Igbo Kwenu: IkengaTTS — a Tone-Preserving Igbo Text-to-Speech via Transfer Learning

Emmanuel Chimezie
Mexkoy Labs
emmanuel@mexkoy.com

Abstract

A strong pre-trained text-to-speech model can be adapted to Igbo with modest fine-tuning. We fine-tune F5-TTS (335M parameters) on 112,521 Igbo clips (255 hours) from the African Voices dataset and obtain intelligible Igbo speech, while multiple from-scratch approaches fail. On a small single-speaker dataset (WAXAL; 934 usable clips, ∼2.7 hours), autoregressive and non-autoregressive models collapse to silence even when model size is reduced from 25M to 1.1M parameters, indicating a data—not capacity—bottleneck. On the larger multi-speaker African Voices set, a from-scratch VITS baseline produces recognizable speech but exhibits severe duration compression. On a held-out test sample ($n = 987$ from 6,937 clips), our fine-tuned model improves ASR-transcribability under MMS, achieving MMS WER 50.3% versus 55.6% on the corresponding recordings (ΔWER $= -5.3$ pp), while the unadapted base model fails (WER 100.5%). We preserve tone-bearing diacritics in the text representation and provide audio samples and an evaluation pipeline for reproduction and further work. Writing and code development were assisted by Claude Code (Anthropic). All scientific claims, experimental design, and analysis remain the sole responsibility of the authors.

1 Introduction

This work contributes to efforts that use modern speech technology to preserve and expand access to African languages. We focus on Igbo, a major language of southeastern Nigeria. Igbo is spoken by tens of millions of people (estimates range from ∼30M to 45M), and the Igbo ethnic group comprises roughly 15–18% of Nigeria's population. Igbo's cultural footprint is global. Chinua Achebe's *Things Fall Apart*, written by an Igbo author and rooted in Igbo society, is among the most influential novels of the twentieth century. Igbo also has a written tradition dating back at least to Pita Nwana's 1933 novel *Omenuko*, alongside a far older oral tradition. Yet, in modern computing, Igbo remains under-served: publicly documented, broadly accessible Igbo text-to-speech (TTS) systems with clear evaluation and reproducible baselines are scarce.

At the same time, there are visible demand signals. Figure 1 shows worldwide Google Trends interest for the "Igbo" spoken-language topic from 2004 to the present. Interest is low and relatively flat through the late 2000s, rises steadily from the mid-2010s, and accelerates sharply in the early 2020s, reaching its highest levels in the most recent period. This measure is normalized and query-based; it is not a measure of speaker population or intergenerational transmission. We use it only

as motivation: attention to Igbo in digital contexts appears to be increasing, making the lack of robust Igbo speech synthesis infrastructure a practical gap.

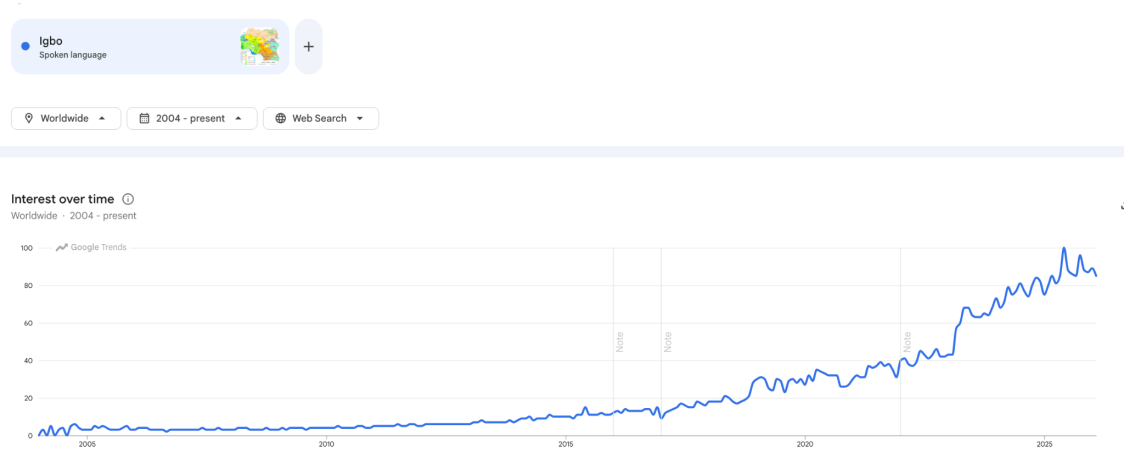

Figure 1: Google Trends "Igbo" (Spoken language topic), Worldwide, 2004–present, Web Search. Values are normalized indices within Google Trends and reflect relative search interest, not absolute search volume or speaker population.

