Towards Inclusive Digital Accessibility Using Multimodal AI and Braille-Aware Language Models

Dr. Ajantha Devi Vairamani
Post Doctoral Researcher, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia

Prof. (Dr.) Anand Nayyar
School of Computer Science and Artificial Intelligence (SCA)
Duy Tan University, Da Nang, Viet Nam

Abstract

Digital access is a vital issue for visually impaired people today, especially in a digital world that is based on multimodal communication (text, voice, visual, infographics, statistics, equations, and formatted documents). Existing devices including screen readers, text systems, Optical Character Recognition, and rule-based Braille translation have enhanced ease of reading from text, but these are few when it comes to complex visual and context-rich content. For many high-level users of technology, this has resulted in a complete lack of tools to help them understand text with visual and contextual information. Those constraints are especially severe in science, technology, engineering and mathematics school education because learning resources in these subjects often rely on visual and spatial representations. Recent advancements in artificial intelligence, natural language processing, computer vision, large language models and multimodal learning offer novel possibilities for intelligent accessibility systems with semantic understanding, multimodal interpretation and personalized content generation.

This paper examines and reviews the available literature on accessibility in the digital realm and assistive digital technologies focusing on AI-based accessibility, multimodal AI, and adaptive Braille generation. The review concludes that: Limited contextual awareness; Poor multimodal processing; Poor personalization; Low adaptive Braille generation support. To close this gap, this paper urges the design of a Multimodal Accessibility Engine , an AI-based methodology to bridge the gap of non-intermezzo-language Braille awareness by harnessing multimodes, including multimodal content comprehension, semantic reasoning, and user-adapted Braille generation for inclusive digital participation by visually impaired users.

Keywords. Digital Accessibility; Assistive Technologies; Artificial Intelligence; Multimodal AI; Braille Generation; Large Language Models; Natural Language Processing; Inclusive Computing; STEM Accessibility; Multimodal Accessibility Engine.

1. **Introduction**.

Nonetheless, those who are visually impaired still face these challenges in the context of using digital media to access information, especially when presented in a visual and various modalities (World Health Organization, 2023; W3C, 2023). While screen readers, text-to-speech systems, Optical Character Recognition tools, and Braille translators have made it easier for individuals to read text, these devices may not be able to support understanding visual images, diagrams, charts, mathematical formulas, tables and complex document structures (Lazar et al., 2017; Harper & Yesilada, 2008). Digital accessibility is the ability to access and use digital systems for people of all abilities. Accessibility of the digital world for visually impaired users needs more than converting text into speech or Braille. Traditional accessibility systems work well in linear text but fail in the case where text requires semantic interpretation, spatial reasoning, or multimodal comprehension (Bigham et al., 2010; Harper & Yesilada, 2008). Making some text inaccessible decreases independent learning, adds up

cognitive load, and reduces equal academic participation for visually impaired learners (Al-Azawei et al., 2016; Brule et al., 2016).

Recent progress in Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, Computer Vision, and Large Language Models have opened up new avenues for Intelligent Accessibility Systems (LeCun et al., 2015; Goodfellow et al., 2016; Devlin et al., 2019). AI systems can read text, speech, and images, extract semantic meaning, write descriptions, and customize content to the users' needs (Baltru aitis et al., 2019; Brown et al., 2020). Multimodal AI is especially important as it will allow systems to integrate information from different modalities to create integrative representations of data (Lu et al., 2019; Antol et al., 2015). This paper presents a literature review of digital accessibility, traditional assistive technologies, AI-based accessibility systems, multimodal AI, language models and adaptive Braille generation.

Objectives of the Paper:

The objectives of this paper include the review of the current state of AI-assisted accessibility technologies for visually impaired users, specifically Multimodal AI, Large Language Model, and adaptive Braille generation. This paper also highlights current research gaps and develops a framework known as Multimodal Accessibility Engine (MADE) to facilitate intelligent and personalized digital accessibility.

