Technoscience Journal for Community Development in Africa 4 (2025) 241–250
ISSN 2659-0573

Research Article

Building a Yoruba text-to-speech engine for automatic reading machines: A concatenative approach

Surajudeen Adewale Yekeen[1*], Abdulazeez Olorundare Ajao[2], Kabirat Oyinlola Raheem[1], Khadijat Oladoyin Mustapha[1], Jumoke Falilat Ajao[3]

[1]Department of Electrical and Electronics Engineering, The Federal Polytechnic, Offa
[2]Department of Computer Engineering, The Federal Polytechnic, Offa
[3]Department of Computer Science, Kwara State University, Malete

Abstract: This research develops a concatenative Text-to-Speech (TTS) system for automatic reading machines (ARMs). TTS system is a major component of an ARM that converts written text to synthetic speech. Corpus concatenation has been the most effective and widely used TTS approach as it is the most efficient in the production of natural and intelligible speech for application in reading aids for the visually impaired, persons with dyslexia, and language learning tools. The abysmal performance of existing TTS in the Yoruba language has resulted in challenges Yoruba speakers face in accessing digital content. This study developed a comprehensive Yoruba speech corpus and implement a concatenative text-to-speech framework, incorporating a Yoruba optical character recognition (YOCR) system, Unicode mapping, syllable segmentation, and speech quality optimization using windowing and pre-emphasis filtering. The developed system achieved Mean Opinion Scores (MOS) of 4.86 for two-syllable words, 4.67 for five-syllable words and at least 4.37 for sentences. The Mel Cepstral Distortion (MCD) metrics showed a maximum mean of 1.58 for concatenated words. The system evaluation using MOS and MCD metrics demonstrates its potential for integration into ARMs and improving digital content accessibility for Yoruba speakers.

Keywords: Yoruba TTS, Concatenative Synthesis, Automatic Reading Machine, Speech Synthesis, Digital Accessibility.

1. Introduction

The rapid advancement of technology has revolutionized access to information. However, for languages with limited digital resources, like Yoruba, this progress can leave a significant portion of the population behind (Akomolede & Olowojebutu, 2025). Yoruba, spoken by over 35 million people primarily in Nigeria and neighbouring West African countries, faces a scarcity of robust Text-to-Speech (TTS) systems that can handle its unique tonal complexities (Yusuff & Osunnuga, 2018). This shortage significantly impedes Yoruba speakers' access to digital content, especially in the realms of education and literacy development (Akomolede & Olowojebutu, 2025). Automatic reading machines (ARMs) offer a promising solution for overcoming these literacy challenges, especially in regions with limited access to traditional reading materials (Ademola, 2024). However, the effectiveness of such machines hinges on the availability of high-quality TTS systems capable of accurately rendering Yoruba text into natural-sounding speech (Adeyemo & Idowu. 2015).

This paper presents the development and evaluation of a concatenative TTS system specifically designed for integration into ARMs. Concatenative synthesis leverages pre-recorded speech segments to generate synthesized speech, offering a well-established approach for achieving high-quality and natural-sounding output. Our research has two primary objectives: (i) Developing a comprehensive Yoruba speech corpus and (ii) Implementing a concatenative synthesis framework.

* Corresponding author
Email: yeqeensirajudeen@gmail.com

KWASU PRESS AND PUBLISHING

By achieving these objectives, this study aims to contribute to the advancement of Yoruba language technology and bridge the digital divide for Yoruba speakers. The developed TTS system, integrated into ARMs, will enhance accessibility of educational and other digital content, fostering literacy development and cultural preservation.

Yoruba, a tonal language, relies on pitch variations to convey meaning. Words with identical spellings can have entirely different meanings depending on the tone used. For example, Ìgbá (locust tree), Igbà (climber), Igba (two hundred), and Igbá (calabash) all share the same orthography but differ in meaning due to tonal variations. Similarly, Àjà (roof) and Ajá (dog) are distinguished solely by tone.

To represent these tonal distinctions, Yoruba utilizes three diacritical marks:

High tone (´) – acute accent (e.g., é)
Mid tone (unmarked) – may be marked with a macron (ẹ) for disambiguation
Low tone (`) – grave accent (e.g., ò)

These diacritical marks are placed on vowel letters within each syllable of a word.

The standard Yoruba orthography, established in 1974 by the Joint Consultative Committee (JCC), consists of 25 characters (Olúmúyìwá, 2013). Seven of these characters represent vowels, while the remaining eighteen function as consonants. Notably, all letters can be written in both uppercase and lowercase forms (Oladiipo, Taiwo, & Emmanuel, 2020).

1.1. Yoruba Orthography (Character Set)

The Yoruba orthography consists of 25 characters, proposed by the Joint Consultative Committee (JCC) in 1974. A few of these characters resemble Roman letters. Among the 25 characters, seven function as vowels while the remaining 18 are consonants. All 25 letters in the Yoruba character set can be written in both uppercase and lowercase forms. The standard Yoruba orthography is shown in Table 1.

Table 1: Standard Yoruba Orthography

Upper case	A B D E Ẹ F G GB I H J K L M N O Ọ P R S Ṣ T U W Y
Lower case	a b d e ẹ f g gb i h j k l m n o ọ p r s ṣ t u w y
Vowels	A E Ẹ I O Ọ U (Upper case) a e ẹ i o ọ u (Lower case)

The letter 'GB' or 'gb' is the only digraph in Yoruba orthography, a combination of two consonants. This is a special case and the only situation where two consonant letters follow each other in Yoruba orthography (Ajao et al., 2015). This combination is taken as a single entity in Yoruba character recognition systems once these two letters follow each other.

