KINYARWANDA TTS: USING A MULTI-SPEAKER DATASET TO BUILD A KINYARWANDA TTS MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

The field of text-to-speech (TTS) technology has been rapidly advancing in recent years, and has become an increasingly important aspect of our lives. This presents an opportunity for Africa, especially in facilitating access to information to many vulnerable demographics, e.g. Illiterates. However, a lack of available quality datasets is a major hindrance. In this work, we create a dataset based on the bible, using an existing Kinyarwanda Speech to Text model we are able to segment and align the voice and the text. then created a multi-speaker Kinyarwanda model.

1 INTRODUCTION

Generally, Africa was traditionally an oral culture, with its history, culture, and values being transmitted orally from one generation to another.While this was a rich and valuable tradition, it was also prone to loss and misinterpretation over time. The introduction of writing allowed African societies to preserve their history, culture, and knowledge in a more reliable and permanent form. The era of writing brought mostly through education was a milestone, however, the literacy rate in Africa is still around 67%, with some areas having less than 30%. This hinders the socio-development of the area, most affected by the problem are the marginalized portion of society, such as people leaving with disabilities, women, people leaving in rural areas, etc. Technology Inclusion, especially with the rise of Artificial intelligence (AI), can be a driving force in mitigating some of the problems and a major driver to benefit those left behind.

Text to speech, the ability of the computer to read out loud written text, is one of those AI tools that can help, with the advent of end-to-end TTS models has made it easier to build speech synthesis, without the need for much more complicated processes such as feature engineering. However, an End-to-end model still requires a lot of data, which low-resource languages do not possess. A good TTS requires studio-quality recording devoid of outside noise and must be of short length (in seconds), which is hard to come by for African languages.

In our work, we build a TTS model leveraging the existing Kinyarwanda Audio bible, where using existing Kinyarwanda[1] Speech to Text (STT) and nemo[2] CTC-Segmentation we align the audio and text datasets. we get small clips of duration ranging from 3 seconds to 47 seconds, we are able to generate 67.84 hours of studio quality dataset, which we use to train the TTS model using the coqui[3] TTS framework.

2 RELATED WORK

2.1 ADVANCES IN TEXT TO SPEECH

With the development of deep learning and artificial intelligence, neural network-based TTS has significantly improved the quality of synthesized speech in recent years. In the near 2010s, as neural network and deep learning have achieved rapid progress, some early neural models such as deep neural network (DNN) based Qian et al. (2014) and recurrent neural network (RNN) basedZen

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_rw_conformer_transducer_large

[2] https://github.com/NVIDIA/NeMo

[3] https://github.com/coqui-ai/TTS

(2015),Fan et al. (2014) are adopted in Statistical Parametric Speech Synthesis (SPSS), an earlier TTS method, to replace HMM for acoustic modeling.

Neural TTS usually consists of three components: a text analysis module, a parameter prediction module (acoustic model), and a vocoder analysis/synthesis module (vocoder). The text analysis module first processes the text, including text normalization Sproat et al. (2001), grapheme-to-phoneme conversion Bisani & Ney (2008), word segmentation, etc, and then extracts the linguistic features, such as phonemes, duration, and POS tags from different granularities. The acoustic models are trained with these linguistic features to generate acoustic features including fundamental frequency, spectrum, or cepstrum Fukada et al. (1992), etc. The vocoders synthesize speech from the predicted acoustic features.

Later, WaveNet Oord et al. (2016) is proposed to directly generate waveforms from linguistic features, which can be regarded as the first modern neural TTS model. Other models like DeepVoice Arik et al. (2017) still follow the three components in statistical parametric synthesis but upgrade them with the corresponding neural network-based models. Furthermore, some spectrogram models (e.g: Tacotron 1/2 Wang et al. (2017), Deep Voice 3 Ping et al. (2018), and FastSpeech 1/2 Ren et al. (2019)Ren et al. (2022)) are proposed to simplify text analysis modules and directly take character/phoneme sequences as input and simplify acoustic features with Mel-spectrograms. Fully end-to-end TTS systems were developed to directly generate waveforms from text, such as ClariNet Ping et al. (2019), FastSpeech 2 Ren et al. (2022), and EATS Donahue et al. (2021). TTS has continued to evolve with multilingual and multi-speaker models Casanova et al. (2022) enabling the training of multiple speakers, and multiple languages, giving advantages to low-resourced languages. The other end of the spectrum is training on a huge dataset Wang et al. (2023) you are able to generate voice using a clone of a 3-second recording.

