1000 African Voices: Advancing inclusive multi-speaker multi-accent speech synthesis

Sewade Ogun, Abraham T. Owodunni, Tobi Olatunji, Eniola Alese, Babatunde Oladimeji, Tejumade Afonja, Kayode Olaleye, Naome A. Etori, Tosin Adewumi

1000 African Voices: Advancing inclusive multi-speaker multi-accent speech synthesis

Sewade Ogun[1,*], *Abraham T. Owodunni*[2,*], *Tobi Olatunji*[2,*], *Eniola Alese*[3,*], *Babatunde Oladimeji*[3,*], *Tejumade Afonja*[4,5,*], *Kayode Olaleye*[6,*], *Naome A. Etori*[7,*], *Tosin Adewumi*[8,*]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France [2]Intron Health [3]Amazethu Research [4]CISPA Helmholtz Center for Information Security [5]AI Saturdays Lagos [6]Data Science for Social Impact Group, University of Pretoria [7]University of Minnesota - Twin Cities, USA [8]Luleå University of Technology [*]Masakhane NLP

sewade.ogun@inria.fr, tobi@intron.io

Abstract

Recent advances in speech synthesis have enabled many useful applications like audio directions in Google Maps, screen readers, and automated content generation on platforms like TikTok. However, these systems are mostly dominated by voices sourced from data-rich geographies with personas representative of their source data. Although 3000 of the world's languages are domiciled in Africa, African voices and personas are under-represented in these systems. As speech synthesis becomes increasingly democratized, it is desirable to increase the representation of African English accents. We present Afro-TTS, the first pan-African accented English speech synthesis system able to generate speech in 86 African accents, with 1000 personas representing the rich phonological diversity across the continent for downstream application in Education, Public Health, and Automated Content Creation. Speaker interpolation retains naturalness and accentedness, enabling the creation of new voices.

Index Terms: text-to-speech, African-accented TTS, accented speech, multi-accent TTS, multi-speaker TTS

1. Introduction

Synthetic voices are used in everyday applications for text reading, content generation, voice-over, etc., and provide audio feedback in applications such as maps, language learning tools, smart speakers, and voice assistants. Synthetic voices have also found wider adoption as a plugin with the proliferation of large language models. With the high naturalness and high quality of synthetic voices [1, 2], they have become more widely used on online platforms and social media.

Synthetic voices are usually generated by speech synthesis systems, and there are several open-source systems available for use, e.g., [3]. Several of these systems are provided in English, and several efforts have been made to cover over 1000 languages of the world, including several African languages [4].

Over 1.4 billion people reside in Africa, and more than 237 million people speak the English language at a bilingual or native level [5], including Nigeria, Uganda, South Africa, etc., and several other African countries speak it as a second or third language. Considering this large demography of speakers, the current representation of personas in synthetic voices does not show this diversity in terms of the typical African accent.

Generating synthetic speech with personas similar to one's demography is paramount for the widespread adoption and acceptability of the technology. For example, a Nigerian content creator would prefer to generate speech in an accent (or language) that is familiar and natural to their audience. Voice cloning methods have been created to generate speech in the voice of a reference speaker, however, these systems still fail to generalize to the diverse African accents from our analysis. This is because speaker representation in speech synthesis systems favors more general accents like the American, British, or Australian accents, etc. Even the widely used English text-to-speech (TTS) datasets are not representative of diverse demographies [6, 7, 8].

Therefore, in this work, we focus on expanding the diversity of synthetic personas in TTS systems to the typical African accent. Our work covers 747 speakers and 86 accents from 9 countries. We focus on improving several TTS systems in generating voices with African-like personas, enabling a larger representation of African voices for use in applications like podcasts and content creation. Finally, we explore a speaker averaging method to create new personas different from those used to train the TTS systems. This was done to ensure that users can generate synthetic speech of diverse accents, without the fallout of overuse of a specific speaker or accent in our dataset.

