Artificial intelligence multilingual image-to-speech for accessibility and text recognition

Rosalina[1], **Hasanul Fahmi**[2], **Genta Sahuri**[3]
[1]Informatics Study Program, Faculty of Computer Science, President University, Bekasi, Indonesia
[2]School of Information Technology, UNITAR International University, Petaling Jaya, Malaysia
[3]Information Systems Study Program, Faculty of Computer Science, President University, Bekasi, Indonesia

Keywords:

Image-to-speech
Multilingual audio descriptions
Natural language processing
Optical character recognition
Text-to-speech

ABSTRACT

The primary challenge for visually impaired and illiterate individuals is accessing and understanding visual content, which hinders their ability to navigate environments and engage with text-based information. This research addresses this problem by implementing an artificial intelligence (AI)-powered multilingual image-to-speech technology that converts text from images into audio descriptions. The system combines optical character recognition (OCR) and text-to-speech (TTS) synthesis, using natural language processing (NLP) and digital signal processing (DSP) to generate spoken outputs in various languages. Tested for accuracy, the system demonstrated high precision, recall, and an average accuracy rate of 0.976, proving its effectiveness in real-world applications. This technology enhances accessibility, significantly improving the quality of life for visually impaired individuals and offering scalable solutions for illiterate populations. The results also provide insights for refining OCR accuracy and expanding multilingual support.

Corresponding Author:

Hasanul Fahmi
School of Information Technology, UNITAR International University
Kelana Jaya, 47301-Petaling Jaya, Selangor, Malaysia
Email: fahmi.zuhri@unitar.my

1. INTRODUCTION

In an increasingly digital world, access to visual content is essential for daily communication, education, and navigation [1], [2]. However, visually impaired and illiterate individuals face significant challenges in interpreting such content, limiting their ability to fully engage with their surroundings and access critical information [3]–[5]. Assistive technologies have made progress in enhancing accessibility, but gaps still exist in providing accurate and real-time solutions that effectively convert visual information into a format accessible to these individuals. Recent advancements in artificial intelligence (AI) [6]–[10], offer promising solutions by combining optical character recognition (OCR), text-to-speech (TTS), and natural language processing (NLP) to transform images into spoken descriptions. This research focuses on implementing and evaluating a multilingual image-to-speech system that amplifies AI to address these accessibility challenges.

A key problem for the visually impaired is the inability to access printed or digital text in images [11]–[13], which is exacerbated when navigating diverse environments or consuming visual content in multiple languages [14], [15]. While existing OCR and TTS technologies provide basic text-to-audio conversion [16], [17], their accuracy often declines in real-world scenarios where text may be distorted, partially visible, or displayed in multiple languages [18]. Furthermore, many current solutions are monolingual, limiting their utility for users in multilingual environments [19]–[21]. Illiterate individuals also face similar barriers in accessing text-based information [22]–[27]. Thus, there is a clear need for a system that can accurately recognize and describe text from images across different languages in real time. Several studies have explored OCR and TTS technologies for accessibility, but few have integrated these technologies into a robust, AI-powered solution capable of providing accurate multilingual descriptions. Previous work by [28] demonstrated that advanced OCR models could improve text recognition rates, especially in noisy or complex image environments. However, combining these technologies with AI-driven NLP and digital signal processing (DSP) techniques remains underexplored. AI advancements now enable more sophisticated and context-aware systems that can enhance both the accuracy and the scope of accessibility solutions for the visually impaired.

The proposed solution integrates state-of-the-art OCR and TTS systems with AI-powered NLP and DSP techniques to create a multilingual image-to-speech technology. By leveraging these technologies, the system is designed to accurately recognize text in multiple languages, even under challenging image conditions, and provide real-time audio descriptions for users. The system not only addresses the challenge of visual text accessibility for visually impaired users but also broadens its scope to support illiterate individuals, enabling them to "hear" text content that they cannot read. This multilingual capability distinguishes it from other existing solutions. The innovative value of this research lies in its approach to combining multiple AI technologies into a cohesive, user-friendly system that enhances real-time accessibility. The system achieves high precision and recall, as demonstrated by an average accuracy rate of 0.976 in tests, making it a reliable tool for real-world applications. Additionally, its multilingual functionality extends the technology's usefulness to diverse populations. The findings contribute to the body of knowledge in assistive technologies by providing a framework for further improvements in OCR accuracy and TTS integration, offering a scalable solution to accessibility challenges faced by visually impaired and illiterate individuals.

