A Kiswahili Dataset for the Development of a Text-to-Speech System

[1]Kelvin Kiptoo Rono, [2]Ciira wa Maina, [3]Elijah Mwangi

[1]School of Engineering, Dedan Kimathi University of Technology, Private Bag 10143, Nyeri, Kenya
[2] Dedan Kimathi University of Technology (DeKUT) Center for Data Science and Artificial Intelligence (DSAIL), Dedan Kimathi University of Technology, Private Bag 10143 Nyeri, Kenya
[3] Faculty of Engineering, University of Nairobi, P.O. Box 30197-00100, Nairobi, Kenya.

***Abstract*: A Text-to-Speech (TTS) system requires adequate data to efficiently build an intelligent and natural system. Based on open-source, free-access data availability, languages are classified into high-resource and low-resource. Kiswahili language, which has a vast population of speakers, is still classified as a low-resource language. It is a low-resource language since there is a limited dataset for building Natural language processing (NLP) systems. This article presents a Kiswahili dataset that contributes to the required language processing resources to build a TTS system or other NLP tasks. A TTS system based on an Artificial Neural Network (ANN) is the current state of the art, and this dataset has been successfully used to build a TTS system. The dataset consists of 7,108 audio clips from a single speaker, with each audio clip varying from 1s to 12.5s. The total audio length is approximately 16 hrs. The dataset created thus meets all the requirements for developing a Kiswahili ANN-based project.**

***Index Terms*: Text-to-Speech (TTS) system, Deep Learning, Natural Language Processing, Kiswahili Dataset, Artificial Neural Networks (ANNs), Language Modelling.**

I. INTRODUCTION

Natural Language Processing helps computers understand and interpret the human language to perform useful tasks. The NLP combines artificial intelligence, linguistics, and cognitive science. The application areas are TTS system and Speech-to-Text (STT) conversion, language modelling, machine translation, content categorization, Name Entity Recognition (NER), and machine translation [1]. These applications have proven to be the current state of the art.

Text-to-Speech and Speech-to-Text systems require a lot of data for training and evaluation. Deep learning algorithms used in training and evaluating a TTS system depend on the data size. The data size used determines the system's success [2,3]. The data size depends on its availability. Thus, based on data availability, languages are classified into high-resource and low-resource. High resource languages have open access data, unlike low resource languages, with limited open access data.

Kiswahili language is a low-resource language. Although the Kiswahili language is spoken and written by over 96 million people [4], there is limited open access data. The article describes the creation of a dataset used in building a TTS system. To efficiently build a TTS system based on ANN, a minimum of 10 hours of audio length is required [3]. This article presents the steps for creating audio in Waveform Audio File Format files (WAVE) and text files for a Kiswahili dataset used in building a TTS system. TTS system built based on ANN produces an audio quality similar to a professional speaker [5].

II. METHODOLOGY

The steps involved in creating the Kiswahili dataset include:

(i). Identifying the sources of data.

(ii). Splitting audio files into 1s-12.5s.

(iii). Segmenting the text and placing the text on a table.

The audio file is assigned a unique ID similar to the text file ID.

Sources of Data

The text corpus comprised 7,108 sentences from Kiswahili audio Bible [6,7], an open-source and non-copyrighted material. The wide variety of data sources helps capture different language properties [8]. The text corpus collected captures the required language parameters.

Segmenting the Texts

The collected words were segmented into an average of 15 words per text file. The number of words per text file was limited because the allowed audio length during TTS system training based on Tacotron 2 model is limited to 12.5s [9]. A Sequence-to-Sequence Neural TTS involves the conversion of input text into a character sequence and feature prediction network that determines a sequence of Mel spectrogram frames from the character sequence [9]. An input text is converted to character embeddings. The system then predicts Mel-spectrograms for each character's embeddings. Finally, a vocoder converts Mel-spectrograms into speech waveforms. Therefore, during the training and evaluation of a TTS system, the audio hyperparameters are the sampling frequency,

number of Mel-frequency bins, and frame shift length. ANN-based model parameters are the size of the encoder & decoder depths and the output per step.

