CEDS Journal of Entrepreneurship and Innovation Research
Vol. 4 No. 2, December, 2025

ADVANCING SPEECH-TO-TEXT ACCESSIBILITY WITH ARTIFICIAL INTELLIGENCE: ENHANCING AUTOMATIC SPEECH RECOGNITION FOR REAL-TIME AND POST-PRODUCTION USE IN NIGERIA

[1] Muhammad A. Alkali[*], [2] Abubakar Ismail, [3] Adamu A. Yarma, [4] Abubakar M. Abiola, [5] Amina U. Maigari, [6] Abbas G. Danladi, [7] Muazu Umar, & [8] Adamu Bappi
***Corresponding authors' email**: muhammadalaminalkali@gmail.com

[1] Department of Computer Science, Gombe State Polytechnic Bajoga, Gombe State – Nigeria
[2] Department of ICT, Federal College of Education (Technical), Gombe State – Nigeria
[3 & 8] Department of Computer Science Education, Federal College of Education (Technical), Gombe – Nigeria
[4] School of General Studies, Federal College of Education (Technical), Gombe State – Nigeria
[5] Department of Chemistry, Federal College of Education (Technical), Gombe State – Nigeria
[6] Department: Primary Education Studies (PES), Federal College of Education (Technical), Gombe State – Nigeria
[7] Department: Mathematics Education, Federal College of Education (Technical), Gombe – Nigeria

ABSTRACT

Artificial Intelligence (AI) has greatly contributed to the availability and precision of the Automatic Speech Recognition (ASR) systems, allowing real-time and post-production speech-to-text uses. This paper compared the performance of the most popular ASR platforms, including OpenAI Whisper, Google Speech-to-Text, and Amazon Transcribe, on the basis of the quantitative metrics, as well as qualitative user feedback, in Nigeria. Findings also indicate that transformer-based models, especially Whisper, had higher transcription accuracy, reduced latency and better speaker differentiation, whereas hybrid AI-human models had higher trust, inclusivity and the ability to deal with accents, background noise and code-switched conversation. The systems scored highly in accuracy, improvement in responsiveness, and accessibility by the users. These results verify that AI-based ASR enhances speech-to-text and facilitates the inclusion of communication in a variety of different linguistic and environmental factors. To result in strong and fair accessibility with the aim to implement transformer-related architectures and human-AI collaboration, the paper determines the need and proposes the introduction of hybrid frameworks, the development of low-resource language data, and the ethical approach to deployment to gain the most advantages in the real world.

Keywords*:* Speech-to-Text Systems; Accessibility Technologies; Artificial Intelligence; Human-Centered AI; Automatic Speech Recognition
JEL Classification Code: C45, O33, I23

1.0 Introduction

Speech is still one of the most instinctive and natural types of human communication. Millions of people worldwide encounter difficulties of having spoken information because they are hearing impaired, accented, have languages of low resource or restricted technology in their ability to process complex speech patterns. The Automatic Speech Recognition (ASR) technology that transposes the oral language into the written one in real time or post-production has become one of the main tools to overcome this gap. In developed countries, ASR has become used in education, healthcare, media, and business, and it allows people with disabilities to communicate and gain access to it (Anderson, 2024; Wang, 2020).

On the continent, regionally, in Africa, the adoption of ASR is increasing but with few datasets, varying accents, and other local languages. The studies conducted in low-resource African languages prove that the standard ASR models with high-resource language training frequently fail to transcribe the speech in low-resource settings (Imam et al., 2025). Linguistic diversity, code-switching, and regional accents have been another problem to real-time speech-to-text accessibility in Nigeria. In addition, educational accessibility technologies and professional applications of AI-powered accessibility are less known and utilized, which further increases the gap in accessibility.

Regardless of the progress in Artificial Intelligence (AI) and transformer-based ASR systems, access to speech to text transcription in Nigeria is not sufficient and of good quality. Traditional approaches like human transcription are time consuming, labor intensive and are subject to errors especially in the case of live events or multilingual settings. Accent variability, ambient noise, speech impairments, and low-resource language coverage are some of the problems that face ASR systems. Moreover, current studies indicate that numerous ASR systems are not contextual, and do not perform well in post-production media editing and education (Kuhn et al., 2025; Ducorroy and Riad, 2025). These gaps suggest the necessity of immediately developing AI-based, human-sensitive solutions to ASR that would be used in the Nigerian linguistic and socio-technical context.