Compared to previous TTS systems based on concatenative synthesis and statistical parametric synthesis, the advantages of neural network-based speech synthesis include higher quality in terms of both intelligibility and naturalness and less requirement on human preprocessing and feature development.

2.2 Text To speech for African languages

There are several works that has focused on building Text to Speech datasets and models for African languages. Bible TTS Meyer et al. (2022), leverages open licensed Bible, Josh, and Al. created Text to Speech datasets for 10 African languages, they used Montreal force Alignment to Align the text with the audio. Ogayo et al. (2022) uses the CMU Flite to build TTS models for 12 African languages, CMU Flite is based on a random forest and does not require any GPU for computation, and a small dataset. The drawback is the quality of the TTS where the naturalness metric is low. Gamayun Öktem et al. (2020) is a data collection platform for machine translation and voice data collection used to collect voice across Africa, easing the creation of new datasets. Gutkin et al. (2020) created an open clean Yoruba dataset. Gakuru et al. (2005) created a parametric Swahili TTS based on the festival speech synthesizer.

2.3 Kinyarwanda

Kinyarwanda is part of the BantuNurse & Philippson (2003) language family spoken in Central-eastern Africa, part of the JD subgroup of Bantu languages alongside Kirundi, Fuliro, Ha, Havu, etc. It is spoken by more than 13 million speakers in Rwanda, and it is one of the official languages of Rwanda, used in official, administrative, education, and as the lingua franca of Rwanda. Kinyarwanda is a tonal languageMuhirwe (2010), although it is written without the tones, as they are implicitly added by the speaker. In Kinyarwanda, 2 words may be written the same but pronounced differently, e.g. family [umuryaango] (family) and door [umuryângo]; the reader must extract the meaning depending on the context. Note, this paper does not cover the scope of the tones.

3 Dataset

To create a viable dataset; we used two separate data sources, we obtained the Kinyarwanda bible audio recordings from Faith Comes By Hearing[4] website and the text from bible.com (BIR: Bibiliya Ijambo ry'Imana) [5] through web-scraping. The audio covers only the 39 books of the Old Testament; consequently, we only used the Old Testament books from the scraped text. We converted numerals to their corresponding Kinyarwanda words using hand-crafted rules.

Each scraped Bible page is stored in its file. A newline in the page's file separates the elements of a page (verses and headers). Since each audio covers a page in the bible, all audio files were longer than thirty seconds, thirty seconds being around the expected audio clip duration needed to train our model. We used the CTC-Segmentation technique Kürzinger et al. (2020) to generate and align audio corresponding to each line on every page of the scraped bible text with their matching audio segment in the audio representing that page. These two data sources mostly overlap, apart from minor differences. After the segmentation, the CTC-segmentation algorithm generates a confidence score. We used the confidence score to measure the overlap and alignment between the audio's content and their corresponding transcript. After finding all audio and transcripts pairs with low alignment confidence scores; we choose which ones to manually correct (add, or remove text in the transcript to match the audio) and which to dismiss depending on the amount of mismatch. Additionally, we ignored all audio files of less than three seconds; since most were erroneous. The result is sixty-seven hours, multi-speaker dataset.