Section 2 reviews some literature particular to this research. Sections 3 and 4 describe the dataset, data preprocessing methods, TTS models, evaluation protocols, and experiments. We discuss the results in Section 5 and conclude with limitations and the summary of our work in Sections 6 and 7 respectively.

2. Literature Review

There have been works extending TTS to other accents. These include works on multilingual, multi-speaker TTS synthesis, e.g., [3], where the accents of different speakers were learned jointly with the language, enabling the model to produce English speech of different accents. Similarly, [9] disentangled the speaker from the accent so that the accent can be assigned to any speaker in the TTS model. Some curated datasets also covered African languages, e.g., CMU Wilderness [10] and MMS [4] datasets, but majorly have single speakers for each language.

In the African context, the authors of [11] curated datasets and built speech synthesis systems for 12 low-resourced African languages. Similarly, BibleTTS [12] contained 86 hours of recordings involving 10 African languages, including Asante Twi, Hausa, and Yoruba. The authors in [13] introduced an open-source corpus of Yoruba, a language spoken by over 22 million people. Other efforts include Lagos-NWU Yoruba corpus [14] involving 16 female and 17 male speakers, and the work by [15] involving 33 speakers. Gamayun [16] introduced multi-speaker TTS for some marginalized languages and code-switched data for four South African languages by [17]. Similarly, a TTS model was implemented in Festival for the Fon language by [18]. Although these works have explored creating TTS datasets and TTS systems for African languages, our work focuses on increasing the representation of African accents in English-based TTS systems.

Table 1: *Afro-TTS dataset statistics. The country column indicates the ISO 3166-1 country code.*

Country	# samples	# speakers	# accents	Duration (h)
NG	25564	549	48	99.85
KE	5307	58	8	16.45
ZA	4279	125	20	16.41
GH	727	4	3	2.28
ZW	47	6	3	0.20
RW	40	1	1	0.15
SL	38	1	1	0.14
UG	26	2	1	0.08
ZM	8	1	1	0.04

3. Methodology

3.1. Dataset: Afro-TTS

Curating a dataset of diverse African-accented English speakers is crucial for this work. Therefore, we initiated our research by online crowd-sourcing of data, following the methodology used in a previous work [19]. Volunteers were asked to read and record a sequence of English texts. The texts contained general domain words and were specifically enriched with African-named entities to accurately represent the diversity of African names and organizations. Table 1 shows the dataset statistics. The data collection process involved 747 paid contributors from 9 countries representing 86 accents. Contributors consented to have their audio used for TTS. The 136-hour, 16-bit, 48 kHz audio files, along with their metadata containing anonymized speaker identity, country, accent, age group, gender, etc., have been open-sourced to facilitate African speech research.[1]

3.1.1. Dataset preprocesing

As the dataset was recorded remotely on different devices, it needed to be processed to be suitable for TTS training. Firstly, the speech samples were denoised using a speech enhancement model [20], which removes various background noise, including stationary and non-stationary noises, and room reverberation. The denoised samples were then passed through a bandwidth-extension model, VoiceFixer [21], to improve the quality of some of the highly degraded utterances, i.e., utterances recorded with low-resolution and clipping distortion. VoiceFixer has three audio processing modes which are suitable for different levels of degradation. To keep the audio file with the best quality per sample, we evaluated the quality of the denoised sample and the three enhanced samples using a quality estimator, WVMOS [22, 23][2], then we selected the sample with the highest predicted MOS score among the samples per utterance. All samples were processed at a sampling rate of 16 kHz. The speech samples were then normalized to a volume level of -27 dB using the RMS-based normalization method in ffmpeg[3] and a voice activity detection tool [24] was used to eliminate long pauses from the speech recordings.

We selected 736 samples covering speakers and accents with more than 20 minutes of data for testing. Finally, the remaining samples were randomly split into training (35042) and development (200) sets. We excluded recordings with a duration of over 50 seconds or samples with character counts exceeding 400 for faster training.

