Building Text-to-Speech Models for Low-Resourced Languages from Crowdsourced Data

Andrew Katumba[1], Sulaiman Kagumire[1], Joyce Nakatumba-Nabende[1], John Quinn[1], and Sudi Murindanyi[1]

[1]Makerere University

November 04, 2024

Abstract

Text-to-speech (TTS) models have expanded the scope of digital inclusivity by becoming a basis for assistive communication technologies for visually impaired people, facilitating language learning, and allowing for digital textual content consumption in audio form across various sectors. Despite these benefits, the full potential of TTS models is often not realized for the majority of low-resourced African languages because they have traditionally required large amounts of high-quality single-speaker recordings, which are financially costly and time-consuming to obtain. In this paper, we demonstrate that crowdsourced recordings can help overcome the lack of single-speaker data by compensating with data from other speakers of similar intonation (how the voice rises and falls in speech). We fine-tuned an English Variational Inference with adversarial learning for an end-to-end Text-to-Speech (VITS) model on over 10 hours of speech from six female Common Voice (CV) speech data speakers for Luganda and Kiswahili. A human mean opinion score evaluation on 100 test sentences shows that the model trained on six speakers sounds more natural than the benchmark models trained on two speakers and a single speaker for both languages. In addition to careful data curation, this approach shows promise for advancing speech synthesis in the context of low-resourced African languages. Our final models for Luganda and Kiswahili are available at https://huggingface.co/marconilab/VITS-commonvoice-females.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Building Text-to-Speech Models for Low-Resourced Languages from Crowdsourced Data

Andrew Katumba[1] | **Sulaiman Kagumire**[1] | **Joyce Nakatumba-Nabende**[2] | **John Quinn**[2] | **Sudi Murindanyi**[1]

[1]Department of Electrical and Computer Engineering, Makerere University, Kampala, Uganda

[2]Department of Computer Science, Makerere University, Kampala, Uganda

Correspondence
Andrew Katumba
Email: andrew.katumba@mak.ac.ug

Abstract

Text-to-speech (TTS) models have expanded the scope of digital inclusivity by becoming a basis for assistive communication technologies for visually impaired people, facilitating language learning, and allowing for digital textual content consumption in audio form across various sectors. Despite these benefits, the full potential of TTS models is often not realized for the majority of low-resourced African languages because they have traditionally required large amounts of high-quality single-speaker recordings, which are financially costly and time-consuming to obtain. In this paper, we demonstrate that crowdsourced recordings can help overcome the lack of single-speaker data by compensating with data from other speakers of similar intonation (how the voice rises and falls in speech). We fine-tuned an English Variational Inference with adversarial learning for an end-to-end Text-to-Speech (VITS) model on over 10 hours of speech from six female Common Voice (CV) speech data speakers for Luganda and Kiswahili. A human mean opinion score evaluation on 100 test sentences shows that the model trained on six speakers sounds more natural than the benchmark models trained on two speakers and a single speaker for both languages. In addition to careful data curation, this approach shows promise for advancing speech synthesis in the context of low-resourced African languages. Our final models for Luganda and Kiswahili are available at https://huggingface.co/marconilab/VITS-commonvoice-females.

KEYWORDS

Text-to-Speech, Low-resourced, Crowdsourced, Common Voice, Luganda, Kiswahili

1 | INTRODUCTION

The naturalness and intelligibility of Text-to-speech (TTS) models have significantly advanced in recent years, primarily driven by the advent of deep neural network models Reddy et al. (2023). These technological advancements have enabled highly effective applications, such as Google's Speech Recognition and Synthesis tools and Amazon's Polly service, which offer a wide range of voices and languages with human-like quality. Despite these achievements, which are prominent in popular languages such as English, French, and Spanish, the development and accessibility of TTS for the majority of low-resourced African languages have not experienced similar progress Gladston and Pradeep (2023). This disparity is because TTS models deployed in such applications traditionally require large amounts of high-quality single-speaker recordings, which are financially costly and time-consuming to collect for low-resourced African languages Ogayo et al. (2022b). This severely limits the accessibility of TTS research for these languages.