1.2. Syllable Structure in Yoruba

Standard Yoruba orthography utilizes 25 characters categorized into vowels (7) and consonants (18) (Adebayo, 2023). Vowel characters with tonal markings define the acoustic sound and meaning of words. Three distinct Yoruba syllable structures can be identified as follows (Orie, 2000)):

(i) Standalone vowel syllable (V): This occurs 21 times in standard Yoruba (e.g., a, i, u).

(ii) Consonant-vowel (CV) syllable: Formed by combining a consonant and a vowel (e.g., ba, pe, so). These are the most frequent syllables, occurring 378 times.

(iii) Nasal vowel syllables (NSV): These involve specific characters (Mm and Nn) combined with vowels to create nasalized sounds (e.g., an, in, un).

Standard Yoruba has five nasalized vowels, totalling 270 occurrences (Adebayo, 2023). Accounting for these variations, the total number of Yoruba phoneme syllables is 669, while the count of grapheme syllables reaches 1,338. Table 1 showcases examples of syllable structures in standard Yoruba.

Table 2: Samples of syllable structure in SY

Words	Yorùbá Syllable Structure		
	C	CV	NSV
Àjà	/À/	/jà/	
Ikán	/I/		/kán/
Adéwálé	/A/	/dé/ /wá/ /lé/	
ìbíyẹmí	/ì/	/bí/ /yẹ/ /mí/	
ìkárùn	/ì/	/ká/	/rùn/

This research contributes to the advancement of Yoruba language technology and holds promise for promoting literacy development for Yoruba cultural preservation. The concatenative approach effectively generates natural and intelligible Yoruba speech. The paper is organised as follows; section one deals with the introduction, followed by related works in section two. Section three elucidates the methodology of the work while section four gives experimental results and discussion, and finally, section five presents the conclusion of the research work.

2. Review of Related Works

Text-to-Speech (TTS) involves two stages: the front-end and the back-end. The front-end involves text preparation, known as text tokenization, normalization, or text recognition (Isewon, 2014). The back-end involves speech synthesis. This section reviews previous works on both optical character recognition (OCR) and text-to-speech synthesis.

3. Methodology

The concept of Text-to-Speech technology combines the capabilities of optical character recognition and speech synthesis systems to enable the conversion of text captured by a camera, scanner, or other means into synthesized speech. This integration offers a convenient and accessible way to convert printed information into an audio format.

3.1. The proposed model

The proposed concatenative Yoruba Text-to-Speech (TTS) model, shown in Figure 1, consists of several modules:

1 Data Acquisition Module: Collects and records Yoruba speech samples.
2. Yoruba Optical Character Recognition (YOCR) Module: Recognizes and processes Yoruba characters from images.
3. Speech Synthesis Module: Synthesizes speech from recognized text.
4. Output Module: Outputs the synthesized speech.

Table 3: Review of related works

S/N	Author(s)	The Work	The Method	Strength of the work	Research gap
1	Katiyar et al. (2017)	Offline Yoruba handwritten character recognition system.	Support Vector Machine (SVM).	Reported 76.7% recognition rate.	Drawbacks of SVM were not mitigated resulting in low performance (Karamizadeh et al., 2014).
2	Oni & Asahiah (2020)	modelling of an OCR system for Yoruba printed text images.	Long Short-Term Memory (LSTM) algorithm.	Reported varying Character Error Rates (CER) due to LSTM's sensitivity to parameter tuning.	High computational costs and large memory requirements.
3	Yu, Kim, & Choi (2023)	OCR in Field Programmable Gate Arrays (FPGA).	Memory-centric and memory-tree-based method.	Reported 34.24µs execution time at 18.59W power consumption.	Memory centric and memory-tree based algorithms are associated with overfitting which usually results in power consumption.
4	Oyeniran et al. (2021)	Developed an improved database of handwritten Yoruba characters.		Reported 66.7% recognition rate.	No classification method reported. Character image enhancement method used was not mentioned.
5	Akinwonmi & Alese (2013)	TTS synthesis using prosodic fusion methods.	Prosodic fusion method.		Voice mismatch, high computational intensity due prosody grouping and intonation differentiation.
6	Iyanda & Ninan (2017)	Developed Yoruba TTS system.	Festival algorithm.	Recorded 55.56% intelligibility and 50% naturalness.	Open-Source TTS system, for less-researched languages like Yoruba, will always result in poor performance for naturalness and intelligibility (Rehm & Uszkoreit, 2012; Kuligowska et al., 2018).