This project is downstream of a growing body of Igbo language technology work. We credit OkwuGbe for early Igbo ASR baselines, IgboAPI for accessible lexical and parallel-text resources, and YarnGPT for demonstrating momentum toward Nigerian language TTS. But these efforts do not yet provide a broadly accessible, reproducible, tone-preserving Igbo TTS system with evaluation.

We make four claims, each supported by the experiments that follow:

First, training from scratch on limited data does not work. On 934 studio-quality clips, an autoregressive decoder collapsed and a non-autoregressive model converged to silence at two parameter scales. On 112,000 clips, a VITS model produced recognizable Igbo but could not learn correct speech duration from diverse spontaneous audio. No from-scratch approach gave us usable output.

Second, a pre-trained model solves the problem. F5-TTS, a flow matching architecture trained on English and Chinese speech, produces intelligible Igbo speech after fine-tuning on 112,000 clips. Pre-training provides the speech production knowledge that Igbo data alone cannot teach from zero.

Third, tone must be preserved in the input representation. Igbo is a register tone language — the word *akwa* means cloth, cry, bed, or egg depending on its tone pattern. A system that drops tonal diacritics does not merely sound unnatural. It says the wrong words.

Fourth, the resulting system generates speech that is at least as intelligible as the original human recordings, as measured by the best available Igbo ASR system. On 987 held-out clips, generated audio achieves a word error rate 5.3 percentage points lower than ground truth.

The organizing insight of this paper is that pre-training is what makes low-resource TTS possible. Everything that follows either demonstrates this or refines it.

2 The Igbo Language

Igbo is a register tone language spoken across southeastern Nigeria and a global diaspora estimated at 40 million people including second-language speakers. Two level tones — HIGH and LOW — distinguish word meaning at the syllable level. The word *akwa* illustrates the stakes: with a high-high tone pattern (*ákwá*) it means "cry," with low-low (*àkwà*) it means "cloth," with high-low (*ákwà*) it means "bed," and with low-high (*àkwá*) it means "egg." These are not subtle pronunciation differences. They are different words entirely.

Igbo orthography is largely phonemic — each letter maps to one sound, and digraphs like *ch*, *gb*, *kp*, *nw*, and *ny* are consistent. The vowel system includes a phonemic distinction between plain vowels (*ị*, *o*, *u*) and their dotted counterparts (*ị̇*, *ọ*, *ụ*), reflecting an Advanced Tongue Root harmony system. This near-perfect orthographic transparency means a grapheme-to-phoneme converter for Igbo can be rule-based — no machine learning required. But most written Igbo omits tone marks. Native readers resolve the ambiguity from context. A TTS system cannot.

The closest precedent for tone-aware TTS comes from Yoruba, another tonal West African language. Ogunremi et al. (2024) built IroyinSpeech, a Yoruba TTS system, and ran a controlled experiment: models trained with diacritical tone marks achieved a Mean Opinion Score of 3.98 out of 5. Models trained without diacritics scored 1.86 — a collapse in perceived quality so severe that the speech was rated worse than robotic. Their conclusion is direct: for tonal languages, diacritics are not optional. Our work confirms and extends this finding for Igbo.

3 Related Work

3.1 Prior Research

The most significant prior work on Igbo speech technology is OkwuGbe (Dossou and Emezue, 2021), the first automatic speech recognition system for Igbo and Fon, using a ResidualCNN-BiGRU architecture with CTC loss on Common Voice data. Their key finding — that including diacritical marks in transcriptions improved accuracy — foreshadowed our own results. The same research group later produced the IgboAPI project (Emezue et al., 2024), providing a dictionary with 5,095 words across 33 Igbo dialects, 17,979 dialectal word pronunciations, and 278,899 parallel Igbo-English sentences. Their machine translation model jumped from 16.87 to 71.95 BLEU after fine-tuning. IgboAPI lists TTS on their roadmap but has released nothing. Our system fills their stated gap.

IroyinSpeech (Ogunremi et al., 2024) is the closest comparable TTS system: a VITS-based model for Yoruba trained on 42 hours of studio recordings from 80 speakers. Beyond the diacritics finding described in Section 2, their work established that VITS can produce intelligible tonal African language speech when given sufficient studio-quality data and tone-marked transcriptions.

F5-TTS (Chen et al., 2024) is our base model. It uses flow matching with a Diffusion Transformer (DiT) to generate speech from text and a reference audio clip, without requiring duration models,

phoneme alignment, or a separate vocoder. On English, it achieves 2.42 percent word error rate on the Seed-TTS test set with a speaker similarity score of 0.664. We adapt it for Igbo by expanding its character vocabulary.

