Grenze International Journal of Engineering and Technology, June Issue

VocalVision: Integrative AI for Multilingual, Multimodal Communication Platform

Rohit V. Dadgal[1], Shruti P. Arsode[2], Kaushal A. Bharade[3], Vihal V. Karhade[4] and Dinesh G. Harkut[5]
[1,5]Prof. Ram Meghe College of Engineering and Management, Amravati, India
Email: rohit.dadgal21@gmail.com, dg.harkut@gmail.com

***Abstract*—VocalVision is an AI system that aims at improving the multilingual and multimedia, text, speech and image communication system. VocalVision has engineered its service using Azure AI services such as Text-to-Speech (TTS), Optical Character Recognition (OCR), Computer Vision, and Translator Text API for enabling natural voice solutions, various language support for text-to-speech services, and precise image description solutions. It includes instant translation messaging for various requirements, and enhancing the usage for blind individuals. It is hoped that this proposed scheme should help address these gaps more effectively and offer more appropriate contextual interactions relevant in education, customer-care, and assistive technology domains.**

***Index Terms*— Image-to-Speech, Image captioning, Image Text extraction, Text Summarization, Multilingual-to-Monolingual Translation.**

I. INTRODUCTION

Cross language and modality communication which is vital in a multi – media society is still a major issue. Today's technologies such as ASR, MT and IC have seen improvements but they are not completely aligned to offer end-to-end solutions. VocalVision solves this inactivity by using Azure AI services which encompass text, speech and image inputs for a production of an accurate output based on a person's personal settings and context. This system improves user engagement through simultaneous translation and natural sounding voices together with, providing descriptions of images in detail making it a useful tool for applications such as educational, customer service, and assistive ones.

The process of creation of VocalVision firstly involves obtaining the inputs from user, which further goes to Azure Cognitive Services for processing text inputs as well as speech and images. The system enables the user to select the most appropriate form of output, thus making the user's communication more effective. Specifically for the identification of speech and for establishing user preference the Speech Recognition API in the Azure suite is deployed.

The Text-to-Speech (TTS) service by Azure AI converts the processed text to human like sound which helps the people in need of audio response especially the blind. The tasks covered here are Automatic Text Summarization using the Microsoft Azure Text Analytics API, Image Caption generation based on Azure Computer Vision & Show, Attend and Tell model, and, Translation by Translator Text API. These outputs are fused using a modality fusion layer that governs real-time responsiveness and a concurrent or simultaneous operation in text, speech, and vision modes for a multicultural society. For main text, paragraph spacing should be single spaced, no space between paragraphs. There is no indention for any paragraph.

Grenze ID: 01.GIJET.11.2.122_20

II. PROJECT OVERVIEW

A. Image-to-Speech, Image Captioning and Image Text Extraction

In the image captioning domain, CNNs, GCNs, and RNNs are used in the VocalVision system because of their effectiveness in feature extraction, pattern recognition and high dimensionality data mining [1]. In our model, Azure directly interconnects with these neural networks and offers a splendid setting that enhances the efficiency and effectiveness of the networks. Building on Azure AI services such as Vision and Custom Vision, VocalVision synthesizes semantically connected natural language descriptions, relationships between objects. Another addition to the proposed approach is the Multi-Level Attention model to help emphasize spatial aspects of images for a more accurate representation [2]. This approach provides an accurate and rich experience to image understanding and its subsequent captioning.

VocalVision also acts to convert visual, text and any form of data into words and vocalization for ease of communication. OCR extract text from image and will convert it to natural and sounding English or any preferred language with the help of Azure TTS service for the better convenience for visually impaired users. Azure Computer Vision coupled with the "Show, Attend and Tell" model produces rich description from datasets such as COCO and ImageNet addressing the semantic gap between primitive and abstract knowledge [3]. Moreover, Azure Translator Text API provides translation to and from hundreds of languages to support communication and a multimodal fusion layer to support real-time interactions to suit the increasing population of users around the world.