2. METHOD

This research aims to develop, implement, and evaluate a multilingual image-to-speech system specifically designed to support visually impaired and illiterate individuals in accessing written information. The system integrates OCR technology to accurately extract text from images, ensuring high-quality text recognition across multiple languages. Once the text is extracted, a TTS synthesis module converts it into natural-sounding speech, allowing users to listen to the content in their preferred language. By combining OCR and TTS, the system enhances accessibility, enabling individuals with visual or reading impairments to interact with textual information more independently. The overall architecture, which outlines the key components and their interactions, is depicted in Figure 1, providing a comprehensive overview of the system's functionality.

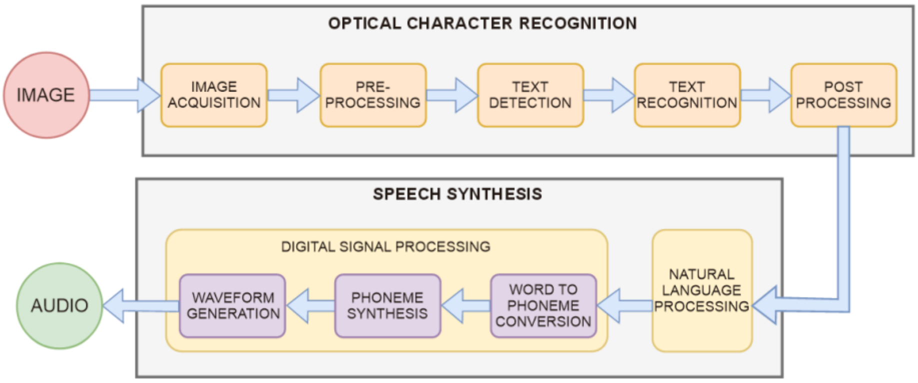

Figure 1. Multilingual image-to-speech system architecture

2.1. Data collection

To develop an AI-powered multilingual image-to-speech technology aimed at enhancing accessibility for visually impaired users, a comprehensive data collection process is essential. The first step involves gathering a diverse image dataset containing text in various languages, including signs, printed documents, and labels. The goal is to ensure that the dataset reflects different fonts, sizes, orientations, and backgrounds to enhance the robustness of the model. For this purpose, we can utilize the international conference on document analysis and recognition (ICDAR) datasets, which provide benchmark images of printed text in multiple languages and formats. These datasets are widely recognized in the field for their comprehensive range of text types and conditions. Additionally, crowdsourcing can be employed to allow users to upload images containing text, further enriching the dataset with real-world examples. This approach not only broadens the dataset but also captures varied conditions under which text appears, enhancing the model's applicability.

The second step involves collecting a high-quality audio dataset that consists of recordings corresponding to the text extracted from images, ensuring accurate pronunciation and intonation for each language. For this purpose, Mozilla's Common Voice project serves as an excellent source. It is an open-source initiative that collects diverse voice samples in multiple languages, allowing contributors to record themselves reading sentences. This dataset provides the variability and richness needed to create effective TTS capabilities. In addition, contracting native speakers or voice actors can enhance the dataset's quality by ensuring accurate pronunciation and emotional expression. The annotation process is crucial for preparing the image dataset for training. It involves manually annotating images with bounding boxes around text areas and providing corresponding text transcriptions in multiple languages. Annotation tools such as LabelImg or RectLabel is used to create bounding boxes around the text, ensuring that various orientations and layouts are captured.

