ÌròyìnSpeech: A multi-purpose Yorùbá Speech Corpus

Tolúlọpẹ́ Ógúnrẹ̀mí [1], **Kọ́lá Túbọ̀sún**[2], **Anuoluwapo Aremu**[2], **Iroro Orife**[3], **David Ifeoluwa Adelani**[4]
[1]Stanford University, [2]Yorùbá Names, [3]Niger-Volta LTI, [4]University College London
tolulope@cs.stanford.edu, project@yorubaname.com, iroro@alumni.cmu.edu, d.adelani@ucl.ac.uk

Abstract

We introduce ÌròyìnSpeech, a new corpus influenced by the desire to increase the amount of high quality, contemporary Yorùbá speech data, which can be used for both Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) tasks. We curated about 23 000 text sentences from news and creative writing domains with the open license CC-BY-4.0. To encourage a participatory approach to data creation, we provide 5 000 curated sentences to the Mozilla Common Voice platform to crowd-source the recording and validation of Yorùbá speech data. In total, we created about 42 hours of speech data recorded by 80 volunteers in-house, and 6 hours of validated recordings on Mozilla Common Voice platform. Our TTS evaluation suggests that a high-fidelity, general domain, single-speaker Yorùbá voice is possible with as little as 5 hours of speech. Similarly, for ASR we obtained a baseline word error rate (WER) of 23.8.

Keywords: Yorùbá language, Automatic Speech Recognition, Speech Synthesis

1. Introduction

Speakers of many African languages have no access to voice-enabled applications in their native languages. One reason is that these technologies all require models for speech recognition and speech synthesis, trained on large datasets of high-fidelity speech and text (Ritchie et al., 2022; Meyer et al., 2022b).

To address this challenge, there have been several efforts to build large-scale multilingual datasets and models by automatically aligning speech and text (Radford et al., 2023; Zhang et al., 2023; Pratap et al., 2023). The results however, are often of poor quality for low-resource African languages due to the dearth of high-quality audio-text pairs and unsatisfactory out-of-domain generalization. Other efforts have focused on building benchmark datasets for over 100 languages using high-quality but small-scale training data (Conneau et al., 2023; Shi et al., 2023a,b).

In this paper, we focus on Yorùbá, a West African language with over 40 million L1 speakers, yet under-represented in contemporary speech research. There have been a number of efforts to build datasets for Yorùbá speech tasks (Odéjobí et al., 2004; Àjàdí, 2007; Akinwonmi and Alese, 2013; Afolabi et al., 2014; Dagba et al., 2016; van Niekerk and Barnard, 2012; van Niekerk et al., 2015; Gutkin et al., 2020a). The datasets are either too small to do speech processing effectively or are single speaker, single domain, as is the case for BibleTTS (Meyer et al., 2022b). Our corpus extends the scope of previous work to address multiple speech application domains.

We introduce the ÌròyìnSpeech— a new dataset created to increase the amount of high quality, contemporary Yorùbá speech. The dataset has a total of 42 hours of audio, recorded by 80 volunteers. We curated text sentences from the news and creative writing domains under an open license, CC-BY-4.0. We also provide 5 000 sentences to the Common Voice (Ardila et al., 2020) platform to crowd-source voice recordings online[1]. We provide extensive baseline experiments using state-of-the-art approaches for TTS and ASR. The code and data will be made freely available on https://github.com/Niger-Volta-LTI/yoruba-voice.

2. The Yorùbá Language

The Yorùbá language is native to south-western Nigeria, Republic of Benin, and Republic of Togo. It is one of the national languages of Nigeria also spoken in other countries like Ghana, Côte d'Ivoire, Sierra Leone, Cuba and Brazil. The language belongs to the Niger-Congo family in the Volta-Niger sub-group, and is spoken by over 40 million native speakers (Eberhard et al., 2019), making it one of the most widely spoken African languages.

