Building a Luganda Text-to-Speech Model From Crowdsourced Data

Sulaiman Kagumire & Andrew Katumba
Department of Electrical and Computer Engineering
Makerere University
Kampala, Uganda
sulaiman.kagumire@gmail.com, andrew.katumba@mak.ac.ug

Joyce Nakatumba-Nabende
Department of Computer Science
Makerere University
Kampala, Uganda
joyce.nabende@mak.ac.ug

John Quinn
Sunbird AI
Kampala, Uganda
jq@sunbird.ai

Abstract

Text-to-Speech (TTS) development for African languages such as Luganda is still limited, primarily due to the scarcity of high-quality, single-speaker recordings essential for training TTS models. Prior work has focused on utilizing the Crowdsourced Common Voice Luganda recordings of multiple female speakers aged between 20-49. Although the generated speech is intelligible, it is still of lower quality compared to their model trained on studio-grade recordings. This is due to the insufficient data preprocessing methods applied to improve the quality of the Common Voice recordings. Furthermore, speech convergence is more difficult to achieve due to varying intonations from multiple speakers, as well as background noise in the training samples. In this paper, we show that the quality of Luganda TTS from Common Voice can improve by training on multiple speakers of close intonation in addition to further preprocessing of the training data. Specifically, we selected six female speakers with close intonation determined by subjectively listening and comparing their voice recordings. In addition to trimming out silent portions from the beginning and end of the recordings, we applied a pre-trained speech enhancement model to reduce background noise and enhance audio quality. We also utilized a pretrained, non-intrusive, self-supervised Mean Opinion Score (MOS) estimation model to filter recordings with an estimated MOS over 3.5, indicating high perceived quality. Subjective MOS evaluations from nine native Luganda speakers demonstrate that our TTS model achieves a significantly better MOS of 3.55 compared to the reported 2.5 MOS of the existing model. Moreover, for a fair comparison, our model trained on six speakers outperforms models trained on a single-speaker (3.13 MOS) or two speakers (3.22 MOS). This showcases the effectiveness of compensating for the lack of data from one speaker with data from multiple speakers of close intonation to improve TTS quality.
Text-to-Speech, Crowdsourced, Luganda, Speech, Text, Multiple Speakers

1 Introduction

Speech synthesis technology has significantly advanced in recent years with deep neural network models, allowing for successful applications such as speech-based virtual assistants. While these

advancements are evident in popular languages such as English, French, and Spanish, many of the low-resourced African languages are still lagging far behind. Notably, major Text-to-Speech (TTS) platforms like Amazon's Polly, NaturalReader, Voice Dream Reader, Speechify, and Google's Speech Recognition and Synthesis, have yet to include support for various African languages. The disparity originates from the fact that TTS models deployed in such platforms require relatively large numbers of high-quality recordings with text transcriptions from a single professional speaker to generate natural-sounding audio Ogayo et al. (2022). For many low-resourced African languages such as Luganda, not only is this type of data difficult to collect but also the existing ones are often proprietary, which limits its accessibility for TTS development. Luganda is a Bantu language spoken by more than 20 million people globally and by over 16.7% of Uganda's population Babirye et al. (2022). Despite having a relatively large number of speakers, the language is arguably overlooked in the development of current TTS systems due to the the lack of high-quality single-speaker TTS corpora.

To reduce the dependency on single speaker datasets in low-resource settings, previous studies Latorre et al. (2019); Luong et al. (2019) have shown that training TTS models on a mixture of available multiple speakers can generate synthetic speech with better quality and stability than a speaker-dependent one. A recent work Owomugisha et al. (2023) has developed a Luganda TTS model[1] from crowdsourced Common Voice Luganda recordings of multiple female speakers aged 20-49. Despite the acceptable intelligibility of their model, it is of lower quality compared to their model trained on single-speaker studio recordings[2]. Additionally, speech convergence was more difficult to attain with Common Voice due to the varying intonations from multiple speakers, as well as the existence of background noise in the training data because Common Voice is primarily collected for Automatic Speech Recognition (ASR), rather than speech generation Ardila et al. (2019). There is therefore a need to enhance TTS models trained on the crowdsourced Common Voice Luganda dataset.