It is, therefore, necessary to utilize crowdsourced speech datasets—originally collected for purposes other than speech synthesis—to build TTS models for low-resourced languages. Although the quality of such datasets may not precisely match that of a standard TTS corpus, previous studies Cooper (2019), Ogun et al. (2023) suggest that they contain usable recordings for producing natural and intelligible TTS voices. Furthermore, given the most contributing single-speaker recordings in crowdsourced datasets may not be extensive, reducing reliance on such data is essential. Existing literature Latorre et al. (2019), Luong et al. (2019) demonstrates that effectively combining recordings from multiple speakers can build TTS models that produce natural-sounding speech, often superior to or equivalent to the quality achieved with single-speaker data.

In this paper, we focus on using Common Voice (CV)[‡] speech data to build TTS models for Luganda and Kiswahili, two indigenous, low-resourced languages widely spoken in East Africa. Luganda, a Bantu language, is spoken by more than 20 million people globally and accounts for over 16.7% of Uganda's population Babirye et al. (2022). Kiswahili, also known as Swahili, another Bantu language, is spoken by over 70 million people across East and Central Africa Elamin et al. (2023). Both languages utilize the Latin Alphabet and belong to the Niger-Congo language family Ogayo et al. (2022a). Despite their significant number of speakers, these languages have been largely overlooked in the development of speech synthesis systems, attributed to the scarcity of good-quality TTS models and large amounts of high-quality TTS corpora.

Using Mozilla's crowdsourced CV speech data, we built TTS models for Luganda and Kiswahili from a mixture of six female speakers. To ensure model convergence and maintain a consistent voice at inference, we selected speakers with a similar intonation determined by subjectively listening and comparing their voice recordings. Using multiple speakers with closely matching intonation guaranteed data uniformity and a sizeable multi-speaker dataset for more reliable model training. We curated the training speech and text data for each language to improve the quality of synthesized speech as CV is primarily collected for Automatic Speech Recognition (ASR), rather than speech generation Ardila et al. (2019). The models trained on six speakers produce a better natural-sounding speech than the benchmark models trained on two speakers and a single speaker with the highest number of speech recordings in CV for both languages.

The remainder of the paper is organized as follows: Section 2 discusses related work in TTS models for Luganda and Kiswahili. Section 3 presents the methodology used in the paper, including the dataset and preprocessing steps. Section 4 discusses the performance of models and evaluation. Section 5 presents the discussion about the findings, and finally, Section 6 concludes the paper.

2 | RELATED WORK

2.1 | TTS for Luganda and Kiswahili

There have been previous efforts towards building TTS models for Luganda and Kiswahili to close the technological gap in mainstream TTS research thus advancing African Natural Language Processing (NLP) research work. When developing a Luganda TTS machine from the MaryTTS Engine, Nandutu and Mwebaze Nandutu and Mwebaze (2020) utilized the Hidden Markov Model-based unit selection speech synthesis method, along with a speech corpus comprising 511 audio recordings from a single female speaker. Recently, Akera et al. Owomugisha et al. (2023) fine-tuned a Tacotron2 TTS model on 5 hours of studio speech recordings in a mixture of Luganda and English from a single female speaker, with an emphasis on code-switching capabilities between the two languages. They also fine-tuned it on 15,000 CV Luganda recordings from all male and female speakers aged between 20 and 49, although thorough preprocessing—particularly in removing background noises—was omitted. This oversight potentially hampered the performances of their models, which relied on the CV dataset.

In the case of the Kiswahili language, a concatenative Kiswahili TTS system Gakuru et al. (2005) was created using the Festival Unit Selection Speech Synthesizer. This system was built using a dataset consisting of 45 minutes of recorded speech delivered by a professional male speaker. The authors in Kelvin Rono et al. (2022) recently developed Kiswahili TTS system using the Tacotron2 architecture and WaveNet vocoder, trained on 7,108 single-speaker audio files of the Kiswahili Audio Bible dataset[§].

2.2 | End-to-End Text-to-Speech

In this work, we utilized VITS Kim et al. (2021), an open-sourced, one-stage end-to-end TTS system, for model training. The end-to-end nature of VITS allows for the direct and fast generation of speech from text, using learned hidden representations. This approach varies with two-stage TTS models, which rely on predefined intermediate features such as mel spectrograms before training a separate vocoder for audio generation. This two-stage process not only affects the quality of the generated speech but also introduces inefficiencies in training and deployment.