Organizations of the Paper:

This paper is structured in the following manner: the second section contains the review of the literature, while the third section highlights the research gaps. The fourth section describes the proposed MADE framework. The final section contains the conclusions drawn from this research.

2. Literature Review.

2.1. Conventional Accessibility and Braille Systems.

Braille standards like Grade 1 Braille, contracted Braille, Unified English Braille, and Nemeth Braille have evolved over the years to accommodate general language, contractions, and mathematical notation (Miesenberger et al., 2014; Ryles, 1996). Traditional Braille translation systems convert printed characters to Braille symbols in a system, which is designed according to pre-defined linguistic and grammatical rules.

Tools like Duxbury Braille Translator and Liblouis have made it possible to change books, documents and textbooks using Braille automatically (WebAIM. 2023, Miesenberger et al., 2014). However, rule-based Braille systems are inherently adapted for linear textual conversion and are not capable of preserving semantic meaning in text, if the text in fact consists of table, diagram, equation or structured layout information. Screen readers and text-to-speech systems are also used frequently by visually impaired users. Text-to-speech systems enable production of synthesized speech on digital text and enhance accessibility to information through educational and work environments. Yet, screen readers frequently need a dependable quality of document organization, alternative text, semantic markup, and accessibility compliance (Lazar et al., 2017). The output from screen readers is vague or incomprehensible, especially when websites, PDFs and applications are ill-structured. These systems can convert text into speech or Braille, but they generally lack context knowledge.

Consequently, there may be limited knowledge and broken data for the visually impaired users rather than informative explanations of content.

2.2 Artificial Intelligence and Accessibility

Artificial Intelligence allows systems to identify patterns, decode language, process pictures, comprehend speech, and produce adaptive results (LeCun et al., 2015; Goodfellow et al., 2016). In the field of accessibility research, AI has been introduced to OCR, speech recognition, image captioning, object detection, scene understanding, document summarization, and adaptive UI (Szegedy et al., 2015; Ren et al., 2015; Redmon et al., 2016). The performance of accessibility systems has improved with Machine Learning and Deep Learning and learning models which do not depend on manually established rules, are able to make sense of data itself.

Convolutional Neural Networks for image recognition have been extensively applied, and Transformer models and Recurrent Neural Networks have been widely applied in speech and language processing (LeCun et al., 2015; He et al., 2016; Vaswani et al., 2017). These advances are beneficial to blind users as they give systems the ability to learn visual content and transform it into textual, auditory or tactile explanation. Through combining computer vision and language rendering techniques, models can be used to automatically generate natural language descriptions of images (Vinyals et al., 2015; Xu et al., 2015., Antol et al. 2015) extend the support to visual question answering platforms to allow a user to ask questions about visual content and get replies in language. These technologies have the potential to support visually impaired users in the interpretation and synthesis of images, diagrams, and educational illustrations. OCR systems can extract text from scanned documents, images, and printed materials, while NLP systems have the ability to summarize, simplify, or restructure content so that it is accessible (Graves et al., 2013; Honnibal & Montani, 2017). However, AI-based systems for accessibility will need to address biases in their data, difficulties in computation for computationally intensive tasks, and reliability when compared with live applications (Goodfellow et al., 2016; Brown et al., 2020).

2.3 NLP, Large Language Models, and Semantic

Accessibility: All aspects are under study. The central issue in accessibility is the use of Natural Language Processing since most assistive systems rely on language comprehension and generation. NLP supports text summarization, question answering, speech interfaces, machine translation, document simplification, and conversational assistance (Honnibal & Montani, 2017; Raffel et al., 2020). While previous NLP systems were mainly rule-based or statistical methods, recent Transformer-based methods provide large improvements regarding the contextual understanding of language. Large Language Models like BERT, GPT and T5 have proven good performance across language understanding, generation, summarization and question answering (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020; Raffel et al., 2020) Furthermore, these models do support accessibility by producing cohesive, comprehensible explanations, generating short prose, boiling down longer materials and generating accurate, context-aware scripts, particularly for speech or Braille output. In turn, semantic accessibility demands that systems understand meaning as opposed to merely converting symbols. For instance, one way to represent a mathematical expression, table or chart is need that we preserve the relationship, hierarchy, and instructional sense. LLMs can support semantic accessibility via the interpretation of the context and generation of

structured descriptions (Devlin et al., 2019; Brown et al., 2020). Conversely, LLMs can generate incorrect or unsupported outputs, therefore validating and adapting their work to the domain are critical for accessibility apps.