For evaluation, we use MMS (Pratap et al., 2023), Meta's Massively Multilingual Speech model, which provides ASR for over 1,100 languages including Igbo via language-specific adapter layers.

3.2 Existing Systems

YarnGPT (Saheed, 2024) is a 360-million-parameter language model that generates speech by predicting discrete audio tokens autoregressively. It is open-source and claims support for multiple Nigerian languages, but publishes no evaluation metrics for Igbo or any other language, provides only three Igbo voices compared to ten English voices, and includes no tone modeling in its pipeline. No independent evaluation exists.

ElevenLabs offers Igbo in its API but publishes no word error rates, no MOS scores, and no quality metrics of any kind. Google does not offer a dedicated Igbo TTS model. Meta's SeamlessM4T v2 supports Igbo for speech recognition and translation but not for speech synthesis.

4 What We Tried

We did not start with F5-TTS. We started with the assumption that a well-chosen architecture, initialized from random weights and trained on Igbo data, would produce intelligible speech. This assumption was wrong. We tested it four times — on 934 clips and on 112,000 clips, across three architecture families — before understanding why.

4.1 Kokoro: Autoregressive encoder–decoder (from scratch)

Our first attempt adapted Kokoro (a simplified StyleTTS2 variant) for Igbo. We built a rule-based grapheme-to-phoneme (G2P) front end with an IPA inventory and explicit tone tokens (HIGH , LOW). To support duration supervision, we generated phoneme alignments with Montreal Forced Aligner on 1,485 clips and saved the resulting TextGrid alignments to avoid re-running alignment during cloud training.

We trained on the WAXAL Igbo TTS dataset (Diack et al., 2026), consisting of 1,552 studio recordings from a single speaker. After filtering clips longer than 15 seconds, 934 clips remained ($\approx$ 2.7 hours).

The model ($\approx$ 30M parameters) learned under teacher forcing but collapsed at inference. Under teacher forcing, predicted mel spectrogram statistics matched training targets (structured output). Under free-running inference, outputs collapsed to near-silence. We attempted aggressive scheduled

sampling and auxiliary free-running losses; this reduced the collapse but did not yield intelligible speech.

Lesson. With limited data, autoregressive mel decoders can fail via exposure-bias collapse: they learn to rely on clean previous frames during training and cannot recover when fed their own imperfect outputs at inference.

4.2 VITS: End-to-end (from scratch)

Our second attempt used VITS (Kim et al., 2021), an end-to-end model that jointly learns text-to-waveform synthesis with adversarial training. We increased data scale substantially: instead of WAXAL's $\approx$ 2.7 hours, we trained on the African Voices Igbo subset collected by Data Science Nigeria (112,521 clips; $\approx$ 255 hours; 447 speakers after filtering to 1–15 seconds).

VITS produced recognizable Igbo speech — our first clear intelligibility signal. However, duration modeling was systematically wrong. Across checkpoints, synthesized utterances were consistently compressed to roughly 40–55% of the ground-truth duration, producing rushed speech even when words were audible. This experiment also showed that Igbo orthography is regular enough that character-level input can work; the custom IPA/G2P pipeline was not required for basic grapheme-to-sound learning at this scale.

Lesson. End-to-end waveform models remove vocoder and alignment engineering, but they do not eliminate duration pathologies when trained from scratch on heterogeneous, multi-speaker speech.

4.3 FastSpeech2: Controlled from-scratch test at two scales

Our third attempt was designed to isolate whether the Kokoro failure was "autoregressive exposure bias" or "data scarcity." FastSpeech2 is non-autoregressive and uses explicit predictors for duration, pitch, and energy, so it should not fail via autoregressive collapse. We returned to the WAXAL filtered setting (934 clips) to stress-test "small, clean, single-speaker" training.

At $\sim$25M parameters, the model overfit: training loss decreased substantially while validation loss remained high, and generated speech was muffled and unintelligible. We verified the vocoder was not the culprit by synthesizing from ground-truth mel spectrograms through the same Griffin–Lim path, which produced intelligible speech.

We then reduced capacity to $\sim$1.1M parameters to test whether the issue was an unfavorable parameter-to-sample ratio. The result was effectively unchanged: similar loss behavior and silent or unusable outputs.

Lesson. In this regime, the binding constraint is data quantity (and diversity), not model capacity. With $\sim$ 2.7 hours of audio, from-scratch mel prediction did not generalize in our setup even when the architecture was non-autoregressive and the model was heavily downsized.