There are cutting-edge, one-stage end-to-end TTS models developed to synthesize human-like speech better than two-stage TTS models. However, the majority of these are not open-sourced. NaturalSpeech Tan et al. (2024) enhances the VITS architecture by introducing an innovative aligner that can be modified during learning and a module for managing bidirectional prior/posterior information. These are intended to minimize the inconsistency between the training phase and actual inference encountered in VITS due to its bijective flow component. Concurrently, the introduction of EfficientTTS 2 Miao et al. (2024) addresses prevalent issues in one-stage TTS systems, such as high computational costs during training. It does so by integrating an aligner capable of differentiation combined with an advanced attention mechanism, along with a predictor for variational alignment, thus enabling the acquisition of a richly expressive, temporally coherent latent representation. It also employs a dual-layer hierarchical variational autoencoder for the generation of waveforms, thereby achieving outputs of better quality than existing one-stage TTS models. Additionally, it

[‡] https://commonvoice.mozilla.org/

[§] https://www.wordproject.org/bibles/audio/05_swahili/index.htm

provides a quicker inference capability and a reduced model size when compared to both NaturalSpeech and VITS Miao et al. (2024).

3 | METHODOLOGY

3.1 | Data

We utilized the free crowdsourced monolingual Luganda (version 12.0) and Kiswahili (version 15.0) speech data obtained from the Mozilla CV Dataset[¶] that is powered by the voices of volunteer contributors around the world. Within the dataset, mp3 audio files are subject to a voting process by volunteers who follow specific guidelines[#]. Recordings receiving more than two positive votes are classified as validated, whereas those with fewer are deemed invalid. The data comes with a CSV file containing audio metadata, 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. We excluded utterances shorter than 1 second or longer than 30 seconds to allow for shorter model training time.

We created three subsets of datasets for each language containing speakers with closely matching intonation to ensure a uniform voice when synthesizing speech. The speaker selection process involved a two-step approach. First, human evaluators anonymously reviewed five audio samples from each of the top 20 speakers with the most recordings. The evaluators focused on key features like intonation, pitch, and rhythm. Audio samples with noticeable inconsistencies in these aspects were discarded, and only the most consistent samples were retained. This process led to the selection of six speakers whose audios were unanimously chosen by all evaluators. To further validate this selection, a K-Means clustering algorithm was applied to the audio data. Key acoustic features such as pitch, intonation, MFCCs, and rhythm were extracted from the audio samples and clustered based on feature similarity. The speakers closest to the cluster centroids were identified, and it was found that the six manually selected speakers matched those in one of the K-Means clusters. The three subsets contain the following number of speakers:

- **One speaker**: The top female speaker with the most recordings.
- **Two speakers**: On top of the first speaker, we selected another female speaker with close intonation and significant audio contributions.
- **Six speakers**: Four more speakers with close intonation to the first two speakers and substantial audio contributions were added.

[¶] https://commonvoice.mozilla.org/en
[#] https://commonvoice.mozilla.org/en/guidelines

3.2 | Audio Preprocessing

A significant problem with CV speech data is the existence of distorted or noisy audio samples. Evidently, there are various types of interference in the recordings, such as background conversations, mouse clicks, and incidental music. In addition, silences at the beginning and end of several utterances can be noticed, which causes misalignment between text and audio, thus degenerating TTS quality. The flowchart in Figure 1 shows the preprocessing steps employed to refine the raw speech data quality for all female speakers.