In this paper, we present an improved Luganda TTS model trained on Common Voice recording of six female speakers with close intonation that is determined through listening and comparing the voices. This approach upgrades speech convergence during synthesis, offering a noticeable improvement over the existing model where individual training voices are quite diverse. On top of trimming silent parts from the beginning and end of each recording, a technique adopted by the existing model, we describe additional data preprocessing steps to enhance the quality of the training speech and text data. These measures not only refine the quality of the training speech data but also improve the generated speech quality. By fine-tuning an English pretrained and end-to-end TTS model on the crowdsourced Common Voice recordings, our model achieves a level of speech naturalness that surpasses that of the existing model. Furthermore, we also compare the performance of the model trained on six female speakers to models trained on a single speaker or two speakers. Evaluation results show that training on a more extensive and diverse dataset sounds more natural than those trained on single speaker or fewer speakers.

By incorporating speech data from multiple speakers with close intonation, our model became more robust to common speech synthesis challenges such as varied pronunciation styles and incorrectly labeled sentences. The model captured a broader range of speech nuances, such as subtle variations in pitch, tone, and emotion, which are characteristics of natural speech. Additionally, by exposing the model to a range of phonetic and prosodic patterns within a controlled intonation range, its ability to generalize and accurately reproduce speech sounds for unseen text inputs was significantly improved.

In Section 2 we discuss related work that utilizes crowdsourced data for Luganda TTS. Section 3 describes the methodology, including the data and the preprocessing techniques. Section 4 presents the Experiments and subjective evaluation results of our model. Section 5 concludes the paper.

[1] https://huggingface.co/Sunbird/sunbird-lug-tts-commonvoice-female
[2] https://huggingface.co/Sunbird/sunbird-lug-tts

2 Related Work

There is previous effort towards building TTS models using ASR crowdsourced data as an alternative to studio-quality datasets. By leveraging the Common Voice English dataset, a multi-speaker GlowTTS model Ogun et al. (2023) was trained on recordings automatically selected using a non-intrusive mean opinion score (MOS) estimator, WVMOS[3]. Their approach improves the overall quality of generated utterances by 1.26 MOS point with respect to training on all the samples and by 0.35 MOS point with respect to training on the LibriTTS dataset[4].

In their recent work Owomugisha et al. (2023), Sunbird AI[5] presented a Tacotron2-based model fine-tuned on 15,000 Common Voice Luganda recordings from female speakers aged between 20 and 49. They employed voice activity detection[6] to trim off the silent portions from the beginning and end of each speech recording. Whilst their model's generated speech is intelligible enough to be usable in some practical applications, it is of lower quality due to existence of background noise in the training data. The model learnt that the noise is an intrinsic part of the speech signals it needs to generate. As a result, the model's output includes similar noise, lowering the overall quality of the generated speech.

3 Methodology

3.1 Data

In this work, we utilized the free crowdsourced monolingual Luganda (version 12.0) speech data obtained from the Mozilla Common Voice platform[7]. The data contains mp3 audio recordings which are up-voted or down-voted by volunteers according to a list of criteria[8]. Utterances with more than two up-votes are marked as validated, otherwise, they are marked as invalidated. The data comes with a CSV file which contains important information about the audio files including the text transcription of each speaker's utterance, name/path of the utterance, gender and age of the speaker. Throughout all experiments, we only considered validated utterances from female speakers.

3.2 Speaker Selection

Among the top 20 contributing female speakers, we identified six voices of speakers with closely matching intonation (the rise and fall of a voice in speaking). We determined closely similar speakers by listening to and comparing the intonation of their respective recordings. This similarity was essential for ensuring that the synthesized speech sounds natural and cohesive, as it minimizes the variability that arises when combining speech data from multiple speakers.