4.4 YarnGPT LoRA attempt (abandoned)

We began a LoRA fine-tuning attempt on YarnGPT but discontinued due to cloud training infrastructure failures unrelated to the architecture. We do not treat this as evidence for or against the model class.

4.5 F5-TTS fine-tune: Transfer learning succeeds

Our final attempt changed the core premise: we stopped training from scratch and instead fine-tuned a large pre-trained TTS model. We used F5-TTS ($\approx$ 335M parameters), pre-trained on large-scale speech corpora (English/Chinese) and designed to generate speech with reference-audio conditioning without explicit duration modeling or phoneme alignment.

We fine-tuned on the same African Voices Igbo dataset used for VITS (112,521 clips). To support Igbo orthography faithfully, we expanded the character vocabulary to include combining diacritics used for tone marks and dot-below vowels (Unicode combining marks for acute/grave and dot-below). After $\sim$ 8 epochs ($\sim$ 235k updates) on a single A100, the model produced intelligible, natural-sounding Igbo speech. Informal stress tests covered proverbs, tone minimal pairs (the four *akwa* variants), common consonant clusters (e.g., *mgbe*, *nke*, *nwányì*, *nghòta*), and numeric constructions.

Lesson. Across all configurations we tested, from-scratch training failed to produce a robust Igbo TTS system: either collapse to silence (small clean data) or systematic duration failure (large heterogeneous data). Fine-tuning a pre-trained model on the same large dataset succeeded. This supports the central hypothesis of the paper: pre-training provides general speech production knowledge that low-resource Igbo data alone does not reliably induce from zero.

5 Method

5.1 Data

We used the African Voices Igbo dataset from Data Science Nigeria (Ogundepo et al., 2025): 112,521 audio clips totaling 255.3 hours of spontaneous speech from 447 speakers across multiple dialect regions, filtered to clips between 1 and 15 seconds. The audio was collected from diverse recording conditions — not studio-quality, but representative of real Igbo speech. We converted from FLAC to 24 kHz mono WAV for compatibility with F5-TTS's Vocos vocoder.

For contrast: the WAXAL dataset (Diack et al., 2026) that we used in our from-scratch experiments contained 1,552 clips from a single studio speaker. Its duration distribution was heavy-tailed — median 15 seconds, 90th percentile 109 seconds, maximum 29.5 minutes — making it hostile to architectures designed for 2-to-10-second utterances. After filtering for length, only 934 clips remained usable. African Voices, by providing two orders of magnitude more clips with moderate durations, made the difference.

5.2 Vocabulary Expansion

F5-TTS uses character-level tokenization. Its base vocabulary of 2,546 tokens covers English and Chinese characters. Igbo requires additional combining diacritics: U+0300 (grave accent, marking LOW tone), U+0301 (acute accent, marking HIGH tone), U+0323 (dot below, distinguishing the vowels ị, ọ, ụ from i, o, u), and U+0304 (macron, used in some orthographic conventions). We expanded the vocabulary to 2,604 tokens and resized the model's text embedding matrix accordingly, initializing new embeddings with small random values.

5.3 Training

We fine-tuned F5-TTS v1 Base (335.8 million parameters) using the following configuration:

Parameter	Value
Learning rate	1×10^{-5}
Batch size	3,200 frames per GPU
Epochs	8
Total gradient updates	235,190
Hardware	1 × NVIDIA A100 SXM 80 GB
Training time	∼20 hours
Tokenizer	Character-level
Save interval	Every 30,000 updates

Training loss dropped to approximately 0.50 after the first epoch, then plateaued at 0.69 from epoch 2 onward. Despite the flat loss curve, subjective audio quality continued to improve through all 8 epochs — the model was refining its Igbo pronunciation and prosody in ways that the aggregate loss did not capture. We archived 37 generated audio samples at intervals from update 500 to update 210,000 to document this progression.

5.4 Tone Handling

Igbo combining diacritics for tone (grave and acute accents) become separate tokens in the expanded character vocabulary. When the model processes the text *àkwà* (cloth), it receives a sequence that includes the grave accent tokens. When it processes *ákwá* (cry), it receives acute accent tokens instead. The tonal distinction is encoded directly in the character sequence.

This is orthographic preservation of tone-bearing characters rather than phonological tone modeling in the linguistic sense. We did not build a separate tone embedding layer or use the phoneme-level G2P with explicit HIGH and LOW tokens that we had designed for the earlier Kokoro approach — though we release that G2P alongside the model for other researchers to use. The character-level approach proved sufficient: the model learns to condition its acoustic generation on the presence or absence of tone-bearing diacritics.