2.2. Optical character recognition

The system employs the Tesseract OCR engine with Python's pytesseract library for extracting text from images. The workflow begins with image acquisition from various sources, including scans and photos. Pre-processing steps such as resizing, noise reduction, and contrast enhancement are used to improve text visibility. Text regions are identified using techniques like edge detection and machine learning, and the detected text is converted into machine-readable format. Post-processing further refines this text, correcting errors and improving formatting.

In the pre-processing phase, we standardize image dimensions to ensure consistent processing across various images. This involves resizing all images to a fixed width and height, maintaining aspect ratio when necessary. The resizing formula as in (1) adjusts the image size while preserving proportions. This standardization simplifies subsequent image analysis by ensuring uniform input dimensions, which helps in maintaining consistency in feature extraction and processing, ultimately improving the accuracy and efficiency of the OCR system.

$$\begin{aligned}&New\ Width = Desired\ Width\\ &New\ Height = Original\ Height\ x\ \left(\frac{Desired\ Width}{Original\ Width}\right)\end{aligned} \tag{1}$$

In addition, we applied filters to remove unwanted artifacts and noise from the image. This involves using Gaussian blur, which smooths the image by averaging the intensities of surrounding pixels. The Gaussian function, given as in (2), this filter reduces noise and enhances image quality by minimizing high-frequency variations, ensuring that the OCR system processes cleaner and more accurate data.

$$I'(x,y) = \frac{1}{2\pi\sigma^2}\iint_{-\infty}^{\infty} I(x',y')exp\ (-\frac{(x-x')^2+(y-y')^2}{2\sigma^2})dx'dy' \tag{2}$$

Contrast enhancement was applied to improve text visibility through contrast stretching. This technique adjusts pixel intensities using the formula as in (3). In this formula, I_{in} represents the original pixel intensity, I_{min} and I_{max} are the minimum and maximum intensities in the image, and L is the number of intensity levels. This adjustment enhances contrast, making text and details more distinct.

$$I_{out} = \frac{I_{in}-I_{min}}{I_{max}-I_{min}} x(L-1) \tag{3}$$

We then identified regions in the image that likely contain text using edge detection. This technique calculates edge strength to locate text boundaries. The edge strength is determined by the formula as in (4). Here, $\frac{\partial I}{\partial x}$ represent the gradients of pixel intensity I in the horizontal and vertical directions, respectively. This calculation highlights areas with high changes in intensity, indicating potential text regions.

$$Edge\ Strength = \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2} \tag{4}$$

After identifying potential text regions, we performed text recognition to convert these regions into machine-readable text using the OCR engine. The Tesseract OCR engine employs pattern recognition algorithms to analyze the detected text areas, matching them to known character patterns. This process involves comparing image segments with a database of character templates to accurately identify and transcribe the text.

2.3. Speech synthesis

The recognized text is converted into speech using the Python Google text-to-speech (gTTS) library. This process begins with NLP, which includes tokenization to break down the text into smaller units, such as words or sentences, and language detection to apply accurate pronunciation rules and phonetic adjustments. In the subsequent phase, DSP techniques are employed. The first step is word-to-phoneme conversion, where text is translated into its phonetic representation. For example, "hello" is converted to phonemes /h/, /ə/, /l/, /oʊ/. This is expressed as in (5).

$$Phoneme = word \rightarrow Phonetic\ Representation \tag{5}$$

Phoneme synthesis converts phonetic representations into speech sounds, where sound waves are generated to represent each phoneme. We implemented WaveNet models to enhance this process. WaveNet uses deep neural networks to produce more natural and human-like sound waves by accurately capturing the complexities of human speech, this is expressed as in (6).

$$Phoneme\ Sound = Phoneme \rightarrow WaveNet\ Model- > Natural\ Sound\ Wave \tag{6}$$

Finally, waveform generation produces a high-quality audio waveform from the synthesized phonemes. This step involves converting the synthesized phoneme sounds into an audio signal that can be played back. The waveform is generated using advanced techniques to ensure clarity and naturalness in the speech output. The waveform is represented as in (7).

