INTERNATIONAL JOUNAL OF: ENGINEERING RESEARCH AND TECHNOLOGY

TIMBOU-AFRICA ACADEMIC PUBLICATIONS
INTERNATIONAL JOURNAL, MAY, 2022 EDITIONS
VOL. 9 NO. 5 ISSN: 2707-5654

SHORT TERM ENERGY ALGORITHM FOR EFFECTIVE CONCATENATIVE SYNTHESIS OF YORUBA SYLLABLES

YEKEEN, SURAJUDEEN ADÉWÁLÉ [1], ÌBÍYẸMÍ, TÚNJÍ SAMUEL[2] OMEIZA, ISAAC O.A.[3] AND AJÍBÓYÈ AYE T.[4]

[1]Department of Electrical/Electronic Engineering, The Federal Polytechnic, Offa, Nigeria. [2,3]Department of Electrical/Electronic Engineering, University of Ilorin, Nigeria. [4]Department of Computer Engineering, University of Ilorin, Nigeria.

ABSTRACT

The existing method of reading the Yorùbá document is manual, thereby limiting the benefit of automatic reading machines and automation of other human endeavours. Most of the existing Text-To-Speech (TTS) applications takes their input from

Introduction

The field of digital signal processing has experienced continuous and significant expansion in development over years (Dastres and Soori 2021). The usefulness of this technology is apparent in many areas of human endeavours, this includes but is not limited to artificial intelligence (AI), machine vision, geological and oil exploration, remote sensing, pattern recognition and human-machine interaction (HCI) as well as image processing.

computer peripherals e.g. keyboard, there is the need for a Yorùbá document reader which takes input from other source and produce TTS with high intelligibility and naturalness. The paper elucidates limitations of concatenative synthesis being the most effective TTS method that produces the best natural and intelligible speech synthesis. Particularly the limitation of co-articulation of concatenated segments. Short Term Energy (STE) of frames in recorded syllables was used to improve the co-articulation of concatenated segments thereby reducing the prosodic discontinuity and improving the naturalness factors (intonation and rhythm). Fifty native speakers were engaged to listen to synthesized concatenated syllables, a Mean Opinion Score (MOS) of 3.8 and 4.2 were achieved for both intelligibility and naturalness respectively.

Keywords: *Concatenative, Reader, speech, STE (Short Term Energy), Text-to-Speech (TTS)*

Automatic document reading is the application of text-to-speech (TTS) synthesis. TTS applications have been found useful in different areas of human endeavours in several languages of the world. While European language has received significant research attention in the area of TTS synthesis and its applications, Yorùbá language which is spoken by over thirty million (30,000,000) people in West African countries (Afolabi, Omidiora, and Arulogun 2013), a huge number of people in the rest of Africa, united kingdom and part of South America has received less attention. Hence, there is a need for research into Yorùbá TTS application systems.

Synthesized speech may be produced using different techniques which include articulatory synthesis, formant synthesis, and concatenative synthesis. All of these have their merit and demerits.

TIJERAT
ISSN: 2707-5654

Articulatory synthesis is based on a model of human vocal tract behaviour during speech mechanism. It possibly has the best synthetic speech when compared to other methods. It is the most difficult method of artificial speech production because it involves complex mathematical equation modeling(Ghosh and Reddy 2011)

Formant synthesis method is a rule-based method. It is based on the rule of the spoken language(Singh and Singh 2012).

Concatenative method is based on a mathematical model. The model coalesced phonemes or syllables to form continuous speech fragments. Speech is synthesized by joining together several pre-recorded units in form of phonemes or syllables or words (Shreekanth, Udayashankara, and Arun 2014).

REVIEW OF RELATED WORKS

Sakai and Shu (Sakai and Shu 2005) proposed a probabilistic approach to corpus-based using unit selection synthesis. The speech features proposed were fundamental frequency contour Fo, duration, spectral characteristics of the units for selection. The target cost and join cost were formulated in a probabilistic framework. The drawback of this is that probabilistic approach to TTS requires the whole words and phrases of the language, then select possible sequence of word match. Hence a very large database is required with high speed processing machine for implementation

Rashad, El-bakry and Isma (Rashad, El-bakry, and Isma 2010) proposed diphone concatenative speech synthesis using MARYTTS and evaluatethe result by the Diagnostic Rhyme Test (DRT) that measures the intelligibility of the synthesized speech and the Categorical Estimation (CE). The drawback, MARY TTS provides script which rely on Linux environment and requires administrative access to running MySQL services, the system easily run into

TIJERAT
ISSN: 2707-5654

unhandled edge error on database management (Sprachsignalverarbeitung et al. 2017).

Patra et al. (Patra et al. 2012) presented the TTS conversion with phonematic concatenation. In the work, the pre-recorded similar English sounds were separated into syllables and concatenate the separated syllables to reconstruct the desired word using Matlab matrix operation. The pre-recorded words were based on similar English words, for example, main, gain, pain, etc. limit the robustness of the system. The drawbacks include robustness of the system, as limited number of word with similar pronunciation are considered.

Afolabi et al. (Afolabi et al. 2013) proposed a Yorùbá TTS by concatenation method but end up using High Level Synthesis (HLS) and Natural Language Processing (NLP) method to actualize the system. The HLS consists of basically two modules, namely the text analysis module and the prosody module. HLS and NLP are open source TTS engine. Open source TTS engine for non-well researched language like Yoruba will always result in poor performance as modules for morphological analysis of part of speech of the language will not be available(Kuligowska, Kisielewicz, and Włodarz 2019)

Pravin, Ghate and Shirbahadurkar (Pravin, M Ghate and Shirbahadurkar 2017) surveyed various method of TTS synthesis for quality synthesized speech for the Marathi language. The authors opined that all methods worked well for the language. The surveyed methods include articulatory synthesis, Formant synthesis, concatenative synthesis, Pitch Synchronous Overlap and Add PSOLA method. Linear prediction-based method, sinusoidal model, and High-Level Synthesis HLS.

Iyanda and Ninan (Iyanda and Ninan 2017) presented the TTS synthesis using open-source festival TTS engine for Yorùbá language with prosodic modelling using Classification and

Regression Tree (CART). The CART Tree returns two levels of break, B and BB, the BB indicates the end of an utterance. The authors were able to record 55.5 % and 50% intelligibility and naturalness at word level respectively. The drawback is that Open source TTS engine for non-well researched language like Yoruba will always result in poor performance as modules for morphological analysis of part of speech of the language will not be available.