Figure 1. shows the input and Figure 2. shows the output caption generated if the user wants to generate captions describing the image, additionally if the user wishes to extract text for the image the model will use Azure Computer vision to extract text with high accuracy.

Figure 1. Input

```
Generating caption for the image...
Description(s) of the image:
'a man sitting under a tree' with confidence 0.58
Generated caption: 'a man sitting under a tree'
```

Figure 2. Output

B. Text Summarization

The applied strategy combines methods of optimization with the GPT-4 API provided by OpenAI to create brief and contextually precise summaries. It starts with input preprocessing where the text goes through cleaning, where the text is transformed into sentences and finally, tokenized before summarization [4]. The identified features involve the terms' frequency and importance of a sentence, and through coverage and redundancy analyses, the objective is aligned with multiple objectives. Prompts used for abstractive summarization are crafted utilizing an optimization-inspired set of metrics by sending information to OpenAI's GPT-4 API: Iterative summarization is used.

The optimization layer does not create the summary; however, it affects the summary generation by smoothing the API outcomes by such parameters as ROUGE and BLEU [4].

Top-p and temperature are two of the controlling parameters that are set to try and control both the creative and the pinpoint vagaries of the results. There is post-processing to correct grammar, make coherent and unify the text and to remove the information which is not necessary. The methodology was tested on standard datasets including CNN/Daily Mail and DUC2002, results of evaluation metrics supported the gains of summary quality and stability. This approach shows that it is possible to incorporate classical optimization techniques with relatively advanced language models for making text summarization fast and effective. The effectiveness of the proposed methodology was tested on benchmark datasets inclusive of CNN/Daily Mail and DUC2002 and based on computed metrics summary quality and stability were established to have been enhanced [4]. This work shows the possibility of integrating classic optimization techniques with advanced language models to develop fast and effective text summarization.

For example, let's assume someone wants to summarize a news article and the input article has 250+ words, thus, to summarize it using VocalVision we were able to summarize the text within 110 words without missing a single important keyword. Figure 3 shows the summarized output for a news article.

```
Summarised Text:
From the beginning of 2023, Odisha recorded 5.20 lakh dog bite cases, with an average of 23,6
47 incidents monthly. In January and February 2024, there were significant spikes, reporting
33,547 and 32,561 cases, primarily involving stray dogs. According to Fisheries and Animal Re
sources Development Minister Gokulananda Mallick, efforts are underway to control the stray d
og population through sterilisation programs, as mandated by the Animal Birth Control Rules,
2023. In 2022-23, 4,605 dogs were sterilised across eight towns. The 20th animal census in 20
19 showed Odisha had 17.34 lakh stray dogs. Additionally, the Orissa High Court previously ad
dressed serious dog attack cases, demanding swift actions to protect citizens.
```

Figure 3. Output of summarized text

C. Translation and multilingual–to–monolingual

The VocalVision platform, for translating and converting multilingual-to-monolingual especially from Hinglish to English uses Microsoft Azure AI services. Azure possesses heavily counter translator capacities, which is provided with the help of the company's Neural Machine Translation (NMT) and Speech-to-Text technologies, which are mostly used to translate spoken Hinglish into readable text, and then into good English [5]. Especially in case of automatic conversion, the system is not actually susceptible to the phenomenon of Hinglish and related aspects of informal language or code switching used in communication [6]. The localization process makes it possible to use a great number of specific phrases correspondent to a specific context, which makes it possible to use specific models that are suitable program to be used in such spheres as educational and customer relations. Further, it is emphasized that by utilizing Azure TTS (Text to Speech), the translated text can be read in natural English which benefits particularly personal assistant, and bots. Microsoft Azure brings to VocalVision translation that will enable the firm to address more than 70 languages for clients [7,8]. Considering the flexibility of the proposed system's structure and scalability, we anticipate the expansion of the system using more training data for other regional Indian languages such as Marathi, Bengali and Tamil [8]. This inclusiveness is not only of value to VocalVision but also of high utility in real-time data processing and timely response across all linguistic populations of the world. Furthermore, by using the MultiUN dataset which enriches the training phase of the model, the platform enhances the possibility to deal with complex multilingual documents providing translations that are later adapted to specific subject areas.

