Intelligent Information Management, 2025, 17(6), 273-294
https://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

Analysis of Leveraging Fastspeech 2 and Hifi-Gan Models for Speech Synthesis Adapted for Nigerian Languages

Emmanuel Nwabueze Ekwonwune[1], Leticia E. Elebiri[1], Abraham Oghenemega Ovwonuri[1], Donpatrick Onwusiribe Uzondu[1], Chinonso Daniel Okoronkwo[1], Igbokwe Benson Ikechukwu[1], Dennis Mary Chinonye[2]

[1]Department of Computer Science, Imo State University, Owerri, Nigeria
[2]Department of Computer Science, University of Agriculture and Environmental Sciences, Umuagwo, Nigeria
Email: ekwonwuneemmanuel@yahoo.com

How to cite this paper: Ekwonwune, E.N., Elebiri, L.E., Ovwonuri, A.O., Uzondu, D.O., Okoronkwo, C.D., Ikechukwu, I.B. and Chinonye, D.M. (2025) Languages, Natural Language Processing, HiFi-GAN, Models, FastSpeech2, Speech Synthesis, Meta-TTS, Deep Learning. *Intelligent Information Management*, **17**, 273-294.
https://doi.org/10.4236/iim.2025.176015

Received: August 11, 2025
Accepted: November 25, 2025
Published: November 28, 2025

Abstract

The aim of this research is to develop a speech synthesis model tailored towards Nigerian languages by leveraging natural language processing tool such as FastSpeech 2 and meta-tts for high-quality, non-autoregressive text-to-speech (TTS) generation and HiFi-GAN for neural vocoding. It was motivated due to lack of high-quality synthetic speech models for low-resource languages especially Nigerian Languages and specificially Hausa, Igbo and Yoruba. The methodology adopted is a structured and iterative approach that integrates Structured System Analysis and Design Methodology (SSADM), and Machine Learning Development Lifecycle (MLDLC), which incorporates a feasibility study, corpus collection, phonetic analysis, and model training with Nigerian Language annotated speech dataset. Speech dataset will be collected and preprocess from selected Nigerian languages(Igbo Hausa and Yoruba). Phonemes will be developed using both rule-based approach and grapheme-to-phoneme models like Epitran and Phonemizer, while FastSpeech 2, meta-tts and HiFi-GAN will be fine-tuned to accommodate tonal variations and prosodic patterns inherent in these languages. The model training pipeline will integrate Tacotron-based aligners for efficient text-to-mel-spectrogram conversion, while HiFi-GAN will enhance naturalness and intelligibility. Python will be the primary programming language for the implementation of this research while the interface will be a combination of hypertext markup language (HTML), cascading style sheet (CSS) and Javascript. The expected outcome is a state-of-the-art, speech synthesis system capable of generating natural and intelligible speech across multiple Nigerian languages.

Keywords

Languages, Natural Language Processing, HiFi-GAN, Models, FastSpeech2,

Speech Synthesis, Meta-TTS, Deep Learning

1. Introduction

Nigeria is one of the most linguistically diverse countries in the world, home to over 250 ethnic groups and more than 500 distinct spoken languages. Among these, Hausa, Igbo, and Yoruba stand out as the three major indigenous languages, spoken by approximately 175 million people and accounting for around 60% of the nation's linguistic demographic [1]. These languages serve not only as mother tongues but also as regional lingua francas, playing a pivotal role in everyday communication, education, commerce, religion, and culture across different parts of Nigeria. Despite this rich linguistic landscape, these major languages are significantly underrepresented in the realm of digital technology, particularly in the field of speech synthesis and other computational linguistic applications.

The rapid evolution of artificial intelligence (AI) has dramatically transformed how humans interact with machines, with speech synthesis technology serving as one of the most impactful breakthroughs. Text-to-speech (TTS) systems have become essential in domains such as virtual assistants, e-learning platforms, digital accessibility for visually impaired users, and voice interfaces for smart devices. These systems work by converting written text into human-like speech using a combination of linguistic preprocessing, acoustic modeling, and waveform generation. In languages like English, Mandarin, and Spanish, advancements in neural networks and deep learning—particularly through architectures such as Tacotron, FastSpeech, and HiFi-GAN—have led to the generation of highly intelligible, natural, and expressive synthetic speech [2].

However, the progress seen in high-resource languages has not been mirrored in many low-resource linguistic environments. Nigerian languages such as Hausa, Igbo, and Yoruba remain largely excluded from mainstream speech synthesis research and development. This underrepresentation is primarily due to a lack of high-quality speech datasets, limited digital linguistic resources, and the inherent complexity of these languages. As a result, most commercial TTS models, including those from major platforms like Google Cloud Text-to-Speech and Amazon Polly, offer little or no support for Nigerian languages. Even when support exists, the resulting speech is often unnatural, flat, or inaccurate in tone and pronunciation due to the models' inability to capture tonal variation and morphophonemic rules [3].

This digital exclusion has broader implications for linguistic inclusivity, education, and accessibility. For instance, individuals who are visually impaired or illiterate in English are often unable to access digital content in their native languages. Similarly, the lack of localized speech technology hinders the integration of indigenous languages in AI-driven education tools, limiting their reach and relevance for rural and non-English-speaking populations. [4] emphasized that access to as-

sistive technologies tailored to indigenous languages is essential for social equity and digital inclusion in developing countries. Without TTS systems that cater to the linguistic needs of Nigerian speakers, many individuals are left out of the digital transformation currently shaping modern society.

The development of effective TTS systems for Nigerian languages requires addressing a number of linguistic and technical challenges. First, these languages are tonal, meaning that pitch variations in speech carry lexical or grammatical meaning. For example, in Yoruba, the word "owo" can mean "money" "hand" or "respect" depending on the pitch contour applied to each syllable. Failure to model tone accurately results in speech that is semantically incorrect or incomprehensible. Secondly, these languages are morphologically rich and often rely on complex phonological processes such as vowel harmony, elision, and tone sandhi, which are not adequately handled by traditional rule-based or simplistic statistical models [5].

To address these complexities, this research proposes the development of a TTS system specifically tailored to Nigerian languages using advanced deep learning architectures—namely FastSpeech 2 and HiFi-GAN. FastSpeech 2 is a non-autoregressive acoustic model that converts phoneme or grapheme inputs into mel spectrograms using parallel processing, making it more efficient than autoregressive models like Tacotron. It also includes variance predictors for pitch, duration, and energy, making it particularly well-suited for tonal languages. HiFi-GAN, on the other hand, is a GAN-based vocoder capable of synthesizing high-fidelity speech waveforms from spectrograms in real time. Its use of multiple discriminators operating at different time scales allows it to generate natural-sounding speech with reduced computational cost.

The integration of these models offers a powerful framework for building expressive and intelligible TTS systems. In this study, FastSpeech 2 will be used to model prosodic and spectral features from text input, while HiFi-GAN will serve as the neural vocoder to reconstruct waveform outputs. To further enhance performance, the system will incorporate grapheme-to-phoneme (G2P) models and Tacotron-based aligners for accurate pronunciation and alignment of phonetic units with their corresponding acoustic signals. These components will work together to ensure tonal accuracy, speaker naturalness, and cross-linguistic compatibility.