YORÙBÁ ORTHOGRAPHY

Yorùbá orthography consists of twenty-five (25) character sets, most of which are similar to Roman characters. Eighteen (18) of these characters set are consonants while the remaining seven (7) are vowels. Among the consonants is a digraph (GB) which combines two consonants(Ajao, Olabiyisi, and Omidiora 2015).
The Yorùbá consonant characters are as follows:
Upper case: B D F G GB H J K L M N P R S Ṣ T W Y
Lower case: b d f g gb h j k l m n p r s ṣ t w y
Yorùbá Vowel characters are as follows:
Upper case: A E Ẹ I O Ọ U
Lower case characters: a e ẹ i o ọ u

YORÙBÁ SYLLABLE STRUCTURE

There are several dialects in the Yorùbá language with each being peculiar to settlement or regions, for effective communication Standard Yorùbá (SY)is adopted. SY is also the language of instruction in education, mass media communication, and adopted character set in textbooks(Eludiora & Odejobi, 2016). There are three basic syllable structures in the SY language(Olmsted and Olmsted 2015); these are:
(i) Vowel (V) syllable: this is a standalone syllable. It always occurs at the beginning of a Yorùbá word, particularly a noun. This

TIJERAT
ISSN: 2707-5654

syllabic structure has twenty-one (21) occurrences in the SY language.

(ii) Consonant Vowel (CV) Syllable: This syllabic structure in the SY orthography consist of a consonant followed immediately with a vowel character. This syllable has three hundred and seventy-eight (378) occurrences in SY.

(iii) Nasal Vowel; this consist of two characters Mm and Nn

(iv) Nasalized syllabic vowel (NSV)

The SY has five (5) Nasalized vowels or phonemes (Akinwonmi and Alese 2013). These nasalized vowel combines with consonant to form a nasalized syllabic vowel. The five nasalized vowels are; an, ẹn, in, ọn, un. The nasalized syllables in SY are two hundred and seventy (270) occurrences. Therefore, there are six hundred and ninety (669) phoneme syllables in the SY language. While grapheme syllables are one thousand three hundred and thirty-eight (1338)

Syllables were chosen as the corpus unit because; Yorùbá is a tonal language, it is the syllable that bears the tone of a word. The tonal marks or diacritics are indicated on the vowel component of the syllable (Iyanda and Ninan 2017),(Odéjobí, Beaumont, and Wong 2006).

Table 1: Examples of the syllabic structures in SY

	Syllabic Structure		
Word	C	CV	NSV
Àjà	/À/	/jà/	
Íkán	/Í/		/kán/
Adewále	/A/	/de/ /wá/ /le/	
Olatunjí	/O/	/jí/ /la/	/tun/

TIJERAT
ISSN: 2707-5654

METHODOLOGY

As mentioned earlier, the Yorùbá document reader is based on TTS synthesis, the processes are basically of two stages; the front end and the back end. The front end involves character recognition and text analysis. This is also called text normalization or text tokenization(Kayte, Mundada, and Kayte 2015). Yorùbá optical character recognition system (YOCRS) was developed to recognize Yorùbá character from the document image.

First, the document image was pre-processed, the pre-processing include; RGB-Grey scale image, This step was carried out using the Luma correction algorithm, then filtering the image using the Weiner filter which is a bandpass filter. Binarization of document images was carried out using Otsu's algorithm for effective edge detection of the characters. De-Skewing of the document for skew correction and finally the character recognition was done using Template matching. Sum of Absolution Difference (SAD) metric was used. The character recognition result was validated using Normalized Cross-Correlation (NCC). The recognized characters were coded and mapped to the Unicode character set.

Unicode is a multibyte character set, a character can be 1byte, 2bytes up to 6bytes. This multibyte attribute gives the flexibility of mapping characters with large code point (CP). Unicode has backward compatibility with ASCII, particularly UTF-8. Unicode character sets have individual code points (CP). Characters with CP greater than 127 but less or equal to 2047 uses 2bytes for coding. There will be a leading byte for characters with multiple bytes and continuation byte(s).

UTF-8 Coding and Decoding of Yoruba Characters are implemented as shown in the following algorithm.

TIJERAT
ISSN: 2707-5654

UTF-8 Character Encoding Algorithm

1. Start: initialize
2. Pre-allocate memory space
3. Identify the character code point (CP)
4. If CP <= 127
5. 1byte memory required
6. Elseif
7. 127 < CP <=2047
8. 2byte memory required
9. Convert decimal CP to binary
10. Fill the first six bits of the last continuation byte with the binary
11. Then fill the first six of each of the next continuation bytes
12. End

UTF-8 Encoding Algorithm Implementation

To encode Yorùbá vowel character **à**
The hexadecimal code point of **à** is U+00E0, and
the decimal code point is 0224,
the binary equivalent is 11100000.
The decimal CP of **à** is 0224,
therefore, 2bytes are required as the CP is greater than 127
à is coded as 1100001110100000

Character Decoding Algorithm

1. Start
2. Determine the number of bytes used for encoding
3. Read the binary code of the leading byte.
4. Read and append the binary code of the subsequent continuation byte
5. Convert the resulting binary code to decimal and hexadecimal code point
6. Map the code point to character set
7. End

Table 2 shows the code points of Yoruba characters, particularly the vowel characters with diacritical (tone) marks, vowels characters were given priority because the remaining consonant character has compatibility with the ASCII code. With the character code points, the memory address of the storage of the character is determined and easily accessed.

TIJERAT
ISSN: 2707-5654

Table 2: Yorùbá vowel syllable characters in UTF-8 Character set with Decimal and Hexadecimal Code Point

	Character	Hexadecimal CP	Decimal CP
22	Ì	00EC	236
23	Ī	012A	298
24	ī	012B	299
25	Ó	00D3	211
26	ó	00F3	243
27	Ò	00D2	210
28	ò	00F2	242
29	Ō	014C	332
30	ō	014D	333
31	Ọ	1ECC	7884
32	ọ	1ECD	7885
33	Ọ́	[1ECC0301]	[7884 769]
34	ọ́	[1ECD0301]	[7885 769]
35	Ọ̀	[1ECC0300]	[7884 768]
36	ọ̀	[1ECD300]	[78845 768]
37	Ú	00DA	218
38	ú	00FA	250
39	Ù	00D9	217
40	ū	016B	363
41	Ū	016A	362
42	ù	00F9	249