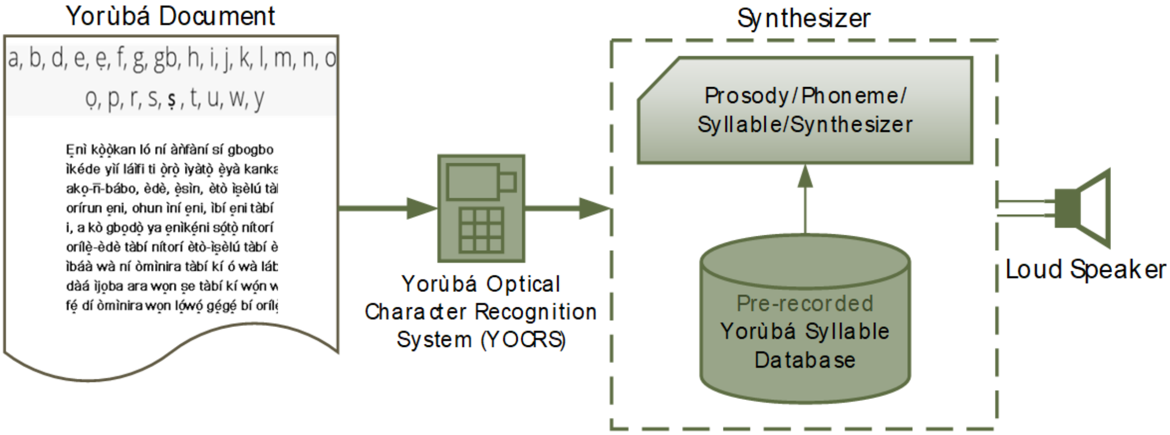

Yorùbá Document
a, b, d, e, ẹ, f, g, gb, h, i, j, k, l, m, n, o
ọ, p, r, s, ṣ , t, u, w, y
Ẹnì kọ̀ọ̀kan ló ní àǹfàní sí gbogbo
ìkéde yìí láìfi ti ọ̀rọ̀ ìyàtọ̀ ẹ̀yà kanka
akọ-ñ-bábo, èdè, ẹ̀sìn, ètò ìṣèlú tàl
orírun ẹni, ohun ìní ẹni, ìbí ẹni tàbí
i, a kò gbọdọ̀ ya ẹnìkẹ́ni sọ́tọ̀ nítorí
orílẹ̀-èdè tàbí nítorí ètò-ìṣèlú tàbí è
ìbáà wà ní òmìnira tàbí kí ó wà láb
dàá ìjọba ara wọn ṣe tàbí kí wọ́n w
fẹ́ dí òmìnira wọn lọ́wọ́ gẹ́gẹ́ bí orílẹ
Yorùbá Optical
Character Recognition
System (YOCRS)
Synthesizer
Prosody/Phoneme/
Syllable/Synthesizer
Pre-recorded
Yorùbá Syllable
Database
Loud Speaker

Figure 1: The proposed Concatenative Yoruba TTS

Each of these modules will be elucidated in the subsequent sections.

3.2. *Optical character recognition*

Optical Character Recognition System (OCRS) typically includes four modules: image acquisition, character pre-processing, segmentation and feature extraction, and recognition. A post-processing module is sometimes included to fine-tune the output for better results. The flowchart for the Yoruba Optical Character Recognition System (YOCRS) is shown in Figure 2. It begins with image acquisition, followed by pre-processing to convert colour images to grayscale to reduce storage space. Image enhancement techniques like filtering and smoothing are applied to remove artefacts and emphasize required details.

The character recognition was achieved using normalized cross correlation template matching method. In Template matching algorithms, the matching block computes match metric values by sliding a template over a region or the entire image and then find the best match (Desai et al., 2014). The template matching metric used was Sum of Absolute Difference (SAD), the equation for SAD template matching metric is given as:

$$SAD = \sum_{i=1}^{n} |X_i - Y_i| \qquad (1)$$

Where Xi and Yi are the elements at index i in arrays X and Y.

The segmented characters were also size-normalized to equal dimensions. The cross-correlation was then carried out on the input images' (document image) segmented characters and the characters in the database. The algorithm for normalized Cross-Correlation (NCC) between two arrays X and Y of the same length n is given in equations 2 and 3:

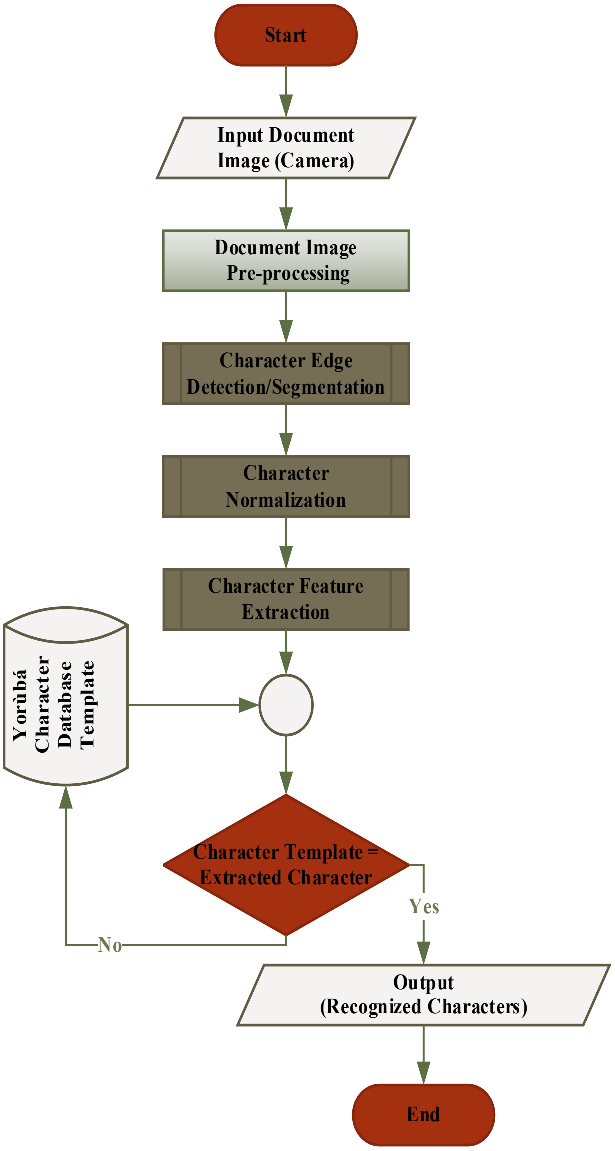


Figure 2: Yorùbá optical character recognition system (YOCRS) flowchart

1. Calculate the mean of array X and array Y:

$$mean_X = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (2)$$

$$mean_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad (3)$$

2. Equation 4 and 5 Calculates the standard deviation of array X and array Y:

$$std_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (X_i - mean_X)^2} \qquad (4)$$

$$std_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i - mean_Y)^2} \qquad (5)$$

