Interspeech 2022
18-22 September 2022, Incheon, Korea

BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus

Josh Meyer[1], *David Ifeoluwa Adelani*[2], *Edresson Casanova*[1,3], *Alp Öktem*[4,5], *Daniel Whitenack*[6] *Julian Weber*[1], *Salomon Kabongo*[7], *Elizabeth Salesky*[8], *Iroro Orife*[9], *Colin Leong, Perez Ogayo*[10], *Chris Emezue*[11], *Jonathan Mukiibi* [12], *Salomey Osei, Apelete Agbolo*[13], *Victor Akinode, Bernard Opoku*[14], *Samuel Olanrewaju, Jesujoba Alabi*[2], *Shamsuddeen Muhammad*

[∀] Masakhane NLP, Africa [1] Coqui, USA [2] Saarland University, Germany [3] University of São Paulo, Brazil [4] CLEAR Global, USA [5] Col·lectivaT, Spain [6] SIL International [7] Leibniz Universität Hannover, Germany [8] Johns Hopkins University, USA [9] Niger-Volta LTI, USA, [10] Carnegie Mellon University, USA [11] Technical University of Munich, Germany [12] Makerere University, Uganda [13] Ewegbe Akademi, Togo, [14] Kwame Nkrumah University of Science and Technology, Ghana

josh@coqui.ai, didelani@lsv.uni-saarland.de, edresson@coqui.ai, alp.oktem@clearglobal.org, dan_whitenack@sil.org

Abstract

BibleTTS is a large, high-quality, open speech dataset for ten languages spoken in Sub-Saharan Africa. The corpus contains up to 86 hours of aligned, studio quality 48kHz single speaker recordings per language, enabling the development of high-quality text-to-speech models. The ten languages represented are: Akuapem Twi, Asante Twi, Chichewa, Ewe, Hausa, Kikuyu, Lingala, Luganda, Luo, and Yoruba. This corpus is a derivative work of Bible recordings made and released by the Open.Bible project from Biblica. We have aligned, cleaned, and filtered the original recordings, and additionally hand-checked a subset of the alignments for each language. We present results for text-to-speech models with Coqui TTS. The data is released under a commercial-friendly CC-BY-SA license.

Index Terms: TTS, text-to-speech, speech synthesis, speech corpora, speech data

1. Introduction

The majority of the world's approximately 7,000 languages [1] do not have open speech datasets, and even fewer have high-quality data with aligned text and speech, which can be used for training text-to-speech (TTS) models. The creation of benchmark datasets such as Librispeech [2], LibriTTS [3], and LJSpeech [4] enabled significant advances through community development on common resources, but these resources cover few languages, and most TTS systems evaluate on English only.

Speech synthesis systems have received significant attention in recent years due to the advances provided by deep learning. These advances enable TTS models to achieve improved naturalness with respect to human speech [5, 6, 7], and improved synthesized speech as driven adoption of virtual assistants [8, 9]. However, neural models often require a non-trivial amount of data for training. This necessity leaves many language communities under-served in the development of speech technologies [10], and it further results in researchers not evaluating models on diverse linguistic phenomena.

In this work, we present the *BibleTTS* corpus, a high-quality aligned speech corpus for ten African languages. This data enables further research and resource creation for these languages and will allow researchers to create meaningful benchmarks against non-English languages.

Creating high-quality aligned datasets typically requires tools not available for most languages, hindering the creation of datasets for lower-resourced languages. Specifically, forced alignment of speech and text typically requires pre-trained acoustic models and grapheme-to-phoneme (G2P) models. This process can be challenging and error-prone without high-quality resources. We demonstrate that it is possible to force-align data without access to any pre-trained models (acoustic or G2P), and still produce quality output.

Additionally, recent corpora that significantly expand linguistic coverage for TTS datasets are often not freely available [11, 12], contain less single-speaker data [13], and/or have lower-quality recordings. BibleTTS stands out in this regard as it is a large, high-fidelity corpus made of single-speaker recordings. The corpus is released under an open CC-BY-SA license. Corpus links and samples created with our TTS models can be accessed from the project website[1].