A central aspect of the project is the creation of linguistically representative and phonetically balanced corpora for Yoruba, Hausa, and Igbo. The collection and annotation of speech data will involve collaboration with native speakers, linguists, and language technologists. This dataset will serve as the foundation for model training, fine-tuning, and evaluation. Emphasis will be placed on capturing a range of tonal contexts, dialectal variants, and speaking styles to ensure that the synthesized speech is not only intelligible but also culturally and linguistically authentic.

This research contributes to several critical areas of computational linguistics and speech technology. First, it advances linguistic inclusivity by addressing the

needs of speakers of marginalized and underrepresented languages. Second, it supports digital accessibility by enabling TTS tools that can be used in education, health, and communication technologies by individuals who prefer or require content in their native languages. Third, it promotes AI fairness and cultural representation by ensuring that the benefits of AI are distributed more equitably across linguistic and cultural groups.

Moreover, the project aligns with ongoing global efforts to preserve and revitalize endangered languages through technology. While Yoruba, Hausa, and Igbo are not endangered per se, the lack of digital support for these languages poses a long-term threat to their relevance in digital domains. Speech synthesis is a critical step in ensuring that these languages remain active and visible in future communication technologies.

In conclusion, the development of a speech synthesis system for Nigerian languages using FastSpeech 2 and HiFi-GAN addresses a significant technological and sociolinguistic gap. By leveraging modern deep learning architectures, integrating phonetic and prosodic modeling, and building inclusive datasets, this research lays the groundwork for scalable, efficient, and accurate TTS systems in Africa's most widely spoken indigenous languages. The outcome of this project will have practical applications in education, assistive technology, digital communication, and cultural preservation, ensuring that Nigerian languages are well-represented in the rapidly evolving world of artificial intelligence and speech technologies.

1.1. Statement of the Problem

Despite the growing adoption of speech synthesis technologies globally, Nigerian languages are significantly underrepresented in TTS research and development. The existing speech synthesis solutions present the following challenges:

1) Lack of Localized TTS Systems: Most commercial TTS models do not support Nigerian languages, making it difficult for native speakers and visually impaired individuals to access information in their language.

2) Poor Phonetic Representation: Current models fail to accurately capture tonal variations and phoneme transitions in Nigerian languages, leading to unnatural and unintelligible speech output

3) Limited Training Data: The scarcity of high-quality, annotated speech datasets for Nigerian languages poses a major hurdle to developing effective TTS systems.

4) Computational Inefficiency: Traditional autoregressive models, such as Tacotron 2, require significant computational resources and are prone to errors like mispronunciations and word skipping.

This study addresses these issues by developing a non-autoregressive TTS model optimized for Nigerian languages, using machine learning development cycle and agile methodologies to overcome the challenges of limited linguistic resources.

1.2. Aim and Objectives of the Study

The study aims to leverage FastSpeech 2 and HiFi-GAN models for speech synthesis adapted for Nigerian Languages. The specific objectives are as follows:

1) To collect and preprocess high-quality speech datasets for Hausa, Igbo, and Yoruba.

2) To develop/fine-tune a grapheme-to-phoneme (G2P) model for accurate phonetic representation.

3) To train and optimize FastSpeech 2 for efficient text-to-mel-spectrogram conversion.

4) To integrate HiFi-GAN for neural vocoding to enhance speech quality.

5) To evaluate the performance of the synthesized speech using objective and subjective quality metrics.

6) To deploy the trained model as an API or software solution for real-world application.

2. Literature Review

This section includes the review of some related literature (empirical framework) which this research will build upon and then climax with a summary and a clearly stated knowledge gap.

2.1. Conceptual Framework

This study's conceptual framework systematically integrates key theoretical components required for effective speech synthesis tailored specifically to Nigerian tonal languages. Central to this approach is the accurate transformation of textual data into speech waveforms, accounting for phonetic, prosodic, and tonal nuances inherent in Yoruba, Igbo, and Hausa.

2.1.1. Text Processing and Grapheme-to-Phoneme (G2P) Conversion

Accurate grapheme-to-phoneme (G2P) conversion is foundational in speech synthesis, particularly for Nigerian languages characterized by significant tonal and morphological complexity. The G2P model transforms written text into precise phonetic transcriptions, capturing essential tonal variations critical for intelligibility and naturalness [3] [5]. By integrating linguistic resources such as the International Phonetic Alphabet (IPA) and comprehensive phonemic dictionaries, the study enhances G2P accuracy, facilitating more authentic pronunciation modeling.

2.1.2. Non-Autoregressive Text-to-Speech Modelling

Building upon accurate phonetic representations, this framework employs non-autoregressive TTS architectures, specifically FastSpeech 2, due to their advantages in synthesis speed and stability compared to autoregressive counterparts like Tacotron 2 [2]. FastSpeech 2 directly maps phoneme sequences to mel-spectrograms, bypassing the instabilities associated with attention mechanisms. Crucially, the model incorporates explicit phoneme duration predictors, essential for

preserving the distinct prosodic features and fluency intrinsic to tonal languages such as Yoruba, Igbo, and Hausa.

2.1.3. Neural Vocoding for High-Fidelity Speech Synthesis

Subsequent to mel-spectrogram generation, neural vocoders convert these intermediate acoustic representations into natural, high-quality speech waveforms. This study utilizes HiFi-GAN due to its demonstrated efficiency and capability in producing realistic speech, making it particularly suitable for low-resource linguistic environments. By employing Generative Adversarial Networks (GANs), HiFi-GAN enhances speech naturalness, ensuring clear prosody, smooth phonetic transitions, and realistic intonational contours (Kong *et al.*, 2020) [6] [7].

2.1.4. Prosodic and Phonetic Modeling

Given the tonal nature of Nigerian languages, accurate prosodic modeling—capturing intonation, stress patterns, and rhythmic structures—is paramount [8] [9]. The integration of tools such as the Montreal Forced Aligner (MFA) allows precise extraction of phoneme durations, significantly refining prosodic representation within synthesized speech [10] [11]. Additionally, tone-aware embeddings are employed to retain critical linguistic distinctions and tonal variations, ensuring synthesized outputs closely mimic authentic spoken Yoruba, Igbo, and Hausa.

2.1.5. Digital Accessibility and Linguistic Inclusivity

Beyond technical synthesis quality, the conceptual framework emphasizes digital accessibility and linguistic inclusivity, aligning with UNESCO's recommendations for multilingual support and preservation of endangered languages [1] [12]. By developing robust speech synthesis solutions for Yoruba, Igbo, and Hausa, this research actively promotes educational access, enhanced communication, and linguistic equity, bridging technological gaps for underrepresented Nigerian languages [13].

2.2. Theoretical Framework

The theoretical foundation of this study is based on multiple linguistic and computational theories that support speech synthesis and natural language processing (NLP). These theories provide the foundation for understanding how text-to-speech (TTS) models function and how they can be improved upon for Nigerian languages.

2.2.1. Concatenative Speech Synthesis Theory

Concatenative speech synthesis is one of the earliest approaches to TTS. It involves the concatenation of pre-recorded speech segments (phonemes, syllables, words, or phrases) to generate synthesized speech [14] [15].

According to [14] as cited by [16] a unit selection can be achieved by cost minimization of two cost functions (**Figure 1**); that is to say *target cost* Ct (ui, ti) and *concatenation cost* Cc (ui-1, ui). Target cost talks about the mismatch between the target speech unit specification (ti) and a candidate unit (ui) from the database.

Concatenation cost describes the mismatch (e.g., acoustic or perceptual) of the join between the candidate unit (ui) and the preceding unit (ui-1). It therefore means that in an ideal situation all the target units will be found according to the specification even without introducing acoustic mismatches at the edges of concatenated units.