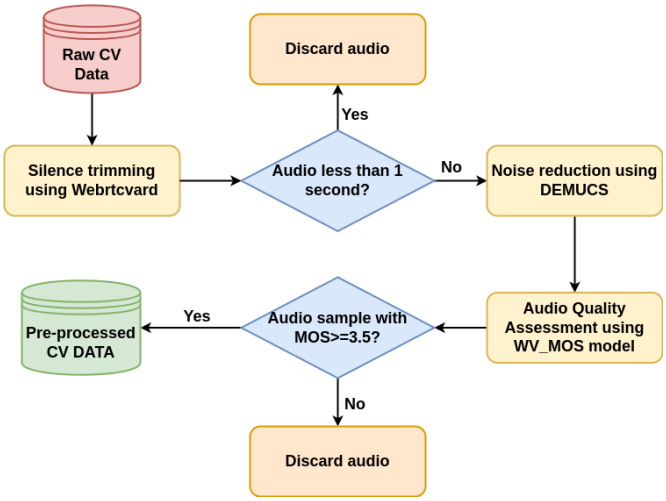

FIGURE 1 Our audio preprocessing workflow. This illustrates the sequential steps of Voice Activity Detection (VAD), Speech Enhancement, and Quality Assessment.

3.2.1 | Silence Trimming

We employed the WebRTC (Web Real-Time Communication) Voice Activity Detection (VAD) to trim out silences at the beginning and end of the recordings. This step is critical for aligning the audio with the corresponding textual transcript and eliminating non-speech segments that do not contribute valuable information for TTS training. Initially, the audio file is divided into frames of 30 milliseconds to enable detailed analysis. A padded sliding window is used over these frames to improve the accuracy of voice activity detection. The process is designed to trigger when the proportion of voiced frames, i.e. those containing speech, exceeds 90%. This high threshold ensures that only segments with a high probability of speech are selected, reducing the chance of including noise. Once triggered, the VAD collector accumulates voiced frames, providing a continuous speech data stream. It remains active until the proportion of voiced frames falls below 90%, indicating a transition back to non-speech or silence. This "detriggering" process ensures that the collected audio segments consist mainly of speech. Significantly, the

analysis window is padded at the beginning and the end with 300 milliseconds. This padding ensures that the beginnings and ends of actual speech sounds are not accidentally clipped.

3.2.2 | Noise reduction

To enhance the quality of the trimmed audio data, we applied a causal speech enhancement model that operates directly on raw waveforms to eliminate various background noises, such as mouse clicks, keyboard typing, and street noise. The model is a real-time adaptation of the DEMUCS architecture designed explicitly for speech enhancement. The model adopts a waveform-to-waveform transformation approach for audio quality purposes, employing a hierarchical generation process incorporating U-Net-like skip connections. This technique enables the model to focus on relevant audio features while effectively eliminating a wide array of background noises. Moreover, the model is optimized to produce a ”clean” version of the speech signal while minimizing the L1 regression loss function, complemented with a spectrogram domain loss.

3.2.3 | Audio Quality Assessment

We employed WV-MOSAndreev et al. (2022), an objective MOS estimation model, to automatically assign MOS scores to the denoised speech data to ensure high-quality speech samples for TTS training. We retained only samples with assigned MOS scores of 3.5. WV-MOS uses the wav2vec 2.0Baevski et al. (2020) architecture, known for its robust, task-agnostic representations obtained through contrastive self-supervised pretraining. Unlike its predecessor, MOSNetLo et al. (2019), WV-MOS incorporates a 2-layer Multilayer Perceptron (MLP) head on top of the pre-trained wav2vec 2.0 model, which optimizes its predictions for speech quality via Mean Squared Error during training. WV-MOS's training utilized the auditory evaluation outcomes from the 2018 Voice Conversion Challenge Lorenzo-Trueba et al. (2018). Empirical results show that WV-MOS achieves a higher Spearman's rank correlation factor of 0.62 on the VCC test data than MOSNet's 0.59Andreev et al. (2022). More importantly, the model was found to generalize better to audio samples beyond VCC 2018Andreev et al. (2022). The resulting training audio samples after employing WV-MOS are shown in Table 1.