2. Related Work

We focus on related work for African languages in the following section. Existing publicly available datasets are typically small. For Yorùbá these include a 2.75 hour corpus [14, 15] and a 4 hour multi-speaker dataset [16]. TWB Gamayun kits [17] include a 6-hour single speaker high-quality Swahili speech corpus optimized for TTS training. Earlier Yorùbá TTS efforts typically used bespoke private data [18, 19, 20, 21, 22]. For isiXhosa, Sesotho, Setswana and Afrikaans, multi-speaker corpora of approximately 2 hours each have been developed for TTS [23, 24]. The CMU Wilderness dataset [11] includes up to 20 hours of high-quality, single-speaker data for several African languages, but it is not publicly available and the alignments can contain noise. TTS systems research for African languages has comprised development efforts in frameworks like Festival [25] or MaryTTS [26] for Yorùbá [27, 28], Ibibio [29], Amharic [30], Fon [31], isiZulu [32], KiSwahili [33]. While many of these systems used concatenative synthesis, in large part because the available corpora were small, there have also been investigations into statistical parametric speech synthesis for Ibibio [34]. Finally, there have been efforts in related tasks,

[1] https://masakhane-io.github.io/bibleTTS/

10.21437/Interspeech.2022-10850

Table 1: *Language, classification and statistics. All language classifications and numbers of speakers are from Ethnologue.*

Language	Classification	African Region	No. of speakers
Éwé [ewe]	Niger-Congo / Kwa	West	5.5M
Hausa [hau]	Afro-Asiatic / Chadic	West	77M
Kikuyu [kik]	Niger-Congo / Bantu	East	8.2M
Lingala [lin]	Niger-Congo / Bantu	Central	40M
Luganda [lug]	Niger-Congo / Bantu	East	11M
Luo [luo]	Nilo-Saharan / Luo–Acholi	East	5.3M
Chichewa [nya]	Niger-Congo / Bantu	South-East	14M
Akuapem Twi [aka]	Niger-Congo / Akan	West	626k
Asante Twi [aka]	Niger-Congo / Akan	West	3.8M
Yorùbá [yor]	Niger-Congo / Volta-Niger	West	46M

such as grapheme-to-phoneme research for Yorùbá [35], intonation modeling [21], and numeral preprocessing [36].

3. Languages represented

Table 1 shows the languages in the BibleTTS corpus, with their language families, the number of speakers[1] and the regions in Africa where they are spoken. The corpus consists of ten languages from the three largest language families in Africa (Niger-Congo, Afro-Asiatic and Nilo-Saharan) and four regions of Africa. All of these languages are tonal and are spoken primarily in sub-Saharan Africa.

3.1. Language Characteristics[2]

Éwé [ewe] uses 35 Latin letters excluding (c, j, q), with 12 additional letters (ɖ, dz, ɛ, *f*, gb, ɣ, kp, ny, ŋ, ɔ, ts, ʋ). Ewe has three tones, and they are marked in text.

Hausa [hau] uses two different writing scripts: Ajami and Boko. The Boko script is the most widely used and is based on the Latin alphabet with 44 letters. The alphabet excludes letters (p, q, v and x) and uses 12 additional letters: ɓ, ɗ, ƙ, ʸ, kw, ƙw, gw, ky, ƙy, gy, sh, ts. Hausa is tonal, but tones are not represented in text.

Kikuyu [kik] uses Latin script with 27 letters excluding (f, l, p, s, v, x, y, z), and including additional nine letters (ĩ, ũ, mb, nd, nj, ng, ng‘,ny, th). Kikuyu uses two tones (high and low) but they are not marked in text.

Lingala [lin] uses the Latin script with 40 letters excluding (j, q, x) and including an additional 17 letters (ɛ, gb, kp, mb, mf, mp, mv, nd, ng, ngb, nk, ns, nt, ny, nz, ɔ, ts). Lingala uses two tones (high and low), but they are not marked in text.