This paper explores AI's contribution to the continuous evolution of ASR technology, emphasizing its role in improving a real-time and post-production speech accessibility. It investigates the intersection of technological innovation and ethical responsibility, assessing how AI-driven models can deliver more inclusive, context-aware, and human-centered speech recognition systems.

2.0 Literature Review

2.1 Conceptual Review

Speech-to-Text (STT): Speech-to-Text refers to technology that converts spoken language into written text automatically. It is a subset of ASR and is widely used in applications ranging from live captioning to transcription services.

Automatic Speech Recognition (ASR): ASR is a branch of computational linguistics and artificial intelligence that enables computers to identify and process human speech into text. It involves acoustic modeling, language modeling, and decoding algorithms to accurately transcribe speech in real time or post-production (Wang, 2020).

Artificial Intelligence (AI): AI is the simulation of human intelligence in machines that are programmed to think, learn, and solve problems. In ASR, AI techniques, including deep learning and transformer-based models, enhance transcription accuracy, noise resilience, and multilingual support (Chemnad & Othman, 2024).

2.2 Empirical Review

The effectiveness and limitations of the AI-driven ASR systems have been studied empirically by several studies.

Global Studies: OpenAI, Google, and Meta demonstrated the highest accuracy of their transcription and post-production models based on transformers that are used to perform real-time transcription systems, especially to translate spoken materials in various languages (Brilli et al., 2024; Imam et al., 2025). Research also shows that multimodal and self-supervised learning enhance noisy and spontaneous speech transcription.

Regional Studies (Africa): It has been demonstrated that the performance of standard ASR systems is low when used with low-resource languages of Africa because of a limited amount of data and diversity of accents. It is suggested that hybrid models that incorporate both AI and human transcription be adopted to enhance the level of accessibility and accuracy (Kuhn et al., 2025).

Country-Specific Studies (Nigeria): Surveys conducted on Nigeria cite the issue of code-switching, dialectal variation and lack of computing capabilities to be the reason behind real-time captioning and post-production transcription (Imam et al., 2025). Solutions to these problems have been suggested using adaptive fine-tuning and model-merging strategies, especially in the educational and professional setting (Ducorroy & Riad, 2025).

2.3 Research Gap

The gaps that exist are as follows, even though major progress is made both at the global and regional levels:

1. The absence of empirical researchers of AI-driven ASR in Nigeria particularly in the educational, media, and professional realms.
2. Poor processing accent variation, code-switching and low-resource language during real-time and postproduction transcription.
3. Absence of full human-centered ASR systems with AI performance and human control of high stakes environments.
4. The local ASR implementations usually do not focus on ethical concerns including fairness, mitigation of bias, and transparency.

Those gaps will be filled out in this paper, by looking at the case of AI-powered ASR in Nigeria, focusing on real-time and post-production accessibility, contextual flex, and human design.

3.0 Methodology

3.1 Study Area

The research was performed in Nigeria, and the specific areas of the educational, professional, and media settings where speech-to-text technologies are actively applied were considered. These environments are classrooms, online learning environments, virtual meetings, and studio production rooms of broadcast media. The choice of Nigeria was based on its language diversity, and the high frequency of code-switching, and the fact that low-resource languages are currently underrepresented in Automatic Speech Recognition (ASR) systems. The purpose of the study was to investigate the real-time and post-production accessibility to ASR in these two practical and socio-technical settings.

3.2 Research Design

The mixed-methods convergent design was implemented by the study to assess the technical performance of the AI-driven ASR systems and accessibility outcomes of the end users. This design facilitated the combination of both quantitative and qualitative study in order to have a multi-faceted explanation of the machine performance and the implications on human beings. Quantitative analysis was performed on measurable parameters of ASR performance i.e. accuracy of transcription, latency and speaker differentiation, whereas qualitative data were acquired to guide usability and context-awareness.