TIJERAT
ISSN: 2707-5654

	Character	Hexadecimal CP	Decimal CP
1	Á	00C1	193
2	á	00E1	225
3	À	00C0	192
4	à	00E0	224
5	Ā	100	256
6	ā	101	257
7	É	00C9	201
8	È	00C8	200
9	Ē	0112	274
10	è	00E8	232
11	é	00E9	233
12	ē	0113	275
13	Ẹ	1EB8	7864
14	ẹ	1EB9	7865
15	Ẹ́	[1EB80301]	[7864 769]
16	ẹ́	[1EB90301]	[7865 769]
17	Ẹ̀	[1EB80300]	[7864 768]
18	ẹ̀	[1EB90300]	[7865 768]
19	Í	00CD	205
20	í	00ED	237
21	Ì	00CC	204

TIJERAT
ISSN: 2707-5654

YORÙBÁ AUDIO SYLLABLE DATABASE

Front-ends and Back-ends are independent components, the camera-based document reader proposed in this study only requires careful integration of both at the final stage. Ten Yorùbá professional native speakers were engaged to read and record Yorùbá phonemes, 669 syllables and some Yorùbá words at TNT FM (102.5MHz) radio station. Each of the recordings took a duration of 1095seconds and memory space of 25.3MB on average. The best speaker with minimum variation in tone and pitch was selected for further pre-processing. The segmentation, pitch marking and annotation of the recordings were carried out with PRAAT version 6.1.47.

Smoothing of the recorded syllables was carried out using Hanning windowing while silence and unvoiced segment were removed using the autocorrelation algorithm.

Figure 2 shows the schematic diagram of the camera-based Yorùbá document reader. The specifications of the digital camera used are 1080P resolution of 12M pixel, with CMOS image sensor, frame rate 30fps, and focal length 8mm - ∞. The image from the camera serves as input for the Yorùbá optical character recognition system (YOCRS).

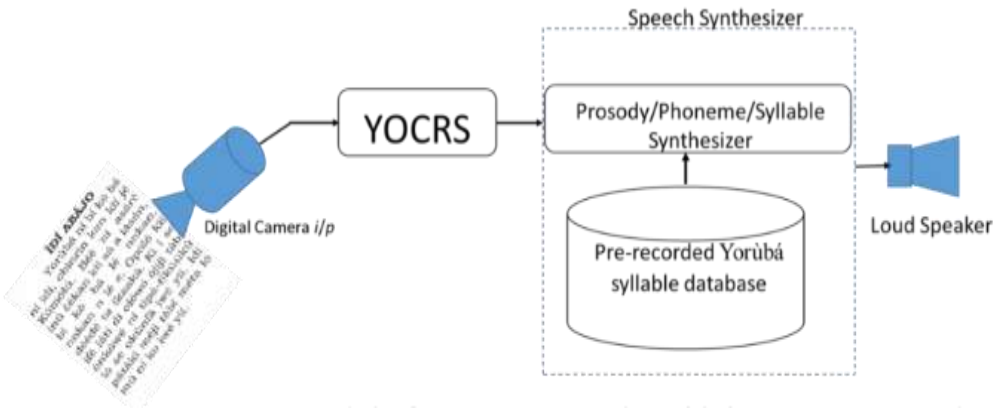

Figure 2: Schematic diagram of camera-based Yoruba document reader

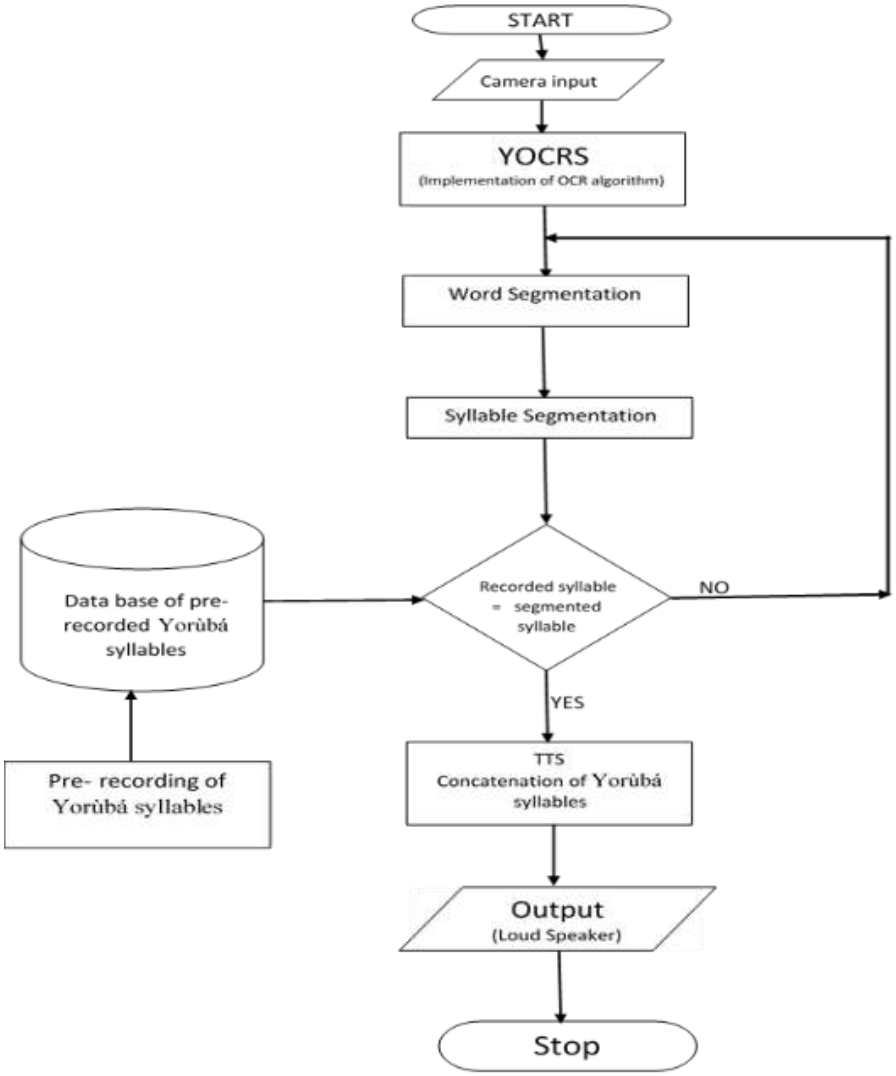

Figure 3: The flow chart of Camera-Based Yoruba Document reader

Figure 3 depicts the flow chart of the camera-based Yorùbá document reader, at the concatenation stage, the syllable signals are divided into frames of 20ms. Frames were made to overlap by 50% as shown in figure 4, the square window of the sides (length) equals the audio syllable frame length for easy estimation and extraction of short term energy (STE).

