. This paper is intended to provide supportive assistance to dyslexic children, offering them innovative and creative ways to learn and understand various concepts. By utilizing visual representations, the study aims to make learning more accessible and engaging for children with dyslexia, helping themovercome some of the traditional barriers they may face in education.**

***Keywords—Dyslexia, image, text, speech, visual learning***

## I. INTRODUCTION

Individuals with dyslexia often encounter difficulties in language-related activities, including reading, writing, and interpreting words. These challenges can hinder the learningprogress of affected children. This project aims to enhance learning outcomes for dyslexic children by converting audio input into realistic images. By providing visual representations of educational concepts, it offers an alternative means of accessing information, aiding young learners in understanding and retaining information more effectively. This approach also provides a creative learning environmentfor children to develop their own ideas. Overall, it presents an innovative method for learning and comprehension, utilizing realistic generated images.

The process initiates with the collection of input through a microphone, capturing the spoken voice. Subsequently, a speech recognition tool is employed to generate precise textual transcriptions of the spoken words. The generated textundergoes text embedding techniques to extract the meaningof the text. Using this extracted text, the project produces a visually alluring representation. In the end, the processed visuals are presented to the user.

To enhance the learning reinforcement, the application includes additional feature to convert images uploaded by theuser into corresponding text. Then the application transformsthe extracted text to generate speech that describes the content of the image.

This project combines speech-to-image and image-to-speech technologies. This framework simplifies the translation of concepts, enabling the application to automatically generate visuals from spoken content and vice versa. The use of AI here represents an innovative approach to enhancing education.

## II. RELATED WORKS

In this section, we examine related research. There are paper on "Direct Speech-to-Image Translation," which focuses on translating speech directly into text, presents broad applications. This approach offers additional features, particularly in scenarios where writing systems are not available. However, a comprehensive investigation into the process and accuracy of direct conversion has yet to be conducted.

As machine learning progresses, image generation hasemerged as a promising field. "Text-to-Image Synthesis for Improved Image Captioning" employs Generative Adversarial Networks (GANs) for this purpose. GANs have garnered significant interest due to their capability to create high-dimensional data. They consist of two components:

a generator, which produces the image, and a discriminator, which evaluates the authenticity of the image.

Another important component is the speech recognition library, a crucial resource for recording and transcribing spoken language, as utilized in "Speech to Image Conversion" journals. This library plays a significant role in recognizing speech patterns and converting them into a textual format.

Optical Character Recognition (OCR) is a technology used to recognize text within a digital image. It is commonlyemployed to recognize text in scanned documents and images. OCR software can convert physical paper documents or images into accessible electronic versions with text.

## III. METHODS

### A. Speech Recogntion

Speech recognition, also known as automatic speech recognition (ASR) is the process of converting spoken language into text. It involves capturing spoken words and converting them into a format that can be understood by computers, typically in the form of text. Speech recognition systems use algorithms and machine learning techniques to recognize and interpret spoken language, enabling users to interact with devices using their voice.

### A. Text Embedding

Word2Vec is a technique used to create word embeddings, which are dense vector representations of words in a high-dimensional space. The main idea behind Word2Vec is to capture the semantic and syntactic similarity between words based on their context in a large corpus of text.

### B. GAN

Generative Adversarial Network is a class of machine learning frameworks designed to generate new data that resembles a given dataset. The framework consists of two neural networks: the generator and the discriminator. 1)Generator: The generator takes random noise as input and generates new data samples. For example, in image generation, the generator takes random noise and outputs an image. 2)Discriminator: The discriminator receives both real data samples from the dataset and fake data samples generated by the generator. Its job is to distinguish between real and fake samples.

### C. OCR

Optical Character Recognition is a technology that enables the conversion of different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera, into editable and searchable data. OCR software recognizes the text within these documents and converts it into machine-readable text. OCR technology has evolved over the years, with advancements in machine learning and artificial intelligence improving its accuracy and efficiency. Modern OCR systems can recognize various languages, fonts, and writing styles, making them versatile tools for handling a wide range of document types.