3. Equation 6 calculates the Normalized Cross-Correlation (NCC) between arrays X and Y:

$$NCC = \frac{\sum_{i=1}^{n} (X_i - mean_X)\,(Y_i - mean_Y)}{n \times std_X \times std_Y} \qquad (6)$$

The NCC value ranges between -1 and 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

3.3. Yoruba Character Unicode Mapping

Segmented and recognized Yoruba characters are encoded and mapped to the Unicode character set. Unicode code points range from U+00000 to U+10FFFF, including standard ASCII and extended ASCII. UTF-8 was selected for Yoruba character coding to conserve memory space, as it represents characters in multiples of eight bits or one byte.

3.4. Yorùbá Character Encoding Algorithm

Algorithm 1 describes the Yoruba character coding steps and the storage in the system memory.

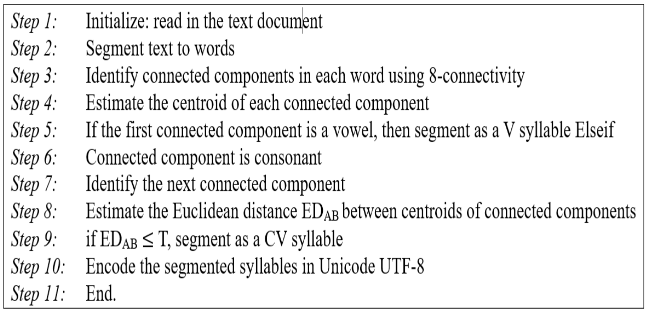
3.5. Grapheme Yorùbá Syllable Segmentation

Yorùbá is a tonal language whose most acoustically and perceptual coherent unit of sound is its syllables (Kumolalo et al., 2017). As mentioned earlier, Yorùbá language has three syllables. The purpose of this section is to segment grapheme Yorùbá words into their syllables. Connected component analysis and Euclidean distance algorithms are being implemented to achieve this. The input Yorùbá text was being returned as tokenized Yorùbá syllables. The algorithm to achieve this is given as Algorithm 2.

Algorithm 1: Yoruba Character Coding Steps

```
Step 1:  Identify the code-point (cp) the character
Step 2:  If the decimal cp ≤ 127, allocate 1 byte to store the character
Step 3:  Else
Step 4:  127< cp ≤ 2047, allocate 2 bytes to store the character
Step 5:  Convert the decimal cp to the binary
Step 6:  Fill the remaining six bits of the last continuation byte with the binary, starting
         with the least significant bit
Step 7:  Then fill the first six bits of the leading byte (if two bytes are used) or next
         continuation byte (if more than two bytes are used) with the remaining binary.
Step 8:  Pad the extra vacant bits with zeros if any
Step 9:  End
```

Algorithm 2: Tokenization of Yorùbá Syllables



```
Step 1:  Initialize: read in the text document
Step 2:  Segment text to words
Step 3:  Identify connected components in each word using 8-connectivity
Step 4:  Estimate the centroid of each connected component
Step 5:  If the first connected component is a vowel, then segment as a V syllable Elseif
Step 6:  Connected component is consonant
Step 7:  Identify the next connected component
Step 8:  Estimate the Euclidean distance $ED_{AB}$ between centroids of connected components
Step 9:  if $ED_{AB} \leq T$, segment as a CV syllable
Step 10: Encode the segmented syllables in Unicode UTF-8
Step 11: End.
```

3.6 Yorùbá Concatenative Speech Synthesis

The concatenative synthesis produces the best synthetic speech with a high degree of naturalness and intelligibility in terms of rhythms and timbre close to the original voice actor (Tan et al., 2021). The computational cost is low and the storage requirement depends on the corpora selected as the concatenation units (Kayte et al., 2015). The front-end, and back-end are independent components, Yorùbá document reader only requires their careful integration at the final stage. The flowchart of the Yorùbá concatenative speech synthesis is shown in Figure 3.

Ten (10) professional Yorùbá speakers were engaged in recording of Yorùbá phonemes, 669 syllables, and some Yorùbá words at Tiwa N Tiwa (TNT) FM (102.5MHz) radio station Ijagbo, Kwara state, Nigeria. Each of the recordings has an average duration of 1095 seconds and a memory space of 25.3MB. The best recordings with minimum variation in tone and pitch were selected for further pre-processing. The segmentation, pitch marking, and annotation of the recordings were carried out with Phonetic Representation and analysis of Speech (PRAAT) version 6.1.47. The annotation of recorded syllables was carried out using Unicode UTF-8 encoding.