Yorùbá has 25 letters without the Latin characters (c, q, v, x and z) and with additional characters (ẹ, gb, ṣ, ọ). There are 18 consonants, seven oral vowels (a, e, ẹ, i, o, ọ, u), five nasal vowels, (an, ẹn, in, ọn, un) and syllabic nasals (m̀, ḿ, ǹ, ń). Yorùbá is a tonal language with three tones: low, middle and high. These tones are represented by the grave ("\"), optional macron ("−") and acute ("/") accents respectively. These tones are applied on vowels and syllabic nasals, but the mid tone is usually

[1]https://commonvoice.mozilla.org/yo

LREC-COLING 2024, pages 9296–9303

ignored in standard Yorùbá orthography. These *diacritics* are important for correct pronunciation and lexical disambiguation.

3. The ÌròyìnSpeech Corpus

3.1. Preparation of text sentences

In contrast to other Yorùbá datasets based on Biblical or religious texts, our goal was to combine news data and fictional texts to create a modern, multi-purpose speech dataset (Gutkin et al., 2020b; Meyer et al., 2022a). The corpus text was obtained from two sources, firstly the MENYO-20k dataset (Adelani et al., 2021), an open-source, multi-domain English-Yorùbá machine translation corpus and secondly, the Yorùbá portion of the MasakhaNER 2.0 dataset (Adelani et al., 2022) (i.e MasakhaNER-YOR) based on the Asejere newspaper[2]. The primary source of the MENYO-20k dataset is the Voice of Nigeria newspaper[3], published by the Nigerian government. We restrict our selection of corpus text to the above published datasets for two reasons (1) they have a non-restrictive license, and (2) the Yorùbá sentences have been further verified for quality issues, for example missing diacritics in the original crawled Asejere and Voice of Nigeria articles. Overall, we obtained 3 048 sentences from Voice of Nigeria, 2 932 sentences from Global Voices, and 5 135 sentences from Asejere. In total, this gives us 11 115 sentences.

In order to obtain more sentences to reach our goal of 40 hours of speech, we added sentences extracted and modified from unpublished short stories previously translated into Yorùbá by the second author. These texts were selected to broaden the domain of the vocabulary used in the dataset. In addition, we split-up long sentences and asked volunteers to manually generate new sentences with similar themes or context as the original seed sentences. They also cross-checked each sentence for errors. In total, we had to manually generate about 12 000 sentences. We then cleaned up the data to create a final script. To ensure the sentences were of high-quality, we verified that diacritics were properly applied on each word and revised offensive or divisive religious terms within the text to reflect a neutral tone. Next, we modified the text for clarity and length to facilitate pronunciation, and localized non-Yorùbá words into Yorùbá. We list below a few examples of the types of names and places which were localized: Kaduna to Ọ̀yọ́, Zamfara to Oǹdó, United States to Ìlú Ọba, Buhari to Bùhárí and Kenya to Kẹ́ńyà.

[2] https://www.asejere.net/
[3] https://yoruba.von.gov.ng/

3.2. Recording of text sentences

3.2.1. Corpus partitions

Table 1 provides the details of the recorded utterances and text preparation. Our text preparation yielded a total of 23 000 sentences which were used to record audio of both ASR and TTS. Our initial setup divided the corpus into two parts: (1) **Partition A**, contains 20 000 sentences, primarily for the recording of ASR audio and (2) **Partition B**, contains 3 000 reserved for TTS recordings.

ASR recording We recorded in-house all 20 000 sentences in Partition A, or some 26 hours for ASR. They were recorded by 80 different volunteers, each recording 250 lines during one-hour studio sessions. Additionally, we added 5 000 sentences to the Mozilla Common Voice crowdsourcing platform, which were recorded by 108 volunteers on the Common Voice website, yielding 6 hours of speech. The sentences used to record on Common Voice were selected from the sentences in Partition A. We note for completeness, that the ASR experiments in this study did not *yet* make use of these 6 hours. We hope that by setting up Common Voice for Yorùbá, native speakers everywhere will be encouraged contribute.

TTS recording We recorded all 3 000 sentences in Partition B by two speakers, one male and one female, ages 25 to 30. This resulted in 3 hours 35 minutes of data, which was below our goal of 10 hours. Since there were no additional curated sentences, we decided to record supplementary sentences obtained from Partition A to reach our 10 hour goal. The final TTS corpus contains 9 000 sentences yielding 10 hours 11 minutes of speech.