For each iteration, the overlapping window is shifted by 50%, this is done to limit the spectral leakage thereby eliminating the high-frequency components at end of the speech signal. Spectral

leakages are artifacts that result in a spectral discontinuity at the end of speech signals.

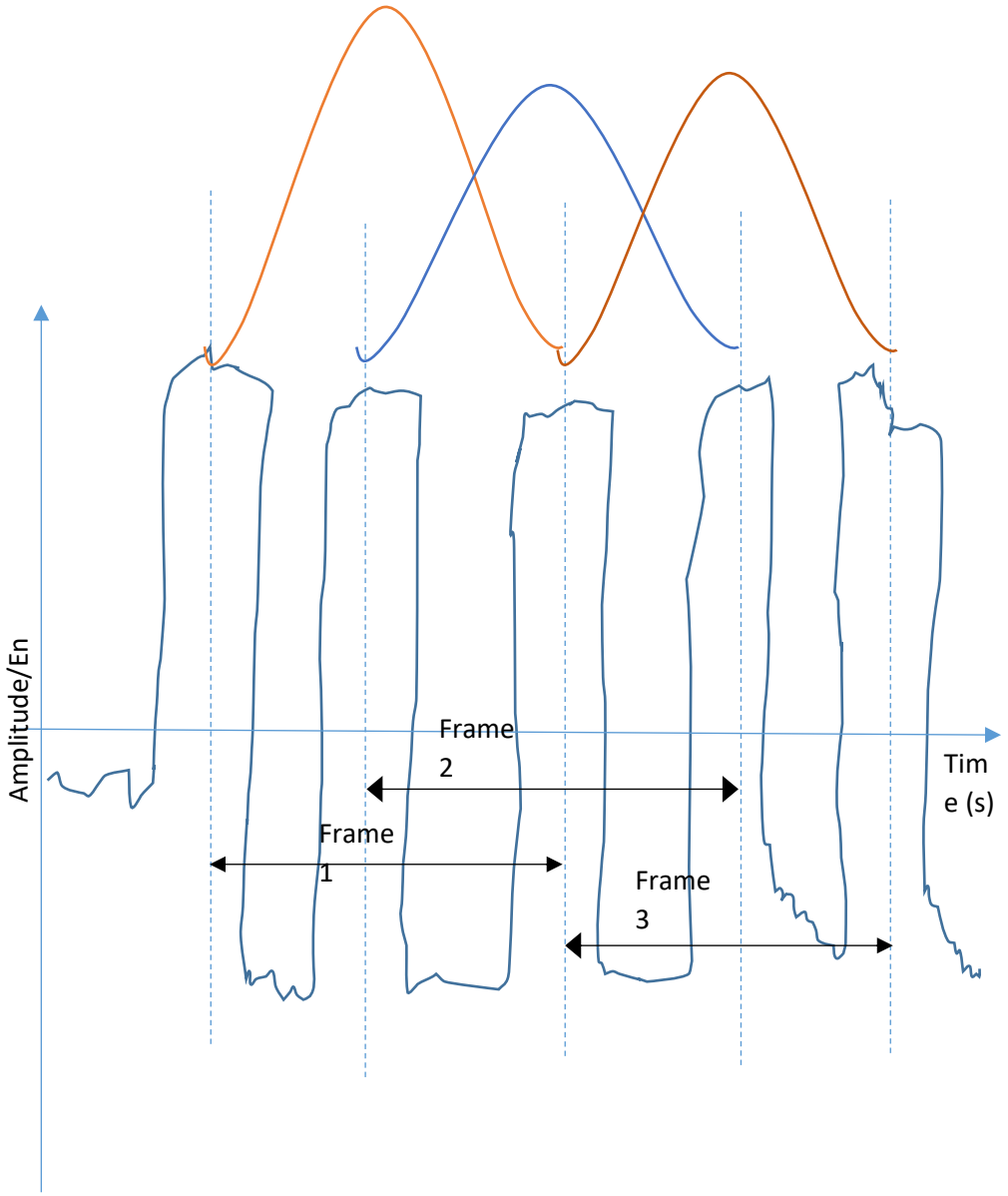

Figure 4: Speech signal division into frames

The sequence speech signal energy is given as in equation (1)

$$S(e) = \sum_{n=-\infty}^{\infty} [x(n)]^2 \quad (1)$$

where S(e) is the signal energy and $x(n)$ is a speech signal of infinite length.

