Evaluating Yoruba Text-to-Speech Systems for Accessible Computer-Based Testing in Visually Impaired Learners

Kausar Moshood[1,2], **Victor Olufemi**[1,3], **Oreoluwa Babatunde**[1], **Emmanuel Bolarinwa**[1], and **Williams Oluwademilade**[1]

[1]LyngualLabs
[2]Oregon State University
[3]Carnegie Mellon University
{kausar,victor,oreoluwa,emmanuel,williams}@lynguallabs.org

Abstract

Text-to-Speech (TTS) technology offers potential to improve exam accessibility for visually impaired learners, but existing systems often underperform in underrepresented languages like Yoruba. This study evaluates current Yoruba TTS models in delivering standardized exam content to five visually impaired students through a web-based interface. Before testing, four Yoruba TTS systems were compared; only Facebook's mms-tts-yor and YarnGPT produced intelligible Yoruba speech. Students experienced exam questions delivered by human voice, Braille, and TTS. All preferred Braille for clarity and independence, some valued human narration, while TTS was least favored due to robotic and unclear output. These results reveal a significant gap between TTS capabilities and the needs of users in low-resource languages. The paper highlights the urgency of developing tone-aware, user-centered TTS solutions to ensure equitable access to digital education for visually impaired speakers of underrepresented languages.

1 Introduction

Computer-Based Testing (CBT) has become a widely adopted format in modern educational assessments, thanks to benefits like scalability, automation, and reduced logistical costs. For example, most exams today from school evaluations to professional certifications use computer-based delivery (Patel et al., 2021). This approach streamlines administration for large candidate numbers while cutting out printing and distribution expenses (Chukwuma-Nosike and Chukwuma, 2023). However, this digital transition also presents serious accessibility challenges for visually impaired learners. Studies have shown that blind or low-vision students often face barriers with standard CBT interfaces, limiting their ability to independently access test content (Patel et al., 2021). These difficulties are especially pronounced in underrepresented language contexts, where assistive tools may not support the local language script or speech (Tubosun, 2023).

Assistive technologies such as screen readers and Braille displays have facilitated access to digital education for many blind users. Screen readers, for instance, convert on-screen text to speech or Braille and are a primary interface for blind computer users (American Foundation for the Blind, 2025). Yet the effectiveness of these tools varies significantly depending on the language and infrastructure involved. Many African languages, including Yoruba, Igbo, and Hausa, have historically been marginalized in technology, resulting in limited support in mainstream screen readers and voice assistants (Tubosun, 2023). Indeed, until recently, no major smart assistant could operate in any indigenous African language, forcing users to rely on English for voice feedback (Tubosun, 2023). This gap means that visually impaired students in such language contexts are at a higher risk of exclusion when exams move to digital formats.

Text-to-Speech (TTS) systems offer a promising path for making CBTs more accessible to blind and low-vision learners. These systems convert written content into spoken output in real time, enabling visually impaired students to independently interact with computer-based questions and answers (Dhaliwal and Sharma, 2024). TTS technology has been hailed for "revolutionizing the world by enabling disabled people to access information and achieve independence" (Dhaliwal and Sharma, 2024). In an exam setting, a reliable TTS voice could allow a blind test-taker to listen to questions and select answers without needing a human reader or proctor, and without the delays of Braille translation. However, most state-of-the-art TTS models today are developed and optimized for high-resource languages like English and Chinese, leaving a significant performance gap when they are applied to low-resource languages with tonal and complex

structures such as Yoruba (Ogunremi et al., 2024).

Yoruba is a major Nigerian language spoken by over 40 million people across West Africa (Eberhard et al., 2019). It is also one of the few indigenous African languages offered as a subject in standardized national exams like those administered by the West African Examinations Council (WAEC) for example, there is a dedicated Yoruba language paper in the WAEC senior secondary exams (waec). Despite its large speaker base and official status in education, little is known about how well existing Yoruba TTS systems can perform in high-stakes CBT environments for visually impaired learners. To date, there has been virtually no published research on applying Yoruba TTS for exam delivery, which indicates a clear knowledge gap that this study aims to fill.