3.7. Yoruba Syllable Speech Signal windowing

Windowing optimizes the computational requirement and analysis of speech signals. Windowing smoothly tapers the signal at its ends, for concatenative synthesis, the tapering reduces the frequency mismatch and spectral leakages at the concatenative points. Segmented Yoruba syllable speech signals were convolved with the Hamming window function in equation (7), the result of the convolution is given in equation (8) which is a windowed speech signal.

$$\omega(n) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi n}{N}\right) & 0 \le n \le N \\ 0 & else \end{cases} \quad (7)$$

In its simplest form, speech signal in frequency domain is given as

$$X[k] = \sum_{k=0}^{N} x[n]e^{-j\frac{2\pi nk}{N}} \quad (8)$$

$where\ X[k] -\ the\ speech\ signal\ in\ frequency$

$x[n] -\ the\ speech\ signal\ in\ the\ time\ domain$

$n - the\ sampling\ time$

$N -\ the\ length\ of\ the\ speech\ signal$

therefore, the windowed syllable speech signal will be given as

$$y[n] = x[n] * \omega[n] \quad (9)$$

3.8 Concatenated Yorùbá Syllable Speech Signal Pre-Emphasis Filtering

Pre-emphasis filters are essentially a Finite Impulse Response (FIR) filter. FIR filters are causal systems whose output depends on past and present input. The causality of the FIR filter is mitigated by shifting and truncation of the impulse response using windowing, and also discretizing the coefficient of its impulse response. Emphasis filter as a non-recursive first-order autoregressive FIR filter is mathematically represented as:

$$y[n] = x[n] \pm \alpha x[n-1] \quad (10)$$

$where\ \ y[n] -\ the\ output\ signal\ at\ the\ curren$

$x[n] -\ the\ input\ signal\ at\ the\ \ current\ sampli$

$\alpha\ is\ the\ emphasis\ coefficent\ that\ determines$

$previous\ sample\ on\ the\ current\ sample$ (Kennedy, 2023).

Pre-emphasis Filter Design Specification (López-Espejo et al., 2024):

1. Passband cutoff frequency ωp <= 4.1kHz
2. Stopband cutoff frequency ωs=> 4.15kHz
3. Stopband deviation δp= 0.01
4. Stopband attenuation = 20logδp = -40dB
5. Transition band Δf = ωs- ωp = 0.05 kHz
6. Sampling frequency f_s = 4.1kHz
7. No of sample N required is given as $\Delta f = \frac{4}{N}$, $\mathrm{N} = \frac{4}{0.05} = 80$
8. Cut-off frequency ωc= $\frac{\omega p\ \ +\ \omega s}{2} = 4.125kHz$
9. Normalized cut-off frequency = $\frac{\omega c}{f_s} = \frac{4.125}{4.1} =$ 1.00609756

The ideal impulse response of low-pass FIR filter is given as Hamming window function for the design as:

$$\omega(n) = 0.54 + 0.46cos\frac{2\pi n}{80} \qquad 0 \le n \le 80 \quad (11)$$

(11)

$$h_I(n) = \begin{cases} 2f_c\frac{\sin(n2\pi f_c)}{n2\pi f_c} & , \quad n \ne 0 \\ & \quad 0 \end{cases} \quad (12)$$

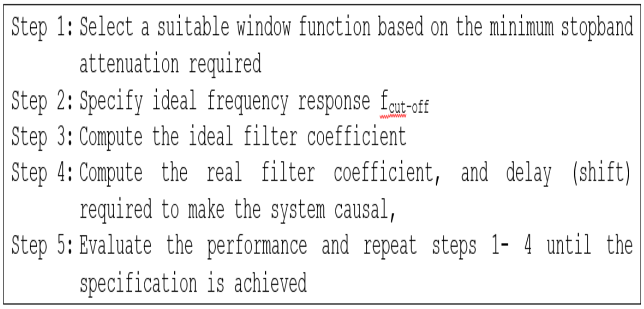
Pre-emphasis Filter Design Algorithm

The algorithm 3 describes the step-by-step implementation of design specification of pre- emphasis filtering.

4. Results

The results showed a high impact of windowing the Yoruba syllable speech signal before concatenation and a positive effect of pre-emphasis filtering of concatenated syllable speech signals. Table 3 shows the comparison of windowed and unwindowed syllable speech signals. The values of signal energy of windowed signal which is approximately zero (0) at the edges i.e. 100 x 10-2ms and (1741 – 1750) x 10-2ms. while the value of signal energy for both windowed and unwindowed signals was approximately the same at the centre of the signal. This results in the smooth joining of the concatenated syllable reducing the spectral mismatch at the concatenative joints.

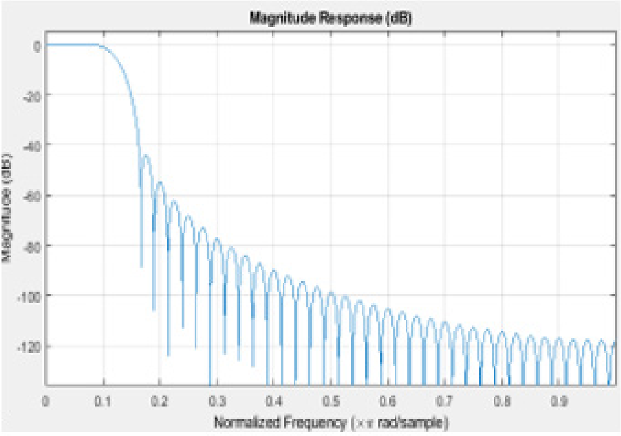
Algorithm 3: Design Specification of Pre- Emphasis Filtering.