TIJERAT
ISSN: 2707-5654

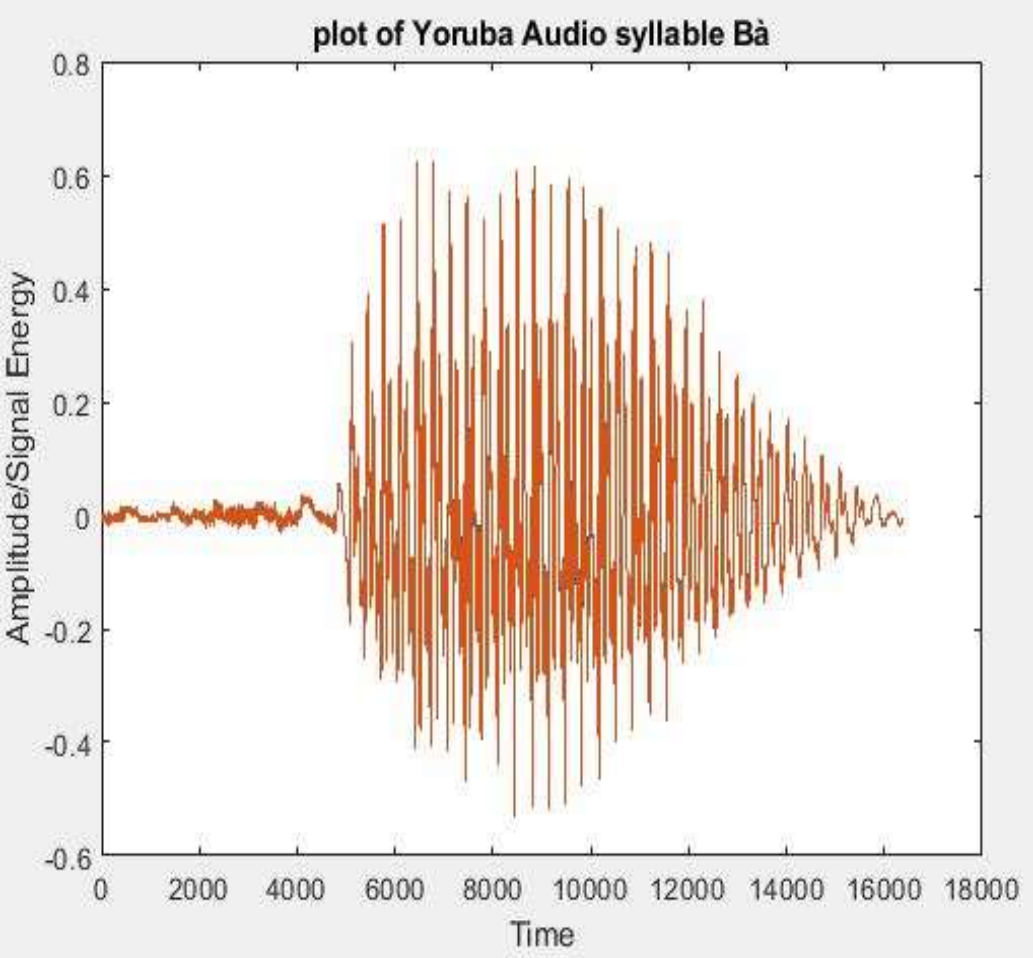

Figure 5 shows the plot of signal infinite length of timeFigure 5 shows the plot of amplitude/signal energy of Yoruba syllable **Bà** speech signal of infinite length from equation (1).
If the speech signal is divided into **N** number of frames, the range of frames will be given as 0: N-1

short term energy of a given frame is given as

$$s_n = \sum_{n=0}^{N-1} [x(n)]^2 \qquad (2)$$

where S_n = *the* short-term energy of a given frame and $x(n)$ is the length of the frame

TIJERAT
ISSN: 2707-5654

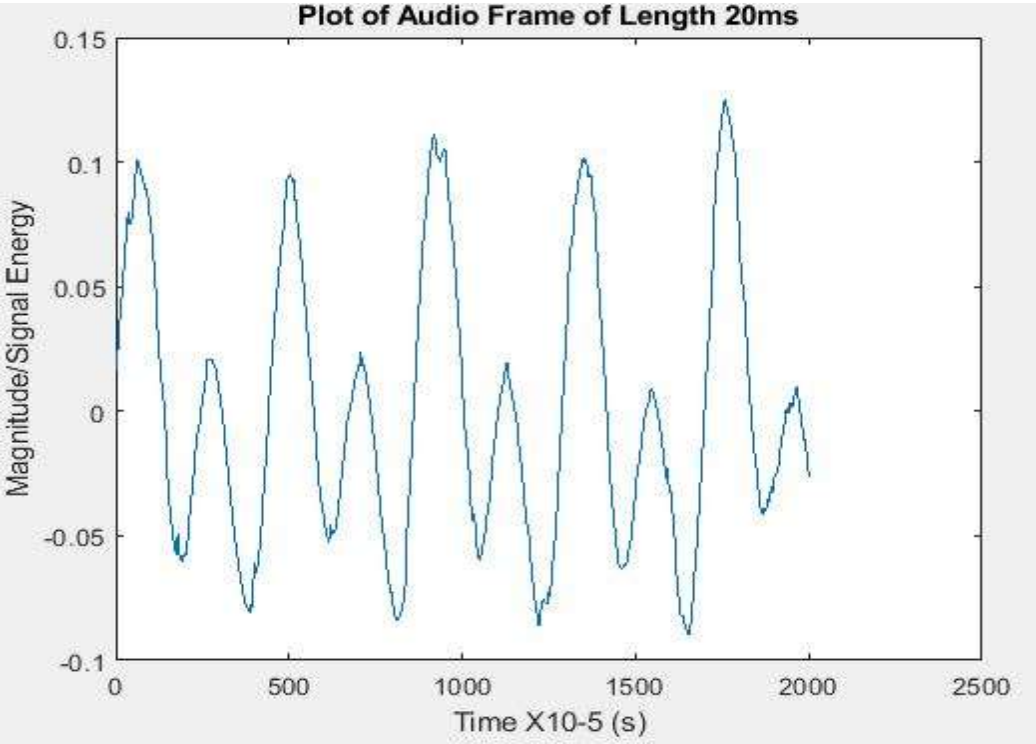

Figure 6 shows the plot of signal of frame length 20ms extracted from syllable **Bà** 00E0

In general, in terms of the window function, the frame sequence of the short term energy is given in equation (3)

$$e(n) = \sum_{n=-\infty}^{\infty} [x(m).\omega\ (n-m)]^2 \qquad (3)$$

Where *e(n)* is the energy sequence of a given frame
x(m) is the speech signal
$\omega(n-m)$ is the window function
and
n is a shift of the window of the frame for each of the iterations, for a square window of dimensions of 1, the value of window shift is given in equation (4).

$$\omega(n) = \begin{cases} 1, 0 \le n \le N\text{-}1 \\ \\ 0\ . \end{cases} \qquad (4)$$

And the window function is given in equation (5)

ω (n-m) = 1 (5)

The framing and windowing procedure is as shown in Figure 4

RESULT AND DISCUSSION

The speech equivalent of syllables obtained from the front-end of the camera-based Yorùbá document reader was concatenated to form words, Figure 7 shows an example of such output. Spectral discontinuities were observed between the syllables and high zero-crossing rate (ZCR), all these result in coarticulation problems and hence reduce the intelligibility and naturalness of concatenated speech signals.