To assess the impact of utilizing multiple speakers over a single speaker or a smaller group of speakers, we trained the model using data from the highest contributing female and then separately with data from the two highest contributing females out of the six selected speakers. Table 1 provides the statistical details for the six speakers, alongside the single speaker and the two speakers.

Table 1: Statistics of the six speakers, one speaker and two speakers

LANGUAGE	SPEAKERS	STATISTICS		
		MAX LENGTH	MIN LENGTH	DURATION (hrs)
Luganda	one	8.49	1.23	8.06
	two	8.49	1.17	10.17
	six	10.72	1.14	19.04

[3] https://github.com/AndreevP/wvmos
[4] https://openslr.org/60/
[5] https://sunbird.ai/
[6] https://github.com/wiseman/py-webrtcvad/blob/master/example.py
[7] https://commonvoice.mozilla.org/en
[8] https://commonvoice.mozilla.org/en/guidelines

3.3 Data Preprocessing

3.3.1 Audio Quality Enhancement

A major problem with Common Voice speech data is the existence of distorted or noisy audio samples. The presence of mouse clicks, low-frequency noise, background speakers and music, among others, can be observed. In addition, silences at the beginning and end of several utterances can eb noticed, which causes misalignment between text and audio that degenerates TTS quality.

To eliminate silence at the start and end of the recordings, we utilized WebRTC Voice Activity Detection (VAD)[9] that was used in the previous work by Sunbird AI. First, each wav file is read, and its audio content is processed in frames of 30 milliseconds duration using a VAD algorithm provided by the webrtcvad[10] library. The VAD algorithm identifies segments of the audio containing speech. Once speech segments are detected, they are collected and concatenated to form continuous speech segments. These speech segments are then saved as new WAV files in a designated directory. Additionally, any speech segments shorter than 1 second are considered problematic and marked for further examination. This process effectively filters out silent parts of the audio files, retaining only segments containing speech.

Furthermore, to enhance the quality of the trimmed audios, we applied a pre-trained speech enhancement model[11] that works directly on the raw audio data to denoise and enhance the audio quality. The model is based on an encoder-decoder architecture with skip-connections Defossez et al. (2020). It is optimized on both time and frequency domains, using multiple loss functions such as the regression loss function (L1 loss), complemented with a spectrogram domain loss Yamamoto et al. (2020; 2019). Empirical evidence shows that it is capable of removing various kinds of background noise including stationary and non-stationary noises, as well as room reverb Defossez et al. (2020).

Additionally, to ensure inclusion of only good-quality training audio samples, we considered an absolute objective speech quality measure based on direct MOS score prediction by a fine-tuned wave2vec2.0 model (WV-MOS) Andreev et al. (2022). WV-MOS is reported to have better system-level correlation with subjective quality measures than the other objective metrics such as Perceptual Evaluation of Speech Quality (PESQ) Rix et al. (2001) and Short-Time Objective Intelligibility (STOI) Taal et al. (2010). We selected denoised recordings with an estimated MOS score above 3.5 by WV-MOS[12].

3.3.2 Text Preprocessing

We eliminated all transcripts with less than three words, alongside their corresponding audio samples. We also replaced special characters with standard representations compatible with TTS input. For example, the **ŋ** character in Luganda was replaced with **ng** characters as shown in Table 2. Additionally, incorrectly written punctuations were also replaced with standard punctuation symbols. For example: double or three full stops were found in the middle or end of the sentence, and these were replaced with one full stop.

Table 2: Sample of a preprocessed sentence with the ŋ charater in the original sentence replaced by the ng characters in the new sentence

ORIGINAL SENTENCE	NEW SENTENCE
yantuma **ŋende** ndeete ekidomola ky'amazzi.	yantuma **ngende** ndeete ekidomola ky'amazzi.