All volunteers, speakers of standard North West Yorùbá[4], were screened for dialect uniformity, and ranged in age from 18 to 69 years. The 9 000 lines for the single speaker (TTS) partition had one male and one female volunteer, while the 20 000 lines multi-speaker (ASR) partition had 80 volunteers, 37 male and 43 female. The studio volunteers were each provided with a token gift, as a gesture of appreciation of their time and efforts recording.

3.2.2. Recording

To create an acoustically suitable environment for recording, we obtained a portable vocal booth. The recording equipment comprised an AT 2020+ USB microphone, USB cables, and a 2022 M1-Series Macbook Pro.

[4] http://www.africa.uga.edu/Yoruba/yorubabout.html

	# hours	# utterances	Corpus partitions used	# Unique sentences used
In-house ASR	26h 00m	20 000	PARTITION A	20 000
Common Voice ASR	6h 00m	5 000	PARTITION A	-
In-house TTS	10h 11m	9 000	PARTITION A & PARTITION B	3 000
Total	42h 11m	34 000	-	23 000

Table 1: A summary of dataset statistics. Some of the utterances used for TTS recording (i.e. 6K utterances) and for Common Voice (5K utterances) are subsets of the PARTITION A.

The first five hours of audio were recorded with Audacity, a free, open-source digital audio editor and recording application. To divide each of these hour-long recordings into a short file for each sentence required many more additional hours of manual post-editing work. To solve this problem, the team developed a custom application for creating speech corpora, dubbed *Yorùbá Voice SpeechRecorder*[5] (Orife et al., 2022).

The app works by reading a prepared text file, usually with 250 sentences and displays each line of text to be read in order. It also provides transport controls to enable recording, playback and file-management or deletion, in the case of multiple takes. Finally, the tool saves individual audio files to the hard-drive, for each sentence, and updates a metadata index, which can be used to programmatically prepare training examples. Over 65% of the total lines recorded in-house were recorded using the SpeechRecorder app.

3.2.3. Post-production

We had four forms of post-processing. Where possible, recordings that had issues which could be manually fixed, were repaired by removing simple audio artifacts and speech disfluencies. In situations where the recording did not correspond with the text but the utterance remained grammatical, we did not rerecord the utterance but rather edited the text sentence to match the audio.

We also fixed tone marking, spelling, or semantic mismatches. Words like “ní ilé” (into the house) or “sí ibẹ̀” (to there) are often contracted to “nílé” and “síbẹ̀” respectively in spoken Yorùbá and were amended in the text sentence accordingly.

If the audio files had any issues which rendered them unusable, then we re-recorded, usually with a different volunteer of the same gender, introducing thusly, a new, different speaker ID. Some of the issues encountered include: (1) *Disfluencies*: hesitations, stammers, clicks, sniffs, etc. (2) *External noises*: paper rustling, microphone touching, intrusive voices, electronic notification beeps, etc (3) *Audio fidelity*: low or uneven audio levels, clipping or distortion, or otherwise unintelligible words (4) *Incorrect dictation* which could not be fixed by changing the script.

4. Speech Synthesis Experiments

We train speech synthesis models using the single-speaker TTS partition of our dataset, resulting in both male and female voices. The 9 000 utterances described in Table 1, are split evenly between the male and female speakers, employing 4 500 utterances for each. We train and evaluate three variants of the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) model[6] as follows:

- Domain adaptation from existing BibleTTS Yorùbá checkpoints (Meyer et al., 2022b).
- Training a VITS model end-to-end from scratch with 5 hours of data
- Training VITS models from scratch without diacritics in the training text
- Training a VITS model end-to-end from scratch with different data scales i.e. number of utterances: 100, 500, 1 000, 2 000 and 4 500.

Model training We trained the VITS models using the Coqui TTS toolkit (Meyer et al., 2022b). We use the AdamW optimizer (Loshchilov and Hutter, 2019) with betas {0.8, 0.99}, weight decay of 0.01, an initial learning rate of 0.0002 decaying exponentially by a gamma of 0.999875. The models were trained with a batch size of 16 using an NVIDIA A10 GPU with 24GB of GPU memory. For the domain adaptation, we fine-tuned for 100K iterations steps, while when training from scratch, 500K iterations steps were required. Finally, for the last experiments, all models were trained for 100K iterations steps since the model performance had typically started to converge.