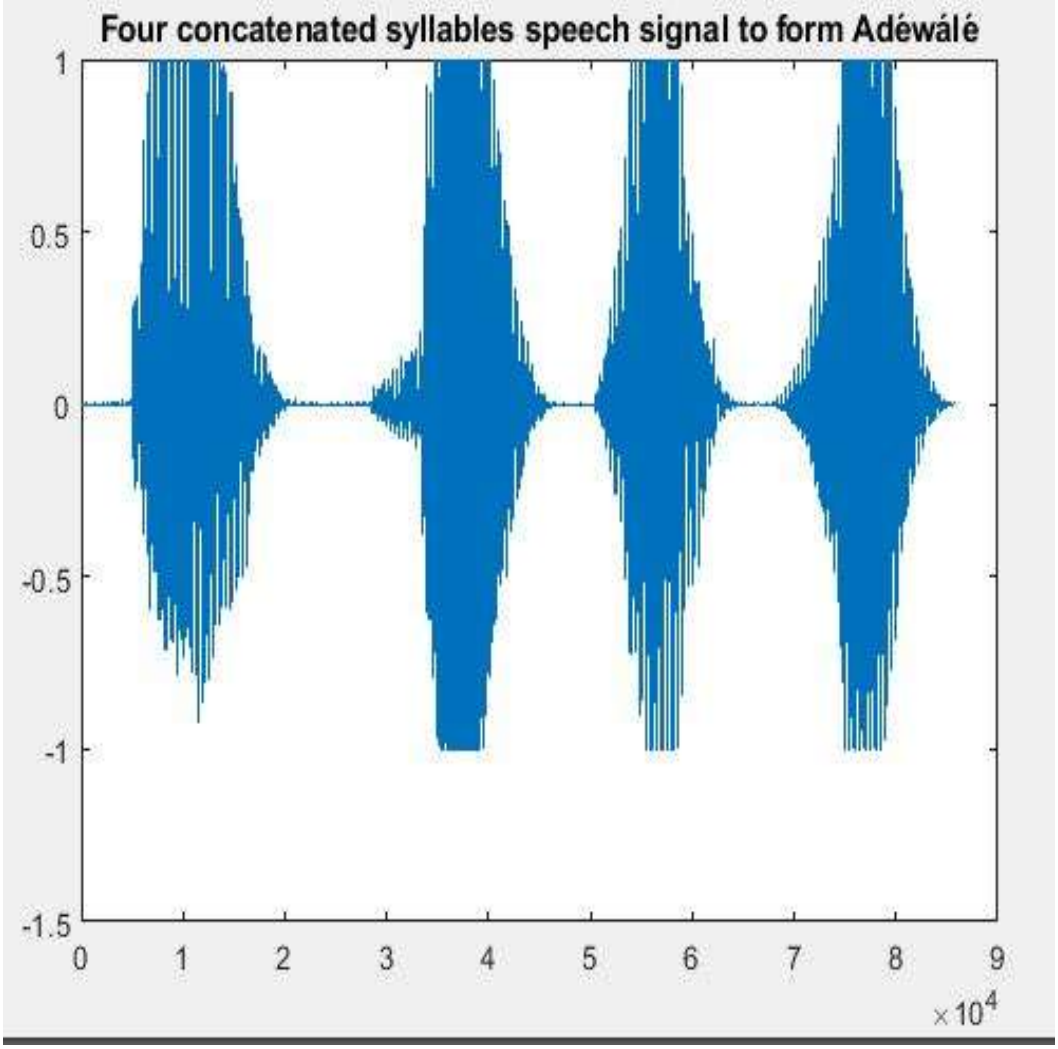

Figure 7: Concatenation of four Yoruba Syllables to form word Adéwálé

The concatenated syllables were divided into frames and the short term energy was estimated as shown in figure 8 and 9

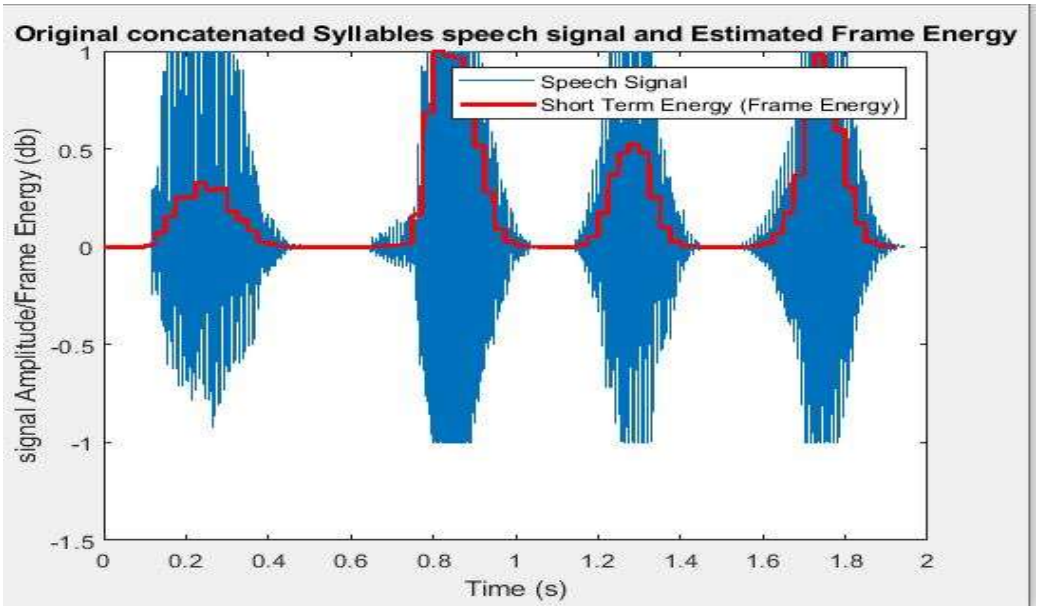

Figure 8: Estimation of energy of concatenated syllable speech signal Short Term Energy (STE)

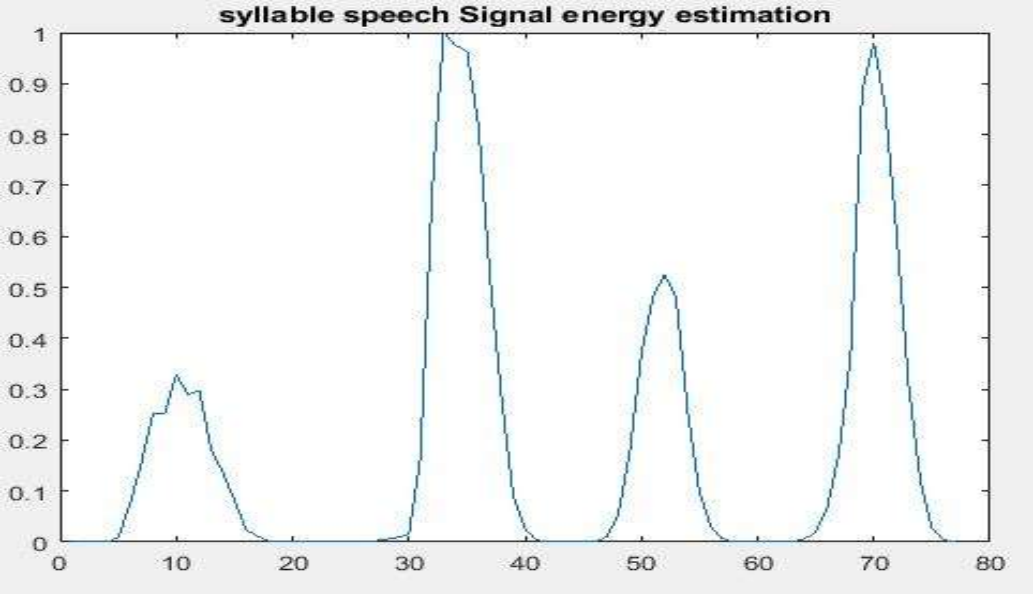

Figure 9: Estimation of energy of concatenated syllable speech Short Term Energy (STE)