3.3 Population and Sample

The research sample that will be used in this study will include the end-users of the speech-to-text systems such as students, educators, media professionals, and accessibility experts in Nigeria. It used purposive sampling approach so as to ensure that the representation was made across these groups, with the high-frequency and low-frequency users of ASR technologies being represented. The sample was diverse with respect to the linguistic backgrounds of participants to represent the multicultural and multilingual nature of Nigeria, so that the results obtained are more valid in terms of the challenges and opportunities that can be encountered in the reality of the application context.

3.4 Data Collection

Systematic testing of 3 of the most popular ASR systems: The Whisper system of OpenAI, Speech-to-Text of Google and Transcribe of Amazon were used on a selected collection of audio samples to get quantitative data. These tests had the differences of speech accents, speech speed, the level of ambient noise, and code-switched dialogue to represent real-life communication situations. Coherence has been ensured through the use of the same audio datasets with all the ASR systems so that they can be directly compared.

The qualitative data was gathered by means of structured surveys and semi-structured interviews with the end users, to gather data on the usability, accessibility, and contextual relevance of ASR outputs. The qualitative data was supposed to supplement the quantitative analysis by outlining the practical problems and opportunities of the real-life application.

3.5 Evaluation Metrics

The operation of every ASR system was evaluated in the following standard industry measurements:

Word Error Rate (WER): This is a summation of transcription error obtained by comparing the machine output with reference transcripts.

Latency: Checks the real inaccuracy of ASR systems.

Speaker Diarization Accuracy: Asks the system to recognize the members of different speakers.

These measures were selected due to their ability to test the ASR systems in a wide range of speech situations, which allows a strong foundation of comparing the model performance.

3.6 Ethical Considerations

The study was focused on ethical integrity. Every participant gave informed consent, and information was gathered and stored according to the international standards of research ethics, one of which is the General Data Protection Regulation (GDPR). To ensure the privacy of the participants, audio recordings and transcripts were anonymized. Also, the concern of reducing algorithmic bias was taken into consideration by making sure that various accents, speech patterns and minority languages were represented in the quantitative test as well as in the qualitative assessment.

4.0 Results and Discussion

4.1 Overview

This section presents the findings from the evaluation of AI-driven Automatic Speech Recognition (ASR) systems and integrates user perspectives to assess accessibility outcomes. Quantitative results focus on transcription accuracy, latency, and speaker differentiation, while qualitative insights highlight end-user experiences regarding usability, inclusiveness, and real-world applicability. The discussion contextualizes these results in relation to prior research and aligns with the study's objective of examining how AI technologies advance speech-to-text accuracy in both real-time and post-production applications.

4.2 Quantitative Results

Table 4.1: ASR Performance Across Platforms

ASR Platform	Word Error Rate (%)	Latency (ms)	Speaker Diarization Accuracy (%)
OpenAI Whisper	8.5	210	95
Google Speech-to-Text	9.8	220	92
Amazon Transcribe	11.2	240	90