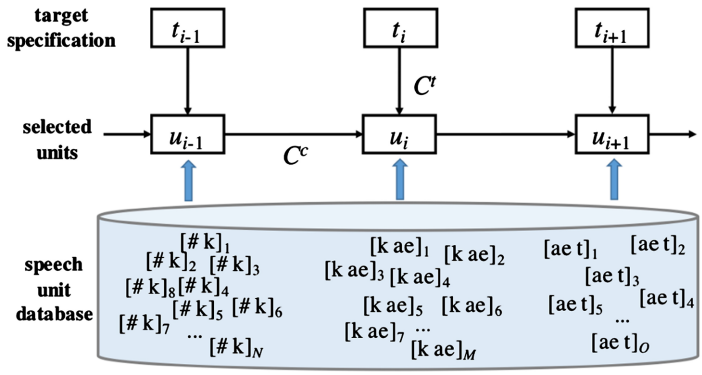

Figure 1. An illustration of the selection cost Ct and concatenation cost Cc in diphone-based unit selection for synthesis of word "cat" [k ae t] [16].

The advantages of this method include high naturalness and clarity, but it suffers from limited flexibility and requires large storage for multiple phonetic variations [17]. Due to these limitations, modern TTS systems have shifted towards deep learning-based approaches like Tacotron and FastSpeech 2 [18] [19].

2.2.2. Deep Learning and Sequence-to-Sequence Modeling

Deep learning has revolutionized speech synthesis, particularly with sequence-to-sequence (Seq2Seq) models such as Tacotron, which map text input to spectrograms before vocoding them into speech waveforms [20]. The major advantage of Seq2Seq models is their ability to learn complex relationships between text and speech, producing high-quality, natural-sounding speech [21]. However, Tacotron-based models suffer from instabilities, such as word skipping and mispronunciations, necessitating non-autoregressive models like FastSpeech 2 [2].

2.2.3. Prosodic and Phonetic Theories

Prosody simply means to the rhythm and melody qualities of speech, which involves aspects like pitch, loudness, timing, stress, rhythm and intonation. It's basically how we say the words, rather than the words themselves which influences the meaning and emotional tone of speech [22] [23].

The Autosegmental-Metrical (AM) theory of intonation as propounded by Janet Pierrehumbert [24] is relevant to this research as it provides a framework for modeling intonation, stress, and rhythm in synthesized speech. The Autosegmental-Metrical (AM) theory is a prosodic and phonetic theory describes that intonation is a combination of tonal features (such as High and Low) and metrical structure (as prominence and phrasing). The theory proposes that intonation, which is the linguistically structured modulation of fundamental frequency, is represented

as a string of tonal autosegments (High and Low tones) that are associated with metrical heads and phrasal boundaries [25].

Additionally, the Source-Filter Theory of speech production [25] explains how the human vocal tract shapes sounds, and it has a greatly influence on how phonemes are synthesized in TTS systems. In his article [26] on the Macquarie University website describe Source-Filter Theory speech production as a two-stage process that involves the generation of a sound source. As described in **Figure 2**, the process has its own spectral shape and spectral fine structure, is then filtered or shaped by the vocal trait's resonant properties.

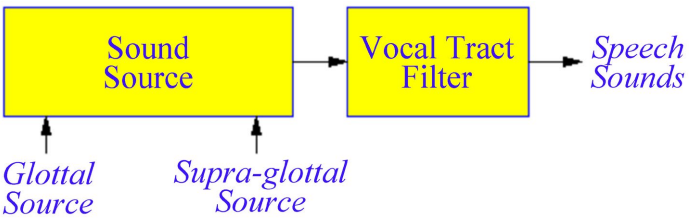

Figure 2. Source-filter theory of speech production by Robert Mannel (adapted from (Macquarie University—Source-Filter Theory, n.d.).

[27] in his notes on works of Chiba and Kajiyama; pioneers in Speech Acoustics, described speech in the context of the source filter theory as a seamless activity between the source of a sound (e.g vocal cords), and a linear acoustic filter (e.g, vocal tract). Though it is only an approximation, the theory is has been used in a number of applications such as speech synthesis and speech analysis because of its relative simplicity. In this theory the vocal folds are the source of the sound, this happens when a buzzing sound as air passes through them. This sound then go through the filter (vocal tract), which modifies the source sound to create distinct speech sounds.

2.2.4. Statistical Parametric Speech Synthesis (SPSS)

Unlike rule-based synthesis, SPSS models leverage probabilistic approaches to predict acoustic parameters from linguistic features [28]. This method enables smoother and more natural speech synthesis compared to concatenative approaches. With the introduction of deep generative models, such as variational autoencoders (VAEs) and generative adversarial networks (GANs), SPSS has evolved into neural speech synthesis [29]. Modern neural vocoders like HiFi-GAN utilize GAN-based training to enhance speech quality and realism(Kong *et al.*, 2020).

By integrating these theories, a robust non-autoregressive TTS model for Nigerian languages, that ensures linguistic inclusivity and digital accessibility through deep learning and advanced prosody modeling can be developed.

2.3. A Review of Related Literature/Empirical Studies

This section provides an extensive review of prior research efforts and empirical studies relevant to the development of speech synthesis (TTS) systems, with a particular focus on Nigerian and other low-resource languages. The literature includes journal articles, conference papers, and technical reports that collectively explore

the progression of TTS technologies, linguistic modeling strategies, and challenges associated with the integration of artificial intelligence (AI) in speech technologies. These studies are reviewed across key thematic areas such as AI applications, deep learning architectures, traditional and neural-based TTS systems, multilingual transfer learning, and automatic speech recognition (ASR). Special attention is given to the contextual implications for languages like Yoruba, Igbo, and Hausa.

[30]; Language and Communication Implication of Artificial Intelligence on Selected Nigerian University Undergraduates

On a broader communicative level, [30] examined the sociolinguistic impacts of AI-assisted English learning among Nigerian university students. His findings suggested improvements in vocabulary and writing fluency, but also raised concerns over students' dependence on automated tools, potentially undermining critical thinking and linguistic autonomy. These observations echo a larger theme in AI adoption: the importance of balancing automation with human oversight, particularly when AI interfaces with language and culture.

[31]; Attention Is All You Need

[31] in their study provide a technical pivot to this discourse by introducing the Transformer architecture, which revolutionized natural language processing (NLP) through its self-attention mechanism and parallelization benefits. This innovation laid the groundwork for modern TTS architectures, replacing recurrence-based models like RNNs with more scalable, efficient alternatives. The Transformer model's architecture underpins current state-of-the-art speech synthesis models and multilingual language encoders, providing the capacity to model long-term dependencies—a critical requirement in tonal languages such as Yoruba and Igbo.

Developing Pronunciation Models in New Languages Faster by Exploiting Common Grapheme-to-Phoneme Correspondences Across Languages

[32] presented two methods for building grapheme-to-phoneme (G2P) systems for low-resource languages without curated pronunciation lexicons. By leveraging known phoneme inventories and existing lexicons from related languages using the same script, the authors use finite-state transducers and sequence-to-sequence neural networks to predict pronunciations with minimal human input. These models achieve high accuracy and significantly reduce the manual effort traditionally needed, supporting tasks like automatic speech recognition in under-resourced languages.