Luganda [lug] uses 24 Latin letters excluding (h, q, x), and including additional two letters (ŋ, ny). Luganda uses three tones, but they are not marked in text.

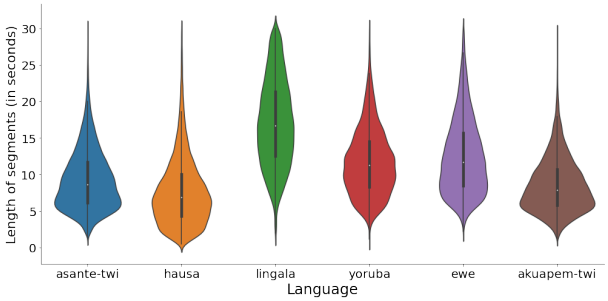
Luo [luo] or Dholuo uses Latin script with 31 letters excluding the letters (c, q, x, v, z), and additional letters (ch, dh, mb, nd, ng’, ng, ny, nj, th, sh). Luo has four tones, but they are not marked in text.

Chichewa [nya] uses the Latin script with 31 letters excluding (q, x, y), and including additional eight letters (ch, kh, ng, ŋ, ph, tch, th, ŵ). Chichewa uses two tones (high and low) but they are not marked in text.

Akan [aka] is a language with multiple dialects (including Fante, Bono, Asante, and Akuapem), and they are collectively known as Twi. In this study, we focus on Asante and Akuapem which are mutually intelligible and share the same alphabets (referred to herein as `aka-Asante` and `aka-Akuapem`). Twi uses 22 Latin letters excluding (c,j,q,v,x,z), and including two additional letters (ɛ, ɔ).

Yorùbá [yor] uses 25 Latin letters without the letters (c, q, v, x and z) and with additional letters (ẹ, gb, ṣ, ọ). Yorùbá is a tonal language with three tones: low, middle, and high. These tones are represented by the grave (e.g. “è ”), optional macron (e.g. “ē”) and acute (e.g. “é”) accents respectively but the mid tone is usually ignored in writing.

Figure 1: *Distribution of the sample length per language. Samples longer than 30s and with fewer than 10 characters were removed, and outlier segments were detected and discarded as described in Section 4.2. Lingala is a slight outlier with the majority of segments between 10 and 20 seconds, while the other five languages have segments centered at 5-10s each.*



4. Corpus creation

The BibleTTS corpus consists of high-quality audio released as 48kHz, 24-bit, mono-channel FLAC files. Recordings for each language are under professional quality, close-microphone conditions (i.e., without background noise or echo). BibleTTS is unique among open speech corpora for the volume of data per speaker and suitability for TTS. The corpus consists of ten languages which are under-represented in today’s voice technology landscape, both in academia and in industry. We release train/dev/test splits for each language, where `dev` is the Book of Ezra, `test` is Colossians, and `train` is all other books.

4.1. Alignment

The BibleTTS corpus contains audio recordings and text transcripts (i.e. “Open Contemporary Bible” translations) which were released by Biblica via the Open.Bible project.[3] The original audio recordings were 48kHz, mono-channel WAV, typically one recording per chapter of the Bible. Each chapter was up to 30 minutes long, which is too long for most modeling tasks. Verses are a natural alternative, as the text already contains verse boundaries. Aligning at the verse level creates more manageable recording lengths of up to 30 seconds (see Figure 1) which are more likely to be consistent across languages than segmentation on voice activity detection or other alternatives.

Potential challenges in alignment include additional content in either the speech or text, such as spoken titles and headings or text annotations, and the availability of pre-trained acoustic models and grapheme-to-phoneme mappings. We have employed various alignment techniques depending on the availability of verse timestamps and resources in each language, and evaluated a subset of the alignments with native speakers.