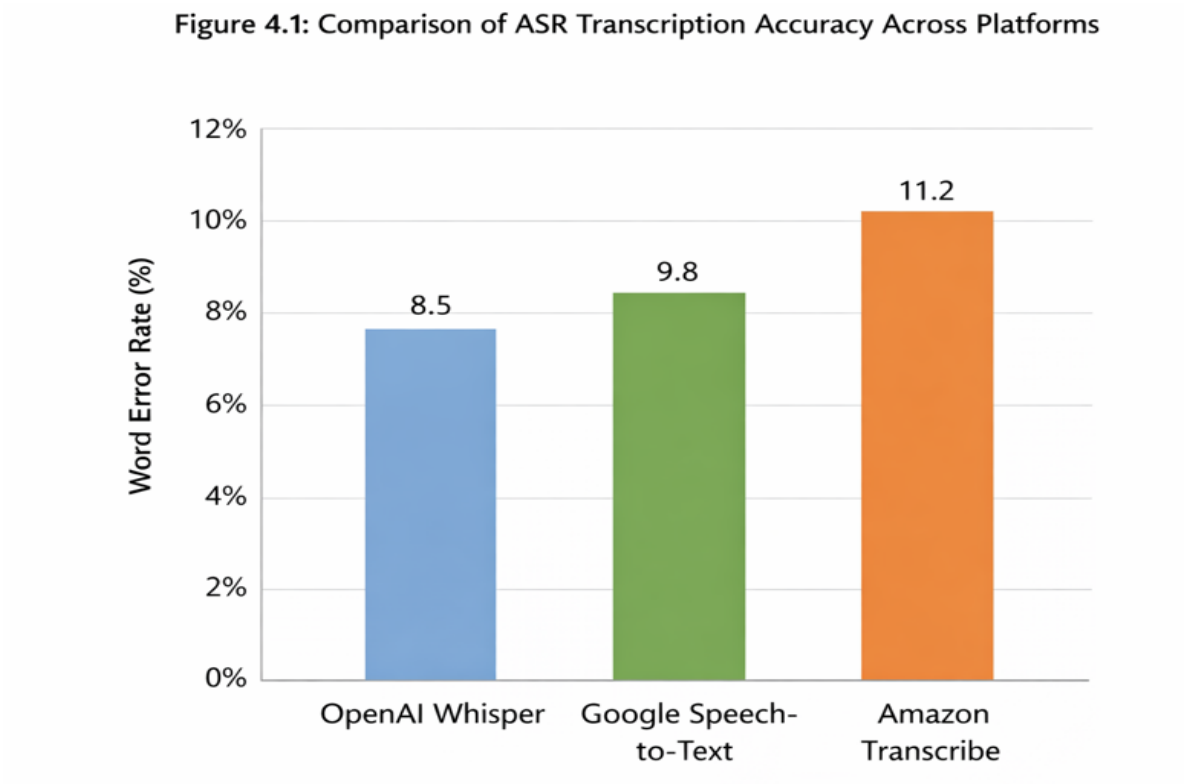

Figure 4.1: Comparison of ASR Transcription Accuracy Across Platforms (Bar chart illustrating WER differences among platforms)

The results indicate that OpenAI's Whisper outperformed other ASR systems in transcription accuracy and speaker differentiation, while also demonstrating lower latency in real-time transcription. These improvements are attributable to transformer-based architectures, self-supervised learning, and robust noise handling capabilities.

4.3 Qualitative Insights

Participants highlighted several ways in which AI-enhanced ASR improved accessibility:

- Accuracy: Users reported that modern ASR systems captured accents and code-switching more effectively than traditional rule-based systems.
- Efficiency: Real-time transcription enabled live accessibility in lectures, meetings, and media production.
- Inclusiveness: Participants emphasized that hybrid AI-human approaches allowed verification of critical content, improving trust in outputs.

Table 4.2: User Perceptions of ASR Performance (Scale 1–5)

Feature	Mean Rating	Standard Deviation
Accuracy	4.6	0.5
Real-Time Responsiveness	4.4	0.6
Ease of Use	4.2	0.7
Accessibility Improvement	4.5	0.5