$$Waveform = Phoneme\ Sounds- > Audio\ Signal \tag{7}$$

2.4. Evaluation parameters for text-to-speech conversion metrics

In developing an effective TTS system, it is essential to establish clear evaluation parameters that measure the quality and intelligibility of synthesized speech. These parameters help determine how well the TTS system can produce audio that is both clear and natural-sounding, particularly for applications aimed at assisting visually impaired users. The following metrics provide a comprehensive framework for assessing the performance of TTS systems:

- Phoneme synthesis quality: this parameter measures the accuracy and precision with which the TTS system generates phonemes the distinct units of sound in speech. A higher value (on a scale of 0 to 1) indicates better performance, suggesting that the system effectively captures the nuances of different languages and dialects. It is typically assessed through subjective listening tests and objective evaluations, such as comparing synthesized phonemes against a reference set.
- Waveform clarity: this qualitative parameter evaluates the overall audio quality of the synthesized speech. A "high" rating signifies that the audio output has minimal distortion, noise, and artifacts, leading to clear and intelligible speech. Waveform clarity is essential for ensuring that listeners can easily understand the generated audio, which is particularly important for applications aimed at aiding visually impaired users.
- Speech naturalness: this parameter assesses the degree to which synthesized speech resembles natural human speech. It evaluates factors such as intonation, rhythm, and expressiveness. An "improved" rating indicates that the TTS system has enhanced it is ability to produce engaging and relatable audio output. This is often determined through user feedback and expert evaluations, as well as comparison with natural speech samples.

2.5. Evaluation metrics for optical character recognition performance and text recognition accuracy

In assessing the effectiveness of OCR systems and text recognition across different languages, several key evaluation metrics are employed. These metrics offer insights into the accuracy and reliability of the systems, ensuring that they meet the needs of various applications.

- Precision is a critical metric that quantifies the accuracy of the OCR system in identifying relevant text. It is defined as the ratio of true positive predictions (correctly identified text) to the total predicted positives (both correct and incorrect identifications). A high precision score indicates that when the system predicts text, it is likely to be accurate, minimizing false positives.
- Recall, also known as sensitivity or true positive rate, measures the OCR system's ability to identify all relevant instances of text in a given dataset. It is calculated as the ratio of true positive predictions to the total actual positives (all instances of text present in the images). A high recall score indicates that the system successfully captures most of the actual text, thereby reducing the likelihood of missing characters or words.

– F1 score serves as a balanced measure that combines both precision and recall into a single metric. It is calculated as the harmonic mean of precision and recall, providing a comprehensive view of the system's overall performance. A high F1 score signifies that the OCR system performs well in both accurately identifying relevant text and capturing as much text as possible from the images.

3. RESULTS AND DISCUSSION

The implementation of the multilingual image-to-speech system, which integrates OCR and TTS synthesis, has produced promising results. The OCR component of the system exhibited high performance in text extraction, achieving an average precision rate of 0.976. This high precision indicates the system's robustness in accurately identifying and extracting text from various types of images, including scanned documents, digital photos, screenshots, and handwritten notes.

The precision metric reflects the proportion of correctly identified text regions out of the total identified text regions. A precision rate of 0.976 signifies that the OCR system is highly effective at minimizing false positives, where non-text areas are incorrectly recognized as text. This high accuracy is crucial for applications where the correct interpretation of text is essential, such as converting printed or handwritten documents into machine-readable formats. The successful implementation of this OCR component underscores the effectiveness of the chosen algorithms and techniques in processing diverse image inputs. By accurately extracting text, the system lays a solid foundation for the subsequent TTS synthesis phase, which relies on the quality of the extracted text to generate clear and coherent spoken output. This integrated approach enhances the accessibility and usability of the image-to-speech system for visually impaired and illiterate individuals across multiple languages.

