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

1. INTRODUCTION

We introduce the ÌròyìnSpeech corpus—a new dataset influenced by a desire to increase the amount of high quality, freely available, contemporary Yorùbá speech. We release a multi-purpose dataset that can be used for both TTS and ASR tasks. We curated text sentences from the news and creative writing domains under an open license i.e., CC-BY-4.0 and had multiple speakers record each sentence. We provide 5 000 of our utterances to the Common Voice [1] platform to crowdsource transcriptions online. The dataset has 38.5 hours of data in total, recorded by 80 volunteers.

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 [2], 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 ignored in writings. The tonal marks are important for correct pronunciation and lexical disambiguation.

3. THE ÌRÒYÌNSPEECH CORPUS

3.1. Preparation of text sentences

Our goal was to combine news data and fictional texts to create a multi-purpose modern speech dataset, as other Yorùbá datasets used utterances from religious texts or biblical data [3, 4]. The corpus texts were obtained from the news article domain of the MENYO-20k dataset [5] (an open-sourced multi-domain English-Yorùbá machine translation corpus) with a non-restrictive license (i.e. CC-BY-4.0) and the Yorùbá portion of the MasakhaNER 2.0 dataset [6] (i.e MasakhaNER-YOR) based on Asejere newspaper[1]. The primary sources of the MENYO-20k dataset are the Voice of Nigeria newspaper[2] (a Nigerian government newspaper that publishes in seven Nigerian or regional languages—Arabic, English, French, Fulani, Hausa, Igbo, and Yorùbá) and the Global Voices newspaper[3](an international, multilingual community of writers, translators, and human rights activists contributing articles in their native language). We restrict our selection of news articles to these 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 50 hours of speech, we added some sentences extracted and modified from unpublished short stories previously translated into Yorùbá [4]. These texts were selected to broaden the domain of the vocabulary used in the dataset. In addition, we divided long sentences from the MENYO-20k and MasakhaNER-YOR into smaller utterances, and asked volunteers to manually generate new utterances with similar story or context as the original seed sentences. They also cross-checked the utterances for errors. In total, we had to manually generate about 8,000 sentences. We then cleaned up the data to create a final script. To ensure the sentences were 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 (e.g names of places; Kaduna to Ọ̀yọ́, Zamfara to Oǹdó, United States to Ìlú Ọba, Buhari to Bùhárí, Kenya to Kẹ́ńyà, etc).

[1] `https://www.asejere.net/`

[2] `https://yoruba.von.gov.ng/`

[3] `https://yo.globalvoices.org/`

[4] Short stories translated by Kọ́lá Túbọ̀sún

3.2. Recording of text sentences

3.2.1. Corpus partitions

Our text preparation yielded a total of 20 000 sentences which was used for the entire recording for both ASR and TTS. 6 000 lines were recorded by two single speakers (one male, one female) in the age group of 25-30, yielding 5 hours of speech which we envision for TTS tasks. 5 000 lines were allocated for the Common Voice crowdsourced platform, which has yielded 6 hours of speech. Finally, we recorded in-house all the 20 000 lines, or some 26 hours for ASR. They were recorded by eighty different volunteers, each recording 250 lines during one-hour studio sessions.

	No. of hours	No. of sentences
TTS	6h 36m	6 000*
In-house ASR	25h 55m	20 000
Common Voice ASR	6h 00m	5 000*
Total	38h 32m	20 000

Table 1. A summary of dataset statistics. The utterances used for TTS recording and for CommonVoice are subsets of the ASR utterances.

All volunteers were speakers of standard North West Yorùbá [5], were screened for dialect uniformity, and ranged in age from 18 to 69 years. The initial 6 000 lines 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 provided with a token gift each as thank you for their time.

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.

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*.

The tool works by reading a prepared text file, usually with 250 sentences and displays each line of text to be read in order. The app 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 for each sentence to the hard-drive as well as a metadata index, which can be used programmatically to prepare training examples. Over 60% of all lines recorded in total were recorded using the SpeechRecorder app.

	No. of hours	No. of sentences
ML-SUPERB Train (1h)	1h	793
ML-SUPERB Train (10 mins)	10m	137
ML-SUPERB Dev	10m	143
ML-SUPERB Test	10m	131
Total	1h 30m	1204

Table 2. A summary of ML-SUPERB dataset subset.

3.2.3. Post-production

We had four forms of post-processing. Where possible, recordings that had issues that could be fixed manually, were repaired by removing simple clicks and disfluenices.

In situations where the recording did not correspond with the text but the utterance remained grammatical, we did not rerecord the utterances but edited the text sentences instead. One example is the possessive extender morpheme vowel e.g. "Ilé wa" (our house) is pronounced as "Ilé e wa" — where the extra 'e' extends from the original root noun to show a self-referential towards the root noun.

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 "síbẹ́" respectively in spoken Yorùbá and are amended in the utterance accordingly. Tone pronunciation mistakes resulting in a change in the word such as "ilé" (house) to "ilẹ̀" (land) also result in utterances being amended.

If audio files had a variety of issues that rendered them unusable we re-recorded them, usually by a different person of the same gender (thus different speaker ID). Some of the issues include:

- Disfluenices: hesitations, stammers, clicks, sniffs, etc.
- External noises: paper rustling, microphone touching, intrusive voices, electronic notification beeps, etc
- Audio fidelity: low or uneven audio levels, clipping or distortion, or otherwise unintelligible words
- Incorrect dictation which could not be fixed by changing the script

3.3. Dataset summary

The resulting audio ended up at 38 hours and 20 minutes in total. We prepared 1hr 30 minutes out of the 36 hours for the ML-SUPERB new language track challenge[6]. ML-SUPERB stands for **M**ulti**L**ingual **S**peech processing

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

[6] https://mlsuperb.netlify.app/

Universal **PER**formance Benchmark, a leaderboard to benchmark the performance of Self-Supervised Learning (SSL) models on over 100 languages in various speech processing tasks. The submitted data to ML-SUPERB is in Table 2.

4. ACKNOWLEDGEMENT

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. Finally, we are grateful to Professor Christopher Manning and Professor Dan Jurafsky for their useful feedback on the draft.

5. REFERENCES

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

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

[3] Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara E. Rivera, and Kólá Túbòsún, "Developing an open-source corpus of yoruba speech," in *Proc. of Interspeech 2020*, October 25–29, Shanghai, China, 2020., 2020, pp. 404–408.

[4] 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, "BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus," in *Proc. Interspeech 2022*, 2022, pp. 2383–2387.

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

[6] 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, "MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 4488–4508, Association for Computational Linguistics.