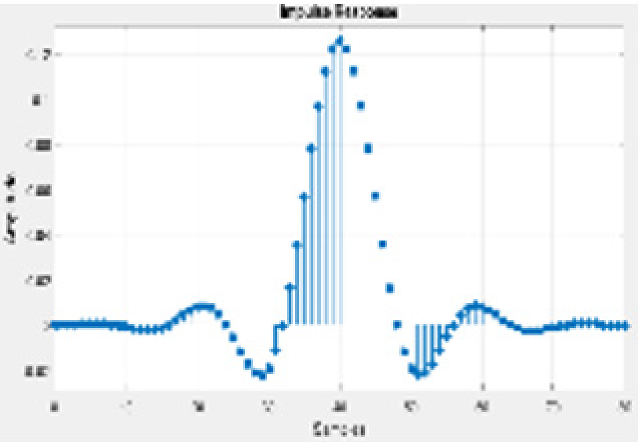
```
Step 1: Select a suitable window function based on the minimum stopband
        attenuation required
Step 2: Specify ideal frequency response $f_{cut-off}$
Step 3: Compute the ideal filter coefficient
Step 4: Compute the real filter coefficient, and delay (shift)
        required to make the system causal,
Step 5: Evaluate the performance and repeat steps 1- 4 until the
        specification is achieved
```

The responses of the designed filter are shown in Figure 4. The magnitude and the impulse response of the designed filter are as expected, the impulse response is symmetrical about the 40th sample, which means that the total number of samples will be eighty (80), and the settling time is also within the desired number of samples.

The designed filter was tested by corrupting some concatenated syllable speech signals with white Gaussian noise of probability density function 0.05. The signal-to-noise ratio (SNR) of the corrupted speech signals was enhanced with a minimum of 3dB. Table 4 shows the comparison between the power spectral density (PSD) of the corrupted speech signal and the filtered speech signal. The PSD of the filtered signal is approximately equal to the PSD of the original (uncorrupted speech signal).



(a) The magnitude response of the designed filter



(b) The impulse response of the designed filter

Figure 4: The responses of the designed filter

4.1. Test and validation of the System

The results showed that the concatenation of two-syllable words has higher MOS compared with three, four, and five-syllable words. Also,

Table 3: Energy of Windowed and Unwindowed speech signal

Time x 10^{-2} (ms)	Energy of Unwindowed Signal	Energy of windowed signal	Time x 10^{-2} (ms)	Energy of Unwindowed Signal	Energy of windowed signal	Time x 10^{-2} (ms)	Energy of Unwindowed Signal	Energy of windowed signal
1	0.3285	0.0263	951	0.1387	0.1380	1741	0.3035	0.0683
2	0.3484	0.0279	952	0.1054	0.1049	1742	0.2381	0.0534
3	0.3635	0.0291	953	0.0767	0.0763	1743	0.1854	0.0413
4	0.3654	0.0292	954	0.0541	0.0538	1744	0.1474	0.0327
5	0.3475	0.0278	955	0.0417	0.0415	1745	0.1175	0.0260
6	0.3148	0.0252	956	0.0363	0.0361	1746	0.0809	0.0178
7	0.2811	0.0225	957	0.0324	0.0322	1747	0.0391	0.0085
8	0.2597	0.0208	958	0.0320	0.0319	1748	-0.0010	-0.0002
9	0.2538	0.0203	959	0.0311	0.0310	1749	-0.0472	-0.0102
10	0.2580	0.0207	960	0.0215	0.0214	1750	-0.1004	-0.0217

Table 4: Comparison of Power Spectral Density of Original, Corrupted and Filtered concatenated speech signal

Frequency (Hz)	Concatenated SSS Power(W)	Corrupted Concatenated SSS Power(W)	Filtered corrupted Concatenated SSS Power(W)
29.4	1.0042	2.7146	1.0002
58.8	1.0012	4.3810	1.0007
88.2	1.0005	4.2500	1.0004
117.6	1.0005	2.1652	1.0003
147.0	1.0003	5.5935	1.0002
176.4	1.0003	2.9154	1.0002
205.8	1.0003	5.5935	1.0002
235.2	1.0005	2.1652	1.0003
264.6	1.0005	4.2500	1.0004
294.0	1.0012	4.3810	1.0007

the results showed lower MOS for phrases, this is due to the increase in the number of concatenative joints. However, windowing and pre-emphasis filtering have improved the concatenation output and eventually the usability of the concatenated Yorùbá syllables. Figures 5 and 6 show the mean opinion scores and respondents' responses for the concatenated speech signals respectively.