Table 1 provides a detailed evaluation of OCR accuracy across different image types, showcasing the system's performance in various scenarios. The table includes metrics such as precision, recall, and F1 score, which are critical for assessing the effectiveness of text extraction. For scanned documents, the OCR system achieves the highest accuracy with a precision of 0.98, recall of 0.97, and an F1 score of 0.975. This indicates that the system reliably extracts text from scanned documents with minimal errors and high completeness. Digital photos follow with a precision of 0.95 and a recall of 0.94, resulting in an F1 score of 0.945. This reflects strong performance, though slightly less accurate than scanned documents due to potential image quality variations. Screenshots show a precision of 0.96 and recall of 0.95, with an F1 score of 0.955. The system performs well in extracting text from screenshots, demonstrating its versatility across different image types. Meanwhile, handwritten notes exhibit the lowest accuracy, with a precision of 0.90, recall of 0.88, and an F1 score of 0.890, while still effective, the system faces more challenges with handwritten text due to its inherent variability and complexity.

Table 1. OCR accuracy across different image types

Image type	Precision	Recall	F1 score
Scanned documents	0.98	0.97	0.975
Digital photos	0.95	0.94	0.945
Screenshots	0.96	0.95	0.955
Handwritten notes	0.90	0.88	0.890

Table 2 provides a comparative analysis of text recognition accuracy across various languages using the OCR system. The metrics presented include precision, recall, and F1 score, which reflect the system's performance in recognizing text from different linguistic contexts. For English, the OCR system demonstrates exceptional accuracy with a precision of 0.97 and a recall of 0.96, resulting in a high F1 score of 0.965. This indicates that the system reliably identifies and extracts English text with minimal errors. Spanish shows strong performance with a precision of 0.95 and a recall of 0.94, leading to an F1 score of 0.945. This reflects the system's capability to handle Spanish text effectively. Mandarin has lower accuracy compared to European languages, with a precision of 0.92, a recall of 0.91, and an F1 score of 0.915. This is attributed to the complexity of Mandarin characters and script. French and Italian also show high performance, with F1 scores of 0.935 for both languages, indicating robust recognition capabilities. Indonesian has a precision of 0.93 and a recall of 0.92, resulting in an F1 score of 0.925, demonstrating effective text recognition. German achieves a precision of 0.96 and a recall of 0.95, with an F1 score of 0.955, highlighting its high accuracy in text recognition. Japanese and Korean have lower scores, with F1 scores of 0.895 and 0.905, respectively, reflecting challenges in recognizing these scripts. Arabic shows the lowest accuracy with a precision of 0.88, a recall of 0.87, and an F1 score of 0.875, due to the intricacies of the Arabic script.

Table 3 presents the performance metrics of the TTS conversion system used in this research. The phoneme synthesis quality is measured at 0.95, indicating a high level of accuracy in generating phoneme sounds from text, which is crucial for producing intelligible speech. Waveform clarity is rated as high, reflecting the effectiveness of the waveform generation process in creating clear and distortion-free audio signals. Lastly, speech naturalness is noted as improved, highlighting that the synthesized speech sounds more natural and human-like, thanks to the implementation of advanced techniques such as WaveNet models. Together, these metrics demonstrate that the TTS system delivers high-quality and realistic speech output, enhancing overall user experience and accessibility.

Meanwhile, Table 4 presents execution times for components of an AI-powered multilingual image-to-speech system, measured in milliseconds (ms). Image preprocessing takes 50 ms, efficiently preparing images for OCR. The OCR phase requires 120 ms to extract text, reflecting the complexity of recognizing various text formats. Following this, text cleaning occurs in 30 ms, where the extracted text is organized for speech conversion. The TTS conversion is the most time-intensive component, taking 200 ms to synthesize natural-sounding speech from the cleaned text, highlighting the computational demands of multilingual synthesis. Lastly, audio playback requires just 15 ms, demonstrating the system's efficiency in delivering audio outputs. The total execution time of 415 ms indicates the cumulative duration from image input to audio output, suggesting the system's responsiveness for real-time applications aimed at assisting visually impaired users.