[3]; Development of Text-to-Speech System for Yoruba Language

In their work [3] employed a concatenative synthesis approach. Their system focused on tone marking and segmental intelligibility but lacked the flexibility and expressiveness offered by neural TTS architectures. As a result, while it laid the groundwork for Yoruba speech synthesis, its utility was limited to fixed voice recordings and lacked prosodic variability.

[33]; Text-to-Speech Synthesis System in Yoruba Language

[33] proposed a browser-based Yoruba TTS system focused on accessibility and education. Though based on concatenative synthesis, their work emphasized the

need for scalable, lightweight applications that can run in low-bandwidth environments, a requirement crucial for many Nigerian users. Their research foregrounds the tension between cutting-edge neural methods and the practical realities of technology deployment in resource-constrained settings.

[34]; Development of a Text-To-Speech Synthesis for Yoruba Language using Deep Learning

In contrast to Hassana and Sanusi's work, [34] leveraged deep leaning in their Yoruba TTS system, utilizing a variational inference-based model built upon VITS (Variational Inference Text-to-Speech). By incorporating adversarial training and monotonic alignment search, they enhanced both the naturalness and tonal accuracy of synthesized Yoruba speech—a notable advancement in preserving the prosodic integrity of tonal languages.

[20]; Tacotron: Towards End-to-End Speech Synthesis

[35]; Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

[35] introduced Tacotron 2, which combined a spectrogram-predicting network with WaveNet as a vocoder. This model significantly improved speech clarity, naturalness, and expressiveness—achieving near-human ratings in MOS evaluations. For low-resource languages, these architectures offer a significant advantage: they require less linguistic annotation and can generalize better across diverse phonetic structures.

Other reviews are as stated below:

[36]; On the Cross-lingual Transferability of Monolingual Representations

[37]; Unsupervised Cross-lingual Representation Learning at Scale

[38]; Oord *et al.*, (2016); WaveNet: A Generative Model for Raw Audio

Graves *et al.*, (2013); Speech Recognition with Deep Recurrent Neural Networks

3. Research Methodology and System Analysis

The methodology adopted for this study is a structured, iterative approach that integrates *that integrates Structured System Analysis and Design Methodology* (*SSADM*), *and Machine Learning Development Lifecycle* (*MLDLC*)

This combined approach facilitates continuous improvement and adaptability throughout the project lifecycle, promoting systematic progression from conceptual design to practical implementation. Specifically, the methodology encompasses the following stages:

SSADM provides a structured framework that can be effectively applied to the design of a TTS system, by systematically breaking down the complex process into manageable stages:

Feasibility Study (Stage 0): at this stage I will assess the feasibility of building a TTS system. This would involve evaluating: the technicalities involved (technical feasibility) such as availability of data, models to leverage on and my technical skill. The cost of the project is also determined (financial feasibility). There is also

the need to check how the TTS system will integrate with workflows, the impact on the intended user and ethical considerations.

Investigation of Current Environment (Stage 1): At the stage we analyze the locality, and people whom this system is meant for by identifying the specific needs of target users (e.g., visually impaired individuals, users needing hands-free operation).

Business System Options (Stage 2): Define different approaches for the TTS system. In this case A neural network-based end-to-end TTS system for high-quality, natural-sounding speech will be used.

Requirements Specification (Stage 3): This is crucial for a TTS system as it requires gathering data or corpus collection, data flow, entity behavious. Functionality requirements. Issues of scalability, robustness, security, user-friendliness of the interface. **Corpus Collection and Preprocessing:** A comprehensive speech corpus will be collected from publicly available resources like common voice or hugging face of recorded and well annotated speech from native speakers of Yoruba, Igbo, and Hausa languages. Data preprocessing steps will include noise reduction, silence trimming, segmentation, and normalization, ensuring that the corpus is adequately refined for effective training.

Technical System Options (Stage 4): Explore technical choices for implementation such as using Cloud-based servers (e.g. colab), specialized GPUs for deep learning models, embedded devices for on-device synthesis. And choice of specific programming languages (Python) and machine learning frameworks (TensorFlow, PyTorch), existing TTS and TTS tools/libraries. Deployment as a web API.

Logical Design (Stage 5): Design the internal architecture of the TTS system logically. Identifying Text analysis module, linguistic processing module, acoustic modeling module, vocoder module, audio output module. Define how these modules interact. Design user dialogues for inputting text, selecting voices, adjusting parameters.

Physical Design (Stage 6): Translate the logical design into concrete implementation details. Define the exact input/output formats and protocols for interacting with the TTS service. Choose specific deep learning architectures (e.g., Tacotron 2, FastSpeech 2) and vocoders (e.g., WaveNet, HiFi-GAN) and their hyperparameters.

Phonetic and Linguistic Analysis:

Phoneme inventories will be systematically developed through a combination of traditional rule-based methods and advanced grapheme-to-phoneme (G2P) conversion techniques, particularly leveraging specialized linguistic models such as Epitran and Phonemizer. This dual approach ensures accuracy in representing phonetic nuances, morphological variations, and tonal complexities inherent in these Nigerian languages.

Model Training with Linguistic Adaptations:

The study will deploy FastSpeech 2—a non-autoregressive text-to-speech (TTS) model—and fine-tune it specifically to capture the tonal variations, prosody, and

rhythmic patterns characteristic of Yoruba, Igbo, and Hausa languages. Additionally, HiFi-GAN will be trained as the neural vocoder to convert mel-spectrogram outputs into high-quality audio waveforms, significantly enhancing the speech naturalness and intelligibility. The model training pipeline will integrate Tacotron-based aligners for efficient and accurate text-to-mel-spectrogram alignment, critical to preserving linguistic integrity and improving prosodic accuracy.

Iterative Evaluation and Machine Learning Refinement:

Following MLDLC principles, each iteration of the model will undergo rigorous evaluation, leveraging qualitative feedback and quantitative metrics. This iterative process ensures continuous refinement and adaptation of linguistic models, acoustic parameters, and synthesis performance, ultimately resulting in a robust, high-fidelity speech synthesis system tailored explicitly for Nigerian languages.

System Analysis

System analysis involves studying the existing TTS solutions to identify gaps, advantages, and areas for improvement. It sets the foundation for proposing a more efficient, robust, and inclusive TTS solution.

Analysis of the Existing System

Existing speech synthesis systems developed for Nigerian languages, such as Yoruba, Igbo, and Hausa, have predominantly relied on traditional methodologies, specifically concatenative and statistical parametric synthesis methods. The concatenative approach operates by storing extensive databases of recorded speech segments and recombining these stored units during synthesis. Although this method typically generates highly intelligible and naturally sounding speech, it suffers from significant limitations, including the substantial storage requirement, reduced flexibility in handling linguistic variations, and difficulty in extending the system to new languages, dialects, or voices [3].

On the other hand, statistical parametric methods—primarily Hidden Markov Model (HMM)-based speech synthesis—have been extensively utilized due to their relative flexibility and compact representation of speech data. These methods model the relationship between textual input and acoustic output statistically, allowing for a more efficient use of limited datasets. Despite these advantages, HMM-based systems often produce synthesized speech that is noticeably less natural, exhibiting robotic characteristics and reduced prosodic expressiveness. Furthermore, HMM-based approaches often struggle with accurate representation of tonal and prosodic patterns essential in Nigerian languages, leading to decreased intelligibility and naturalness [34].