Model evaluation For subjective evaluation, we ran Mean Opinion Score (MOS) and MUSHRA tests (ITU-RBS.1534, 2014) via an online web application, with 70 participants. We also did objective

[5] https://github.com/Niger-Volta-LTI/yoruba-voice-speech-recorder

[6] But also release Tacotron2-DCA and Glow TTS models trained with the same data

evaluation of the models' output, measuring the mean Mel Cepstral Distortion (Kominek et al., 2008) of selected utterances for each model. Our evaluation data is also based on the *news domain* but on samples *unseen* during training.

4.1. Does training with diacritics affect synthesised voices?

The importance of tone, represented orthographically by diacritics in Yorùbá led us to question whether there would be a difference between the speech produced by models trained with or without diacritics. For the male and female voices, we train TTS models from scratch with and without diacritics, presenting the results in Table 2.

Model	MOS↑	MUSHRA↑	MCD↓
With diacritics, Male	**3.98**	**60.91**	**7.47**
No diacritics, Male	1.86	18.97	8.93
With diacritics, Female	**2.82**	**48.24**	**6.26**
No diacritics, Female	1.50	19.02	11.88

Table 2: Results of experiments training TTS models with and without diacritics. The Mean Opinion Scores (MOS), MUSHRA scores and Mel-Cepstral Distortion (MCD) is measured. Across the board, models trained with diacritics perform better than those not trained with diacritics.

Across voices and evaluation metrics, we find that training models with diacritics leads to more natural sounding speech. The female voice does not have as high a MOS or MUSHRA score as the male voice, indicating a less natural sounding voice in comparison.

4.2. Does continued pre-training result in a more natural voice?

The availability of BibleTTS models (Meyer et al., 2022b) in Yorùbá provided the opportunity to use our dataset to continue pre-training TTS models which already produce natural sounding speech. The BibleTTS Yorùbá voice uses a single male voice. We wanted to observe whether starting from a trained TTS model would lead to a more natural sounding voice than using our data alone. To test this hypothesis, we trained models from scratch and continued pre-training the BibleTTS Yorùbá checkpoint. Results of the model evaluation are in Table 3.

For the female voice, we see that training from scratch leads to the best performance across objective and subjective metrics. This is likely due to the BibleTTS voice being male, and thus a more difficult adaptation during continued pre-training. For the male voice, the results are mixed. Although the MOS score is higher for the continued pre-training voice, the MUSHRA and MCD scores are higher for the model where we train from scratch. This means that when compared to the continued pre-training model, the from-scratch model is rated more natural.

Model	MOS↑	MUSHRA↑	MCD↓
Male	3.98	**60.91**	**7.47**
BibleTTS → Male	**4.22**	52.59	8.86
Female	**2.82**	**48.24**	**6.26**
BibleTTS → Female	2.57	41.63	8.93

Table 3: Subjective {MOS, MUSHRA} and Objective {MCD} evaluation results of voices trained from scratch (Male, Female) and continued pre-training (BibleTTS → Male, BibleTTS → Female).

4.3. How much data is required to train a synthetic voice?

Given the low-resource setting we work in, we experiment how far we the push the limit of a few resources. We test how many utterances are required to train a model that produces natural speech. Due to the number of models already trained, we measure this solely through objective evaluation with MCD.

Number of utts.	MCD↓
100 utterances	7.49
500 utterances	6.99
1000 utterances	7.11
2000 utterances	7.09
4500 utterances	**6.85**

Table 4: MCD of TTS models trained with varying numbers of utterances, ranging from 100 to 4 500. The model trained with the most utterances has the best performance.

The results for this experiment are in Table 4. The Mel Cepstral Distortion is highest for the model trained with the fewest utterances (7.49) and lowest for the model trained with the most utterances (6.85). Although the MCD does not decrease monotonically as the number of utterances increases, there is evidence that more data is better. Overall, based solely on this objective evaluation, one may reason that only 500 utterances are necessary to train a satisfactory Yorùbá VITS model.

5. Automatic Speech Recognition Experiments

To evaluate our corpus for speech recognition tasks, we train several baseline ASR models, with the following **data split of 15 000 / 1 000 / 4 000** for training, development and test respectively. We explore training a conformer model end-to-end and

finetuning self-supervised speech representations. The results of these experiments are in Table 5.