The posture of the series of steps in implementing the VocalVision is depicted below from the Data Flow Process of the Cross-lingual speech inputs into English speech as presented in Figure. 4. The formal language translator is the system, which in its turn helps the informal language and also the mixed Hinglish-English input through the Microsoft Azure tools. The like it has translation of Hinglish words and phrases like: accessed with the help of its search box – " इस unknown number का call ना उठाये " which, translated in an accurate context means – 'Do not pick up the call of this unknown number.' Every part is scripted to run in live mode, with auto-translation and an Auto Language Detection – there never was a need for a Hinglish filter. Speech recognition Real time text to natural English speech Available modes: speech recognition; text input; rapid conversion of speech from text; real-time conversion of text to natural English from text to speech. As the system's ability is generalizable to industries by using custom terminology, it can be applied to the other regional Indian languages and has more scope. Therefore, this system offers translation through multiple languages at the high velocity with a satisfactory level of precision and has a great effectiveness in the customer relations, learning, and virtual secretaries. In this it utilizes Microsoft Azure as a way to address this problem and relay a far more profound solution to the normal flow of information across languages.

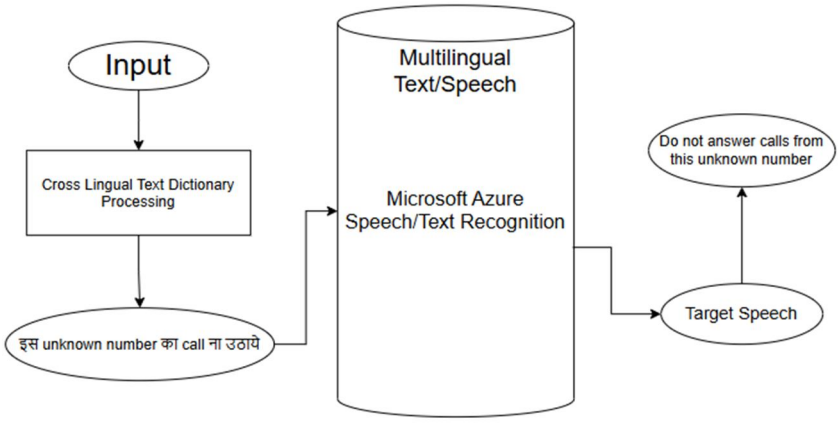

Figure 4. Input Speech / Text extraction and translation.

The flow of generating capacitor in multiple languages in context with an input image is depicted in Figure 5. The system can extract textual descriptions of what is in the image and can translate it into multiple languages as will be evident from the captioning in multiple languages. According to user preferences, the above captions can be easily translated into speech using a TTS feature of the platform. This adaptable design guarantees convenient delivery in the preferred language, either textual or the natural voice output, overall improving the performance.