Table 2. Text recognition accuracy with different languages

Language	Precision	Recall	F1 score
English	0.97	0.96	0.965
Spanish	0.95	0.94	0.945
Mandarin	0.92	0.91	0.915
French	0.94	0.93	0.935
Indonesian	0.93	0.92	0.925
German	0.96	0.95	0.955
Italian	0.94	0.93	0.935
Japanese	0.90	0.89	0.895
Korean	0.91	0.90	0.905
Arabic	0.88	0.87	0.875

Table 3. Performance metrics of TTS conversion

Metric	Value
Phoneme synthesis quality	0.95
Waveform clarity	High
Speech naturalness	Improved

Table 4. Execution times of different components in the ai-powered multilingual image-to-speech system

Component	Execution time (ms)	Notes
Image preprocessing	50	Time taken to load and preprocess the image
OCR	120	Time taken for the OCR model to extract text
Text cleaning	30	Time taken for cleaning and formatting the extracted text
TTS conversion	200	Time taken to convert the cleaned text to speech
Audio playback	15	Time taken to play the generated audio
Total execution time	415	Sum of all execution times for the complete process

The high OCR accuracy achieved by the system underscores its effectiveness in accurately extracting text from images. The rigorous pre-processing techniques, such as contrast stretching and noise reduction, played a crucial role in enhancing text visibility and thus improving the performance of the OCR. Contrast stretching adjusted pixel intensity levels to make text stand out more distinctly against its background, while noise reduction minimized artifacts that could hinder text recognition. Together, these techniques facilitated more accurate text extraction by the OCR engine. Further enhancement in the system's performance is attributed to the use of advanced models like WaveNet for phoneme synthesis. WaveNet, a deep generative model for creating raw audio waveforms, significantly improved the naturalness and quality of the synthesized speech. Unlike traditional speech synthesis methods, WaveNet models generate more natural and human-like speech by modeling the audio waveform at a finer level of detail. This advancement is reflected in the improved phoneme synthesis quality, where the generated speech closely resembles natural human speech in terms of fluidity and expressiveness.

The system's ability to perform well across different languages and text types demonstrates it is robustness and reliability in TTS conversion. It effectively handles multiple languages, including those with complex scripts and phonetic structures, providing accurate and clear spoken output. The integration of WaveNet for waveform generation further enhances the realism of the synthesized speech, making it more engaging and easier to understand for users. Overall, the combination of high-accuracy OCR and advanced TTS technologies offers a powerful solution for enhancing accessibility for visually impaired and illiterate individuals. By bridging the gap between visual and auditory information, the system makes text-based content more accessible through spoken output. This integration represents a significant advancement in assistive technology, enabling users to access information that was previously less accessible due to visual impairments.

Future work could focus on expanding multilingual support to include a broader range of languages and dialects, as well as improving real-time TTS conversion capabilities. Enhancements in these areas would further increase the system's versatility and applicability in various contexts, making it a more inclusive tool for users with diverse needs. Additionally, improving real-time TTS conversion capabilities is crucial for enhancing the system's usability. A real-time conversion would allow users to receive spoken output instantly as text is recognized, making the system more responsive and practical for dynamic environments. This enhancement would be particularly beneficial in applications such as live events or real-time document reading, where immediate feedback is essential. By addressing these areas, the system's versatility would be significantly increased. Users with varying linguistic and accessibility needs would benefit from a more adaptable and efficient tool. This progress would ensure that the system remains relevant and useful in diverse contexts, making it a more inclusive solution for individuals with visual impairments or literacy challenges worldwide.

4. CONCLUSION

The implementation of the multilingual image-to-speech system effectively aligns with the expectations outlined in the introduction. The integration of OCR and TTS synthesis demonstrated high accuracy in text extraction and improved speech generation quality, validating the system's capability to bridge visual and auditory information for enhanced accessibility. The successful use of advanced models like WaveNet for phoneme synthesis and waveform generation has resulted in natural and high-quality speech output, meeting the initial goal of providing a reliable and inclusive tool for visually impaired and illiterate individuals. The results and discussion highlight the system's strong performance across various languages and text types, affirming it is versatility and practical utility. Looking ahead, future research could focus on expanding multilingual support to include a broader range of languages and dialects, thereby increasing the system's global applicability. Additionally, enhancing real-time TTS capabilities could further improve the system's responsiveness and user experience. The prospect of further development includes refining these features to make the system more adaptable and useful in diverse contexts. By addressing these areas, future studies can build on the current research to enhance accessibility technologies, ensuring they meet the evolving needs of users worldwide.