In this study, we investigate the feasibility and effectiveness of using AI-driven Yoruba TTS systems for administering CBT-style exam questions to visually impaired secondary school students. By engaging five blind students in an experimental setup, we evaluate their experience using two publicly available Yoruba TTS models, comparing the results with traditional Braille and human narration modes. The study provides both technical insights and human-centered evidence to inform the development of inclusive, linguistically aware educational technologies. Ultimately, our goal is to highlight whether current TTS technology is up to the task of delivering Yoruba exam content accessibly, and what improvements are needed to ensure equitable access to digital assessments for visually impaired learners.

1.1 Research Questions

This study carried out an experiment on the use of TTS systems to assist visually impaired students during computer-based tests (CBTs). The aim was to understand how effective current Yoruba TTS models are when used in real exam conditions. The following research questions guided the study:

- **Clarity of TTS Output:** Can existing Yoruba TTS systems read out standardized exam questions clearly and correctly for blind students (i.e., with sufficient intelligibility and accuracy)?
- **Comprehension Compared to Braille/Human Voice:** Do visually impaired students comprehend Yoruba TTS-delivered questions as well as they do with Braille or human narration during test scenarios?
- **Student Experience and Preference:** How do students feel about using TTS in a CBT setting, particularly regarding comfort, focus, and sense of independence?

2 Background and Related Work

2.1 TTS Systems and Accessibility

TTS systems play a central role in accessibility, particularly for blind and visually impaired users. These systems are widely used in screen readers, digital assistants, and educational tools to provide real-time auditory access to written content (American Foundation for the Blind, 2025; Rella, 2023). In educational settings, TTS enables learners with visual impairments to independently engage with computer-based content, including assessments, without relying on human assistance (Dhaliwal and Sharma, 2024).

A typical TTS pipeline includes text normalization, linguistic analysis, acoustic modeling, and waveform generation (Rella, 2023). Advances in deep learning have led to the development of end-to-end neural models like Tacotron 2, FastSpeech, and VITS, which have demonstrated near-human speech quality in high-resource languages such as English and Mandarin (Shen et al., 2018; Ren et al., 2019; Kim et al., 2021). These models have significantly improved voice quality in screen readers and accessibility tools, providing smoother and more natural reading experiences.

However, deploying TTS in high-stakes scenarios like exams requires more than fluency. The system must produce highly intelligible, accurate, and well-pronounced speech, especially in tonal languages where pitch can change word meaning (Ogunremi et al., 2024). For languages like Yoruba, which require tone sensitivity and language-specific modeling, this remains a major challenge.

2.2 Challenges in TTS for Underrepresented Languages

Developing TTS systems for underrepresented languages such as Yoruba presents significant challenges due to limited linguistic resources. Yoruba is a tonal language where pitch accents (e.g., á, à, a) influence meaning. When TTS systems fail to model tone accurately, they risk generating unnatural or misleading speech (Ogunremi et al., 2024;

Tubosun, 2023). The lack of large high-quality Yoruba datasets, particularly diacritically marked text further complicates model training and tone learning.

While multilingual efforts like Meta's Massively Multilingual Speech (MMS) project and Mozilla's Common Voice have introduced Yoruba into their data pools (Pratap et al., 2023; Ardila et al., 2020), the models built on these datasets still struggle with tonal fluency and pronunciation. Early Yoruba TTS systems based on festival or rule-based engines achieved syntactic correctness but lacked expressive prosody and were rarely tested in applied contexts like exam delivery (Gutkin et al., 2020). Additionally, many existing corpora are domain-limited, such as religious readings, and do not reflect diverse use cases like standardized assessments.

As a result, current Yoruba TTS tools are often intelligible at the sentence level but unreliable in scenarios that demand tonal precision, such as multiple-choice exams where slight tonal differences can change answer meanings.

2.3 Speech vs. Braille in Educational Testing

Braille remains a trusted method for blind students during standardized testing. It offers silent, tactile interaction with exam content, supporting independence, concentration, and accuracy (Willings, 2017). However, Braille literacy rates remain low globally due to limited access to training and materials. In Nigeria and similar contexts, many visually impaired students lack the infrastructure or instruction required to become fluent Braille users (National Federation of the Blind, 2009).

Moreover, digital Braille solutions such as refreshable Braille displays are costly and scarce in low-resource environments (Perkins School for the Blind, 2024). Paper Braille, while helpful, requires advance preparation and lacks flexibility for real-time or adaptive CBT systems. This limits its practicality for digital-first education systems and national e-assessment platforms.