2.4 Multimodal AI Approaches for Accessibility.

Multimodal AI means systems that process and integrate information from various modalities like text, speech, images, video, and structured documents (Baltrušaitis et al., 2019; Lu et al., 2019). Educational resources, websites, presentations and professional documents frequently utilize written explanation accompanied by diagrams, charts, images, sound and interaction. Multimodal learning allows AI systems to recognize relationships between textual and visual content. An image captioning model, for example, may extract information of visual features to generate a textual description, while a visual question answering model may merge image understanding with language reasoning (Vinyals et al., 2015; Xu et al., 2015; Antol et al., 2015). Multimodal AI can provide visual access by offering availability of diagrams, charts, infographics, scanned documents, and educational videos for the visually impaired. Computer Vision models can detect objects, extract text and layout structures, and interpret visual relationships (Szegedy et al., 2015; He et al., 2016; Ren et al., 2015). NLP models have the potential to convert these outputs into interpretations, summaries, or Braille-friendly models (Devlin et al., 2019; Raffel et al., 2020). However, the implementation of multimodal accessibility is technically challenging. Systems will need to maintain semantic consistency between modalities, not produce misleading information, and return short and meaningful outputs (Baltru aitis et al., 2019; Lu et al., 2019). This becomes more complex in STEM domains, where diagrams and equations need to be rendered accurately and representatively.

2.5 Adaptive Braille Generation and STEM Accessibility

Adaptive Braille generation is Braille output production where context, content structure and user needs are taken into consideration. Conventional Braille translation systems translate text into Braille symbols, but adaptive systems work to save semantic content and create customized translations dependent on user's ability, reading speed and cognitive habits (Brule et al., 2016; Miesenberger et al., 2014). Personalization is crucial because impaired users have various degrees of expertise in Braille systems. An inexperienced user might prefer the use of uncontracted Braille and simplified explanation and a more advanced user might want contracted Braille since it's a faster reading speed.

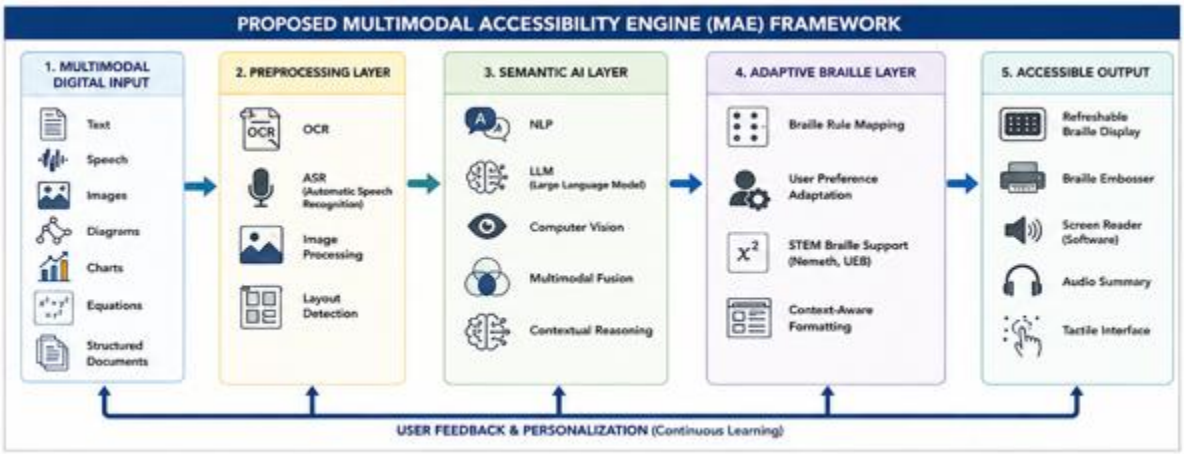
Adaptive systems modify the contraction, formatting, depth of description, and navigation structure based on user needs. Special focus is needed in the accessibility area of STEM content, where technical information often contains numerical notation, symbolic expressions, spatial plans and representations in diagrams. While Nemeth Braille and Unified English Braille can help with mathematical representation, they struggle with automatic conversion of STEM content where equations, graphs, and diagrams appear embedded in a convoluted layout (Ryles, 1996; Bigham et al., 2010). When providing structured explanations of STEM content, AI-based systems are often helpful by integrating OCR, layout analysis, understanding images, and providing language generation (Ren et al., 2015; Xu et al., 2015; Devlin et al., 2019). So, to bridge multimodal AI to tactile accessibility, adaptive Braille generation can be the solution. Thus, future systems can generate Braille outputs that are not only technically correct, but also meaningful and readable via semantic understanding and user-centered personalization.