6 Evaluation

6.1 The Entanglement Problem

Evaluating TTS for a low-resource language creates a measurement problem that does not exist for English or Chinese. The standard approach — generate speech, run it through an ASR system, and measure word error rate — conflates two sources of error: the TTS model and the ASR model. When the best available ASR system achieves only 40 to 57 percent word error rate on authentic Igbo recordings, absolute WER on generated speech is meaningless. A WER of 50 percent could indicate terrible TTS or mediocre TTS evaluated by a mediocre ASR — or, as we found, it could indicate good TTS evaluated by limited ASR.

We resolve this by reporting WER *delta*: the difference between the word error rate on generated speech and the word error rate on the corresponding ground truth recording, measured by the same ASR system on the same text. A negative delta means the generated speech is more intelligible to the ASR than the original recording. This controls for the ASR system's baseline competence on Igbo and isolates the TTS contribution.

6.2 Test Data

We evaluate on the African Voices development-test set: 6,937 held-out Igbo audio clips that were never seen during training. For each test clip, we select a reference speaker from a pool of 20 evaluation speakers (10 female, 10 male) drawn from 3 dialect batches in the dataset. Speakers are gender-matched to the test clip and assigned by round-robin rotation.

We evaluated on a stratified random sample of 1,000 clips from this test set (seed 42), yielding 987 valid results after 13 synthesis failures (1.3% failure rate).

6.3 Metrics

We report three metrics:

WER delta measures intelligibility, controlled for ASR quality. We transcribe both the generated audio and the ground truth audio using MMS-1B-all with the Igbo language adapter (Pratap et al., 2023), then report the difference. Negative values indicate the generated speech is more intelligible to the ASR system.

SIM-o measures speaker similarity. We extract speaker embeddings from both generated and reference audio using WavLM-base-plus-sv (Chen et al., 2022) and compute their cosine similarity. Values range from 0 (no similarity) to 1 (identical).

UTMOS predicts perceived naturalness on a 1-to-5 scale using the UTMOS22-strong model (Saeki et al., 2022). We report it for both generated and ground truth audio but note a critical caveat:

UTMOS was trained exclusively on English speech and has never been validated on Igbo or any tonal African language. The absolute scores are not interpretable on the English MOS scale. We include UTMOS for directional comparison between generated and ground truth only.

6.4 Statistical Approach

We report bootstrap 95 percent confidence intervals computed with 10,000 resamples. For paired comparisons between generated and ground truth metrics (WER, UTMOS), we use the Wilcoxon signed-rank test.

Our evaluation uses 987 samples drawn from a 6,937-clip held-out test set. For context: F5-TTS evaluated on 1,127 samples, StyleTTS 2 on 80, and NaturalSpeech 3 on 40. Our evaluation set is comparable to the largest published TTS evaluations.

7 Results

7.1 Main Results

System	MMS WER ↓	ΔWER (gen − GT) ↓	SIM-o ↑	UTMOS ↑
Ground truth (recordings)	55.6% (54.5–56.8)	—	—	1.28 (1.28–1.28)
F5-TTS fine-tuned (ours)	50.3% (49.1–51.5)	−5.3 pp ($p < 10^{-18}$)	0.963 (0.960–0.965)	1.68 (1.66–1.70)
F5-TTS base (no fine-tuning)	100.5% (99.8–101.3)	+44.9 pp ($p < 10^{-159}$)	0.958 (0.956–0.961)	1.55 (1.53–1.56)

$n = 987$. 95% confidence intervals from 10,000 bootstrap resamples. p-values from Wilcoxon signed-rank tests. The same test clips (sampled from the 6,937-item held-out set with seed 42) and the same 20 reference speakers were used across systems, enabling paired comparisons.

The fine-tuned model's synthesized audio has a lower MMS WER (ΔWER $= -5.3$ pp, $p < 10^{-18}$) than the original recordings. This suggests the synthesized audio is more ASR-transcribable than the ground truth under MMS, not necessarily "more intelligible than humans" in terms of human perception.

The unadapted F5-TTS model failed as an Igbo TTS system, with an MMS WER over 100% and a ΔWER of +44.9 pp ($p < 10^{-159}$). Fine-tuning was crucial; it caused a 50.2 pp shift in ΔWER, improving it from +44.9 pp to −5.3 pp. The base model's SIM-o remains high (0.958) despite its linguistic failure. This suggests SIM-o primarily reflects the preservation of speaker/voice characteristics under reference conditioning, not the correctness of the spoken language.

UTMOS scores are low for all conditions (ground truth 1.28; fine-tuned 1.68; base 1.55). As UTMOS is trained on English MOS and not calibrated for Igbo, we use it only directionally. Within this limitation, synthesized audio is rated higher than recordings.