[2] `https://omniglot.com/writing/<language-name>.htm`

[3] `https://open.bible/resources`

4.1.1. Verse timestamps

Three languages (aka-Akuapem, aka-Asante, and lin) were straightforwardly segmented using verse-level timestamps released by the Open.Bible project. The timestamps show the start time of every verse, as well as when the book and chapter titles were spoken. With these timestamps, verses were isolated and saved as individual audio files using sox. These alignment scripts can be found on Github at coqui-ai/open-bible-scripts. [4]

4.1.2. Forced alignment using pre-trained acoustic models

Forced alignment is the process of extracting timestamps given an audio and a transcript pair, and requires either a pre-trained acoustic model or training one from scratch. For Hausa (hau), we opted to use the Montreal Forced Aligner (MFA) [37] for which there is a pre-trained Hausa model [38]. The code is open-sourced on Github[5]. The process is as follows:

1. Audio of each chapter of each book is downloaded together with their script in the form of an XML file,
2. XML script is parsed and converted into a plain normalized text file. Normalization entails: (a) adding the chapter title at the beginning of the script as "Sura <chapter-no>", (b) converting numbers into written form using a dictionary prepared with Hausa linguists, and (c) adding a new line after every sentence ending punctuation mark (e.g. .?!").
3. A grapheme-to-phoneme (G2P) dictionary is created from the word list extracted from the transcripts using the Hausa G2P model.
4. Alignment is performed for each chapter using the audio and normalized script with a beam length of 1000.
5. The time-aligned TextGrid file is processed in parallel with the sentence-segmented transcript to partition the chapter audio into sentence-level audio chunks with their transcriptions.

4.1.3. Forced alignment from scratch

Two languages (ewe and yor) were aligned via forced alignment from scratch. Using only the found audio and transcripts (i.e., without a pre-trained acoustic model), an acoustic model was trained and the data aligned with the Montreal Forced Aligner. Graphemes were used as a proxy for phonemes in place of G2P data. The code used to generate alignments can be found in the coqui-ai/open-bible-scripts repository.[6]

After forced alignment, we used regular expressions to pull out whole verses which were aligned such that silence occurred both at the beginning and the end of a verse. Segmenting out audio at the verse-level instead of splitting on silence may allow downstream TTS models to capture higher-level prosody.

4.2. Outlier detection

Following the alignment stage, we detected and removed outliers using the data-checker toolkit together with human judgments. The relevant code is open-sourced on Github at coqui-ai/data-checker.[7] First, all segments longer than 30 seconds, or less than 10 characters in the aligned transcript, were removed. Then, the removal of outliers was performed and fine-tuned for each language independently until the major offending samples[8] were no longer encountered, as described below.

Every pair of <audio,transcript> was assigned an "outlier score", and the most extreme outliers were removed. First, the ratio of transcript length (characters) to audio length (seconds) was calculated for each sample. Then a Gaussian distribution was estimated for all samples in a given language. Lastly, the number of standard deviations from the mean was calculated for each sample. Outliers were excluded if they existed more than N standard deviations away from the mean, where N was fine-tuned per language with an iterative human-in-the-loop approach, until minimal offending samples were encountered. For most languages, it was sufficient to exclude samples more than 3 standard deviations from the mean (or .2% of the data). However, yor notably required more outliers removed to attain a quality dataset. The resulting distribution of segment lengths per language is shown in Figure 1.

4.3. Human evaluation of alignment quality

We facilitated human evaluation of both the alignment and the output of the TTS models. In total, we collected labels from 15 annotators (three per language) for ewe, hau, lin, aka-Asante, and aka-Akuapem and an additional five annotators for yor. To judge the quality of <audio,transcript> pairs from our alignments, we randomly sampled 50 example pairings of aligned transcripts and the corresponding audio clips across the train, dev, and test sets. Annotators selected the one option that best described the quality of the alignment:

1. Audio contains EXTRA words not in the transcript
2. Audio is MISSING words that are in the transcript
3. Audio is MISSING words AND includes EXTRA words
4. No missing or extra words

In cases where the labels corresponding to various annotators disagreed, we took the majority vote label. In cases where the number of labels was spread evenly among different choices, we noted these as "conflicting." Results of human evaluation are shown in Table 3.