TTS systems offer an alternative that can scale across devices without special hardware. When properly designed, they can render questions in local languages on standard laptops or mobile phones. However, TTS for exams must be accurate, fast, and linguistically tuned. In tonal languages like Yoruba, incorrect pronunciation or flat prosody can compromise understanding and fairness. Despite TTS's potential, few studies have directly compared its performance with Braille in timed, exam-like scenarios, particularly in African languages. This study addresses that gap using real-world testing with Yoruba-speaking students.

3 Methodology

3.1 Study Design

This study used a three-phase approach to compare the effectiveness of human voice, Braille, and TTS in delivering standardized exam content to visually impaired students. The same set of WAEC Yoruba questions was presented to each participant using all three formats. The responses and experiences of the students across these three modes formed the basis of the findings.

3.2 Participants

All five students (three males and two females) who participated in this experiment were visually impaired, fluent in Yoruba language, and between the ages of 15 and 18. They were selected from an educational institution that supports students with special needs and each of them had prior experience using Braille but limited or no experience with TTS systems.

3.3 Question Selection

Ten multiple-choice questions were selected from the 2024 WAEC Yoruba exam paper. The questions were standardized and covered common themes such as comprehension, tone-sensitive vocabulary, and culturally rooted expressions. Each question had four answer options (A–D) and was stored in a structured CSV file for use during the TTS phase.

3.4 Phase One: Human Voice Delivery

In the first phase, each question was read aloud to the students by a fluent Yoruba speaker in a quiet room. The speaker maintained a consistent pace and tone to minimize variability while students listened carefully and selected their answers verbally. This phase served as a familiar benchmark, as human narration is often used in assisted testing environments.

3.5 Phase Two: Braille Delivery

The second phase involved presenting the same 10 questions to each student in Braille format. Each student was given enough time to read the questions and options independently and then respond verbally. This phase reflected current best practices in inclusive exam administration and provided a direct comparison point for evaluating TTS usability.

3.6 Phase Three: TTS Web Interface Delivery

The third phase involved presenting the same WAEC Yoruba questions to participants using a custom-built web interface. This approach tested whether modern Yoruba TTS models could serve as a viable delivery method for computer-based tests (CBTs) designed for visually impaired students.

3.6.1 Model Evaluation

Four Yoruba TTS models were evaluated to determine their suitability for delivering exam content to visually impaired learners. The models tested were: Facebook's MMS-TTS-Yor, YarnGPT, Tacotron 2 (Google)(Shen et al., 2018), YorubaTTS (Túbsún and Olúòkun, 2017). Each model was assessed using *Mean Opinion Score (MOS)*, a standard subjective metric used in speech quality testing (ITU-T, 1996). Each model was rated by four native Yoruba speakers using a 5-point Likert scale, where 1 = Bad(very unnatural), 2= Poor (unnatural), 3=Fair (somewhat unnatural), 4=Good (mostly natural), 5=Excellent (very natural).

The MOS for each model was calculated using the formula:

$$MOS = \frac{1}{N}\sum_{i=1}^{N} r_i$$

where r_i represents the rating assigned by the ith evaluator, and N is the total number of raters. This method, standardized by the International Telecommunication Union, ensures a consistent and interpretable quality metric across speech systems (ITU-T, 1996).

Models with a MOS score of 3.0 or higher were considered acceptable for inclusion in the live exam testing phase. The final selection was based on both quantitative MOS results and qualitative listener feedback on pronunciation, fluency, and tone handling.

3.6.2 Model Selection

Of the four TTS models evaluated, two were selected for further testing based on their intelligibility, tone accuracy, and integration ease for Yoruba exam content.

- **Facebook MMS-TTS-Yor** is a neural TTS model released by Meta AI as part of the *Massively Multilingual Speech* project. It was trained on public Yorùbá Bible recordings and produced moderately natural, intelligible speech suitable for exam content (Pratap et al., 2023).
- **YarnGPT** is a lightweight, open-source Yorùbá TTS model hosted on Hugging Face (Azeez, 2025). Though less natural than MMS-TTS-Yor, it remained intelligible and was easy to integrate into the test interface.