3.4 Model

For this work, we fine-tuned a Variational Autoencoder with Adversarial Learning (VITS) TTS model Kim et al. (2021) pre-trained on the LJ Speech, a single-speaker English dataset Ito & Johnson (2017). VITS is an end-to-end speech synthesis model that predicts a speech waveform conditional on an

[9] https://github.com/wiseman/py-webrtcvad/blob/master/example.py
[10] https://github.com/wiseman/py-webrtcvad
[11] https://github.com/facebookresearch/denoiser
[12] https://github.com/AndreevP/wvmos

input text sequence. It is a conditional variational autoencoder (VAE) Kingma & Welling (2013) comprised of a posterior encoder, decoder, and conditional prior encoder.

VITS is trained end-to-end with a combination of losses derived from variational lower bound and adversarial training. To improve the expressiveness of the model, normalizing flows are applied to the conditional prior distribution. During inference, the text encodings are up-sampled based on the duration prediction module and then mapped into the waveform using a cascade of the flow module and HiFi-GAN decoder. VITS's end-to-end approach combines the GlowTTS encoder Kim et al. (2020) and HiFiGAN vocoder Kong et al. (2020), which simplified our TTS training pipeline, making it more streamlined and easier to implement.

3.5 Evaluation

To assess the performance of our model trained on six speakers, we conducted mean opinion score (MOS) tests. These tests involved synthesizing audio samples using each model and subsequently evaluating their perceived naturalness on a 5-point scale. We generated 100 audio samples from the model and presented them to 10 native Luganda speakers, who were tasked with evaluating the naturalness of the synthesized speech.

The MOS is calculated as:

$$\text{MOS} = \frac{\sum \text{scores}}{N}$$

Where:

- MOS: Mean Opinion Score.
- $\sum$ scores: Sum of individual scores provided by all raters.
- N: Total number of individual scores provided by all raters.

4 Experiments and Results

Model training experiments were carried out using the Coqui TTS Library[13]. For all experiments, we fine-tuned an English VITS model provided by Coqui. All models were trained on a single NVIDIA GeForce GTX 1080 Ti GPU for 1,000,000 steps at a learning rate of 0.001 and a batch size of 16. The AdamW optimizer was used to update the model's parameters with a weighting decay of 0.01. The dataset was split with a 90% training set and a 10% validation set. The best training parameter settings that were found to give the best possible synthesis performance on the Common Voice data are shown in Table 3. These underwent a couple of quality checks including examining for the noise level of the clips by checking spectrograms as well as finding suitable audio processing parameters. We also reduced the sample rate to 22.05 kHz to speed up data-loading and consequently accelerate model training.

Table 3: Hyperparameters tuned to adapt the pre-trained VITS model to Common Voice

HYPERPARAMETER	VALUE
preemphasis	0.98
ref_level_db	20
mel_fmax	8000
log_func	np.log
spec_gain	1
use_phoneme	False
phoneme_language	False

MOS evaluation results by ten native Luganda speakers of our model are shown in Table 4. We compare our results to the existing model MOS results as reported in their paper Owomugisha et al. (2023). They fine-tuned a Tacotron2 model on Luganda Common Voice of female speakers within the age range of 20-49. Despite the existing model being trained on a broader range of speakers, our model achieved a higher quality score. Our results highlights the importance of not just the quantity

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

Table 4: MOS comparison of our VITS model trained on six female Luganda speakers to the existing Tacotron2 model

MODEL	MOS
Tacotron2-based (existing)	2.50
VITS-based (ours)	**3.55**

but also the selection of speakers with close intonation. Our approach focused on selecting speakers with closely aligned intonations, which proved to be more effective for maintaining a more consistent synthesis voice than the broader age-based selection used in the Tacotron2 model. Additionally, the quality of our model is influenced by the data preprocessing steps taken to clean the training speech recordings. By reducing background noise and selecting recording with an estimated MOS of above 3.5, we ensure that the model learns from higher quality speech samples than only trimming silents parts, which allows distorted samples in the training data. This approach allows our model to concentrate more effectively on learning the characteristics of the speech itself, resulting in more natural-sounding synthesized speech.