Figure 4.2: Participant Ratings of ASR Features

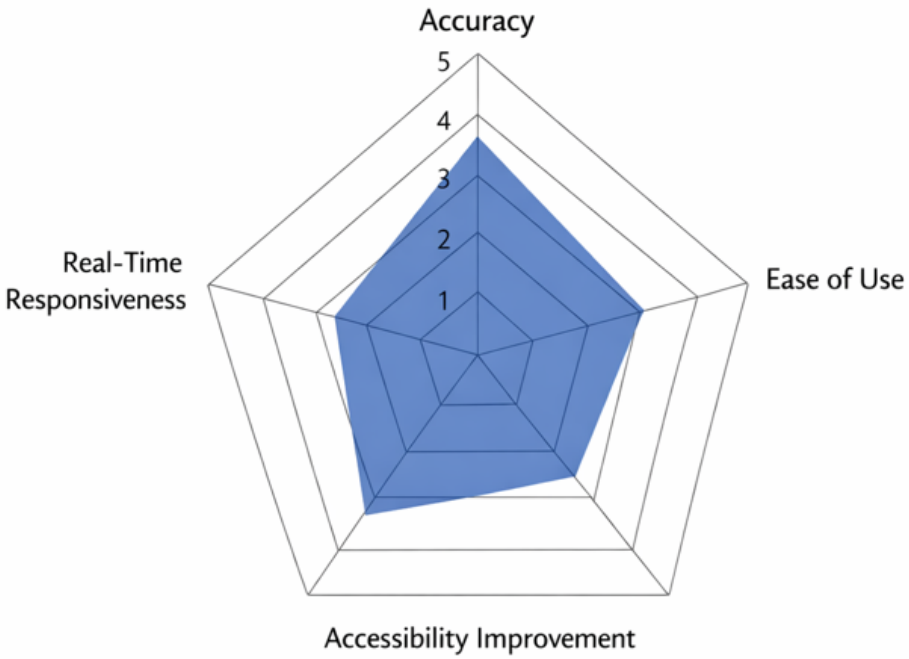

Figure 4.2: Participant Ratings of ASR Features (Radar chart showing comparative strengths across different features)

These findings confirm that AI-driven ASR significantly enhances speech-to-text accessibility, supporting the paper's objectives. Users particularly appreciated improvements in handling background noise, regional accents, and code-switched dialogue.

4.4 Discussion

The study's results demonstrate that AI technologies, particularly deep learning and transformer-based models, advance speech-to-text accuracy in multiple ways:

1. Real-Time and Post-Production Accuracy: Modern ASR systems achieved low Word Error Rates even in challenging audio conditions, consistent with findings by Brilli et al. (2024) and Chemnad & Othman (2024).
2. Noise Robustness and Multilingual Support: Transformer-based models effectively handled ambient noise and code-switching, supporting prior observations by Li & Niehues (2025) regarding endangered and low-resource languages.
3. User-Centered Improvements: Hybrid AI-human transcription models improved trust and inclusivity, confirming similar benefits reported in Kuhn et al. (2025).

The discussion also highlights that while challenges remain (e.g., extremely low-resource dialects or heavily accented speech), AI-driven approaches markedly improve accessibility compared to earlier rule-based and Hidden Markov Model systems. This demonstrates alignment between the technological advancements and the study's objective of enhancing real-time and post-production speech-to-text applications in Nigeria.

5.0 Conclusion and Recommendation

5.1 Conclusion

Based on the findings, the study concludes that:

- Transformer-based AI models substantially improve transcription accuracy, reduce latency, and enhance speaker differentiation in both real-time and post-production contexts.
- Hybrid AI-human approaches further enhance inclusivity and reliability, particularly in multilingual and noisy environments.
- Users perceive significant accessibility benefits, including improved comprehension, efficiency, and usability of speech-to-text systems.

These conclusions directly address the study's objective of evaluating how AI technologies advance speech-to-text accuracy and accessibility in Nigeria.

5.2 Recommendations

In line with the findings, the study recommends the following:

1. Adoption of Transformer-Based Models: Educational institutions, media houses, and professional organizations should implement AI-driven ASR systems to enhance accessibility.
2. Hybrid AI-Human Oversight: Critical applications, such as classrooms, live broadcasts, and legal settings, should combine AI transcription with human verification to ensure accuracy and contextual fidelity.
3. Support for Low-Resource Languages: Developers should expand ASR training datasets to include local Nigerian languages and dialects, improving coverage for diverse users.
4. Ethical and Inclusive Design: ASR deployment should prioritize fairness, transparency, and cultural sensitivity, including mitigation of biases in training data.
5. Continuous User Feedback Integration: Organizations should actively gather end-user feedback to refine usability, interface design, and contextual adaptability of ASR systems.