TIJERAT
ISSN: 2707-5654

Short energy contours of the frames were extracted as shown in figure 10,

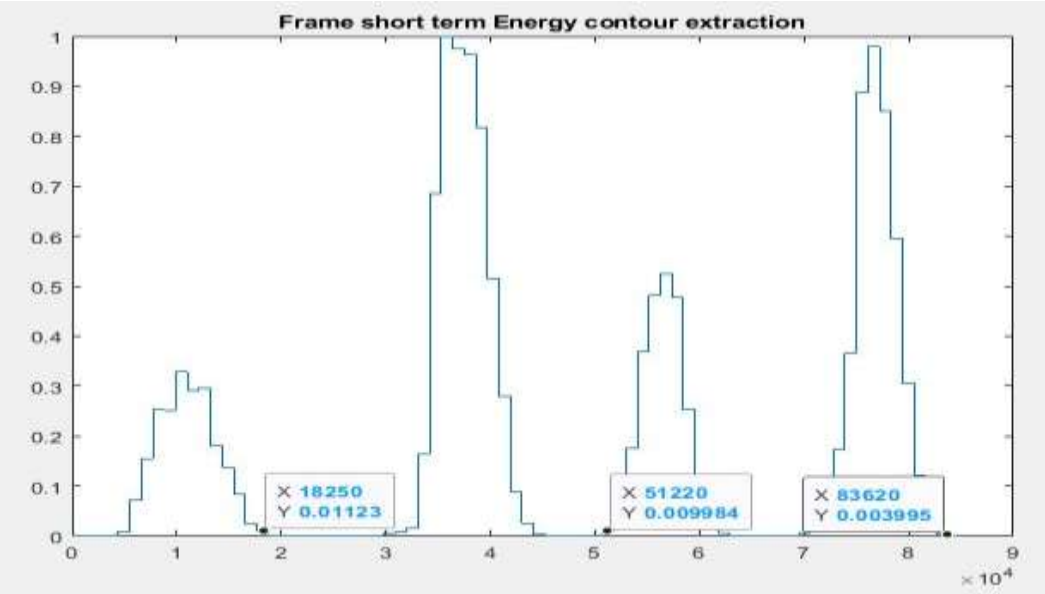

Figure 10: Frame short term energy extraction of concatenated syllable speech signal

`The range of STE 0.004 to 0.01 was considered as the range of the concatenated speech syllable signal with spectral discontinuity and was removed as this range of energy will constitute artifacts that reduce the naturalness and intelligibility of synthesized speeches.

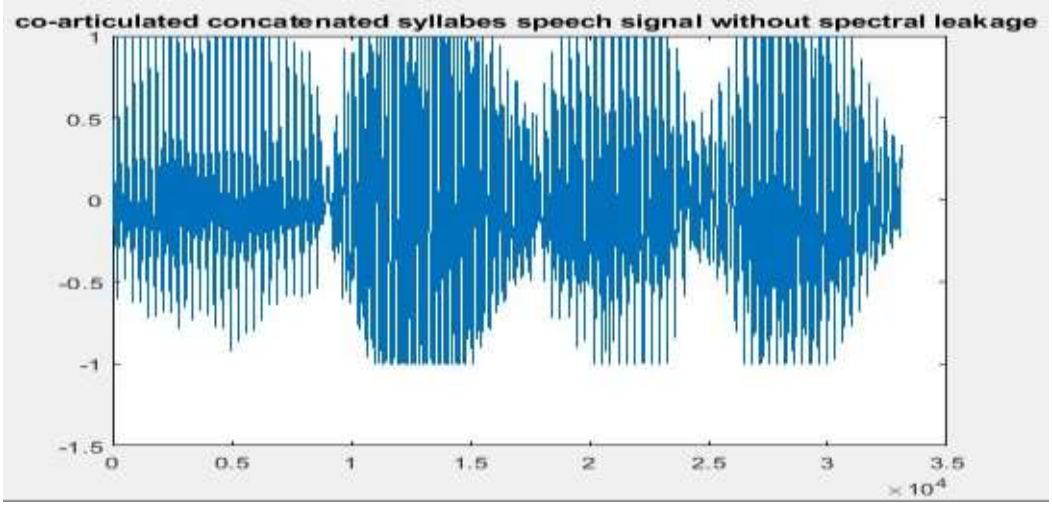

Figure 11: Co-articulated concatenated syllables speech signal without spectral leakage

TEST AND VALIDATION OF RESULTS

Mean Opinion Score, MOS was used to validate the reliability and useability of the document reader. MOS is a five-point scale using simple language to describe the quality of communication or speech signal, it is based on a subjective listening test(Viswanathan and Viswanathan 2005). MOS is an ITU-T P.800 recommendation for subjective determination of communication transmission or speech quality(Viswanathan and Viswanathan 2005).

The five-point rating are

Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Fifty (50) native Yorùbá speakers were selected as responders for listening test to rate the output of the Camera-based Yoruba document reader on the concatenation of Yorùbá syllables to form words, phrases, and sentences. The ratings are based on intelligibility and naturalness. The summary of the MOS test is as shown in Tables 3 and 4.

Table 3: Summary of responders' response on Naturalness

Yorùbá Word/Phrase	Meaning	Opinion score					
		5	4	3	2	1	MOS
Adéwálé	Crown returned home	24	22	3	1	0	4.38
Ìbíyẹmí	Lineage fit me	26	19	4	1	0	4.40

Motigbo	I have heard	23	19	5	3	0	4.24
Gbogbo ọ̀rọ̀ àbáláyé	All antique words	21	20	6	2	1	4.16
Ọnà tí èdè Yorùbá pín sí	Categories of Yorùbá language	22	20	4	1	3	4.14
Èdè Yorùbá gbajú-gbájà ni Nàìjíríà	Yorùbá language is popular in Nigeria	20	18	6	2	4	3.96
Bí èdè Yorùbá ṣe di kíkọ sílẹ̀	How Yorùbá become a written language	19	19	6	4	2	3.98
Overall MOS	4.18						

This result showed that 83.6% of the responder agreed to the naturalness of the synthesized speech from concatenated syllables.