4.1 Comparison by Number of Speakers

We also trained two models on a single speaker and two speakers with the highest number of samples from the six speakers. As shown in Table 5, we compare their performance to our model trained on all the six speakers, which performs better. We also noticed that the model trained on two speakers of close intonation as well performed better than the one trained on a single speaker. Our results

Table 5: MOS comparison by different number of speakers

SPEAKERS	MOS
One	3.13
Two	3.22
Six	**3.55**

highlight the significant impact of utilizing multiple speakers over a single speaker or fewer speakers. The observed improvement in naturalness with multiple speakers can be attributed to the benefits of training on a more extensive and diverse dataset. The incremental improvements in MOS from one to two speakers and then to six speakers suggest that while even a small increase in speaker diversity can enhance speech quality, larger and more diverse datasets offer more significant improvements. By incorporating speech data from many speakers with close intonation, the model became more robust against common speech synthesis challenges such as varied pronunciation styles and incorrectly labeled sentences. This means that training on multiple speakers of close intonation compensates for the lack of data from a single speaker, a common challenge for underrepresented languages in speech synthesis research. The broader range of speech samples from speakers provides the model with a comprehensive understanding of the language's phonetic and prosodic nuances, enabling it to generalize better to new inputs and maintain a more consistent voice quality during synthesis.

5 Conclusion and Future Work

In this paper, we presented a Luganda TTS model trained from crowdsourced Common Voice data of multiple speakers of closely matching intonation determined through listening and comparing their voices. Additionally, we presented preprocessing techniques aimed at reducing noise from Luganda Common Voice speech and text data, crucial for effective TTS training. MOS evaluations results show that our model (3.55 MOS) trained on six speakers of similar intonation can produce better quality than the existing model (2.50 MOS) that was trained on multiple speakers based on age range. Our results show that by carefully selecting multiple speakers of close intonation, the model can generate a more consistent and better quality voice at synthesis than training on speakers of different intonations despite being from a certain age range. Furthermore, our results highlight the importance of training on a diverse and multi-speaker dataset, as it consistently produces better quality outputs compared to models trained on a single speaker or fewer speakers. We show that this is true for our model trained on six speakers and for models trained on a single speaker and two speakers. Future

work will focus on applying our methodology to build models from Common Voice data of other low-resourced African languages.

References

Pavel Andreev, Aibek Alanov, Oleg Ivanov, and Dmitry Vetrov. Hifi++: a unified framework for neural vocoding, bandwidth extension and speech enhancement. *arXiv e-prints*, pp. arXiv–2203, 2022.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, and Davis David. Building text and speech datasets for low resourced languages: A case of languages in east africa. 2022.

Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*, 2020.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540. PMLR, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, Srikanth Ronanki, and Viacheslav Klimkov. Effect of data reduction on sequence-to-sequence neural tts. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7075–7079. IEEE, 2019.

Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, and Nobuyuki Nishizawa. Training multi-speaker neural text-to-speech systems using speaker-imbalanced speech corpora. *arXiv preprint arXiv:1904.00771*, 2019.

Perez Ogayo, Graham Neubig, and Alan W Black. Building TTS systems for low resource languages under resource constraints. In *Proc. 1st Workshop on Speech for Social Good (S4SG)*, 2022.

Sewade Ogun, Vincent Colotte, and Emmanuel Vincent. Can we use common voice to train a multi-speaker tts system? In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 900–905. IEEE, 2023.

Isaac Owomugisha, Benjamin Akera, Ernest Tonny Mwebaze, and John Quinn. Multilingual model and data resources for text-to-speech in ugandan languages. In *4th Workshop on African Natural Language Processing*, 2023.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pp. 4214–4217. IEEE, 2010.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Probability density distillation with generative adversarial networks for high-quality parallel waveform generation. *arXiv preprint arXiv:1904.04472*, 2019.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203. IEEE, 2020.