As discussed in Section 4.1, some languages (aka-Asante, aka-Akuapem, lin) were segmented using existing verse-level timestamp files. Interestingly, annotators labeled these languages as having a high percentage of samples where the audio contains additional words not present in the aligned text. Aligning from scratch (ewe, yor) produced a greater proportion of segments with exact matches between speech and text than using forced alignment with a pre-trained acoustic model (hau). However, it should be noted that significantly more data was removed due to outliers for yor, and less data overall was aligned with ewe, yor (see the statistics for unaligned vs. aligned hours in Table 2).

5. TTS Models

To experimentally validate the quality of our dataset, we train the VITS end-to-end speech synthesis model [7] with the sampling rate 22050 Hz in the six aligned languages. The chosen languages are Akuapem Twi, Asante Twi, Éwé, Hausa, Lingala, and Yorùbá. We chose VITS for its state-of-the-art naturalness

[4] https://github.com/coqui-ai/open-bible-scripts
[5] https://github.com/alpoktem/bible2speechDB
[6] https://github.com/coqui-ai/open-bible-scripts
[7] https://github.com/coqui-ai/data-checker
[8] "Major offending samples" was not explicitly defined, but refers to samples labeled by a non-native speaker of these languages as containing obvious mismatches between transcripts and speech.

Table 2: *Corpus statistics. The corpus consists of data for ten languages, of which six have been aligned and formatted for immediate use to train TTS models.*

Language	Unaligned Hours	Unaligned Samples	Aligned Hours	Aligned Samples
Éwé	100.1	1,167	86.8	24,957
Hausa	103.2	1,189	86.6	40,603
Kikuyu	90.6	1,189	–	–
Lingala	151.7	1,189	71.6	15,117
Luganda	110.4	1,189	–	–
Luo	80.4	1,189	–	–
Chichewa	115.9	1,162	–	–
Akuapem Twi	75.7	1,189	67.1	28,238
Asante Twi	82.6	1,189	74.9	29,021
Yorùbá	93.6	1,189	33.3	10,228

and also for its robust alignment mechanism [39]. The model takes characters as input and does not require a phonemizer.

To accelerate training we use transfer learning. We start from a model pre-trained on LJSpeech [40] for 1M steps, which is available via the Coqui TTS repository[9]. We continue training for approximately 110K steps for each one of the languages. We use the AdamW optimizer [41] with betas 0.8 and 0.99, weight decay 0.01, and an initial learning rate of 0.0002 decaying exponentially by a gamma of 0.999875 [42]. The models were trained using an NVIDIA A100 SXM4 80GB with a batch size of 100. All models are released in the Coqui TTS toolkit[10].

5.1. Results and Discussion

We evaluated the synthesized speech using subjective judgments, averaged across multiple speakers. We randomly sampled 50 segments from the in-domain test set, as well as out-of-domain corpora to test the models' ability to generalize to non-Bible contexts. The out-of-domain sentences were obtained from the NEWS corpus[11] (except for Akuapem Twi). Annotators rated the quality of synthesized speech in terms of naturalness of voice and appropriateness of pronunciation for the particular language or dialect. Annotators selected from a 5-point Likert rating for each sample: 1 (bad), 2 (poor), 3 (fair), 4 (good), and 5 (excellent). The mean opinion scores (MOS) are shown in Table 4. We additionally use mel cepstral distortion (MCD) [43], an automatic edit distance metric, to assess quality for the in-domain segments where we have reference speech, with dynamic time warping (DTW) to align the segments. MCD largely follows MOS: languages with better human judgments (higher MOS) typically have better (lower) MCD scores, though MCD can be misleading, as in Lingala.

The MOS judgments seem related to the goodness of alignment evaluations. That is, the language with the best alignments (`ewe`) was also rated the best MOS for speech synthesized from the resulting model. Similarly, the language with the worst alignments (`hau`) resulted in the TTS model with the lowest out-of-domain MOS scores. To improve MOS, it may be necessary to either improve the alignments or apply more stringent outlier exclusion criteria, which should be possible, as the training data size remains significantly larger than many available TTS corpora for these languages or others.