Abbreviations and name titles were expanded, e.g., "Alh $\rightarrow$ Alhaji", "Maj $\rightarrow$ Major" and numbers were converted to their word forms using the NeMo text normalization toolkit [25]. In addition, some punctuation marks (open bracket, closed bracket, colon, and semicolon), were also expanded to their word form, as they were read out in the dataset according to annotation instructions.

3.2. TTS models

Two state-of-the-art, open-source, end-to-end TTS models, *VITS* [1] and *XTTS* [26], were used in our experiments. VITS is an end-to-end model that adopts variational inference augmented with normalizing flows and an adversarial training process. XTTS is a recent multilingual TTS system with cross-language voice cloning capabilities comprising three modules; a VQ-VAE module, a GPT module, and an audio decoder. VITS (86.6 M parameters) was trained on the VCTK dataset (44 h of 109 native English speakers) for 500k iterations while XTTS (version 2, 750 M parameters) had been trained on a dataset of over 16k hours comprising 16 languages,[4] including English.

The models were fine-tuned on the training set as *VITS-FT* and *XTTS-FT*[5] respectively. We also trained a randomly initialized VITS model from scratch as *VITS-O* on the training set to validate the quality of the data for TTS experiments. Lastly, we modified the VITS model as *VITS-EXT* to take an external 256-dimensional l2-normalized speaker embedding vector as speaker conditioning. The speaker embeddings were extracted from Resemblyzer [27], a speaker embedding extractor. The weights of incompatible layers were re-initialized during fine-tuning, e.g., the speaker embedding module.

3.3. Speaker interpolation

Speaker interpolation has been typically used in Hidden Markov Model (HMM)-based TTS systems for changing speaker characteristics [28]. Interpolation can be done in several ways including interpolating between different Gaussian distributions of speakers or between speaker representations. Given speakers $S1$ and $S2$, for example, a new speaker $S3$ can be generated using a linear interpolation of their speaker representations:

$$S3 = \alpha * S1 + (1 - \alpha) * S2 \quad (1)$$

The interpolation ratio α enables the speaker characteristics to be changed from one speaker to another along a spectrum. Therefore, to generate more personas with African accents, we interpolated speakers with the same gender, country, and accent, using their speaker embeddings, creating over 200 additional speakers. We filtered by country because there are regional differences in the accent from different countries even for the same language, e.g., Swahili and Hausa language speakers in different countries.

3.4. Evaluation protocol

Several subjective and objective metrics were used to compare the performance of the TTS systems. For objective evaluation, we computed the overall quality **(WV-MOS)** and the naturalness **(NISQA)** [29] of the utterances using model-based quality estimators. The cosine similarity **(cos-sim)** of the target speaker

[1] models and dataset can be accessed via `https://huggingface.co/intronhealth/afro-tts`

[2] `https://github.com/AndreevP/wvmos`

[3] `https://github.com/slhck/ffmpeg-normalize`

[4] `https://docs.coqui.ai/en/latest/models/xtts.html`

[5] Only the GPT module was fine-tuned in our experiments.

Table 2: *Objective and subjective evaluation results (with 95 % confidence interval) for pre-trained and fine-tuned TTS models. Mean opinion score (MOS), naturalness MOS (Nat-MOS), model-based MOS (WV-MOS), model-based naturalness (NISQA), cosine similarity (cos-sim), accentedness MOS (Accent-MOS), % word error rate (% WER), and preference rankings are provided.*