3.3 | TTS Model

For this work, we fine-tune the VITS model Kim et al. (2021), pre-trained on the LJSpeech dataset Ito and Johnson (2017). We selected VITS because of its one-stage, end-to-end framework that directly synthesizes speech from text. This capability is particularly advantageous for achieving high-quality, natural-sounding speech without the complexities and inefficiencies existing in traditional open-sourced two-stage TTS systems. A key feature of the VITS model is its implementation of normalizing flows Rezende and Mohamed (2015) in adjusting the conditional prior distribution, combined with adversarial training techniques applied directly to the waveform domain. This combination enhances the model's ability to generate high-quality speech waveforms. The structure of VITS comprises several principal components, including posterior and prior encoders, a decoder, a discriminator, and a stochastic duration predictor. It is noteworthy that the posterior encoder, which employs non-causal WaveNet residual blocks Prenger et al. (2019), along with the discriminators, are only active during the model's training phase and not during actual inference. The role of the posterior encoder is to ascertain the mean and variance of the speech's normal posterior distribution. Meanwhile, the prior encoder processes the input phonemes and enhances the prior distribution's adaptability. The decoder's function is inspired by the HiFi-GAN vocoder Kong et al. (2020), crucial for converting mel-spectrograms into finely detailed speech waveforms.

3.4 | Evaluation

To evaluate the performances of the models trained on the three data subsets for both languages, we conducted mean opinion score (MOS) tests that involved synthesizing audio samples using each model and then conducting evaluations to gauge their perceived naturalness on a 5-point scale. For Luganda, we generated 100 audio samples from each model. These samples were presented to 10 native Luganda speakers, who were tasked with evaluating the naturalness of the synthesized speech. Similarly, for Kiswahili, 30 audio samples per model were synthesized and subjected to evaluation by a separate group of 10 native Kiswahili speakers.

The MOS is calculated as:

$$\text{MOS} = \frac{\sum \text{scores}}{N} \tag{1}$$

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

We finetuned a pre-trained English VITS model on the one, two and six female speakers data subsets for each language. We utilized a single NVIDIA GeForce GTX 1080 Ti GPU to train the models, executing 900k iterations with an initial learning rate of 0.001, a batch size of 16, and the AdamW optimizer which applied weighting decay of 0.01. Each data subset was split with a 90% training set and a 10% validation set. The best training parameter settings identified to give the best possible synthesis performance on the CV data are shown in Table 2. These parameters underwent several quality checks, including examining the

TABLE 1 Statistics of speakers selected for model training from Luganda and Kiswahili Common Voice data

Language	No. of Speakers	Total Utterances	Max Length	Min Length	Total duration (hrs)
Luganda	one	3677	8.49	1.23	3.83
	two	6937	8.49	1.17	7.22
	six	17201	10.72	1.14	19.04
Swahili	one	4557	11.56	1.22	5.12
	two	7045	11.56	1.22	7.84
	six	13132	13.25	1.46	15.52

noise level by checking spectrograms and finding suitable audio processing parameters. We checked whether spectrograms appear cluttered, particularly during silent parts, to determine whether the data may or may not be suitable for TTS, even after applying the audio preprocessing techniques.

TABLE 2 Hyperparameters tuned to adapt the pre-trained VITS model to Luganda and Kiswahili 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

4.1 | Results

4.1.1 | Subjective Evaluation

We compare the naturalness of our models trained on one, two and six speakers for each language through subjective listening and MOS ratings as shown in Table 3. For Luganda, training on utterances of six female speakers with close intonation achieved a better MOS of 3.55 than training on two speakers (3.22 MOS) or a single speaker (3.15 MOS). In the case of Kiswahili, training on six speakers achieved a better MOS of 4.05 than training on two speakers (3.93 MOS) or a single speaker (3.84 MOS). Interestingly, training on two speakers achieved more naturalness than training on a single speaker for both languages despite the minimal difference in total hours for both datasets.

TABLE 3 Comparison of Evaluated MOS on Common Voice for different number of speakers

No. of Speakers	Luganda	Kiswahili
One	3.13	3.84
Two	3.22	3.93
Six	**3.55**	**4.05**

We also compared the naturalness of our Luganda TTS model trained on six female speakers against the existing Tacotron2-based model trained on female speakers within the age range of 20-49 from CV. The results are as initially reported in their work Owomugisha et al. (2023). Despite being trained on a limited range of speakers, our model is perceived to sound more natural than the existing one, as illustrated by the higher attained MOS in Table 4.