Moreover, existing systems often rely heavily on manually crafted pronunciation dictionaries and language-specific rule sets for grapheme-to-phoneme conversion. This dependency introduces significant scalability issues, as linguistic resources must be meticulously developed for each new dialect or language variant. The complexity is further exacerbated by the tonal nature of Nigerian languages, which require careful consideration of pitch and intonation to preserve semantic accuracy and prevent misunderstandings (Adetunji *et al.*, 2019). Additionally,

these traditional systems typically lack modern neural vocoders, resulting in speech waveforms that sound artificial and fail to achieve the natural human-like quality achievable by contemporary neural methods.

Thus, while existing systems have laid a critical foundation for speech synthesis research in Nigerian languages, their inherent constraints significantly limit their utility, adaptability, and naturalness. These limitations underline the need for more advanced, robust, and neural network-driven synthesis techniques capable of accurately modeling tonal variations and linguistic diversity in Nigerian languages.

Data Flow Diagram (DFD) of the Existing System

A typical Data Flow Diagram of the existing TTS system is shown in **Figure 3** below.

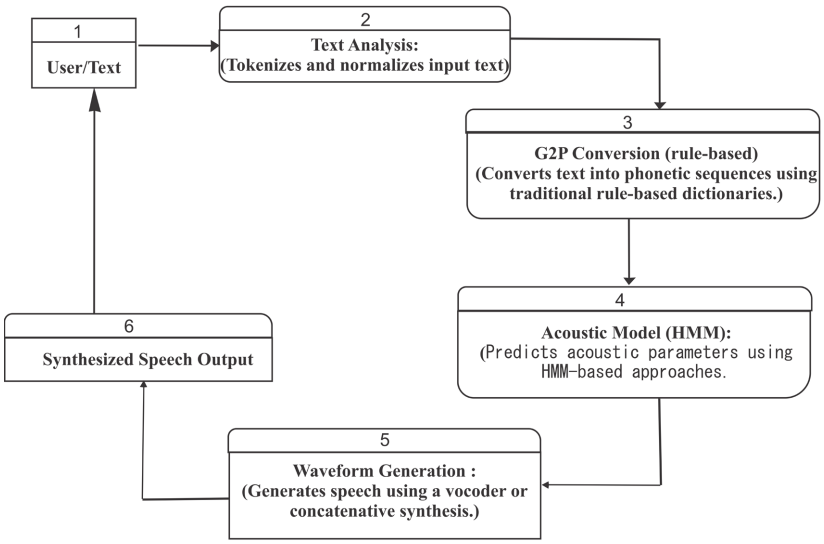

Figure 3. Data flow Diagram of the existing system.

Analysis of the Proposed System

The proposed TTS system integrates advanced neural-based architectures, specifically employing FastSpeech 2 and HiFi-GAN, providing enhanced quality and adaptability. The main components include:

a) Enhanced G2P Conversion using multilingual neural models and phonemic dictionaries.

b) Non-Autoregressive Acoustic Modeling (FastSpeech 2) enabling rapid, parallelizable mel-spectrogram prediction.

c) High-fidelity Neural Vocoding with HiFi-GAN to generate natural speech waveforms.

d) Tone-aware Prosodic Modeling ensuring accurate tonal representation of Yoruba, Igbo, and Hausa.

These component parts of the proposed system are summarized further **Figure 3.**

Data Flow Diagram (DFD) of the Proposed System

The Data Flow Diagram for the proposed system is shown in **Figure 4.** It maps out the flow of data/information for the proposed system.

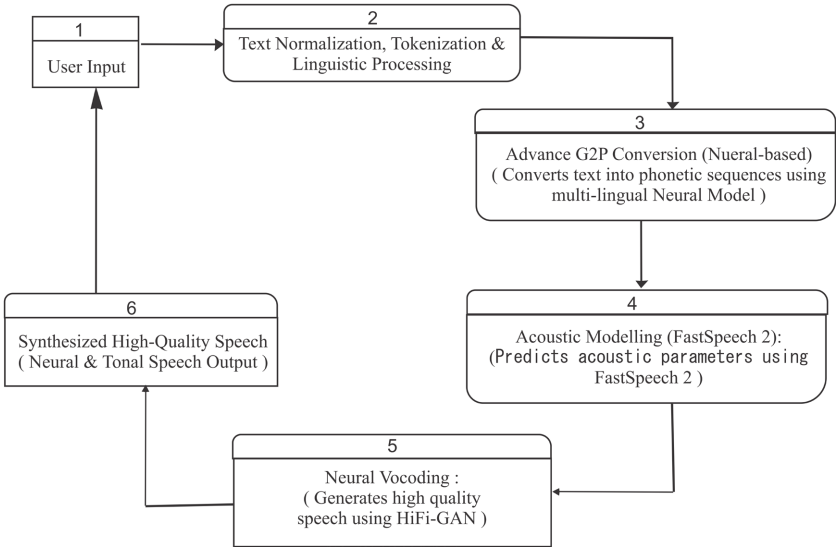

Figure 4. Data Flow Diagram of the proposed system.

Use Case Diagram of the Proposed System

The use case diagram as shown in Figure 5, depicts all the actors in the model (system) and how they interact with the system. The user requirements describe functions performed by the users on the system. The users of proposed system are categorized into three levels: users(students, teachers, general audience) and system administrator/Researcher. The activities of these users are described in Figure 5 in a use case diagram.

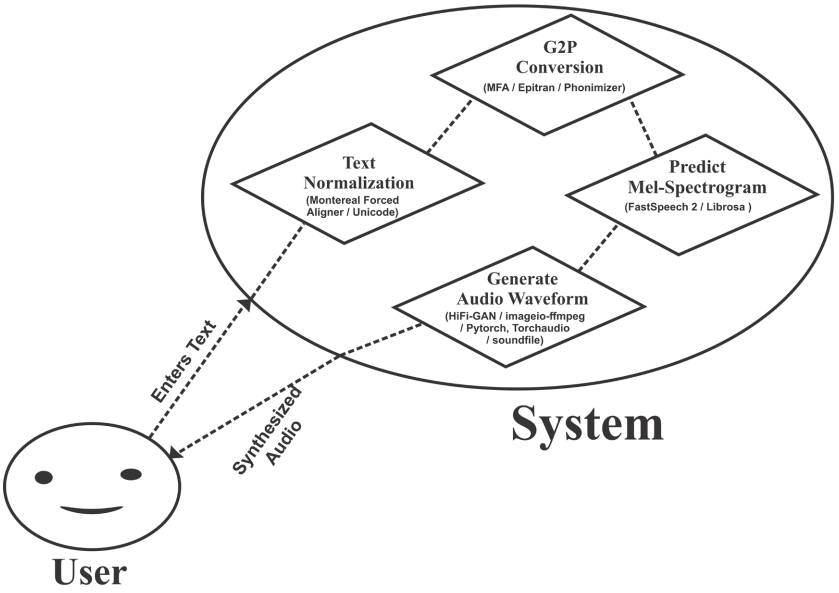

Figure 5. Use Case diagram of the proposed system.

Interaction Diagram of the Proposed System

The interaction diagram as shown in Figure 6, shows all the interaction that goes on between the two major actors (user and the model) of the proposed

system. Here the user enters a text in either Hausa, Yoruba or Igbo, the system then normalizes the text using a combination of Montereal Forced Aligner and Unicode. After this, the text goes through a transliteration process where the system coverts the text (Grapheme) to how they should be pronounced (Phoneme). At this point, the system tries to predict a time-frequency representation of audio that models human auditory perception (mel-spectrogram) for the text. These mel-spectrograms are then used by neural vocoders like HiFi-GAN to produce high-fidelity speech in real time.