ACKNOWLEDGEMENTS

The author thanks UNITAR International University and President University for their support and resources throughout this research. Special appreciation is extended to the faculty and staff for their guidance and encouragement. Their financial and institutional backing was essential to the success of this study.

FUNDING INFORMATION

This research was supported by funding from UNITAR and President University.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hasanul Fahmi	✓	✓		✓	✓	✓			✓	✓			✓	✓
Rosalina	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Genta Sahuri	✓		✓	✓		✓			✓		✓			

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has complied with all relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] Y. K. Dwivedi, *et al.*, “Setting the future of digital and social media marketing research: perspectives and research propositions,” *International Journal of Information Management*, vol. 59, no. 1, pp. 1–37, 2021, doi: 10.1016/j.ijinfomgt.2020.102168.

[2] B. Kuriakose, R. Shrestha, and F. E. Sandnes, “Tools and technologies for blind and visually impaired navigation support: a review,” *IETE Technical Review*, vol. 39, no. 1, pp. 1-16, Sep. 2020, doi: 10.1080/02564602.2020.1819893.

[3] S. Klauke, C. Sondocie, and I. Fine, “The impact of low vision on social function: the potential importance of lost visual social cues,” *Journal of Optometry*, vol. 16, no. 1, May 2022, doi: 10.1016/j.optom.2022.03.003.

[4] M. Fayyad and A. R. Al-Sinnawi, “Challenges of achieving financial inclusion for individuals with visual impairments,” *Heliyo*n, vol. 10, no. 16, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35573.

[5] F. Fuentes, A. Moreno, and F. Díez, “The usability of icts in people with visual disabilities: a challenge in spain,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 17, Aug. 2022, doi: 10.3390/ijerph191710782.

[6] J. Wang, S. Wang, and Y. Zhang, “Artificial intelligence for visually impaired,” *Displays*, vol. 77, Apr. 2023, doi: 10.1016/j.displa.2023.102391.

[7] R. C. Joshi, N. Singh, A. K. Sharma, R. Burget and M. K. Dutta, “AI-sensevision: a low-cost artificial-intelligence-based robust and real-time assistance for visually impaired people,” *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 3, pp. 325-336, Jun. 2024, doi: 10.1109/THMS.2024.3375655.

[8] R. Shendge, A. Patil, and S. Kadu, “Smart navigation for visually impaired people using artificial intelligence,” *ITM Web of Conferences*, vol. 44, 2022, doi: 10.1051/itmconf/20224403053.

[9] S. Selvan, J. Stella, K. B and N. V. G. S. Nikitha, “Smart shopping trolley based on iot and ai for the visually impaired,” *in International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS*), Coimbatore, India, 2024, pp. 132-138, doi: 10.1109/ICC-ROBINS60238.2024.10533927.

[10] A. Kuzdeuov, O. Mukayev, S. Nurgaliyev, A. Kunbolsyn and H. A. Varol, “ChatGPT for visually impaired and blind,” *in International Conference on Artificial Intelligence in Information and Communication (ICAIIC),* Osaka, Japan, 2024, pp. 722-727, doi: 10.1109/ICAIIC60209.2024.10463430.

[11] C. Cherotich, K. P. Cheptoo, and R. M. Obare, “Challenges in accessing digital resources among visually impaired (VI) students at the university of nairobi library,” *Information Development*, Jun. 2024, doi: 10.1177/02666669241259083.

[12] Y. Abdelaal and D. Al-Thani, “Accessibility first: detecting frustration in web browsing for visually impaired and sighted smartphone users,” *Universal Access in the Information Society*, Oct. 2023, doi: 10.1007/s10209-023-01053-3.

[13] A. Baumgartner, T. Rohrbach, and P. Schönhagen, “If the phone were broken, I’d be screwed’: media use of people with disabilities in the digital era,” *Disability & Society*, pp. 1–25, May 2021, doi: 10.1080/09687599.2021.1916884.