	Overall quality and naturalness				Speaker & accent similarity		Intelligibility	Preference
Model	MOS	Nat-MOS	WV-MOS	NISQA	cos-sim	Accent-MOS	% WER	Ranking
GT denoised	3.75 ± 0.04	4.55 ± 0.03	2.85 ± 0.04	4.55 ± 0.03	-	4.49 ± 0.03	-	-
VITS	3.80 ± 0.04	2.84 ± 0.06	4.26 ± 0.02	3.84 ± 0.04	-	1.81 ± 0.05	32.75	-
XTTS	3.92 ± 0.04	3.31 ± 0.06	3.71 ± 0.02	3.00 ± 0.03	0.828	2.31 ± 0.06	13.81	-
Trained/fine-tuned models								
VITS-O	3.02 ± 0.05	4.00 ± 0.04	2.93 ± 0.03	2.94 ± 0.02	0.834	4.02 ± 0.04	66.77	-
VITS-FT	3.33 ± 0.04	4.18 ± 0.04	3.03 ± 0.03	2.97 ± 0.03	0.834	4.16 ± 0.04	51.77	(1192) 2rd
VITS-EXT	3.14 ± 0.05	4.07 ± 0.04	3.02 ± 0.03	$\mathbf{3.06 \pm 0.03}$	**0.914**	4.07 ± 0.04	57.31	(1168) 3rd
XTTS-FT	$\mathbf{3.77 \pm 0.04}$	$\mathbf{4.39 \pm 0.03}$	$\mathbf{3.31 \pm 0.03}$	$\mathbf{3.07 \pm 0.04}$	**0.889**	$\mathbf{4.35 \pm 0.03}$	**19.20**	**(1235) 1st**

embedding to the reference speaker embedding was also computed to measure the speaker similarity. To measure the intelligibility of the generated utterances, we computed the word error rate **(WER)** between the ground truth text and text predicted by Whisper-large-v3 [30] from the Hugging Face library.

For subjective evaluation, we performed listening tests on the generated utterances using the Intron Speech Vault Platform.[6] We asked participants to evaluate the overall quality **(MOS)**, naturalness **(Nat-MOS)**, and accentedness **(Accent-MOS)** of the utterances on a Likert scale of 1–5 with 1.0 intervals. MOS evaluates the lack of noise and correct pronunciation while Nat-MOS measures how natural or human-like the speech sounds, regardless of intelligibility or clarity. We also asked participants from the same country as the reference accent to rate the closeness of the generated utterance to the reference accent **(Accent-Match)**, reference country **(Country-Match)**, and reference gender **(Gender-Match)** on a scale of 1–5. Finally, we performed a **preference** test[7] on our accented models using a modified Hugging Face TTS Arena tool [31]. Here, volunteers were shown a pair of utterances generated by two TTS models among the finetuned TTS models at every turn, then they were asked to select the more natural utterance between the pair.

In addition, the naturalness of utterances generated by the newly created speakers was evaluated objectively and through listening tests. We also computed the equal error rate **(EER)** of the new speaker's utterance against the utterances of the averaged speakers to validate that they are indeed different.

4. Experiments

4.1. Training and finetuning hyper-parameters

All the VITS-like models, VITS-O, VITS-FT, VITS-EXT, were trained with mixed-precision training on 4 GPUs for 350k iterations, with a global batch size of 64. We used the VITS implementation from the authors[8] for training and inference, with the default training hyper-parameters. Similarly, XTTS was fine-tuned on 1 GPU using the Coqui library[9] for about 250 iterations with default finetuning hyper-parameters, using a batch size of 2 and gradient accumulation of 128. The VITS-like models take phonemes (without stress markers) interspersed with a blank token as input and generate audio at a 16 kHz sampling rate while the XTTS-FT model was fine-tuned at 24 kHz by upsampling the 16 kHz processed training data. In addition,

[6] https://speech.intron.health

[7] https://huggingface.co/spaces/eniolaa/AfroTTS-Evaluation

[8] https://github.com/jaywalnut310/vits

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

Table 3: *Country-level MOS results (with 95 % confidence interval) for the best model (XTTS-FT) showing naturalness (Nat-MOS), accentedness (A-MOS), country-match (Country-M), and accent-match (Accent-M).*

	Nat-MOS	A-MOS	Country-M	Accent-M
KE	4.44 ± 0.09	4.37 ± 0.09	3.93 ± 0.52	3.90 ± 0.48
NG	4.37 ± 0.04	4.33 ± 0.04	4.24 ± 0.11	3.54 ± 0.15
ZA	4.46 ± 0.11	4.47 ± 0.10	3.40 ± 0.49	2.93 ± 0.53

to mitigate regional bias during model training, we computed the average duration per speaker on the original dataset, and then we duplicated all the samples corresponding to a speaker to equal the average duration of 10 minutes if the speaker's duration was less than the average.