Table 4: Summary of responders' response on the intelligibility

Yorùbá Word/Phrase	**Meaning**	**Opinion score**					
		5	4	3	2	1	MOS
Adéwálé	Crown returned home	21	18	6	3	2	4.06
Ìbíyẹmí	Lineage fit me	22	17	4	4	3	4.02
Motígbọ́	I have heard	19	17	7	5	2	3.92
Gbogbo ọ̀rọ̀ àbáláyé	All antique words	21	18	6	3	2	4.06
Ọnà tí èdè Yorùbá pín sí	Categories of Yorùbá language	18	20	6	3	3	3.94
Èdè Yorùbá gbajú-gbájà ni Nàìjíríà	Yorùbá language is popular in Nigeria	14	13	4	8	11	3.22
Bí èdè Yorùbá ṣe di kíkọ sílẹ̀	How Yorùbá become a written language	17	14	5	4	10	3.48
Overall MOS							3.81

This result also showed that 76.2% of the responder agreed to the intelligibility of the synthesized corpus-based synthesis using syllable as corpus unit.

CONCLUSION

This paper presented a camera-based document reader using Yoruba syllable as corpus unit selection method. The speech signal

is a continuous time-varying signal, the digital signal processing tools are better used on blocks of a speech signal. This is evident in the result obtained in the co-articulation at the concatenative points of the syllables. The extraction of short-term energy has played a great role in removing the spectral discontinuity which is one of the major limitations of the speech signal concatenative synthesis.

REFERENCES

Afolabi, Akin, Elijah Omidiora, and Tayo Arulogun. 2013. "Development of Text to Speech System for Yoruba Language." *Innovative Systems Design and Engineering ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online) Vol.4, No.9, 2013- Special Issue - 2nd International Conference on Engineering and Technology Research* 4(9):1–8.

Ajao, Jumoke F., Stephen O. Olabiyisi, and Elijah O. Omidiora. 2015. "Yoruba Handwriting Word Recognition Quality Evaluation of Preprocessing Attributes Using Information Theory Approach." 9(1):18–23.

Akinwonmi, Akintoba Emmanuel, and Boniface Kayode Alese. 2013. "A Prosodic Text-to-Speech System for Yorùbá Language." Pp. 630–35 in *2013 8th International Conference for Internet Technology and Secured Transactions, ICITST 2013*.

Dastres, Roza, and Mohsen Soori. 2021. "A Review in Advanced Digital Signal Processing Systems." *International Journal of Electrical and Computer Engineering* 15(3):122–27.

Eludiora, Safiriyu I., and Odetunji A. Odejobi. 2016. "Development of an English to YorùbáMachine Translator." *International Journal of Modern Education and Computer Science* 8(11):8–19. doi: 10.5815/ijmecs.2016.11.02.

Ghosh, Krishnendu, and Ramu Reddy. 2011. "Development of Syllable-Based Text to Speech Synthesis System in Bengali." 167–81. doi: 10.1007/s10772-011-9094-4.

Iyanda, Abimbola Rhoda, and Olufemi Deborah Ninan. 2017. "Development of a Yorúbà Text-to-Speech System Using Festival." *Innovative Systems Design and Engineering ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online) Vol.8, No.5, 2017* 8(5):1–9.

Kayte, Sangramsing, Monica Mundada, and Charansing Kayte. 2015. "A Marathi Hidden-Markov Model Based Speech Synthesis System." *IOSR Journal of VLSI and Signal Processing Ver. I* 5(6):34–39. doi: 10.9790/4200-05613439.

Kuligowska, Karolina, Paweł Kisielewicz, and Aleksandra Włodarz. 2019. "Speech Synthesis Systems : Disadvantages and Limitations Speech Synthesis Systems : Disadvantages and Limitations." (March). doi:

10.14419/ijet.v7i2.28.12933.

Odéjobí, Odétúnjí A., Anthony J. Beaumont, and Shun Ha Sylvia Wong. 2006. "Intonation Contour Realisation for Standard Yorùbá Text-to-Speech Synthesis: A Fuzzy Computational Approach." *Computer Speech and Language* 20(4):563–88. doi: 10.1016/j.csl.2005.08.006.

Olmsted, David L., and David L. Olmsted. 2015. "The Phonemes of Yoruba == WORD ==." 7956(1951):245–49. doi: 10.1080/00437956.1951.11659409.

Patra, Tapas Kumar, Patra Biplab, and Mohapatra Puspanjali. 2012. "Text to Speech Conversion with Phonematic Concatenation." *International Journal of Electronics Communication and Computer Technology (IJECCT)* 2(November 2015):223–26.

Pravin, M Ghate and Shirbahadurkar, S. D. 2017. "A SURVEY ON METHODS OF TTS AND VARIOUS TEST FOR EVALUATING THE QUALITY OF SYNTHESIZED SPEECH." 07:15236–39.

Rashad, M. Z., Hazem M. El-bakry, and Islam R. Isma. 2010. "Diphone Speech Synthesis System for Arabic Using MARY TTS." 2(4):18–26.

Sakai, Shinsuke, and Han Shu. 2005. "A Probabilistic Approach to Unit Selection for Corpus-Based Speech Synthesis." 81–84.

Shreekanth, T., V. Udayashankara, and Kumar. .. Arun. 2014. "An Unit Selection Based Hindi Text To Speech Synthesis System Using Syllable as a Basic Unit." *IOSR Journal of VLSI and Signal Processing* 4(4):49–57. doi: 10.9790/4200-04424957.

Singh, Deepika, and Parminder Singh. 2012. "Removal of Spectral Discontinuity in Concatenated Speech Waveform." *International Journal of Computer Applications* 53(16):9–12. doi: 10.5120/8504-2250.

Sprachsignalverarbeitung, Konferenz Elektronische, Eveloping New, Language Tools, T. H. E. Case Of, Uxembourgish Ingmar Steiner, Le Maguer, Judith Manzoni, and Peter Gilles. 2017. "D m Tts: L." 186–92.

Viswanathan, Mahesh, and Madhubalan Viswanathan. 2005. "Measuring Speech Quality for Text-to-Speech Systems: Development and Assessment of a Modified Mean Opinion Score (MOS) Scale." *Computer Speech and Language* 19(1):55–83. doi: 10.1016/j.csl.2003.12.001.