5.1. End-to-End Conformer model

We use ESPNet to train a 12-layer Conformer model end-to-end with an RNN language model (LM) for decoding. We use unigram tokenization and have a perplexity of 54.0 on the held-out validation set.

5.2. Finetuning wav2vec 2.0

We finetune wav2vec 2.0 XLSR-300m (Babu et al., 2022) and train an end-to-end Conformer model (Gulati et al., 2020). For wav2vec 2.0 XLSR-300m, a massive multilingual model, pretrained for speech tasks on 128 languages, we finetune for 20 000 steps, equating to 10.67 epochs.

Model	WER↓
Conformer + RNN LM	69.7
wav2vec 2.0 finetuned	40.6
+bigram model	27.6
+trigram model	**23.8**

Table 5: Word-error-rate (WER) for end-to-end Conformer model and finetuned wav2vec 2.0. Finertuning wav2vec 2.0 outperforms training an end-to-end Conformer model from scratch.

We observe that finetuning wav2vec 2.0 leads to significantly better performance versus training the Conformer model end-to-end. The addition of an n-gram language model further lowers the error-rate, with the trigram LM model prevailing.

Overall, a more substantial evaluation, beyond these initial baselines, will be required to better understand the benefits of finetuning a multilingual model versus training a simpler monolingual model. Finally, we hope that these initial results will encourage the inclusion of the Yorùbá language in more multilingual ASR evaluation benchmarks.

6. Conclusion

In this work we present an open dataset of 42 hours of high quality Yorùbá speech data to be used for both Speech Synthesis and Automatic Speech Recognition research. For TTS, we remark that models trained with diactrics generate speech that is perceived as more natural than those trained without diacritcs, while models continually trained from existing models may not always sound more natural than those trained from scratch. In ASR, we see that finetuning wav2vec 2.0 with a trigram model leads to the lowest word error rate. This data will be made freely available in the hope that it will invigorate speech research and accelerate the development of technology for the Yorùbá language.

7. Ethics Statement

For this project, we obtained the consent of all the volunteers that contributed their voice recording to the ÌròyìnSpeech project. Also, our recording does not include private or sensitive conversations that can violate our volunteers privacy since the utterances are mostly from the news domain.

8. Acknowledgements

We would like to thank the reviewers for their comments and suggestions. This work was carried out with support from Imminent Translated, whose 2022 funding helped support the project. Special thanks to Mr. Bode Adedeji and Miss Aguobi Nkechinyere Faith of the Department of Linguistics, University of Lagos, for permission to place our booth in their shared office while we did some of our recordings. David Adelani acknowledges the support of DeepMind Academic Fellowship programme. This research was funded in part by a Stanford School of Engineering Fellowship to TO. Finally, we are grateful to Professor Christopher Manning and Professor Dan Jurafsky for their useful feedback on the draft.

9. Bibliographical References

Tunde Adegbola and Lydia U. Odilinye. 2012. Quantifying the effect of corpus size on the quality of automatic diacritization of yoruba texts. In *Spoken Language Technologies for Under-Resourced Languages*.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi,

Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

Akingbemisilu Abiola Afolabi, Elijah Olusayo Omidiora, and O. T. Arulogun. 2014. Development of text to speech system for yoruba language.

Odéjobí Odétúnjí Àjàdí. 2007. A quantitative model of yorùbá speech intonation using stem-ml. *INFOCOMP Journal of Computer Science*.

Olúgbénga O Akinadé and Ọdẹ́túnjí A Ọdẹ́jọbí. 2014. Computational modelling of yorùbá numerals in a number-to-text conversion system. *Journal of Language Modelling*.

Akintoba Emmanuel Akinwonmi and B. K. Alese. 2013. A prosodic text-to-speech system for yorùbá language. *ICITST*.

John OR Aoga, Theophile K Dagba, and Codjo C Fanou. 2016. Integration of yoruba language into marytts. *International Journal of Speech Technology*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

F. O. Asahiah, O. A. Odejobi, and E. R. Adagunodo. 2017. Restoring tone-marks in standard yoruba electronic text: Improved model. *Computer Science*, 18(3):301–315.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Alan W Black. 2019. Cmu wilderness multilingual speech dataset. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Théophile K. Dagba, John O. R. Aoga, and Codjo C. Fanou. 2016. Design of a yoruba language speech corpus for the purposes of text-to-speech (tts) synthesis. In *ACIIDS*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the world. twenty-second edition.