Stratified by duration. The WER improvement is largest for long clips (8–15 seconds), where the model achieves a 11.7 percentage point advantage over ground truth. Short clips (< 4 s) show

a smaller advantage of 1.5 points.

Duration	n	WER (gen)	WER (GT)	ΔWER	SIM-o
Short (< 4 s)	10	48.9%	50.5%	−1.5 pp	0.953
Medium (4–8 s)	938	50.5%	55.6%	−5.1 pp	0.963
Long (8–15 s)	39	46.5%	58.2%	−11.7 pp	0.968

Stratified by gender. The WER improvement holds for female speakers (−6.5 pp, $n = 827$) but not for male speakers (+0.7 pp, $n = 160$). The male result is directionally flat, not reversed: male ground truth WER is already lower (58.2% vs 55.1% for female), leaving less room for the generated speech to improve on. The gender imbalance in the test set (83% female) reflects the underlying dataset composition.

Gender	n	WER (gen)	WER (GT)	ΔWER	SIM-o
Female	827	48.7%	55.1%	−6.5 pp	0.964
Male	160	58.9%	58.2%	+0.7 pp	0.958

7.2 Qualitative analysis: tone minimal pairs and Igbo-specific phonotactics

In informal listening checks, the model produces distinct pitch contours for tone-marked minimal pairs, including the four *akwa* variants introduced in Section 2. It also renders sequences with dotted vowels and tone marks (e.g., *Onye ajụjụ anaghị efu ụzọ*) cleanly. Common Igbo consonant clusters (e.g., *mgbe*, *nke*, *nwanyị*, *nghọta*) are articulated without obvious distortion, and multi-word numeric constructions (e.g., *narị iri abụo na asatọ*) are synthesized without breaking prosodic continuity.

8 Discussion

8.1 The Data Threshold — And Why Pre-Training Crosses It

From-scratch training failed to produce a robust Igbo TTS system, unlike transfer learning. Even when VITS was trained and F5-TTS was fine-tuned on the same African Voices Igbo dataset (112,521 clips; 447 speakers), only the pre-trained model yielded consistently usable synthesis. The unadapted F5-TTS baseline had a 100.5% MMS WER, much worse than recordings (+44.9 pp). After fine-tuning, the WER dropped to 50.3%, outperforming recordings under MMS (−5.3 pp). This shows that while pre-training provides fluent speech (high SIM-o), fine-tuning is essential to teach the model Igbo's lexical and phonological competence, including diacritics.

Training TTS models from scratch requires a "data threshold." With only ≈2.7 hours of single-speaker audio, both autoregressive models (collapsed) and non-autoregressive models (failed to

learn/silence) failed, even with significantly smaller models. Using ≈255 hours of varied multi-speaker audio, VITS generated recognizable speech but showed consistent duration issues, suggesting difficulty in learning stable alignment and duration from scratch amidst high variability. F5-TTS, having a speech prior from pre-training, only required the 112k clips of Igbo data to learn the language-specific content.

For low-resource languages, fine-tuning a strong pre-trained Text-to-Speech (TTS) model is better than training from scratch. Our fine-tuning (≈20 hours on an A100) shows compute is not the main issue; the primary obstacle is starting without the speech prior a foundation model provides.

8.2 Tone preservation

We treat tone marks as part of the text, using character-level tokenization. Igbo tone marks (acute/grave) and dotted vowels are represented as explicit Unicode combining marks and distinct tokens. This makes tonal information available, allowing the model to learn different pitch contours from varying diacritic sequences.

Current evidence for tone correctness is qualitative. A controlled ablation will quantify the impact of removing combining diacritics on MMS metrics and failure rates. However, since most Igbo text lacks tone marks, a crucial next step is a diacritization front end to restore them from unmarked text. This requires sequence labeling with supervision from marked corpora, but dialect variation and orthographic inconsistency necessitate careful normalization and evaluation.

8.3 Why synthesized audio scores better than recordings under ASR

Synthesized audio having lower MMS WER than recordings seems counterintuitive as "TTS is better than humans," but the metric actually means MMS finds synthesized speech easier to transcribe than spontaneous recordings in our protocol.

Synthesized speech is often more canonical and cleaner than spontaneous recordings. TTS produces consistent pacing, stable acoustics, and standardized articulation, making it more ASR-friendly, whereas spontaneous speech contains noise, disfluencies, and variation that degrade ASR performance.

We emphasize ΔWER as recognizer-conditional ASR-transcribability. Negative ΔWER means better transcribability under MMS, not superior human intelligibility or naturalness. Human listening tests with native Igbo speakers are still required to directly measure perceived quality and semantic correctness.