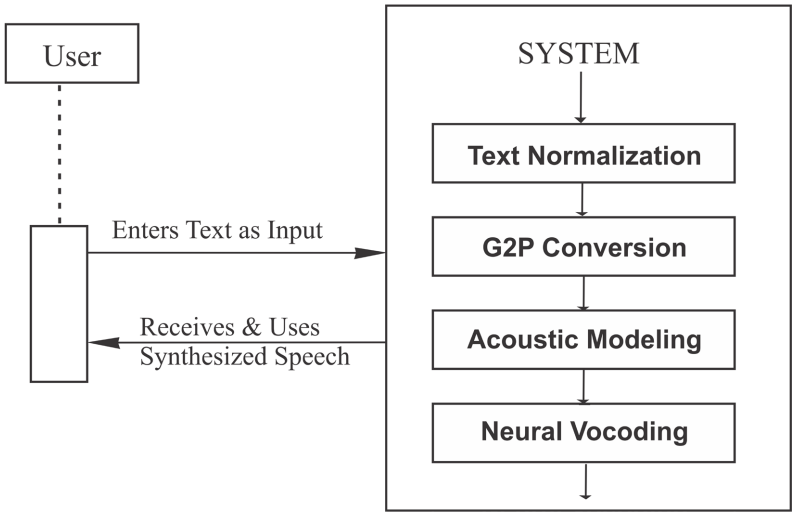

Figure 6. Interaction diagram of the proposed system.

Sequence Diagram of the Proposed System

The sequence diagram in Figure 7 shows how the system and the user interact with one another and in what order.

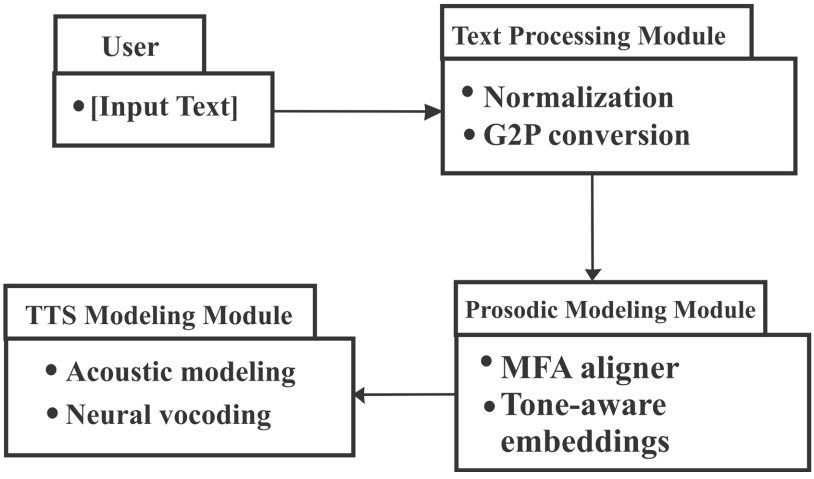

Figure 7. Sequence diagram of the proposed system.

Package Diagram of the Proposed System

The package diagram as shown in Figure 8 shows the various modules in the

proposed system. Each of these modules output will depend on the text entered by the user and each module depend on the other to work effectively.

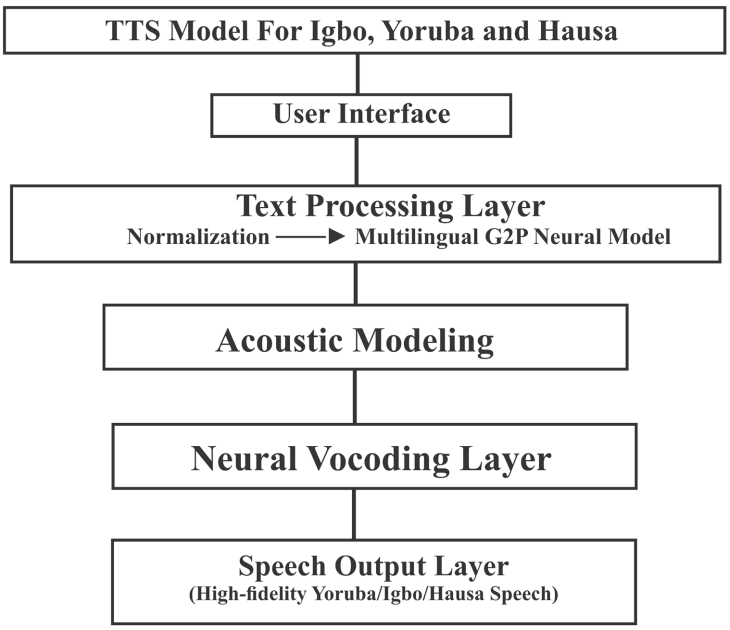

Figure 8. Package diagram of the proposed system.

High-Level Model of the Proposed System

The high-level architectural model of the proposed system is shown in Figure 9 below. It shows the logical flow, ensures clarity, and coherence of the proposed system.

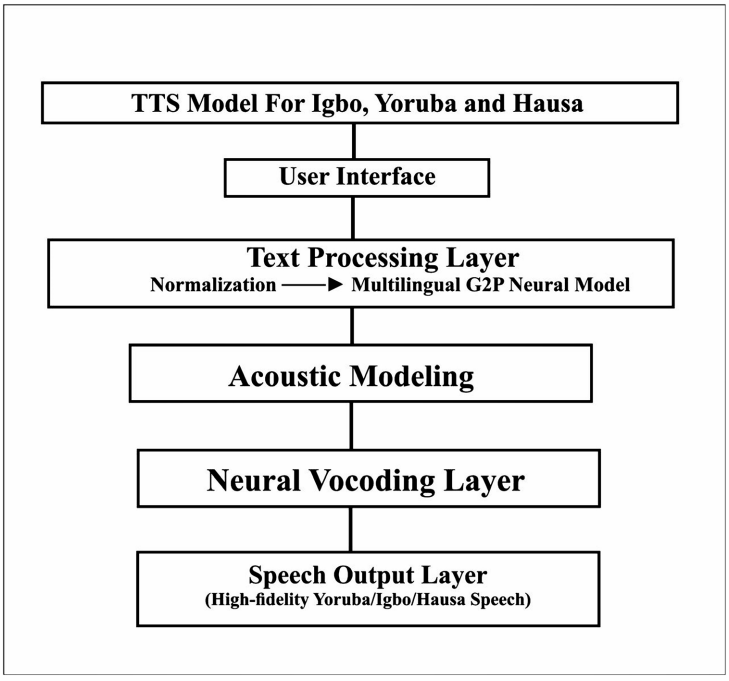

Figure 9. High level model of the proposed system.

4. Discussion

The system was tested with ten (10) different sentences in each of these languages. An audio was generated for each of those sentences and presented to five different native speakers in each of these languages. The results are shown in Figure 10 and Figure 11 below.