TABLE 4 Performance 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**

4.1.2 | Text Length

We also compared the perceived naturalness of synthesized speech as a function of the length of the input text across the three models. This evaluation is pivotal, as some TTS models excel in generating high-quality speech for shorter sentences but face challenges in maintaining that quality as sentence length increases. To quantify this, we averaged MOS across all models based on sentence length, where sentences with fewer than 10 words are labelled as **short**; otherwise, they are **long**. Results for this comparison are shown in Table 5. Training on six speakers performs well both on short and long sentences compared to one speaker and two speakers in both languages.

TABLE 5 Average MOS scores by sentence length

Language	No. of Speakers	Short	Long
Luganda	one	3.12	3.19
	two	3.26	3.09
	six	**3.57**	**3.47**
Kiswahili	one	3.86	3.87
	two	3.92	3.93
	six	**4.06**	**4.05**

5 | RESULTS DISCUSSION

5.1 | MOS Evaluation

The subjective human evaluation results in Table 3 highlight a significant impact of utilizing multiple speakers over a single speaker or fewer speakers for Luganda and Kiswahili. We observed that models trained on a mixture of six female speakers with closely matched intonation exhibit better naturalness and intelligibility compared to those trained on a single speaker or two speakers. Models trained on two speakers also sound more natural than those trained on single speakers for both languages.

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 for both languages 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, our models became more robust against common speech synthesis challenges such as varied pronunciation styles and incorrectly labelled sentences. This robustness was reflected in more accurate and consistent alignments between the text input and the synthesized speech, evident in the enhanced naturalness of the output.

Furthermore, 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 similar speakers provides the model with a comprehensive understanding of the language's phonetic and prosodic characteristics, enabling it to generalize better to new inputs and maintain a consistent voice quality during synthesis.

5.2 | Comparison with Previous Work

Our approach to Luganda speech synthesis involves a careful selection of speakers based on their intonation, contrary to the existing Tacotron2-based modelOwomugisha et al. (2023) that uses age-based criteria. Specifically, we fine-tuned using our Luganda model on data from six speakers selected for their closely matched intonations. This strategic choice was grounded in the hypothesis that intonation homogeneity among speakers would yield a more coherent and naturally sounding synthesized voice. The empirical results from our experiments validate this hypothesis as shown in Table 4, demonstrating an improvement in voice consistency and naturalness.

Consequently, the combination of the data preprocessing techniques we employed and the end-to-end nature of VITS significantly impacts the quality of the synthesized speech produced by our models. Removing silences at the start and end of the recordings aligned the text and the spoken words which is essential for producing speech that sounds natural and fluid. Misalignments can lead to awkward pauses, abrupt speech, or unnatural intonations, which diminish the listener's experience. The denoiser model enhanced the training data quality by reducing background noise which could potentially degrade the model's learning efficiency. Moreover, using an objective MOS estimation model, WV-MOS, helped to enhance the quality of our dataset by ensuring that only speech samples with high MOS scores were selected for training.

5.3 | MOS Variations Between Luganda and Kiswahili

We observed that Kiswahili consistently outperformed Luganda in perceived naturalness and intelligibility, despite having a smaller data size. This trend is potentially attributed to the higher quality of the Kiswahili CV dataset compared to that of Luganda. According to estimations by the WV-MOS model, the average predicted MOS for Kiswahili was 3.92, higher than the 3.56 MOS predicted for Luganda. These estimated scores suggest that the Kiswahili CV recordings likely contain minimal background noise and feature more consistent pronunciation, factors which contribute to more effective model training and superior synthesized speech.

6 | CONCLUSION

This paper presents Luganda and Kiswahili TTS models built using the crowdsourced Common Voice speech datasets. With careful data curation, we show that models trained on a mixture of speakers with close intonation are perceived to be more natural and intelligible than those trained on one or fewer speakers. We show this is true for models trained on six speakers compared to using two or one speaker for both languages. Training on six speakers achieved better MOS of 3.55 than 3.22 and 3.13 when using one or two speakers for Luganda. Training on six speakers for Kiswahili achieved a better MOS of 4.05 than 3.84 and 3.93 when using a single or two speakers. Our Luganda TTS model, trained on six speakers, outperforms an existing Tacotron2-based model trained on all female speakers aged 20-49 by a 1.05 MOS margin. Our results show that choosing speakers of a closely matching intonation, alongside proper audio preprocessing, can compensate for the lack of high-quality single-speaker TTS data to build TTS models for the majority of low-resourced African languages.