[9] `https://github.com/coqui-ai/TTS`
[10] `https://github.com/coqui-ai/TTS`
[11] `http://github.com/masakhane-io/lacuna_pos_ner`

Table 3: *Human evaluation of alignment. Shown are percentages of <audio,transcript> samples with an exact match (EM), added words, missing words, or both.*

Language	EM	Add.	Miss.	Both	Conflict. Labels
Éwé	92%	2%	4%	0%	2%
Hausa	32%	68%	0%	0%	0%
Lingala	74%	12%	0%	0%	14%
Akuapem Twi	88%	0%	8%	2%	2%
Asante Twi	78%	2%	6%	6%	8%
Yorùbá	76%	24%	0%	0%	0%

Table 4: *Evaluation of TTS model outputs using both human judgments (MOS) and an automatic metric (MCD). In-Domain texts are Bible verses, and Out-of-Domain is news.*

Language	MOS In-Domain	MOS Out-of-domain	MCD
Éwé	4.34	3.87	5.8
Hausa	3.42	2.34	7.6
Lingala	3.31	2.40	5.6
Akuapem Twi	2.79	—	7.5
Asante Twi	3.07	2.44	6.8
Yorùbá	4.06	2.93	5.8

6. Conclusions

The BibleTTS corpus is the first of its kind in many respects. The quality and volume of the data is extremely rare in open speech corpora – these are professional, studio quality, 48kHz recordings, with up to 86 hours of verse-aligned data per language, for 10 languages spoken in sub-Saharan Africa. The BibleTTS license is research and commercial friendly: CC-BY-SA. We hope that this corpus will enable advances in speech technology for African languages and also will unlock new techniques in TTS, which require more, higher-quality data.

We described our approach to verse and sentence-level alignment of the original found data with a variety of different resources. We used human evaluation to assess the quality of the resulting alignments, and validate the resulting data by training high-quality speech synthesis models with Coqui TTS.

There are two clear and immediate avenues for future work: (1) verse-level alignment of the remaining four languages (`kik`, `lug`, `luo`, and `nya`), and (2) improvement of the quality of existing alignments. Given the volume of data per language, it may well be the case that we can be more conservative with outlier removal, keeping only 20 or 30 hours of the best data, and obtain even better resulting TTS models. Nevertheless, we have shown that the data can already be used to produce high-quality TTS models (as with Ewe), on both in and out of domain text. We plan to update BibleTTS such that we have high-quality verse-level alignments for all ten languages.

7. Acknowledgements

We are very grateful to the volunteers, Richard J. Bonnie, Komlanvi D. Akoly, Komlanvi M. Klove, Ibrahim Haruna, Oluwabusayo O. Awoyomi, Emmanuel Anebi, Christian Kilapi, Pacifick Taba, who helped with human evaluation and the Masakhane community.

8. References

[1] D. M. Eberhard, G. F. Simons, and C. D. F. (eds.), "Ethnologue: Languages of the world. twenty-third edition." SIL International, 2022. [Online]. Available: http://www.ethnologue.com

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015.

[3] 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.

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

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018.

[6] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," *arXiv preprint arXiv:2005.05957*, 2020.

[7] 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.

[8] A. Purington, J. G. Taft, S. Sannon, N. N. Bazarova, and S. H. Taylor, "" alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017.

[9] P. Dempsey, "The teardown: Google home personal assistant," *Engineering & Technology*, 2017.

[10] E. Casanova, A. C. Junior, C. Shulby, F. S. d. Oliveira, J. P. Teixeira, M. A. Ponti, and S. Aluísio, "Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese," *Language Resources and Evaluation*, 2022.

[11] A. W. Black, "Cmu wilderness multilingual speech dataset," in *ICASSP*. IEEE, 2019.