The maximum audio length required during training and evaluation is given by [4]:

$$\text{Audio length} = \text{maximum number of iterations} \times \text{output for each step} \times \text{frame shift} \quad (1)$$

The maximum duration of the audio is shown in Table 3 for the implementation of the ANN-based model [5]. Therefore, varying the parameters determines the maximum length of audio. To ensure that each speech waveform is assigned to the corresponding word, then it is recommended the maximum length of the audio be set to 12.5s [9]. This eliminates a word assigned a noise speech waveform or non-corresponding speech waveform

Table 1. ANN-based Model parameters for the implementation of ANN-based TTS system

Maximum number of iterations	200
Output for each training step	5
Frame shift(milliseconds)	12.5

$$Maximum\ audio\ length = 200 \times 5 \times 12.5 = 12{,}500\ ms \quad (2)$$

Therefore, the maximum audio clip length in the Kiswahili dataset was 12.5s, and the minimum audio clip length was 1s.

Normalizing the Kiswahili NSWs

Non-Standard Words (NSWs) cannot be pronounced without expanding the word into the standard form. NSW's involves expanding all the abbreviations, numbers, letter-sequence, and other NSWs [10]. The text file was divided into unique IDs, transcribed words, and Normalized words. NSWs in Kiswahili were classified into Alphabetical tokens, numerical, and combined tokens (for mixed tokens requiring further expansion), as shown in Table 2.

Table 2. Classification of NSWs consisting of Numbers, abbreviations, letter-sequence, and combination of numbers & letters

Class	Description	Examples
Alphabetical Tokens	Abbreviation	$ (dola), n.k.(na kadhalika)
	letter sequence	*KBC(K B C)*
Numbers	cardinal number	28(ishirini na nane)
	Ordinal number	10.5(kumi nukta tano)
		1/5(thuluthi tano)
	Number range	10-30(kumi hadi thelathini)
	Telephone number	911 (Tisa tisa moja)
	Time	2:45(saa nane arubaini na tano)
	Date	2nd july(julai mbili)
	Year	2021(elfu mbili ishirini na moja)
	Money	$50(dola hamsini)
	Percentage	*10 %(asilimia kumi)*
Combined tokens	Mixed	*K24(K ishirini na nne)*

Creating the Audio File Data

The audio files downloaded were more than 12.5s in length. The audio files were split into short audio clips of lengths between 1s to 12.5s. The audio files were saved as a single channel 16 Pulse Code Modulated WAVE file with a sampling rate of 22.05 kHz. These was the required properties for an audio file in developing a TTS system [11]. Each audio file was assigned a unique ID matching the corresponding text file. A total of 7,108 audio files were created.

Creating the Text File Data

The text files were created and saved in a CSV format. Each text file has one record. The text file contains three parts which are separated by a pipe character. The parts are divided into a unique ID for each line, transcribed words, and normalized

texts in their expanded forms, as shown in Table 3. A unique ID is a number assigned to each text file. The transcribed words are the text spoken by a reader. Normalized texts are the expansion of abbreviations, numbers, and the letter sequence into full words.

Table 3. A sample of Text Files comprising a unique ID, Transcribed words, and Normalized words.

Unique ID	Transcribed Words	Normalized Words
KISWA-00464	Tabia zako zimemfanya rafiki yako kuanza kufikiria kuwa wewe si mwaminifu.	Tabia zako zimemfanya rafiki yako kuanza kufikiria kuwa wewe si mwaminifu.
KISWA-00465	Jaribu juu chini umuondolee shaka hiyo kwa kuwa naye mara nyingi upatapo nafasi.	Jaribu juu chini umuondolee shaka hiyo kwa kuwa naye mara nyingi upatapo nafasi.
KISWA-00467	Mei 21 – Juni 21	Mei ishirini na moja hadi Juni ishirini na moja
KISWA-00132	Bw Kimani anaongea maneno mazuri.	Bwana Kimani anaongea maneno mazuri.

Each text file has a unique ID matching each audio file ID. The unique ID was assigned from KISWA-0001 to KISWA-1570.

III. RESULTS

The dataset comprises 7,108 text files and audio clips from a single speaker. Each audio clip length varies between 1s to 12.5s. The total audio length is approximately 16 hours. A total of 7,108 text files were created and saved in a CSV format. Each text file has one record. The text file contains three parts which are separated by a pipe character. The parts are divided into a unique ID for each line, transcribed words, and normalized texts in their expanded forms, as shown in Table 4

Table 4. A text file consists of a Unique ID, Transcribed text, and normalized text

Unique ID	Transcribed Text	Normalized Text
KISWA-00473	Julai 23-Agosti 22	Julai ishirini na tatu hadi Agosti ishirini na mbili

The CSV file containing the texts can be accessed in the link provided below:
https://data.mendeley.com/datasets/vbvj6j6pm9/1
Each text file has a unique ID matching each audio file ID. The audio file samples can be downloaded in the link provided below:
https://data.mendeley.com/datasets/vbvj6j6pm9/1

The dataset contains approximately 106,000 Kiswahili words consisting of 7,108 WAVE files. The mean number of words per audio file was 14.96. The minimum length of each audio clip is 1s, while the maximum is 12.5s. Each audio file's properties include a sampling rate of 22.05 kHz and 16-bit single-channel unsigned Pulse Code Modulation (PCM).