### D. Denoising

Denoising refers to the process of removing noise from a signal. In the context of images, denoising typically involves removing unwanted artifacts or imperfections (noise) from the image to improve its quality and clarity. Noise in images can come from various sources, such as electronic interference in digital images or imperfections in the capturing process. Denoising is an important preprocessing step in various image processing tasks, such as image restoration, image enhancement, and feature extraction, as it helps improve the quality and reliability of subsequent processing steps.

### E. Open AI

Whisper.AI specializes in conversational analytics and voice-based solutions. It is an AI-driven technologies to analyze and improve customer interactions through voice, speech, and natural language processing. Its solutions are used in various industries, including customer service, sales, and support, to enhance customer experience and operationalefficiency.

## IV. MODELS

### A. Speech-to-Image Conversion

The speech-to-image model begins by capturing and processing the user's speech input, aiming to reduce external noise and enhance clarity. This processed speech is then passed through a Speech Recognition

system (Fig 1), which converts the audio signal into textual form. The resulting text undergoes text embedding, a process that converts the text into numerical vectors. These vectors represent the semanticmeaning of the text, enabling the model to understand the content.

The embedded text serves as the input to a Generative Adversarial Network (GAN) algorithm, which comprises two components: the Generator and the Discriminator. The Generator utilizes the text embedding vector to produce images that closely resemble real images. Meanwhile, the Discriminator distinguishes between real images from a dataset and the generated images from the Generator. It is trained to correctly classify the origin of the images, further refining the Generator's output.

The generated images undergo post-processing, whichincludes tasks such as noise reduction, clarity enhancement, and overall visual quality improvement. Finally, the processed image is displayed to the user, completing the speech-to-image conversion process.

*B. Image-to-Speech Conversion*

To further elaborate on the process, after acquiring an image (Fig 2), typically in formats like JPEG or PNG, the next step is to preprocess it to optimize its quality and then send it for Optical Character Recognition (OCR). This preprocessing stage often involves several operations such as noise removal, enhancing the image, and resizing it if necessary. These steps are crucial to ensure that the image is clear and suitable for accurate OCR results.

Once the image is preprocessed, it is passed through an OCR system to extract text from it. The OCR system analyzes the image and converts the text it detects into a readable format. This extracted text can then be used to generate a textual description of the object or scene depicted in the image. Finally, the extracted text is converted into speech using technologies like OpenAI's Whisper AI. This synthesized speech can then be presented to the user, providing them with a spoken description of the content of the image. This entire process enables the transformation of visual information into accessible and understandable spoken content. This speech-to-image can be beneficial for understanding complex concepts or instructions, as somedyslexic individuals may find it easier to comprehend information visually. This can be particularly helpful for learning complex concepts or following instructions.