[12] M. Harper, "The IARPA Babel multilingual speech database," 2011. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/babel

[13] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "Multilingual tedx corpus for speech recognition and translation," in *Proceedings of Interspeech*, 2021.

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

[15] D. R. van Niekerk and E. Barnard, "Tone realisation in a yorùbá speech recognition corpus," in *SLTU*, 2012.

[16] A. Gutkin, I. Demirsahin, O. Kjartansson, C. E. Rivera, and K. Túbòsún, "Developing an open-source corpus of yoruba speech," 2020.

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

[18] T. K. Dagba, J. O. R. Aoga, and C. C. Fanou, "Design of a yoruba language speech corpus for the purposes of text-to-speech (tts) synthesis," in *ACIIDS*, 2016.

[19] A. A. Afolabi, E. O. Omidiora, and O. T. Arulogun, "Development of text to speech system for yoruba language," 2014.

[20] A. E. Akinwonmi and B. K. Alese, "A prosodic text-to-speech system for yorùbá language," *ICITST*, 2013.

[21] O. O. Àjàdí, "A quantitative model of yorùbá speech intonation using stem-ml," *INFOCOMP Journal of Computer Science*, 2007.

[22] O. À. Odéjobí, A. J. Beaumont, and S. H. S. Wong, "A computational model of intonation for yorùbá text-to-speech synthesis: Design and analysis," in *TSD*, 2004.

[23] 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.

[24] E. Barnard, M. H. Davel, C. van Heerden, F. De Wet, and J. Badenhorst, "The nchlt speech corpus of the south african languages." Workshop Spoken Language Technologies for Under-resourced Languages (SLTU), 2014.

[25] A. W. Black, P. A. Taylor, and R. Caley, "Festival speech synthesis system," 1998.

[26] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, "Open source voice creation toolkit for the mary tts platform," in *INTERSPEECH*, 2011.

[27] A. R. Iyanda and O. D. Ninan, "Development of a yorúbà texttto-speech system using festival," *Innovative Systems Design and Engineering (ISDE)*, vol. 8, no. 5, 2017.

[28] J. O. Aoga, T. K. Dagba, and C. C. Fanou, "Integration of yoruba language into marytts," *International Journal of Speech Technology*, 2016.

[29] M. E. Ekpenyong, E.-A. Urua, and D. Gibbon, "Towards an unrestricted domain tts system for african tone languages," *International Journal of Speech Technology*, 2008.

[30] S. H. Mariam, K. Prahallad, A. W. Black, R. Kumar, and R. Sangal, "Unit selection voice for amharic using festvox," in *SSW*, 2004.

[31] T. K. Dagba and C. Y. Boco, "A text to speech system for fon language using multisyn algorithm," in *KES*, 2014.

[32] J. A. Louw, M. H. Davel, and E. Barnard, "A general-purpose isizulu speech synthesizer," *South African Journal of African Languages*, 2005.

[33] M. Gakuru, F. K. Iraki, R. C. F. Tucker, K. B. Shalonova, and K. Ngugi, "Development of a kiswahili text to speech system," in *INTERSPEECH*, 2005.

[34] M. E. Ekpenyong, E.-A. Urua, O. Watts, S. King, and J. Yamagishi, "Statistical parametric speech synthesis for ibibio," *Speech Communication*, 2014.

[35] A. R. Ìyàndá, O. A. Odejobi, F. A. Soyoye, and O. O. Akinadé, "Development of grapheme-to-phoneme conversion system for yorùbá text-to-speech synthesis," 2014.

[36] O. O. Akinadé and d. A. Ọdéjọbí, "Computational modelling of yorùbá numerals in a number-to-text conversion system," *Journal of Language Modelling*, 2014.

[37] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, Stockholm, Sweden, 2017.

[38] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text amp; speech database in 20 languages," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[39] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, 2020.

[40] K. Ito *et al.*, "The lj speech dataset," 2017.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, 2019.

[43] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Spoken Languages Technologies for Under-Resourced Languages*, 2008.