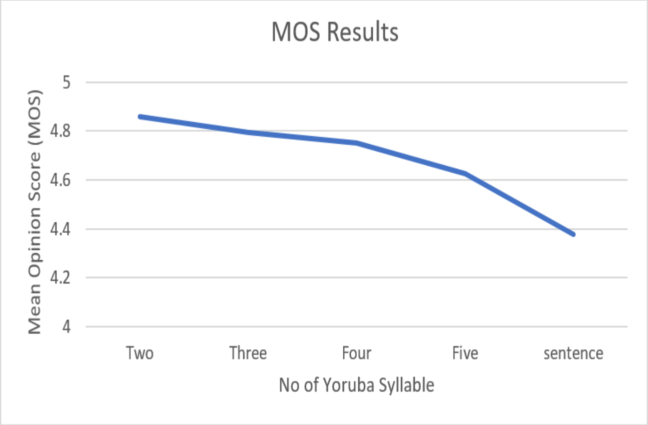


Figure 5: Mean Opinion Scores

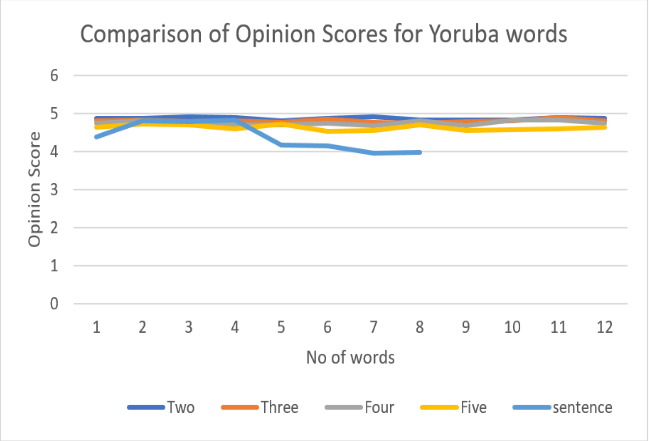


Figure 6: Respondents' Response to the Concatenated Speech Signals.

The system developed achieved a maximum intelligibility MOS of 4.86 for two-syllable Yoruba words and the least MOS of 4.37 for concatenated phrases/sentences. This showed maximum intelligibility of 97.2% and least intelligibility of 87.4% which is better when compared with 55.6% reported in the literature using Festival algorithm. An MCD value of 4.0 or lower is considered excellent quality synthetic speech, the developed system achieved a mean MCD of 1.58. This signifies minimum distortion between the recorded speech and the synthesized speech.

5.Discussion

The performance of the developed concatenative TTS system represents a significant advancement over existing Yoruba speech synthesis models. Specifically, the system achieved a maximum Mean Opinion Score (MOS) of 4.86 for two-syllable words and 4.37 for full sentences. When compared to the work of Iyanda and Ninan (2017), who utilized the open-source Festival algorithm and reported an intelligibility rate of 55.56% (approximately 2.78 MOS), our approach demonstrates a substantial increase in clarity and naturalness. This improvement validates the research premise that generic, open-source TTS systems often yield poor results for less-researched, tonal languages like Yoruba.

Unlike the "prosodic fusion" methods proposed by Akinwonmi and Alese (2013), which often suffered from voice mismatch and high computational intensity, our syllable-based concatenative framework utilizes Hamming windowing to smoothly taper signals. As shown in Table 3, windowing reduced signal energy at the edges to approximately zero, effectively mitigating the spectral leakages and frequency mismatches that typically occur at concatenative joints. Furthermore, the achieved Mean Mel Cepstral Distortion (MCD) of 1.58 is well below the industry benchmark of 4.0 for excellent quality speech, indicating significantly lower distortion than

Table 5: MCD of concatenated word Ìbíyemí and Adewale

	Ìbíyemí				Adéwálé		
S/N	mc_i^r	mc_i^c	MCD	S/N	mc_i^r	mc_i^c	MCD
1	16.3355	15.1096	1.2259	1	16.2083	15.7605	0.4478
2	16.0065	14.3055	1.7010	2	16.2848	15.6107	0.6741
3	15.4593	14.1543	1.3050	3	16.1636	15.2799	0.8837
4	15.3090	14.5436	0.7654	4	16.2052	15.1441	1.0611
5	15.4873	14.6931	0.7942	5	16.1057	14.9320	1.1737
6	15.8187	14.8139	1.0048	6	16.1131	14.8639	1.2492
7	15.9126	14.4878	1.4248	7	15.9626	14.3978	1.5648
8	15.8986	13.9162	1.9824	8	15.9903	13.8907	2.0996
9	15.9240	13.2048	2.7192	9	16.0573	12.7362	3.3211
10	16.0900	13.1314	2.9586	10	16.0700	12.8065	3.2635
Mean MCD			1.58813	Mean MCD			1.57386

previous efforts.

The integration of the Yoruba Optical Character Recognition (YOCRS) module addresses the "abysmal performance" of previous digital tools noted by Akomolede and Olowojebutu (2025). By employing Normalized Cross-Correlation (NCC), the system avoids the low-performance drawbacks associated with Support Vector Machines (SVM) reported by Katiyar et al. (2017), who achieved only a 76.7% recognition rate. This robust front-end ensures that the tonal diacritics - essential for distinguishing words like Àjà (roof) and Ajá (dog) - are accurately mapped to Unicode before synthesis.

6. Conclusion

This paper has addressed the critical issue of digital accessibility for speakers of the Yoruba language through the development and evaluation of a concatenative Text-to-Speech (TTS) system tailored for use in Automatic Reading Machines (ARM). By leveraging concatenative synthesis and carefully constructing a comprehensive Yoruba speech syllable database, we have successfully generated natural-sounding Yoruba speech, thereby overcoming the lack of robust TTS solutions for less researched languages such as Yoruba.

The implementation of a concatenative synthesis framework optimized for Yoruba phonetics, including consideration of Diacritical Tonal Marks (DTM), has been detailed. Through rigorous evaluation using both objective metrics, Mel Cepstral Distortion (MCD), and subjective measures Mean Opinion Score (MOS), we have demonstrated the intelligibility, naturalness, distortion-free, and robustness of the synthesized Yoruba speech across various linguistic contexts.