At inference time, we set the noise scale of the VITS-like models to 0.667, with the duration noise scale set to 0.8 following the original work. For speaker and accent representation, we either used the learned speaker embedding, the speaker embedding extracted from the speaker extractor model (for VITS-E), or the test utterance (for XTTS models). XTTS-based models generate speech at 24 kHz while VITS-like models generate speech at 16 kHz. Therefore, all the generated samples were resampled to 16 kHz for evaluation. Also, the speaker interpolation parameter α was set to 0.5, and at most three speakers were interpolated. Additionally, statistical significance tests were carried out on the presented results using the bootstrap method. The best model or best models with similar scores among the trained and fine-tuned models are highlighted in bold.

5. Results and Discussion

Aggregated results for objective and subjective evaluations are presented in Table 2 excluding results for speaker interpolated speakers. Furthermore, in Tables 3 and 4, we show a breakdown of the ratings for the best fine-tuned model where the raters' country/accent match the utterance country/accent enabling us to measure the similarity in accent, country, and gender of the utterances. Here, we only show results for three counties where we had significant evaluators to measure the similarity in accent, country, and gender of the utterances. Also, these three countries make up 97.8 % of the dataset.

We generated 735 utterances per model, along with 162 speaker-interpolated utterances from VITS-based models. Also, 427 unique participants (64.6 % female, 60 accents, from 9 countries) provided 26,974 human ratings representing roughly 5 ratings per utterance. Additionally, 26,116 ratings were provided across models and 858 ratings were provided for speaker-interpolated utterances. *GT denoised* are the Ground truth (GT) reference test samples.

Table 4: *Accent-level: Best model (XTTS-FT) results for ratings where the utterance accent is matched to the rater's accent.*

Accent	Country-Match	Accent-Match
Afrikaans	4.80 ± 0.56	2.67 ± 5.17
Hausa	4.23 ± 0.22	3.93 ± 0.25
Igbo	4.12 ± 1.37	2.25 ± 1.24
Swahili	3.89 ± 0.56	3.77 ± 0.57
Tswana	3.75 ± 3.01	3.50 ± 1.59
Yoruba	4.25 ± 0.14	3.30 ± 0.20
Zulu	3.09 ± 0.59	2.88 ± 0.62

5.1. Naturalness and overall quality results

Table 2 shows that although participants rated the pre-trained models higher in overall quality (MOS), the best fine-tuned model (XTTS-FT) was rated 1.08 MOS points higher than its pre-trained version, and 1.55 MOS points higher than the VITS baseline in terms of naturalness, probably due to the better pronunciation of African named entities in the reference text. Our results demonstrate that we are indeed able to generate speech that is more relatable to an African audience. However, model-based quality metrics (WV-MOS and NISQA) show inverse results (pre-trained models are better) possibly because underlying models lack exposure to African-sounding speech.

5.2. Accentedness and speaker similarity results

Most significantly in Table 2, participants rated XTTS-FT only 0.14 MOS points lower in Accentedness (Accent-MOS) than the GT in contrast to the XTTS baseline which was rated 2.18 MOS points below the GT. This validates that our approach generates natural-sounding accented speech, bridging the current gap in the representation of African voices in speech synthesis. Speaker similarity results (cos-sim) also showed that generated utterances from fine-tuned models are closer to reference utterances than generated utterances from pre-trained models.

5.3. Preference scores

Preference test scores in Table 2 show that raters prefer utterances generated by XTTS-FT. Preference test scores aligned well with MOS metrics where the XTTS-FT model outperformed the VITS-based models. Also, the pre-trained XTTS model had a higher MOS score on average than the pre-trained VITS model likely because of its multilingual pretraining.

5.4. Intelligibility

Utterances by XTTS models had lower WER than VITS-based models perhaps as a result of the greater diversity and quantity (363x) of its pretraining data. A higher WER after fine-tuning of XTTS was also due to noise artifacts in the Afro-TTS dataset.