[14] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal and B. Schiele, “Translating video content to natural language descriptions,” *in IEEE International Conference on Computer Vision,* Sydney, NSW, Australia, Mar. 2013, pp. 433-440, doi: 10.1109/ICCV.2013.61.

[15] D. Jindal, C. Kaur, A. Panigrahi, B. Soni, A. Sharma and S. Singla, “Multilingual cross-modal image synthesis with text-guided generative ai,” *in Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT),* Sonepat, India, 2024, pp. 576-582, doi: 10.1109/CCICT62777.2024.00096.

[16] S. K. Singla and R. K. Yadav, “Optical character recognition based speech synthesis system using labview,” *Journal of Applied Research and Technology*, vol. 12, no. 5, pp. 919-926, Oct. 2014, doi: 10.1016/s1665-6423(14)70598-x.

[17] S. Faizullah, M. S. Ayub, S. Hussain, and M. A. Khan, “A survey of ocr in arabic language: applications, techniques, and challenges,” *Applied sciences*, vol. 13, no. 7, Apr. 2023, doi: 10.3390/app13074584.

[18] M. E. Matre and D. L. Cameron, “A scoping review on the use of speech-to-text technology for adolescents with learning difficulties in secondary education,” *Disability and Rehabilitation: Assistive Technology*, pp. 1–14, Nov. 2022, doi: 10.1080/17483107.2022.2149865.

[19] L. Orynbay, B. Razakhova, P. Peer, B. Meden, and Ž. Emeršič, “Recent advances in synthesis and interaction of speech, text, and vision,” *Electronics,* vol. 13, no. 9, Apr. 2024, doi: 10.3390/electronics13091726.

[20] Sa. Kasmaiee and M. Tadjfar, “Elliptical pressure swirl jet issuing into stagnant air,” *Physics of Fluids*, vol. 36, no. 7, Jul. 2024, doi: 10.1063/5.0198105.

[21] Z. Cai, Y. Yang, and M. Li, “Cross-lingual multi-speaker speech synthesis with limited bilingual training data,” *Computer Speech & Language*, vol. 77, Jan. 2023, doi: 10.1016/j.csl.2022.101427.

[22] X. Peng, H. Cao, S. Setlur, V. Govindaraju, and P. Natarajan, “Multilingual OCR research and applications,” *Proceedings of the 4th International Workshop on Multilingual,* 2013, doi: 10.1145/2505377.2509977.

[23] D. Purmayanti, “The challenges of implementing digital literacy in teaching and learning activities for efl learners in Indonesia,” *BATARA DIDI: English Language Journal*, vol. 1, no. 2, pp. 101-110, Oct. 2022, doi: 10.56209/badi.v1i2.38.

[24] P. Williams, “Exploring the challenges of developing digital literacy in the context of special educational needs communities,” *Innovation in Teaching and Learning in Information and Computer Sciences*, vol. 5, no. 1, pp. 1-16, Jan. 2006, doi: 10.11120/ital.2006.05010006.

[25] P. Reddy, K. Chaudhary, and S. Hussein, “A digital literacy model to narrow the digital literacy skills gap,” *Heliyon*, vol. 9, no. 4, Apr. 2023, doi: 10.1016/j.heliyon.2023.e14878.

[26] S. Kasmaiee, M. Tadjfar, S. Kasmaiee, and G. Ahmadi, “Linear stability analysis of surface waves of liquid jet injected in transverse gas flow with different angles,” *Theoretical and Computational Fluid Dynamics*, vol. 38, pp. 107–138, Feb. 2024, doi: 10.1007/s00162-024-00685-2.

[27] Sa. Kasmaiee and M. Tadjfar, “Non-circular pressure swirl nozzles injecting into stagnant air,” *International Journal of Multiphase Flow*, vol. 175, May 2024, doi: 10.1016/j.ijmultiphaseflow.2024.104798.

[28] J. Gao, A. Zongwen, and B. Xuezong, “A new representation method for probability distributions of multimodal and irregular data based on uniform mixture model,” *Annals of Operations Research,* Apr. 2019, doi: 10.1007/s10479-019-03236-9.