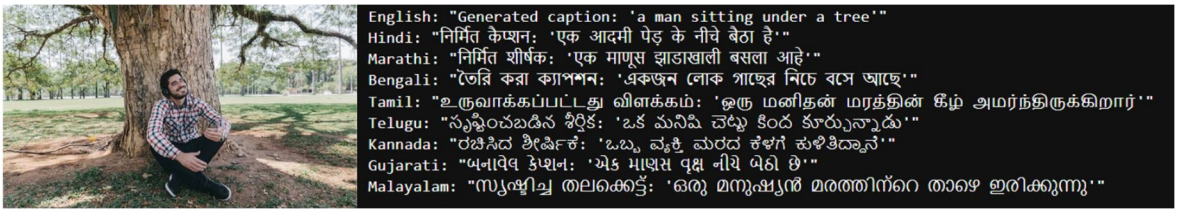

Figure 5. Translated text for the Image caption generated

III. METHODOLOGY

In this work, the VocalVision platform uses a multi-faceted approach to amplify multilingual and multimodal outcomes for personal data intended to solve tasks. User commands can be vocal or voice assisted text or graphics based which subsequently decide on the input form which can be text or image required output in text or speech. For text outputs, there is a natural and clear-spoken Text-to-Speech (TTS) component in the Azure module especially helpful for visually impaired citizens. Automated text summarization, image description and translation are all achieved through the use of various AI services from Azure. For image inputs, it offers text OCR, caption to image, and translating those captions, guaranteeing, culturally sensitive output. The Multimodal Fusion Layer integrates the textual, the audio and the image so that though they are different they are in one form.

Real time processing makes the platform capable of fast, adaptive and instant decision making in response to dynamic conditions prevailing in the application environment. Availability for more than 70 languages with the options for Marathi, Bengali and Tamil and also can add further training, using the MultiUN dataset. This approach is effective in education, customer relations and any form of dealing. It also supports the use of custom models in industries for the right and accurate results to be achieved. Envisioned for broad application, high versatility, and flexibility, VocalVision is a needed requirement for supporting an assortment of diverse linguistic user groups, allowing for the conversion of text, speech, and sign input information effectively and quickly. The platform-generated data flow diagram in Figure 6. Illustrates a user interacting with the platform has the vocal/written/visual material content support simplifying the interaction with easy and quicker maneuvering.

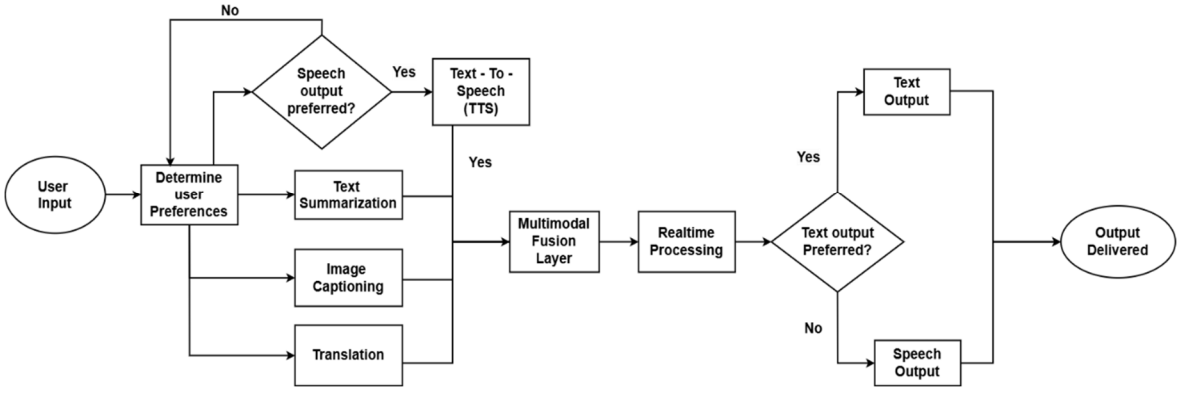

Figure 6. Overall Flowchart of the model

IV. RESULTS

The findings of the VocalVision platform show how multilingual and multimodal communication is possible. The evaluation of text summarization by using standard datasets such as CNN and DM, image captioning datasets like COCO, and translation tasks achieved high accuracy, precision and recall and f1 score. It proved that the performance of the given system from the solution of the proposed problem set was of high quality by

evaluating the performance using commonly known performance measures such as ROUGE and BLEU. The comparative analysis showed better results than existing solutions and reference models for VocalVision. Features like the case studies where a full outline of a news article over the 250 word length would distilled down to 110 words while retain central information proved the capacity of the platform as in Figure 3. Users of the tool also expressed the positive evaluations such as effortless and natural. Concerning error analysis, some difficulties with accepting complex text inputs were observed on occasional bases while the performance of all other aspects was commendably high. The findings were further supported by data flow diagrams. Cohort analysis further affirmed the importance of the effects, corroborating the platform's usability in learning, relations with customers or partners, and affairs.