Anmol Gulati, Chung-Cheng Chiu, James Qin, Jiahui Yu, Niki Parmar, Ruoming Pang, Shibo Wang, Wei Han, Yonghui Wu, Yu Zhang, and Zhengdong Zhang, editors. 2020. *Conformer: Convolution-augmented Transformer for Speech Recognition*.

Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara E Rivera, and Kólá Túbòsún. 2020a. Developing an open-source corpus of yoruba speech.

Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara E. Rivera, and Kólá Túbòsún. 2020b. Developing an open-source corpus of yoruba speech. In *Proc. of Interspeech 2020*, pages 404–408, October 25–29, Shanghai, China, 2020.

ITU-RBS.1534. 2014. Method for the subjective assessment of intermediate quality level of.

Abimbola Rhoda Iyanda and Olufemi Deborah Ninan. 2017. Development of a yorúbà texttospeech system using festival. *Innovative Systems Design and Engineering (ISDE)*, 8(5).

Abímbólá Rhoda Ìyàndá, Odetunji Ajadi Odejobi, Festus Ayodeji Soyoye, and Olúbénga O. Akinadé. 2014. Development of grapheme-to-phoneme conversion system for yorùbá text-to-speech synthesis.

C.D. Jones, A.B. Smith, and E.F. Roberts. 2003. Article title. In *Proceedings Title*, volume II, pages 803–806. IEEE.

John Kominek, Tanja Schultz, and Alan W. Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Proc. Speech Technology for Under-Resourced Languages (SLTU-2008)*, pages 63–68.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo Kabenamualu, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022a. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. In *Proc. Interspeech 2022*, pages 2383–2387.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, et al. 2022b. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. *arXiv preprint arXiv:2207.03546*.

Odétúnjí Àjàdí Odéjobí, Anthony Joseph Beaumont, and Shun Ha Sylvia Wong. 2004. A computational model of intonation for yorùbá text-to-speech synthesis: Design and analysis. In *TSD*.

Perez Ogayo, Graham Neubig, and Alan W Black. 2022. Building african voices. In *23rd Annual Conference of the International Speech Communication Association (InterSpeech 2022)*, Incheon, Korea.

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

Iroro Orife, Aremu Anuoluwapo, Kọ́lá Túbọ̀sún, , David Ifeoluwa Adelani, and Tolúlọpẹ́ Ógúnrẹ̀mí. 2022. Yorùbá voice speech recorder.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sandy Ritchie, You-Chi Cheng, Mingqing Chen, Rajiv Mathews, Daan van Esch, Bo Li, and Khe Chai Sim. 2022. Large vocabulary speech recognition for languages of africa: multilingual modeling and self-supervised learning. *ArXiv*, abs/2208.03067.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady ElSahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Peng Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, A J Kallet, Ilia Kulikov, Janice Lam, Shang-Wen Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, M.L. Ramadan, Abinesh Ramakrishnan, Anna Sun, Ke M. Tran, Tuan Q Tran, I. A. Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bo Yu, Pierre Yves Andrews, Can Balioglu, Marta Ruiz Costa-jussà, Onur Çelebi, Maha Elbayad, Cynthia Gao, Francisco Guzm'an, Justine T. Kao, Ann Lee, Alexandre Mourachko, Juan Miguel Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *ArXiv*, abs/2308.11596.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023a. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Proc. INTERSPEECH 2023*, pages 884–888.

Jiatong Shi, William Chen, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023b. Findings of the 2023 ml-superb challenge: Pre-training and evaluation over more languages and beyond.

Daniel van Niekerk, Etienne Barnard, Oluwapelumi Giwa, and Azeez Sosimi. 2015. Lagos-nwu yoruba speech corpus.

Daniel R. van Niekerk and Etienne Barnard. 2012.

Tone realisation in a yorùbá speech recognition corpus. In *SLTU*.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara N. Sainath, Pedro J. Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Franccoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *ArXiv*, abs/2303.01037.