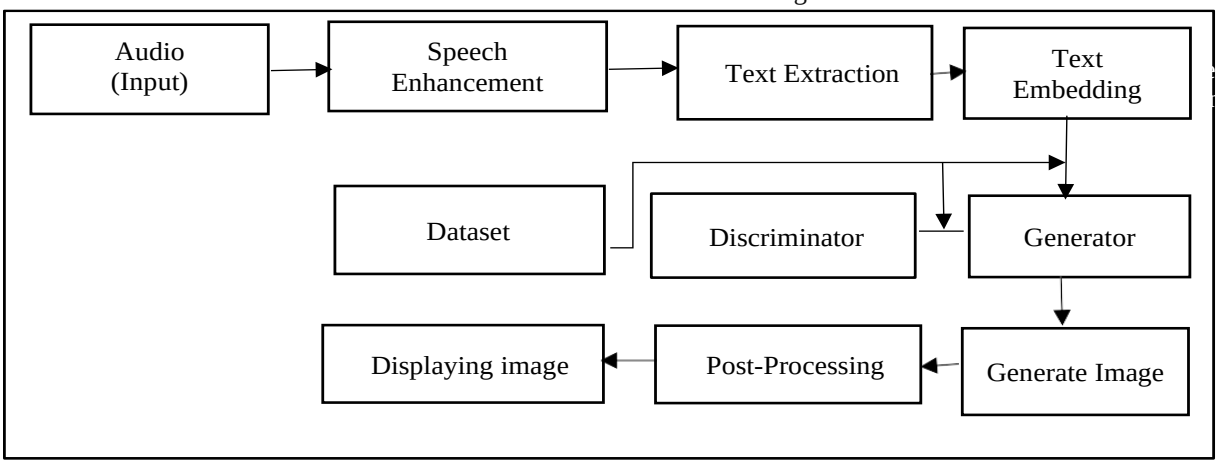


Fig. 1. Speech-to-Image Conversion

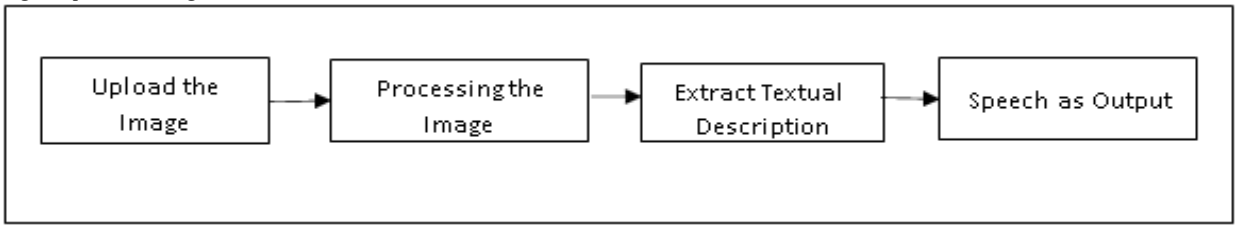


Fig. 2. Image-to-Speech Conversion

Also image-to-text can help with reading difficulties, allowing individuals to listen to text instead of struggling to read it themselves which, can help improve reading fluency and comprehension, as well as reduce the cognitive load associated with decoding written text. Overall, these technologies can enhance the learning experience for people with dyslexia by providing alternative and more accessible ways to consume and interact with information.

## V. Result And Discussion

By integrating, the accuracy of translating spoken words into text is significantly improved. This improvement is particularly evident in the resulting images, which are not only relevant but also adept at capturing the essence of the spoken word. With the assistance of OpenAI's advanced algorithms, sophisticated, high-quality images are generated effectively encapsulating the meaning and context of the spoken content. This enhancement in accuracy and relevance contributes greatly to enhance the overall user experience.

Moreover, the integration of these technologies enables the application to offer a more inclusive learning experience, catering to individuals with diverse learning styles. By leveraging AI, the application not only improves the accuracy of translation but also allows for the customization of learning materials based on individual preferences and needs. This customization plays a crucial role in enhancing the effectiveness of the learning process, as it ensures that the content is tailored to suit the unique requirements of each learner.

Furthermore, this fusion of technologies represents a significant advancement in educational tools, offering innovative ways to improve learning outcomes. The ability toaccurately translate spoken words into text and generate relevant images greatly enhances the educational experience,making it more engaging and interactive. This, in turn, leads to improved comprehension and retention of information, ultimately resulting in better learning outcomes for users. Overall, the integration of these technologies has the potentialto revolutionize the field of education, offering newpossibilities for enhancing the learning process and making it more accessible to a wider audience.

In the end, the integration of text embedding process, computer vision, and AI technologies has led to significant advancements in educational tools. These advancements havenot only improved the accuracy of translating spoken words into text and generating relevant images but also offer a moreinclusive learning experience. By catering to diverse learningstyles and preferences, these technologies have the potential to revolutionize the field of education, offering innovative ways to enhance learning outcomes and make education moreaccessible to all.