Table 1: Dataset specification

Specification	**value**
Total number of speakers	11
Duration of a single clip (seconds)	3 - 47
Total number of clips	39951
Total number of hours	67.84

4 Model

YourTTSCasanova et al. (2022) is a multilingual, multi-speaker TTS model with zero-shot capabilities, it is based on the VITSKim et al. (2021) model. The VITS model is a single-stage end-to-end text-to-speech model. YourTTS builds upon the work on VITS and adds multi-speakers and multilingual capabilities. In this work, we trained a YouTTS model using sixty-seven hours of bible data, using a single A-100 GPU, we used raw text instead of phonemes since we did not have a grapheme-to-phoneme converter. After a hundred epochs and sixteen hours of training, we reached a reasonable quality voice. The model was deployed on hugginface[6]

5 Result and Analysis

The model was provided to 10 native Kinyarwanda testers, it was found that the model performed better with words related to the Bible than out-of-domain words, such as current words that were not present during the writing of the bible in 1979. Kinyarwanda is a tonal language, thus testers would easily pick words that are not well pronounced, associating the voice with the northern accent of Rwanda. The model has the advantage of being lightweight when running inference, thus can be deployed on CPU servers which is affordable in the African context where GPU-based servers are quite expensive and nonexistent in local clouds.

[4] https://www.faithcomesbyhearing.com/audio-bible-resources/recordings-database

[5] https://www.bible.com/bible/395

[6] https://huggingface.co/DigitalUmuganda/Kinyarwanda_YourTTS

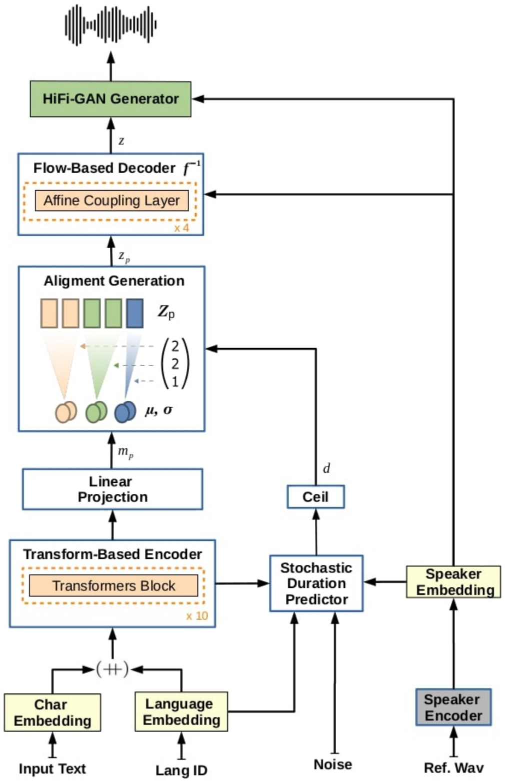

Figure 1: YourTTS training Architecture (Casanova et al., 2022)

Table 2: Training Hyperparameters

Hyperparameter	**Value**
Epochs	100
Sampling rate	22050
Learning rate	0.001
Weight decay	0.01
Betas	[0.8,0.99]
Batch size	12
Num. of CPUs	72
Num. of GPUs	1
Num. of MELs	80
Use phonemes	False

6 CONCLUSION

In this paper, we created and used a Bible Kinyarwanda dataset to build a TTS model, leveraging an existing Kinyarwanda STT model; we used the CTC-Segmentation technique to align the voice and the text from different sources. We generated 67.84 hours dataset and trained it on the yourTTS model, which we tested on in-domain and out-of-domain text. The model is light and performs well, even though the tones are not always correct.

Future work will involve training a multilingual model using other Bantu languages such as Luganda and Swahili, and investigating ways to improve the tonal quality, using techniques such as creating a tonal dictionary, and adding a studio-quality dataset that uses modern and frequently used sentences.

References

Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep Voice 2: Multi-Speaker Neural Text-to-Speech, September 2017. URL http://arxiv.org/abs/1705.08947. arXiv:1705.08947 [cs].

Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, May 2008. ISSN 0167-6393. URL https://www.sciencedirect.com/science/article/pii/S0167639308000046.

Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone, February 2022. URL http://arxiv.org/abs/2112.02418. arXiv:2112.02418 [cs, eess].

Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. End-to-End Adversarial Text-to-Speech, March 2021. URL http://arxiv.org/abs/2006.03575. arXiv:2006.03575 [cs, eess].