V. Conclusions

The VocalVision platform is the next leap in multilingual and multimodal communication to handle text, voice, and image data in one convenient space through Microsoft Azure's AI Services. The client-oriented conception adapts it for real-time usage and individual performance, it will be highly effective in various spheres including education to help people dive into the material in an easier way; customer service, providing fast and accurate multilingual support and business interactions, allowing for smoother communication between people from different countries. With availability in over 70 languages and able to accommodate specific regions, VocalVision improves global language barriers. Future releases will introduce support for more languages, improve translation and voice quality, and add new features such as higher quality image recognition capabilities and methods to identify the tone of the text. With this aggressive adaptability, VocalVision aims to develop the appliance of communication technology as people's needs change over time across the world.

References

[1] Y. Kesavan, V. Muley, and M. Kolhekar, "Automated image caption generation using deep neural networks," in 2019 Global Conference for Advancement in Technology (GCAT), Bangalore, India, Oct 18-20, 2019, pp. 1-6. doi: 978-1-7281-3694-3/19/$31.00 [IEEE].

[2] J. Effendi, S. Sakriani, and S. Nakamura, "End-to-End Image-to-Speech Generation for Untranscribed Unknown Languages," IEEE , vol. 9, pp. 55144–55156, 20, doi: 10.1109/ACCESS.20.3071541

[3] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to Text: Survey of Image and Video Captioning," IEEE Trans. Emerg. Topics Comput. Intell., vol. 1, pp. 1–8, January 2019. DOI: 10.1109/TETCI.2019.2892755 [Supported by National Science Foundation Award 1651902 and U.S. Army Research Office Award W911NF-17-1-0367].

[4] M. H. H. Wahab, N. Hafiza Ali, N. A. W. Abdul Hamid, S. K. Subramaniam, R. Latip, and M. Othman, "A Review on Optimization-Based Automatic Text Summarization Approach," IEEE Access, vol. 12, pp. 4892-4905, Jan. 2024, doi: 10.1109/ACCESS.2023.3348075.

[5] Neural Machine Translation Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In 3rd International Conference on Learning Representations (ICLR).

[6] [6] Hinglish Language Analysis Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). "I am borrowing ya mixing?": An Analysis of English-Hindi Code Mixing in Facebook. In Proceedings of the First Workshop on Computational Approaches to Code Switching, 116–126.

[7] Microsoft Azure Documentation Microsoft. (2023). Azure Cognitive Services – Speech-to-Text and Text-to-Speech.

[8] Zhang X, Peng Y, Xu X. An Overview of Speech Recognition Technology. In: 2019 4th International Conference on Control, Robotics and Cybernetics (CRC). IEEE. 2019; p. 81–85. Available from: https://doi.org/10.1109/CRC.2019.00025.

[9] S. Liu, L. Bai, Y. Hu, and H. Wang, "Image captioning based on deep neural networks," MATEC Web of Conferences, vol. 232, p. 01052, 2018 [EITCE 2018].

[10] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning based on object relations using geometric attention," Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, pp. 1-10, December 2019 [arXiv:1905.12345].

[11] Z. Yuan, X. Li, and Q. Wang, "Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning," IEEE Access, vol. 8, pp. 2608–2610, January 2020. DOI: 10.1109/ACCESS.2019.29695 [Supported by National Key Research and Development Program of China under Grant 2017YFB1002202].

[12] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "Visual-Relationship Attention for Image Captioning," in Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14-19 July 2019, pp. 1-8. DOI: 10.1109/IJCNN.2019.8852044 [Supported by University of Technology Sydney].

[13] Rudrappa NT, Reddy MV, Hitek. HiTEK: Multilingual Speech Identification Using Combinatorial Model. In: International Conference On "Advances in Computer Vision and Artificial Intelligence Technologies ACVAIT. 2022.