3. Research Gaps

First of all, most current technologies used for making content accessible only provide for linear text processing. While screen readers and Braille translators can process text, they cannot interpret images, diagrams, charts, equations, and layout (Harper & Yesilada, 2008; Bigham et al., 2010; Lazar et al., 2017). Rule-based solutions allow for transforming content into another form but do not provide any interpretation of relationships between visual, textual, and structural content (Bigham et al., 2010; Baltru aitis et al., 2019). This is a considerable limitation when it comes to educational or technical information where the relationship between symbols, labels, and other elements is crucial. The issue with existing technology is that they generate the same output for all the users regardless of their proficiency level in Braille, reading speed, their preferred mode of thinking, and the context of studying. Content needs different methods of processing and it is quite hard to combine them into a semantic whole (Baltru aitis et al., 2019; Lu et al., 2019, Brule et al., 2016; Miesenberger et al., 2014). Fifth, there are no multimodal Braille-aligned datasets. However, datasets related to accessibility with regard to connecting images, semantics, and Braille are scarce (Goodfellow et al., 2016; Brown et al., 2020; Chen et al., 2015; Antol et al., 2015; Baltrušaitis et al., 2019). This impedes the construction of reliable AI systems that could work with Braille in a meaningful way. It would be helpful to create systems that would combine all four aspects mentioned above.

4. Proposed Multimodal Accessibility Engine Framework

Such gaps in research prompt the introduction of the Multimodal Accessibility Engine (MADE) – a conceptual framework based on AI technologies and aimed at creating intelligent semantically relevant and adaptive representations of multimodal digital content in Braille. In contrast to traditional approaches focused on conversion of the text either into speech or Braille, MADE uses Natural Language Processing (NLP), Computer Vision (CV), speech processing, multimodal fusion, semantic reasoning, and adaptive Braille generation (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020). The first stage of this framework involves multimodal input processing when information is extracted from text, images, scanned documents, diagrams, tables, equations, and audio materials. Information extraction is performed by Computer Vision technologies, such as OCR, document layout analysis, and object detection, as well as by speech processing (Szegedy et al., 2015; Ren et al., 2015; He et al., 2016). At the next stage of the framework, the extracted information is further analyzed through the semantic understanding layer where NLP and Large Language Models determine contextual relationships between the elements of the source content.



Proposed Multimodal Accessibility Engine framework for inclusive digital accessibility.