ACKNOWLEDGMENTS

This work was carried out with support from Google.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

REFERENCES

Andreev, P., Alanov, A., Ivanov, O. & Vetrov, D. (2022) Hifi++: a unified framework for neural vocoding, bandwidth extension and speech enhancement. *arXiv e-prints*„ arXiv–2203.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J. et al. (2019) Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*,.

Babirye, C., Nakatumba-Nabende, J., Katumba, A., Ogwang, R., Francis, J.T., Mukiibi, J. et al. (2022) Building text and speech datasets for low resourced languages: A case of languages in east africa,.

Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449–12460.

Cooper, E. (2019) *Text-to-speech synthesis using found data for low-resource languages*. : Columbia University.

Elamin, M., Chanie, Y., Ewuzie, P. & Rutunda, S. Multilingual automatic speech recognition for kinyarwanda, swahili, and luganda: Advancing asr in select east african languages. In: *4th Workshop on African Natural Language Processing, 2023*.

Gakuru, M., Iraki, F.K., Tucker, R., Shalonova, K. & Ngugi, K. Development of a kiswahili text to speech system. In: *Ninth European Conference on Speech Communication and Technology, 2005*.

Gladston, A.R. & Pradeep, K. Exploring solutions for text-to-speech synthesis of low-resource languages. In: *2023 4th International Conference on Signal Processing and Communication (ICSPC). IEEE, 2023*, pp. 168–172.

Ito, K. & Johnson, L. (2017) *The lj speech dataset*. https://keithito.com/LJ-Speech-Dataset/.

Kelvin Rono, K., Ciira, W.M. & Mwangi, E. (2022) Development of a kiswahili text-to-speech system based on tacotron 2 and wavenet vocoder,.

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

Kong, J., Kim, J. & Bae, J. (2020) *Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis*. https://arxiv.org/abs/2010.05646.

Latorre, J., Lachowicz, J., Lorenzo-Trueba, J., Merritt, T., Drugman, T., Ronanki, S. et al. Effect of data reduction on sequence-to-sequence neural tts. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019*, pp. 7075–7079.

Lo, C.C., Fu, S.W., Huang, W.C., Wang, X., Yamagishi, J., Tsao, Y. et al. (2019) Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*,.

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T. et al. (2018) The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*,.

Luong, H.T., Wang, X., Yamagishi, J. & Nishizawa, N. (2019) Training multi-speaker neural text-to-speech systems using speaker-imbalanced speech corpora. *arXiv preprint arXiv:1904.00771*,.

Miao, C., Zhu, Q., Chen, M., Ma, J., Wang, S. & Xiao, J. (2024) Efficienttts 2: Variational end-to-end text-to-speech synthesis and voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 1650–1661. doi:10.1109/TASLP.2024.3369528.

Nandutu, I. & Mwebaze, E. (2020) *Luganda text-to-speech machine*. https://arxiv.org/abs/2005.05447.

Ogayo, P., Neubig, G. & Black, A.W. (2022) Building african voices. *arXiv preprint arXiv:2207.00688*,.

Ogayo, P., Neubig, G. & Black, A.W. Building TTS systems for low resource languages under resource constraints. In: *Proc. 1st Workshop on Speech for Social Good (S4SG), 2022b*.

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

Owomugisha, I., Akera, B., Mwebaze, E.T. & Quinn, J. Multilingual model and data resources for text-to-speech in ugandan languages. In: *4th Workshop on African Natural Language Processing, 2023*.

Prenger, R., Valle, R. & Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019*, pp. 3617–3621.

Reddy, V.M., Vaishnavi, T. & Kumar, K.P. Speech-to-text and text-to-speech recognition using deep learning. In: *2023 2nd International Conference on Edge Computing and Applications (ICECAA). IEEE, 2023*, pp. 657–666.

Rezende, D. & Mohamed, S. Variational inference with normalizing flows. In: *International conference on machine learning. PMLR, 2015*, pp. 1530–1538.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y. et al. (2024) Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*„ 1–12.doi:10.1109/TPAMI.2024.3356232.