Yuchen Fan, Yao Qian, Fenglong Xie, and F. Soong. TTS synthesis with bidirectional LSTM based recurrent neural networks. 2014. URL https://www.semanticscholar.org/paper/TTS-synthesis-with-bidirectional-LSTM-based-neural-Fan-Qian/c217905bc98f00af747e8e9d5f6b79fb89a90886.

T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai. An adaptive algorithm for mel-cepstral analysis of speech. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 137–140 vol.1, March 1992. doi: 10.1109/ICASSP.1992.225953. ISSN: 1520-6149.

Mucemi Gakuru, Frederick Iraki, Roger Tucker, Ksenia Shalonova, and Kamanda Ngugi. Development of a kiswahili text to speech system. pp. 1481–1484, 09 2005. doi: 10.21437/Interspeech.2005-522.

Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara E. Rivera, and Kólá Túbòsún. Developing an open-source corpus of yoruba speech. In *Proc. of Interspeech 2020*, pp. 404–408, October 25–29, Shanghai, China, 2020., 2020. URL http://dx.doi.org/10.21437/Interspeech.2020-1096.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, June 2021. URL http://arxiv.org/abs/2106.06103. arXiv:2106.06103 [cs, eess].

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. CTC-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer*, pp. 267–278. Springer International Publishing, 2020. doi: 10.1007/978-3-030-60276-5_27. URL https://doi.org/10.1007%2F978-3-030-60276-5_27.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Samuel Olanrewaju, Jesujoba Alabi, and Shamsuddeen Muhammad. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus, July 2022. URL http://arxiv.org/abs/2207.03546. arXiv:2207.03546 [cs, eess].

Jackson Muhirwe. Morphological Analysis of Tone Marked Kinyarwanda Text. In Anssi Yli-Jyrä, András Kornai, Jacques Sakarovitch, and Bruce Watson (eds.), *Finite-State Methods and Natural Language Processing*, Lecture Notes in Computer Science, pp. 48–55, Berlin, Heidelberg, 2010. Springer. ISBN 978-3-642-14684-8. doi: 10.1007/978-3-642-14684-8_6.

Derek Nurse and Gérard Philippson. *The Bantu Languages*. Routledge, London, July 2003. ISBN 978-0-203-98792-6. doi: 10.4324/9780203987926.

Perez Ogayo, Graham Neubig, and Alan W. Black. Building African Voices, July 2022. URL http://arxiv.org/abs/2207.00688. arXiv:2207.00688 [cs, eess].

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, September 2016. URL http://arxiv.org/abs/1609.03499. arXiv:1609.03499 [cs].

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, February 2018. URL http://arxiv.org/abs/1710.07654. arXiv:1710.07654 [cs, eess].

Wei Ping, Kainan Peng, and Jitong Chen. ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech, February 2019. URL http://arxiv.org/abs/1807.07281. arXiv:1807.07281 [cs, eess].

Yao Qian, Yuchen Fan, Wenping Hu, and Frank K. Soong. On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3829–3833, May 2014. doi: 10.1109/ICASSP.2014.6854318. ISSN: 2379-190X.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, Robust and Controllable Text to Speech, November 2019. URL http://arxiv.org/abs/1905.09263. arXiv:1905.09263 [cs, eess].

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, August 2022. URL http://arxiv.org/abs/2006.04558. arXiv:2006.04558 [cs, eess].

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christina D. Richards. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333, July 2001. ISSN 0885-2308. doi: 10.1006/csla.2001.0169. URL https://www.sciencedirect.com/science/article/pii/S088523080190169X.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers, January 2023. URL http://arxiv.org/abs/2301.02111. arXiv:2301.02111 [cs, eess].

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis, April 2017. URL http://arxiv.org/abs/1703.10135. arXiv:1703.10135 [cs].

Heiga Zen. Acoustic Modeling in Statistical Parametric Speech Synthesis - From HMM to LSTM-RNN. In *Proc. MLSLP*, 2015.

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, and Grace Tang. Gamayun - language technology for humanitarian response. In *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, pp. 1–4, 2020. doi: 10.1109/GHTC46280.2020.9342939.