8.4 Limitations

Evaluation coverage. We evaluated 987 clips from a 6,937-clip test set, enough for tight bootstrap intervals on the main effect, but limiting detailed stratification (dialect, noise, device). A full-set

evaluation would lower variance and enable stronger subgroup analysis.

Single training run. We report a single fine-tuning configuration without a hyperparameter sweep. Performance could improve with different schedules, learning rates, or conditioning.

No human evaluation. All reported metrics are automated. UTMOS is not calibrated for Igbo, and ASR-based metrics are recognizer-conditional. A native-listener study is needed to validate naturalness, intelligibility, and tone correctness.

Tone-marked input requirement. Accurate tonal output requires tone-marked input. Without automatic tone restoration, real-world application necessitates manual tone marking or reliance on already tone-marked sources.

ASR evaluation floor. High MMS error rates on authentic Igbo speech limit the utility of WER-based proxies. While ΔWER offers partial control, it doesn't eliminate reliance on the recognizer's biases.

Speaker similarity inflation. Using reference speakers from the training set likely inflates SIM-o. A more robust evaluation requires unseen reference speakers and various reference conditions.

9 Future Work

Three extensions follow directly from the current limitations and from what the results suggest is most leverage.

Human evaluation (native listeners). Current metrics are limited (WER, UTMOS). Future work includes a listener study with native Igbo speakers to measure intelligibility, naturalness, and speaker similarity. We will also pilot a tone-focused measure using curated minimal pairs and sentences, where raters judge if the intended lexical tone pattern is preserved. This diagnostic aims to assess tonal correctness, a key challenge in tonal-language TTS, complementing existing MOS/MUSHRA evaluations.

Automatic tone and diacritic restoration. As most Igbo text lacks tone marks, the main deployment barrier for our synthesis pipeline (which requires tone-marked input) is the lack of a diacritization front end to restore tones and missing dotted vowels. This could be implemented via sequence labeling (e.g., CRF) or token classification on an adapted language model. Training data can come from tone-marked texts and lexicons, but dialect and orthography variations will necessitate careful normalization and evaluation.

Joint TTS–ASR bootstrapping. Pre-trained TTS models offer scalable data synthesis. A promising approach is using these models to expand paired data for Igbo, where labeled speech is limited. For TTS, generate audio for large unlabeled text corpora, then filter and score it for training. For ASR, transcribe unlabeled Igbo audio (radio, podcasts, etc.), then filter and correct transcripts. This bootstrapping is vital for Igbo, where unlabeled text and audio significantly outweigh labeled speech.

10 Conclusion

Building usable Igbo text-to-speech under low-resource constraints proved challenging. Attempts with ≈2.7 hours of single-speaker studio data failed to yield a robust system, with models either collapsing or failing to learn. An end-to-end model trained on ≈255 hours of heterogeneous multi-speaker data produced recognizable speech but suffered from systematic duration issues.

Fine-tuning a large, pre-trained TTS model on the ≈255-hour Igbo dataset succeeded where training from scratch failed. The system generates intelligible, natural Igbo speech and shows improved ASR-transcribability (negative ΔWER) compared to the original recordings under MMS. This suggests the synthesized speech is cleaner and more canonical.

We release code, weights, and the evaluation pipeline for reproduction and extension. The key takeaway for low-resource settings is that pre-training enables a usable baseline, turning repeated failure into a foundation for future work like human evaluation, tone restoration, and cross-dialect testing.

References

References

[1] Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., and Chen, X. (2024). F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. *arXiv:2410.06885*.

[2] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., and Wei, F. (2022). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6).

[3] Diack, A., Nelson, P., Agbesi, K., et al. (2026). WAXAL: A Large-Scale Multilingual African Language Speech Corpus. *arXiv:2602.02734*.

[4] Dossou, B. F. P. and Emezue, C. C. (2021). OkwuGbe: End-to-End Speech Recognition for Fon and Igbo. *Proceedings of the 2nd Workshop on African Natural Language Processing, EACL 2021. arXiv:2103.07762*.

[5] Emezue, C. C., Okoh, E., Mbonu, C., et al. (2024). The IgboAPI Dataset: Empowering Igbo Language Technologies through Multi-dialectal Dictionary and Speech Data. *arXiv:2405.00997*.

[6] Kim, J., Kong, J., and Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*.

[7] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proceedings of Interspeech 2017*.

[8] Ogundepo, O., et al. (2025). African Voices: A Large-Scale Multilingual Speech Dataset for African Languages. *Data Science Nigeria*.