3.6.3 System Architecture and Web Deployment

To make the models usable in a controlled testing environment, we built a web interface using Streamlit, a Python-based framework for building interactive data applications.

- Frontend Interface: Each question was presented one at a time along with audio playback controls. Yoruba ordinal numbering (e.g., Keta, Kewàá) was used to guide question progression.
- TTS Processing: For each question, the text was passed into the Facebook mms-tts-yor model via Hugging Face's AutoTokenizer and AutoModelForTextToWaveform classes. The resulting waveform was saved using the soundfile library and played directly in the browser.

The system is deployed on Streamlit Community Cloud (`https://yoruba-cbt-tts.streamlit.app/`), and the source code is available at `https://github.com/Moshood-Kausar/Yoruba_CBT_TTS`.

4 Results and Findings

This section presents the feedback from five visually impaired students who participated in the experiment. Each student engaged with the same set of Yoruba WAEC questions through three delivery methods: human voice, Braille, and Text-to-Speech (TTS) via a web-based interface. After the sessions, they were asked to identify which method they preferred and why.

4.1 Delivery Method Preferences

- All five students chose Braille as their top preference. It gave them a sense of control, quietness, and independence during the test.
- Three students found the human voice method helpful and clear but also pointed out that depending on someone else was less ideal. The remaining two preferred working alone.

Table 1: Mean Opinion Scores (MOS) of four tested Yoruba TTS models.

Yoruba TTS Model	MOS (1–5)	Remarks
MMS-TTS-Yor	3.7	Best overall; moderately natural
YarnGPT	3.2	Moderate
YorubaTTS	2.5	Partially intelligible
Tacotron 2	2.1	Poor clarity and hard to follow

Table 2: Preferences expressed by five students after using all three delivery methods.

Delivery Method	Preferred (n=5)	Key Feedback
Human Voice	3	Easy to understand but made students feel dependent. Two preferred not to rely on others.
Braille	5	Most comfortable and familiar. Allowed for full independence, no distractions, and no background noise.
TTS System (Facebook mms-tts-yor and YarnGPT)	1	Liked the idea of using technology but found the speech robotic and harder to understand.

- Only one student liked the idea of using TTS, mainly for its modern approach, but raised concerns about poor voice quality, mispronunciation, and tonal inaccuracies.

5 Discussion

The current generation of Text-to-Speech (TTS) models shows promising capabilities in high-resource languages, but remains largely inadequate for underrepresented languages like Yoruba, especially in educational contexts such as Computer-Based Testing (CBT) for visually impaired learners. While the use of TTS offers a scalable and digital alternative to traditional methods, this study reveals critical shortcomings in its ability to support comprehension, independence, and user comfort in real exam settings.

5.1 Gaps in TTS for Accessibility

The results indicate that existing TTS systems, particularly mms-tts-yor and YarnGPT, still struggle with producing clear and natural speech. All five participants preferred Braille over TTS, citing better independence, fewer distractions, and stronger understanding. The main issues with TTS included flat, robotic delivery, mispronunciation, and inconsistent tone handling.

Yoruba's tonal structure makes it more complex than many high-resource languages, a single word can have different meanings based on tone alone. Unfortunately, the TTS models tested were not trained with sufficient tonal context or high-quality Yoruba data, which made them unreliable in a CBT scenario where clarity is critical. This gap mirrors broader concerns about the digital marginalization of African languages in AI systems. Despite the rise of multilingual models, training data imbalances remain a major barrier to inclusive performance.

5.2 Technology vs. User Experience

From a usability perspective, the TTS web interface was functional, accessible, and easy to navigate. However, good design alone did not make up for the weak audio quality. While one student appreciated the idea of using technology for independent testing, the others found the synthesized speech difficult to follow. These findings reinforce the idea that accessibility is not just about adding speech output, it must be accurate, culturally aware, and linguistically appropriate. Visually impaired learners need tools that help them feel in control and confident, especially in exam settings. At present, the TTS systems tested fall short of delivering that experience.

6 Conclusion

This study examined how effective current Text-to-Speech (TTS) models are for delivering computer-based exam content in an underrepresented language, Yoruba, for visually impaired students. Using real WAEC Yoruba questions and feedback from five blind students, we compared the experience of listening to TTS-generated speech with human narration and Braille.