5.5. Regional diversity considerations

Table 3 reveals that although the naturalness and accentedness of generated utterances from our best model are close to the GT, regional differences surface. South Africans (ZA) rated South African-generated utterances lower in Accent-Match than West Africans (NG) rated the generated utterances with West African accents. Although most participants agreed that the generated utterances represent the reference country from Table 4, the generated accents do not always match the reference accent, e.g., generated speech in Afrikaans accent may sound like Zulu, and Igbo generated accent may sound like Yoruba.

Table 5: *Speaker interpolation results showing MOS, naturalness (Nat-MOS), and accentedness (A-MOS) of utterances generated using novel speakers from speaker interpolation.*

Model	MOS	Nat-MOS	A-MOS	%EER
VITS-EXT	3.17 ± 0.16	4.05 ± 0.15	4.10 ± 0.14	20.42
VITS-FT	3.47 ± 0.15	4.22 ± 0.14	4.37 ± 0.11	16.20
VITS-O	3.18 ± 0.17	4.08 ± 0.16	4.20 ± 0.13	16.20

Table 6: *Overall speaker interpolation MOS results showing how much synthetic utterances from interpolated speakers match the expected accent, country, and gender of the source speakers. Accent-match (Accent-M), gender-match (Gender-M), and country-match (Country-M) are provided.*

Country	Accent-M	Gender-M	Country-M
KE	4.17 ± 0.21	4.79 ± 0.13	4.20 ± 0.21
NG	3.65 ± 0.13	4.62 ± 0.08	3.91 ± 0.13
ZA	3.47 ± 0.19	4.65 ± 0.09	3.64 ± 0.18

Indeed, in multilingual countries like NG and ZA, speaker accents are difficult to classify into binary accent classes [32], as many speakers have dual accents. Notably, although East African speakers have lower representation in the dataset compared to Southern Africans, East Africans (e.g., Swahili) generally rated Accent-Match higher than South Africans (e.g., Zulu, Afrikaans). These inconsistencies may reflect the accent imbalance in the Afro-TTS dataset and require further investigation.

5.6. Effects of speaker interpolation

Table 5 shows MOS results on speaker-interpolated utterances from fine-tuned VITS models. Although %EER shows interpolated speakers have a high correlation with source speakers, our results show that speaker interpolation is indeed a viable approach for creating novel synthetic speakers that sound African (i.e., natural and accented). Furthermore, Table 6 shows that the generated utterances' gender, accent, and country match that of the reference interpolated speakers. In the future, speaker interpolation outside of the same accents could facilitate the exploration of novel or multilingual accents.

6. Limitations

Although we included 86 distinct African accents, this is a small fraction of more than 3000 languages and accents across the continent. Additionally, imbalanced accent representation in our dataset may yield biased performance favoring majority accents. Lastly, we acknowledge the privacy risk of releasing multi-speaker TTS systems that mimic voices in the source data increasing the risk of voice cloning or voice theft. We mitigate this by removing any speaker identifiers, making it more challenging to identify individuals. Finally, disentangling of speaker and accent characteristics is left for future work.

7. Conclusion

We developed an African-accented TTS system that achieves near-GT MOS for naturalness and accentedness using the pan-African TTS dataset, a 136-hour dataset containing 747 speakers with 86 African accents from 9 countries. Although open questions remain, our work greatly improves the representation of African voices in speech synthesis.

8. Acknowledgements

We appreciate the over 700 African contributors whose voices made this work possible. We appreciate the invaluable support from Intron Health for contributing the datasets for this work, the pan-African platform for data collection, developing custom UIs for human evaluation (MOS), quality reviews, multi-currency contributor payments, and compute for experiments. Experiments presented in this paper were partly carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`). Tejumade Afonja is partially supported by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. We appreciate the support provided by the BioRAMP researchers, whose collaboration and insights have been fundamental to our research.

9. References

[1] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[2] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[3] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[4] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[5] Wikipedia contributors, "List of countries by english-speaking population — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=List_of_countries_by_English-speaking_population&oldid=1211690147, 2024, [Online; accessed 9-March-2024].