[9] Ogunremi, T., Tubosun, K., Aremu, A., Orife, I., and Adelani, D. I. (2024). IroyinSpeech: A Multi-purpose Yoruba Speech Corpus. *Proceedings of LREC-COLING 2024*. *arXiv:2307.16071*.

[10] Pratap, V., Tjandra, A., Shi, B., Tober, P., Babu, A., Kunber, S., et al. (2023). Scaling Speech Technology to 1,000+ Languages. *arXiv:2305.13516*.

[11] Saheed, N. (2024). YarnGPT: Language Model-Based Text-to-Speech for Nigerian Languages. *GitHub repository*.

[12] Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., and Saruwatari, H. (2022). UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. *Proceedings of Interspeech 2022*.

[13] Wagner, P., et al. (2025). Good Practices for Evaluation in TTS. *arXiv:2503.03250*.

A Igbo Phoneme Inventory

Igbo has 8 vowels with Advanced Tongue Root (ATR) harmony: +ATR {i, e, o, u} and −ATR {, , } plus open {a}. The consonant inventory includes 28 phonemes, among them labio-velar plosives (b, kp), a palatal nasal (), labialized consonants (, k, ŋ), and postalveolar affricates (t, d). Two level tones — HIGH () and LOW () — and a phonological downstep () complete the prosodic system.

We designed a 50-token IPA phoneme vocabulary and rule-based grapheme-to-phoneme converter for the Kokoro and FastSpeech2 approaches described in Section 4. The converter processes digraphs before single characters (handling *ch*, *sh*, *gb*, *kp*, *gh*, *gw*, *kw*, *nw*, *ny* correctly), maps dotted vowels to their IPA equivalents (ị→, ọ→, ụ→), and encodes tone marks as explicit tokens. The final F5-TTS model uses character-level tokenization instead and does not use this converter. Both are available in the released code at `igbo_tts/phonemes.py` and `igbo_tts/g2p.py`.

B Training Progression

We archived 37 generated audio samples at intervals from update 500 to update 210,000. At update 500, the model produces English-accented babbling — the pre-trained weights impose English phonotactics on the Igbo text. By update 5,000, it produces speech with an Igbo-like prosodic contour but no recognizable words. By update 30,000 (end of epoch 1), individual Igbo words are audible. By update 90,000 (epoch 3), ChatGPT's voice mode can transcribe and respond to the generated Igbo. By the final checkpoint at update 235,000, all stress tests pass.

These samples are available on the demo page and illustrate a finding worth noting: training loss plateaued at 0.69 after epoch 2, but perceptual quality continued improving through all 8 epochs. Loss is a poor proxy for what a listener hears.

C Failed Approaches — Detailed Technical Notes

The four from-scratch approaches described in Section 4 produced a set of practical lessons that may be useful to other researchers working on low-resource TTS:

WAXAL duration distribution. The WAXAL dataset has a heavy-tailed duration distribution unsuitable for architectures designed for short utterances. The median clip is 15 seconds, the 90th percentile is 109 seconds, and the maximum is 29.5 minutes. After filtering to clips under 15 seconds with text under 320 characters, only 934 of 1,552 clips remained usable. Of the full set, 303 clips exceeded 60 seconds.

BatchNorm corruption under bfloat16. The Kokoro model's PostNet uses BatchNorm1d layers that accumulate running statistics. Under bfloat16 mixed precision training, these statistics diverged to NaN — silently corrupting the model's refined mel output path while leaving the coarse path intact. The coarse mel loss continued decreasing normally, masking the failure. This was diagnosed only when inference produced all-NaN output from the PostNet.

Scheduled sampling insufficiency. Our initial scheduled sampling configuration — 5,000 warmup batches (equivalent to 167 epochs on the small dataset), 0.5 maximum self-feeding probability — meant the model spent most of training in pure teacher forcing. Reducing warmup to 500 batches and raising the probability to 0.8 improved free-running mel standard deviation from 0.05 to 1.4, but did not cross the threshold to intelligibility. The exposure bias problem in autoregressive TTS decoders may require fundamentally different training objectives, not just schedule adjustments.

Duration predictor weakness. In the FastSpeech2 experiments, the 3-layer MLP duration predictor achieved only 54 percent of target mel frame coverage even with duration loss weight increased from 0.05 to 0.2. The model could produce decent mel spectrograms given ground truth durations (L1 loss of 259) but degraded sharply with predicted durations (L1 loss of 467). This coupling between duration accuracy and acoustic quality — where small duration errors cascade into large spectral errors — suggests that duration prediction may be the hardest subproblem in from-scratch non-autoregressive TTS for new languages.