This interdisciplinary effort not only addresses the immediate need for digital accessibility in educational, legal, and other human endeavours contexts but also contributes to the broader goals of literacy development and cultural preservation among Yoruba speakers. Moving forward, continued research and development in this area will be vital for further enhancing digital accessibility and promoting the richness of the Yoruba language and culture.

References

Adebayo, T. (2023) Emerging grammars in contemporary Yoruba phonology. *Canadian Journal of Linguistics/ Revue canadienne de linguistique*, *68*(2), 250-303.

Ademola, E. O. (2024) Reading strategies in the AI age: Enhancing comprehension and engagement with advanced technologies. In *Proceedings of the 38th iSTEAMS Multidisciplinary Bespoke Conference*.

Adeyemo, O. O., & Idowu, A. (2015) Development and Integration of Text to Speech Usability Interface for Visually Impaired Users in Yoruba Language. *African Journal of Computing and ICT*, *8*(1), 87-94.

Ajao, J. F., Olabiyisi, S. O., Omidiora, E. O., & Odejobi, O. O. (2015) Yoruba handwriting word recognition quality evaluation of preprocessing attributes using information theory approach. *International Journal of Applied Information Systems*, *9*(1), 18-23.

Akinwonmi, A. E., & Alese, B. K. (2013) A prosodic text-to-speech system for yorùbá language. In *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)* (pp. 630-635). IEEE.

Akomolede, K. K., & Olowojebutu, A. O. (2025) Enhancing Digital Inclusion through AI-Based Yorùbá Language Localization: Challenges, Solutions, and Future Prospects. *Tech-Sphere Journal for Pure and Applied Sciences*, *2*(1).

Desai, B. K., Potdar, D., & Patel, M. P. (2014) Template matching technique using enhanced SAD technique. *International Journal of Engineering Research & Technology*, *3*(5), 1371-1376.

Isewon, I., Oyelade, J., & Oladipupo, O. (2014) Design and implementation of text to speech conversion for visually impaired people. *International Journal of Applied Information Systems*, *7*(2), 25-30.

Iyanda, A. R., & Ninan, O. D. (2017) Development of a Yoruba Text-to-Speech System Using Festival. Innovative Systems Design and Engineering, 8(5), 1–9.

Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J., & javad Rajabi, M. (2014) Advantage and drawback of support vector machine functionality. In *2014 international conference on computer, communications, and control technology (I4CT)* (pp. 63-65). IEEE.

Kasthuri, E., & James, A. P. (2012) Perceptive Speech Filters for Speech Signal Noise Reduction. *International Journal of Computer Applications, 55(18).*

Katiyar, G., Katiyar, A., & Mehfuz, S. (2017) Off-line handwritten character recognition system using support vector machine. *American Journal of Neural Networks and Applications*, *3*(2), 22-28.

Kayte, S. N., Mundada, M., Kayte, C. N., & Gawali, D. (2015) Approach of Syllable Based Unit Selection Text-To-Speech Synthesis System for Marathi Using Three Level Fall Back Technique. IOSR Journal of VLSI and Signal Processing, 5(6), 31–35. Retrieved from www.iosrjournals.org

Kennedy, H. L. (2023) Recursive and non-recursive filters for sequential smoothing and prediction with instantaneous phase and frequency estimation applications (extended version). *arXiv preprint arXiv:2311.07089*.

Kuligowska, K., Kisielewicz, P., & Włodarz, A. (2018) Speech synthesis systems: disadvantages and limitations. *Int J Res Eng Technol (UAE)*, *7*, 234-239.

Kumolalo, F. O., Adagunodo, E. R., & Odejobi, O. A. (2010) Development of a syllabicator for Yorùbá language. *Proceedings of OAU TekConf*, 47-51.

López-Espejo, I., Joglekar, A., Peinado, A. M., & Jensen, J. (2024) On speech pre-emphasis as a simple and inexpensive method to boost speech enhancement. *arXiv preprint arXiv:2401.09315*.

Oladiipo, A. F., Taiwo, O. M., & Emmanuel, A. A. (2020) Spelling error patterns in typed yorubá text documents. *International Journal of Information Engineering & Electronic Business*, *12*(6).

Olúmúyìwá, T. (2013) Yoruba writing: standards and trends. *Journal of Arts and Humanities*, *2*(1), 40-51.

Oni, O. J., & Asahiah, F. O. (2020) Computational modelling of an optical character recognition system for Yorùbá printed text images. *Scientific African*, *9*, e00415.

Orie, O. O. (2000) Syllable asymmetries in comparative Yoruba phonology. *Journal of Linguistics*, *36*(1), 39-84.

Oyeniran, O. O., Oyeniyi, J. O., Omotosho, L. O., & Ogundoyin, I. K. (2021) Development of an improved database for yoruba handwritten character. *Journal of Engineering Studies and Research*, *27*(4), 84-89.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021) A Survey on Neural Speech Synthesis. arXiv:2106.15561 [eess.AS]. Retrieved from http://arxiv.org/abs/2106.15561

Yu, K., Kim, M., & Choi, J. R. (2023) Memory-Tree Based Design of Optical Character Recognition in FPGA. *Electronics*, 12(3), 1–16.

Yusuff, A., & Osunnuga, O. (2018) Issues and challenges of adopting digital technologies by African language media: The Yorùbá example. *African language digital media and communication*, 209-215.