There are some outputs generated by the model by givinginput as speech to the system for lion going to school (Fig. 3),boy carrying apples (Fig. 4) and man interacting with robot (Fig 5). Here the speech is given as input and the corresponding output is generated respectively as an image. Similarly, the image taken by the model will also generate itsspeech description.

converting spoken words into visual representations, the tooloffers a unique way to access and understand information, catering to the specific needs of dyslexic learners. Furthermore, the ability to convert images back into speech provides a comprehensive learning experience, enabling children to engage with and comprehend content in a manner that suits their learning style. Overall, this project has the potential to greatly enhance the educational experience for dyslexic children, offering new avenues for learning and understanding.


Fig. 3. Image generated for a lion going to *school*

Fig. 4. Image generated for a boy carrying bag of apples

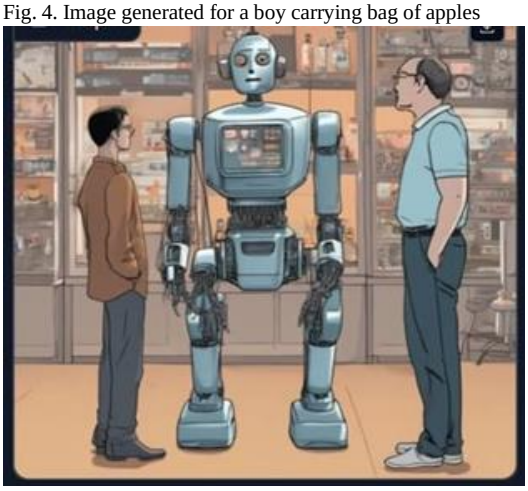

Fig. 5. Man interacting with Robot

## VI. CONCLUSION

In conclusion, the development of a visual learning aid for dyslexic children through a speech-to-image generator represents a significant step forward in educational technology. This innovative approach leverages advanced technologies such as AI to create a more inclusive and effective learning environment for children with dyslexia.

## ACKNOWLEDGMENT

We would like to express our sincere gratitude to the management of our college "Ponjesly college of Engineering", for providing us with the necessary resources and support throughout the duration of this project. We are immensely thankful to our esteemed professors and faculties whose guidance and encouragement have been invaluable in shaping our understanding and execution of this project. Their expertise and mentorship have played a pivotal role in our academic journey.

Special thanks to our project guide, Assistant Professor Mrs. Sowmiya, whose unwavering support, insightful feedback, and dedication have been instrumental in navigating through the challenges and achieving our goals. Her mentorship has not only enriched this project but has alsoleft a lasting impact on our personal and professional growth. We are truly grateful for her guidance and inspiration. Additionally, we extend our heartfelt appreciation to all those who contributed directly or indirectly to the success of this endeavor. We would like to thank everyone involved for their collective efforts in making our project a success.

## REFERENCES

[1] E. Bergelson and D. Swingley, "At 6–9 months, human infants know the meanings of many common nouns," Proceedings of the National Academy of Sciences, vol. 109, no. 9, pp. 3253–3258, 2012.

[2] H. Zhang, T. Xu, and H. Li, "Stackgan: Text to photo- realistic image synthesis with stacked generative adversarial networks," in 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017, pp. 5908–5916.

[3] T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked genera tive adversarial networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2018

[4] D. Tannen, Spoken and written language: Exploring orality and literacy. ABLEX Publishing Corporation,1982, vol. 32.

[5] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672 2680.

[6] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in Proceedings of the onThematic Workshops of ACM Multimedia 2017. ACM,2017, pp. 349–357. [Online]. Available:

https://arxiv.org/pdf/1704.08292.pdf

[7] T.H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," arXiv preprint arXiv:1905.09773, 2019.

[8] A. Duarte, F. Roldan, M. Tubau, J. Escur, S. Pascual, A. Salvador, E. Mohedano, K. McGuinness, J. Torres, and X. Giro-i Nieto, "Wav2pix: speech-conditioned face generation using generative adversarial net works," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, 2019.

[9] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2017, pp. 506–517.