Our findings show that:

- **Braille was the most preferred method**, offering clarity, ease of navigation, and a strong sense of independence throughout the test.
- Human voice was partially helpful, but introduced dependency.
- TTS models, though promising, were not yet usable for CBT due to poor tone handling and robotic delivery.

The performance limitations of even the best available models (like Facebook's 'mms-tts-yor') reveal a clear accessibility gap in speech technology for low-resource languages. While tools like Braille remain reliable, they are not scalable in digital-first environments, making TTS a critical area for future development.

6.1 Future Work

There is still a significant gap between what current TTS systems can do and what visually impaired students actually need, especially in languages that are underrepresented in AI development. Moving forward, one key step is the creation of high-quality, open-source Yoruba speech datasets that reflect the natural rhythm, tone, and variation of the language. This would make it possible to train models that not only produce intelligible speech but also capture the linguistic richness that matters in real exam settings.

It will also be important to design and evaluate these systems with direct input from users. Many TTS models today are built and tested using automated metrics, but this study shows that actual student experience tells a very different story. Working closely with educators, accessibility centers, and Yoruba speakers can help ensure that future tools are truly usable in classrooms and testing environments. As these systems improve, there will also be a need to think about how they fit into educational policies and standardized testing frameworks. With the right support, TTS technology can become a reliable, inclusive tool not just for Yoruba, but for many languages that have been left behind in the digital space.

References

American Foundation for the Blind. 2025. Screen readers. Accessed: 2025-02-22.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Saheed Azeez. 2025. YarnGPT: Nigerian-accented English text-to-speech model.

Chika Chukwuma-Nosike and Peace Chukwuma. 2023. Computer-based test (CBT) evaluation innovation: Prospects in curriculum implementation in Nigeria. *Journal of Agriculture and Food Sciences*, 21(2). Special Issue: Man, Environmental Safety and Sustainability the Role of Research, Chapter 24, pp. 331–343.

Manpreet Kaur Dhaliwal and Rohini Sharma. 2024. Improving accessibility and independence for blind/visually impaired persons based on speech synthesis technology. *International Journal of Computer Applications*, 186(28):12–20.

David Eberhard, Gary Simons, and Chuck Fennig. 2019. *Ethnologue: Languages of the World*, 22nd edition. SIL International.

Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara E. Rivera, and Kólá Túbọ̀sún. 2020. Developing an open-source corpus of Yoruba speech. In *Proc. of Interspeech 2020*, pages 404–408, Shanghai, China.

ITU-T. 1996. Method for subjective determination of transmission quality. Technical Report Recommendation P.800, International Telecommunication Union - Telecommunication Standardization Sector. Retrieved from https://www.itu.int/rec/T-REC-P.800-199608-I/en.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Preprint*, arXiv:2106.06103.

National Federation of the Blind. 2009. *The Braille Literacy Crisis in America: Facing the Truth, Reversing the Trend, Empowering the Blind*. NFB Jernigan Institute, Baltimore, MD.

Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iroro Orife, and David Ifeoluwa Adelani. 2024. 'Ir'oy'inSpeech: A multi-purpose Yorùbá speech corpus. *Preprint*, arXiv:2307.16071.

Pawan Kumar Patel, Amey Karkare, and Gaurav Raheja. 2021. Inclusive accommodations for persons with visual impairments in computer-based tests. *Studies in Health Technology and Informatics*, 282:219–237.

Perkins School for the Blind. 2024. An overview of Braille devices. https://www.perkins.org/resource/overview-braille-devices/.

Vineel Pratap, Adithya Tjandra, Bowen Shi, Jing Huang, Qiantong Xu, Aravindh Krishnaswamy Babu, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Sirisha Rella. 2023. Exploring unique applications of text-to-speech technology.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, robust and controllable text to speech. *Preprint*, arXiv:1905.09263.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Kola Tubosun. 2023. Making machines speak Yorùbá.

Klá Túbsún and Adéday Olúòkun. 2017. Yorùbá text-to-speech system. https://www.ttsyoruba.com. Accessed: 2025-02-22.

waec. The west african examinations council. Accessed: 2025-02-22.

Carmen Willings. 2017. Educational assessments for students who are blind or visually impaired.