IV. Conclusion

The data were collected from various bible books to capture different language parameters. Also, the data consisted of Non-Standard Words (NSWs), which included abbreviations, numbers, and symbols. The NSWs were expanded in the text file.

The dataset was used to build a Kiswahili TTS system based on Tacotron 2 model, which was successful. The dataset can also build an STT system based on ANN. The dataset also contributes to open-source, available data for language processing tasks, especially for Kiswahili, a low-resource language. Finally, the dataset will assist researchers in using the data to develop language processing models in the NLP field.

V. ACKNOWLEDGEMENTS

I am thankful to Dedan Kimathi University of Technology for awarding me a scholarship to study M.sc Telecommunication Engineering. I am also grateful to my research supervisors, Dr. Ciira Maina, Director of the Center for Data Science and Artificial Intelligence, and Prof. Elijah Mwangi, Faculty of Engineering, University of Nairobi, who have been insightful throughout this research project. Their counsel, guidance, and expert opinions I received are invaluable.

REFERENCES

[1] J. Brownlee, “7 Applications of Deep Learning for Natural Language Processing,” Machine Learning Mastery, 07 Aug 2019. [Online]. Available: https://machinelearningmastery.com/applications-of-deep-learning-for-natural-language-processing/. [Accessed: 13 August 2021].

[2] J.Latorre, J.Lachowicz, J.Lorenzo-Trueba, T.Merritt, T.Drugman, S.Ronanki, and K.Viacheslav, ”Effect of Data Reduction on Sequence-to-Sequence Neural TTS,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics,

Speech, and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 7075-7079, doi: 10.1109/ICASSP.2019.8682168.

[3] K.Sodimana, P.D.Silva, S.Sarin, K.Pipatsrisawat, O.Kjartansson, M.Jansche, and L.Ha, " A Step-by-Step Process for Building TTS Voices Using Open Source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced s (SLTU), Gurugram, India, 2018, pp.66-70, doI: 10.21437/SLTU.2018-14

[4] D. Doochin, "How Many People Speak Swahili, and Where Is It Spoken?" *Babbel Magazine*. [Online]. Available: https://www.babbel.com/en/magazine/how-many-people-speak-swahili#:~:text=There%20are%20about%2016%20million,of%20eastern%20and%20southeastern%20Africa. [Accessed: 13 November 2021].

[5] K. Simonyan, S. Dieleman, A.V.D.Oord, S.Dieleman, H.Zen,O.Vinyals, N.Kalchbrenner, A. Senior, and A. Graves, "Wavenet Generative Model for Raw Audio," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, California, United States, pp. 1–15, 2016.

[6] WordProject,Swahili Audio Bible, n.d[online].Available https://www.wordproject.org/bibles/audio/05_swahili/index.htm. [Accessed: 29-Nov-2021]

[7] WordProject,Swahili Audio Bible, n.d[online].Available https://www.wordproject.org/contact/new/copyrights.htm. [Accessed: 29-Nov-2021]

[8] M.Gakuru, F.K.Iraki, R.Tucker, K.Shalonova, and K.Ngugi," Development of a Kiswahili Text-to-Speech System.", Interspeech 2005, Lisbon, Portugal, September 2005.

[9] GitHub, An implementation of Tacotron speech synthesis in TensorFlow, 2017[online].Available: https://github.com/keithito/tacotron. [Accessed: 20 November 2021]

[10] E. Flint, E. Ford, O. Thomas, A. Caines, and P. Buttery, "A Text Normalisation System for Non-Standard English Words," Proceedings of the 3rd Workshop on Noisy User-generated Text, Copenhagen, Denmark, pp. 107–115, 2017.

[11] J.Shen, R.Pang, R.J.Weiss, M.Schuster, N.Jaitly, Z.Yang, Z.Chen, Y.Zhang, Y.Wang, R.Skerry-Ryan, R.A.Saurous, Y.Agiomyrgiannakis, and Y.Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368