The third component is multimodal fusion (Baltrušaitis et al., 2019; Lu et al., 2019), which combines information from multiple modalities through linking captions to images, equations to explanations and diagrams to text around them. The fourth component is personalized Braille generation (Ryles, 1996; Miesenberger et al., 2014; Brule et al., 2016), which produces personalized Braille outputs based on document structure, semantic context and preferences of users, e.g., their Braille proficiency and reading speed. Finally, the user feedback and personalization component constantly improves generated output through learning from user interaction and reading behavior. With the combination of multimodal AI and Braille-aware semantic generation, the presented MADE system takes digital accessibility beyond just format conversion to a new level of intelligent and personalized access to complex digital documents

Conclusion

This paper is a review of the evolution of digital accessibility devices for visually-impaired persons, concentrating on classic assistive technology. In addition, it also studies AI-driven accessibility, multimodal AI, NLP, Large Language Models and adaptive Braille generation. Existing tools such as screen readers, text-to-speech systems, OCR applications, and Braille translators have increased access to text, but they are still poorly capable of dealing with multimodal, structured and STEM-related materials (Lazar et al., 2017; Harper & Yesilada, 2008; Bigham et al., 2010). AI, Deep Learning, Computer Vision, NLP, and multimodal learning provide robust frameworks for next-generation accessibility systems, as we see in the review above (LeCun et al., 2015; Goodfellow et al., 2016; Baltrušaitis et al., 2019). Large Language Models and Transformer-based architectures (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020) that are able to leverage semantic reasoning, summarization, and context-aware content generation. There are still large gaps in semantic preservation, multimodal integration, personalization, dataset availability, and adaptive Braille generation. To overcome these limitations, in this paper, the work has suggested the model of Multimodal Accessibility Engine (MADE) that can incorporate multimodal content processing, semantic reasoning and adaptive Braille generation. Such a framework has the potential to increase access to challenging digital content by producing individualized, meaningful, and Braille-comparable representations. Future studies should consider building multimodal Braille data, assessing user experience for the visually impaired learners, improving STEM content representation, and ethical, reliable, and inclusive deployment of AI-based accessibility systems.

Future Scope: The future scope of research is directed towards the creation of large Braille multimodal datasets, fine-tuning of Braille aware LLMs and the validation of our approach by conducting user tests with visually challenged users. Furthermore, the framework could be extended for multi-language Braille synthesis, advanced STEM materials and trustworthy AI.

References

Al-Azawei, A., Serenelli, F., & Lundqvist, K. (2016). Universal Design for Learning (UDL): A content analysis of peer reviewed journals from 2012 to 2015. *Journal of the Scholarship of Teaching and Learning*, *16*(3), 39-56. doi: 10.14434/josotl.v16i3.19295

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425-2433). doi: 10.1109/ICCV.2015.279.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. https://doi.org/10.48550/arXiv.1409.0473

T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., & Yeh, T. (2010). VizWiz: Nearly real-time answers to visual questions. *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, 333–342. https://doi.org/10.1145/1805986.1806020

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Brule, E., Bailly, G., Brock, A., Valentin, F., Denis, G., & Jouffrais, C. (2016). MapSense: Multi-sensory interactive maps for children living with visual impairments. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 445–457. https://doi.org/10.1145/2858036.2858375

Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*. https://doi.org/10.48550/arXiv.1504.00325

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. https://doi.org/10.48550/arXiv.1303.5778

Harper, S., & Yesilada, Y. (2008). *Web accessibility: A foundation for research*. Springer. https://doi.org/10.1007/978-1-84800-050-6

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Explosion AI. https://doi.org/10.5281/zenodo.1000885

Lazar, J., Goldstein, D. F., & Taylor, A. (2017). *Ensuring digital accessibility through process and policy*. Morgan Kaufmann.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32. https://doi.org/10.48550/arXiv.1908.02265

Miesenberger, K., Fels, D., Archambault, D., Peňáz, P., & Zagler, W. (Eds.). (2014). *Computers helping people with special needs*. Springer. https://doi.org/10.1007/978-3-319-08596-8

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI Technical Report. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67. https://doi.org/10.48550/arXiv.1910.10683

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28. https://doi.org/10.48550/arXiv.1506.01497

Ryles, R. (1996). The impact of Braille reading skills on employment, income, education, and reading habits. *Journal of Visual Impairment & Blindness*, 90(3), 219–226. https://doi.org/10.1177/0145482X9609000312

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1706.03762

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

WebAIM. (2023). *Screen reader user survey*. Web Accessibility in Mind. https://webaim.org/projects/screenreadersurvey10/