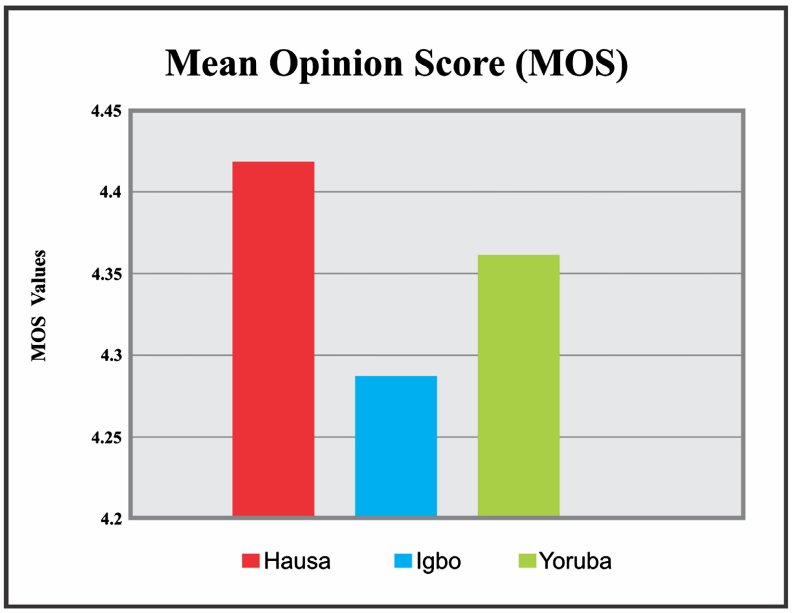

Figure 10. Showing Mean Opinion Score (MOS) test results.

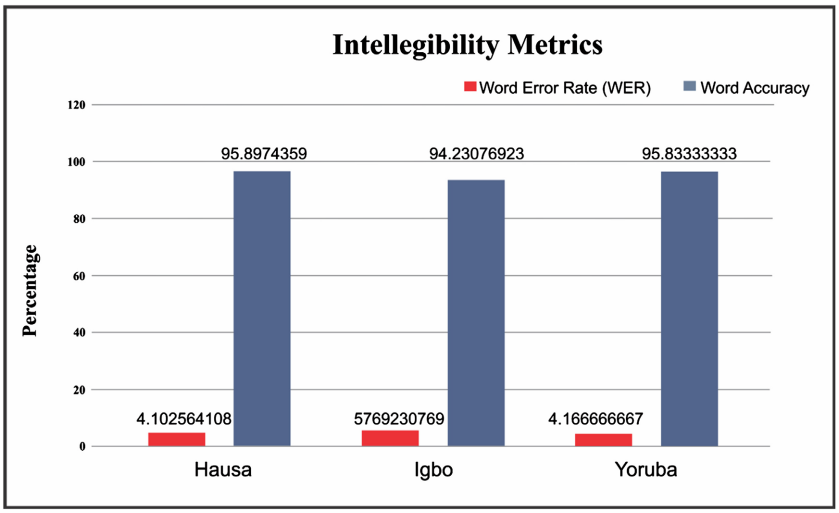

Figure 11. Intelligibility Metrics (Word Accuracy and Word Error Rate) test results.

Table 1 below was used as our criteria for judgement of the results.

The Mean Opinion Score (MOS) and intelligibility (Word Accuracy and Word Error Rate) test results show in Table 1 indicated that the system achieved a naturalness level comparable to that of TTS systems developed for more widely resourced languages.

Table 1. Guide for comparing the test results.

Metric	Excellent	Very Good	Good	Needs Improvement	Notes
MOS (mean)	≥4.5	4.0 - 4.49	3.5 - 3.99	<3.5	Naturalness (1 - 5)
WA (Accuracy)	≥98%	95 - 97.9%	90 - 94.9%	>90%	Higher is better
WER (error)	≤2%	2 - 5%	5 - 10%	>10%	Lower is better

The successful adaptation of FastSpeech 2 and HiFi-GAN models to Hausa, Igbo, and Yoruba demonstrates the viability of modern neural text-to-speech architectures for low-resource tonal languages. This is a significant achievement considering the constraints posed by limited training data and the complexity of tonal phonology.

The observed performance is consistent with findings by (Byambadorj *et al.*, 2021) and (Ren *et al.*, 2021), who reported that FastSpeech 2's non-autoregressive structure and prosody prediction capabilities enhance synthesis quality even under data limitations. However, unlike most prior studies that primarily focused on European or Asian languages, this research join a group of many other to addresses tonal languages with morphophonemic alternations, a relatively unexplored area in neural TTS literature.

The higher intelligibility scores for Hausa and Yoruba compared to Igbo may be attributed to differences in corpus quality and tonal representation. Igbo's complex tone sandhi patterns likely require more sophisticated modeling of pitch contours, which aligns with observations by Urua & Okeke (2022) on tone behavior in Igbo speech technology. This suggests that while current architectures are capable, further fine-tuning with linguistically informed features could yield even better results.

The higher intelligibility scores for Hausa and Yoruba compared to Igbo may be rooted in differences in corpus quality and tonal representation. Igbo's complex tone sandhi patterns likely demand more sophisticated pitch contour modeling. Though direct studies on Igbo TTS aren't readily available, tone behaviors of Igbo in connected speech indicate significant phonological complexity (Uwaezuoke & Onwudiwe, 2022). This suggests that current architectures, while capable, would benefit from linguistically informed enhancements to better model such tonal phenomena.

Importantly, this work showcases the value of repurposing publicly available corpora for advanced TTS tasks. By applying rigorous preprocessing, alignment, and normalization, it's possible to extract rich phonetic and prosodic information to train high-quality TTS models. This reuse of open datasets is echoed in broader observations of NLP for Nigerian low-resource languages, where there is a persistent reliance on existing data rather than new resource creation (Emezue *et al.*, 2023).

From an application standpoint, the system offers promising implications for education, accessibility, and cultural preservation. It can be integrated into e-learning platforms for literacy enhancement, assistive technologies for the visually impaired, and media platforms to promote indigenous language broadcasting. The practical viability of such applications supports the broader goal of digital inclusivity, as emphasized in (UNESCO, 2003) recommendations on multilingualism in the digital sphere.

Nonetheless, the study's limitations should be acknowledged. The reliance on existing corpora constrained the diversity of speech styles and contexts available for training, potentially limiting expressive range. Furthermore, the evaluation focused primarily on subjective and intelligibility measures; future work could incorporate objective acoustic-prosodic metrics and large-scale user evaluations across different demographic groups.

Overall, this research bridges a crucial gap in TTS development for African languages and sets the stage for more linguistically informed, data-efficient approaches. The combination of cutting-edge machine learning methods with careful linguistic consideration provides a template for similar initiatives targeting other low-resource languages worldwide.

5. Conclusions

This study has successfully demonstrated that state-of-the-art non-autoregressive architectures such as FastSpeech 2, when combined with high-fidelity neural vocoders like HiFi-GAN, can be effectively adapted to low-resource tonal languages. By incorporating G2P modeling, tonal embeddings, and prosodic alignment tools, the proposed TTS system overcame critical limitations in existing models—particularly poor tonal representation, low data availability, and computational inefficiency.

The developed system not only advances AI research for Nigerian languages but also contributes to bridging the digital divide, ensuring that speakers of Hausa, Igbo, and Yoruba can access technology in their native tongues. Its modular design allows for future scalability to other African languages, aligning with global efforts in linguistic preservation and cultural representation in AI systems.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Akujobi, O.S. (2019) The English Language Coalescence and Multilingualism in Nigeria. *IGWEBUIKE*: *An African Journal of Arts and Humanities*, **5**, 1-16.