[6] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[7] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[8] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," *Interspeech 2019*, 2019.

[9] R. Badlani, R. Valle, K. J. Shih, J. F. Santos, S. Gururani, and B. Catanzaro, "Multilingual multiaccented multispeaker TTS with RADTTS," in *Proc. ICASSP*. IEEE, 2023.

[10] A. W. Black, "CMU wilderness multilingual speech dataset," in *Proc. ICASSP*. IEEE, 2019, pp. 5971–5975.

[11] P. Ogayo, G. Neubig, and A. W Black, "Building African Voices," in *Proc. Interspeech 2022*, 2022, pp. 1263–1267.

[12] J. Meyer, D. I. Adelani, E. Casanova, A. Öktem, D. W. J. Weber, S. Kabongo, E. Salesky, I. Orife, C. Leong, P. Ogayo *et al.*, "BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus," in *Proc. Interspeech*, 2022, pp. 2383–2387.

[13] A. Gutkin, I. Demirsahin, O. Kjartansson, C. E. Rivera, and K. Túbòsún, "Developing an open-source corpus of yoruba speech," in *Proc. Interspeech*, October 25–29, Shanghai, China, 2020., 2020, pp. 404–408. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1096

[14] D. van Niekerk, E. Barnard, O. Giwa, and A. Sosimi, "Lagos-NWU Yoruba speech corpus," 2015.

[15] D. Van Niekerk and E. Barnard, "Tone realisation in a Yoruba speech recognition corpus," 2012.

[16] A. Öktem, M. A. Jaam, E. DeLuca, and G. Tang, "Gamayun-language technology for humanitarian response," in *IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE, 2020, pp. 1–4.

[17] D. van Niekerk, C. van Heerden, M. Davel, N. Kleynhans, O. Kjartansson, M. Jansche, and L. Ha, "Rapid development of TTS corpora for four south african languages," 2017.

[18] T. K. Dagba and C. Boco, "A text to speech system for Fon language using multisyn algorithm," *Procedia Computer Science*, vol. 35, pp. 447–455, 2014.

[19] T. Olatunji, T. Afonja, A. Yadavalli, C. C. Emezue, S. Singh, B. F. Dossou, J. Osuchukwu, S. Osei, A. L. Tonja, N. Etori *et al.*, "AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1669–1685, 2023.

[20] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech 2020*, 2020, pp. 3291–3295.

[21] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "VoiceFixer: Toward general speech restoration with neural vocoder," 2021.

[22] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "HiFi++: a unified framework for bandwidth extension and speech enhancement," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[23] S. Ogun, V. Colotte, and E. Vincent, "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.

[24] J. Wiseman, "py-webrtcvad," https://github.com/wiseman/py-webrtcvad, 2016.

[25] Y. Zhang, E. Bakhturina, and B. Ginsburg, "NeMo (Inverse) Text Normalization: From Development to Production," in *Proc. Interspeech*, 2021, pp. 4857–4859.

[26] Eren Gölge, "Xtts v1 — techincal notes," https://medium.com/@erogol/xtts-v1-techincal-notes-eb83ff05bdc, 2023, [Online; accessed 1-March-2024].

[27] C. Jemine, "Resemblyzer," https://github.com/resemble-ai/Resemblyzer, 2019.

[28] T. Yoshimura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. of Eurospeech*, 1997, pp. 2523–2526.

[29] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *Proc. Interspeech*, 2020.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[31] S. Vaibhav, F. Clémentine *et al.*, "TTS Arena: Benchmarking text-to-speech models in the wild," https://huggingface.co/blog/arena-tts, 2024, [Online; accessed 1-March-2024].

[32] A. Owodunni, A. Yadavalli, C. Emezue, T. Olatunji, and C. Mbataku, "AccentFold: A journey through African accents for zero-shot ASR adaptation to target accents," in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2146–2161. [Online]. Available: https://aclanthology.org/2024.findings-eacl.142