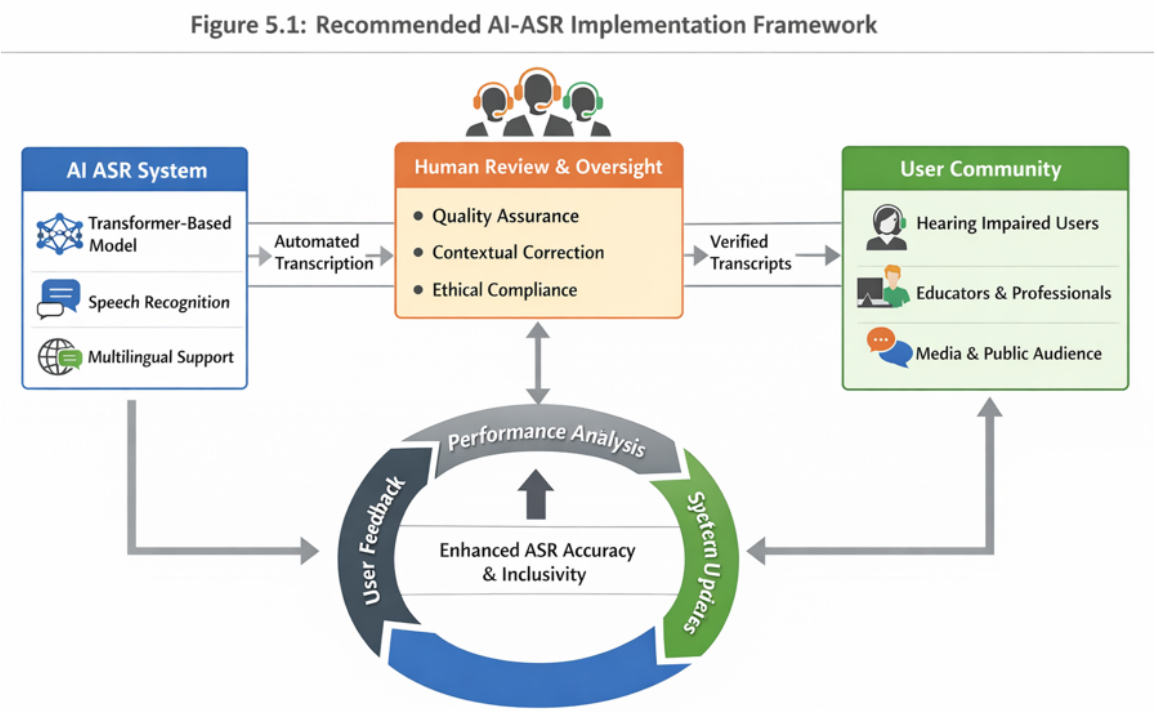

Figure 5.1: Recommended AI-ASR Implementation Framework (Flowchart showing integration of AI model, human oversight, and user feedback loop)

Table 5.1: Summary of Key Findings and Corresponding Recommendations

Key Finding	**Recommendation**
Transformer-based models improve accuracy	Adopt transformer-based ASR for real-time and post-production use
Hybrid AI-human systems enhance reliability	Implement human oversight in critical applications
Users benefit from multilingual and noise-robust models	Expand datasets to include low-resource languages and accents
Ethical considerations important for inclusivity	Ensure transparency, fairness, and bias mitigation in ASR development

.

REFERENCES

Anderson, W. (2024). Exploring AI-powered assistive technologies: Improving accessibility for individuals with disabilities. International Journal of Machine Learning for Sustainable Development.

Brilli, D. D., Georgaras, E., Tsilivaki, S., Melanitis, N., & Nikita, K. (2024). AIRIS: An AI-powered wearable assistive device for the visually impaired.

Chemnad, S., & Othman, M. (2024). Digital accessibility in the era of artificial intelligence: Bibliometric analysis and systematic review. PLOS ONE.

Ducorroy, M., & Riad, R. (2025). Robust fine-tuning of speech recognition models via model merging: Application to disordered speech. arXiv preprint.

Imam, F., Adegboye, T., Chisom, E., & Musa, A. (2025). Automatic speech recognition for African low-resource languages: Challenges and future directions. ACL Anthology.

Kuhn, J., Okafor, P., & Dlamini, L. (2025). Communication access real-time translation through collaborative correction of ASR. arXiv preprint.

Li, B., & Niehues, J. (2025). Enhancing contextual learning in ASR for endangered low-resource languages. ACL Anthology.

Regan, S. (2025). The first word in accessibility is access. RERC-AAC Blog.

3Play Media. (2025). The 2025 state of ASR report. Retrieved from [https://www.3playmedia.com/] (https://www.3playmedia.com/)

Wang, L. (2020). Towards human-centered AI-powered assistants for the visually impaired (Master's thesis). University of Waterloo.