[2] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y. (2020) FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. http://arxiv.org/abs/2006.04558

[3] Afolabi, A., Omidiora, E. and Arulogun, T. (2013) Development of Text to Speech System for Yoruba Language. *Innovative Systems Design and Engineering*, **4**, 1-7. https://www.iiste.org

[4] Ekpenyong, M.E., Urua, E. and Gibbon, D. (2008) Towards an Unrestricted Domain TTS System for African Tone Languages. *International Journal of Speech Technology*, **11**, 87-96. https://doi.org/10.1007/s10772-009-9037-5

[5] Oyelade, J.O., Isewon, I., Famade, A. and Oyelade, J. (2022) Foundation of Computer Science FCS. *International Journal of Applied Information Systems*, **12**.

[6] Kong, J., Kim, J. and Bae, J. (2020) HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. http://arxiv.org/abs/2010.05646

[7] Ping, W., Peng, K., Zhao, K. and Song, Z. (2019) WaveFlow: A Compact Flow-Based Model for Raw Audio. http://arxiv.org/abs/1912.01219

[8] Ngor, C.I.-A. (2024) Tone Nature of Nigerian English. *African Journal of Humanities and Contemporary Education Research*, **15**, 399-415. https://doi.org/10.62154/e2bnwx92

[9] Salau, A.O., Olowoyo, T.D. and Akinola, S.O. (2020) Accent Classification of the Three Major Nigerian Indigenous Languages Using 1D CNN LSTM Network Model. In: Jain, S., *et al.*, Eds., *Advances in Computational Intelligence Techniques*, Springer, 1-16. https://doi.org/10.1007/978-981-15-2620-6_1

[10] Fromont, R., Clark, L., Black, J.W. and Blackwood, M. (2023) Maximizing Accuracy of Forced Alignment for Spontaneous Child Speech.

[11] Wu, H., Yun, J., Li, X., Huang, H. and Liu, C. (2023) Using a Forced Aligner for Prosody Research. *Humanities and Social Sciences Communications*, **10**, Article No. 429. https://doi.org/10.1057/s41599-023-01931-4

[12] UNESCO (2021) Towards Sustainable Preservation and Accessibility of Documentary Heritage. https://unesdoc.unesco.org/ark:/48223/pf0000380171

[13] Tan, Y. and Jehom, W.J. (2024) The Function of Digital Technology in Minority Language Preservation: The Case of the Gyalrong Tibetan Language. *Preservation, Digital Technology & Culture*, **53**, 165-177. https://doi.org/10.1515/pdtc-2024-0021

[14] Hunt, A.J. and Black, A.W. (1996) Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. 1996 *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 1, 373-376. https://doi.org/10.1109/icassp.1996.541110

[15] Rabiner, L.R. and Schafer, R.W. (2007) Introduction to Digital Speech Processing. *Foundations and Trends® in Signal Processing*, **1**, 1-194. https://doi.org/10.1561/2000000001

[16] Bäckström, T., Räsänen, O., Zewoudie, A., Zarazaga, P.P., Koivusalo, L., Das, S., *et al.* (2022) Introduction to Speech Processing: 2nd Edition. https://doi.org/10.5281/ZENODO.6821775

[17] Kuligowska, K., Kisielewicz, P. and Włodarz, A. (2018) Speech Synthesis Systems: Disadvantages and Limitations. *International Journal of Engineering & Technology*, **7**, 234-239. https://doi.org/10.14419/ijet.v7i2.28.12933

[18] Hande, S.S. (2014) A Review of Concatenative Text to Speech Synthesis. *International Journal of Latest Technology in Engineering, Management & Applied Science*, **3**, 12-15. https://www.academia.edu/download/34937201/12-15.pdf

[19] Khan, R.A., and Chitode, J.S. (2016) Concatenative Speech Synthesis: A Review. *International Journal of Computer Applications*, **136**, 1-6. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=dcb1aefcc8d80c90392fa9b6f2740b4516e8ec44

[20] Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., *et al.* (2017) Tacotron: Towards End-to-End Speech Synthesis. *Interspeech* 2017, Stockholm, 20-

24 August 2017, 4006-4010. https://doi.org/10.21437/interspeech.2017-1452

[21] Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., *et al.* (2018) Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. http://arxiv.org/abs/1806.04558

[22] Ohala, J.J. and Kawasaki, H. (1984) Prosodic Phonology and Phonetics. *Phonology Yearbook*, **1**, 113-127. https://doi.org/10.1017/s0952675700000312

[23] Peterson, G.E. and Shoup, J.E. (1966) A Physiological Theory of Phonetics. *Journal of Speech and Hearing Research*, **9**, 5-67. https://doi.org/10.1044/jshr.0901.05

[24] Pierrehumbert, J.B. (1980) The Phonology and Phonetics of English Intonation.

[25] Fant, G. (1981) The Source Filter Concept in Voice Production. STL-QPSR, 22, 021-037. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=647bee8e1ea9b5fcaa27dd8c0937a165a8f5f717

[26] Ma, R., Qian, M., Fathullah, Y., Tang, S., Gales, M. and Knill, K. (2025) Cross-Lingual Transfer Learning for Speech Translation. *Proceedings of the* 2025 *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 2, 33-43. https://doi.org/10.18653/v1/2025.naacl-short.4

[27] Fant, G. (2001) T. Chiba and M. Kajiyama, Pioneers in Speech Acoustics. *Journal of the Phonetic Society of Japan*, **5**, 4-5. https://doi.org/10.24467/onseikenkyu.5.2_4

[28] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (1999) Simultaneous Modeling of Spectrum, Pitch and Duration in Hmm-Based Speech Synthesis. 6*th European Conference on Speech Communication and Technology*, *EUROSPEECH* 1999, Budapest, 5-9 September 1999, 2347-2350.

[29] Ping, W., Peng, K. and Chen, J. (2018) ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech. http://arxiv.org/abs/1807.07281

[30] Atoi, N.E. (2024) Language and Communication Implication of Artificial Intelligence on Selected Nigerian University Undergraduates. *UJAH*: *Unizik Journal of Arts and Humanities*, **25**, 109-154. https://doi.org/10.4314/ujah.v25i1.5

[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010. http://arxiv.org/abs/1706.03762

[32] Bleyan, H., Ritchie, S., Mortensen, J.F. and Esch, D.V. (2019) Developing Pronunciation Models in New Languages Faster by Exploiting Common Grapheme-to-Phoneme Correspondences across Languages. *Interspeech* 2019, Graz, 15-19 September 2019, 2100-2104. https://doi.org/10.21437/interspeech.2019-1781

[33] Hassana, I.L. and Sanusi, M. (2019) Text to Speech Synthesis System in Yoruba Language. *International Journal of Advances in Scientific Research and Engineering*, **5**, 180-191. https://doi.org/10.31695/ijasre.2019.33568

[34] Olaniyan, O.M. and Akinode, V. (2023) Development of a Text-to-Speech Synthesis for Yoruba Language Using Deep Learning. *Technology and Innovation*, **2**, 1-7.

[35] Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., *et al.* (2017) Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. http://arxiv.org/abs/1712.05884

[36] Artetxe, M., Ruder, S. and Yogatama, D. (2020) On the Cross-Lingual Transferability of Monolingual Representations. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, July 2020, 4623-4637.

https://doi.org/10.18653/v1/2020.acl-main.421

[37] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., *et al.* (2020) Unsupervised Cross-Lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, 8440-8451. https://doi.org/10.18653/v1/2020.acl-main.747

[38] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., *et al.* (2016) WaveNet: A Generative Model for Raw Audio. http://arxiv.org/abs/1609.03499