[10] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in The European Conference on Computer Vision (ECCV),
September 2018

[11] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson,
Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," arXiv preprint arXiv:1904.06037, 2019.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and
S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in Advances in Neural Information Processing Systems, 2017, pp. 6626–6637.

[13] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2439–2448.
[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros,"Image-to-image translation.

[14] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in Automatic Speech Recognition and Understand ing (ASRU), 2015 IEEE Workshop on. IEEE, 2015, pp. 237–244.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEEconference on computer vision and pattern recognition, 2014, pp. 580–587

[16] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[17] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual sematic embeddings of untranscribed speech," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2018, pp. 4969–4973.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[19] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 7893– 7897

[20] A. Waibel and I. R. Lane, "Enhanced speech-to-speech translation system and methods for adding a new word,"Mar. 3 2015, uS Patent 8,972,268.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.\ [36] A.

[22] B´erard, O. Pietquin, C. Servan, and L. Besacier, "Listenand translate: A proof of concept for end-to-end speech-to-text translation," arXiv preprint arXiv:1612.01744, 2016.

[23] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos,
E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coateset al., "Deep speech: Scaling up end-to-end speech recognition," arXiv preprint arXiv:1412.5567, 2014

[24] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng,
G. Chen et al., "Deep speech 2: End-to-end speechrecognition in english and mandarin," in International Conference on Machine Learning, 2016, pp. 173–182.

[25] S. Hochreiter, Y. Bengio, P. Frasconi, J.

Schmidhuber etal., “Gradient f low in recurrent nets: the difficulty of learning long-term dependencies,”

[26] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014

[27] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in Advances in neural information processing systems, 2014, pp. 487–495.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset,” California Institute of Tech nology, Tech. Rep.CNS-TR-2011-001, 2011.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in European conference oncomputer vision. Springer, 2014, pp. 740–755.

[30] D. Harwath, A. Torralba, and J. Glass, “Unsupervised learning of spoken language with visual context,” in Advances in Neural Information Processing Systems, 2016, pp. 1858–1866.

[31] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W.Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi- speaker neural text-to speech,” arXiv preprint arXiv:1705.08947, 2017. Chen et al., “Deep speech 2: End-to-end speech recognition in english and mandarin,” in International Conference on Machine Learning, 2016,pp. 173–182.

[32] S. Morishima, and H. Harashima, “Speech-to-Image Media Conversion based on VQ and Neural Network,” In Acoustics, Speech, and signal Processing, IEEE International Conference on IEEE Computer Society, pp. 2865-2866, 1991.

[33] H. Yang, S. Chen, and R. Jiang, “Deep Learning-Based Speech-to-Image Conversion for Science Course,” In INTED2021 Proceedings, pp. 2910-2917, 2021.

[34] Jiguo Li et al., “Direct Speech-to-Image Translation,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 3, pp. 517-529, 2020.

[35] Stanislav Frolov et al., “Adversarial Text-to-Image Synthesis: A Review,” Neural Networks, vol. 144, pp. 187-209, 2021.

[36] Xinsheng Wang et al., “Generating Images from SpokenDescriptions,” IEEE/ACM Transactions on Audio,Speech, and Language Processing, vol. 29, pp. 850-865,2021.

[37] Lakshmi Prasanna Yeluri et al., “Automated Voice-to- Image Generation Using Generative AdversarialNetworks in Machine Learning,” In E3S Web of Conferences, 15th International Conference on Materials Processing and Characterization (ICMPC 2023), vol.430, 2023.

[38] Uday Kamath, John Liu, and James Whitaker, Deep learning for NLP and Speech Recognition, Springer Nature Switzerland, 2019.

[39] Santosh K. Gaikwad, Bharti W. Gawali, and Pravin Yannawar, “A Review on Speech RecognitionTechnique,” International Journal of Computer Applications, vol. 10, no. 3, pp. 16-24, 2010.

[40] Dong Yu, and Li Deng, Automatic Speech Recognition, A Deep Learning Approach, Springer-Verlag London, 2015.