AFRODIGITS: A COMMUNITY-DRIVEN SPOKEN DIGIT DATASET FOR AFRICAN LANGUAGES

Chris Chinenye Emezue[1*], **Sanchit Gandhi**[1], **Lewis Tunstall**[1], **Abubakar Abid**[1], **Josh Meyer**[2],

Quentin Lhoest[1], **Pete Allen**[1], **Patrick von Platen**[1], **Douwe Kiela**[1], **Yacine Jernite**[1],

Julien Chaumond[1], **Merve Noyan**[1], **Omar Sanseviero**[1]
[1]Hugging Face,[2] Coqui

ABSTRACT

The advancement of speech technologies has been remarkable, yet its integration with African languages remains limited due to the scarcity of African speech corpora. To address this issue, we present AfroDigits, a minimalist, community-driven dataset of spoken digits for African languages, currently covering 38 African languages. As a demonstration of the practical applications of AfroDigits, we conduct audio digit classification experiments on six African languages [Igbo (ibo), Yoruba (yor), Rundi (run), Oshiwambo (kua), Shona (sna), and Oromo (gax)] using the Wav2Vec2.0-Large and XLS-R models. Our experiments reveal a useful insight on the effect of mixing African speech corpora during finetuning. AfroDigits is the first published audio digit dataset for African languages and we believe it will, among other things, pave the way for Afro-centric speech applications such as the recognition of telephone numbers, and street numbers. We release the dataset and platform publicly at `https://huggingface.co/datasets/chrisjay/crowd-speech-africa` and `https://huggingface.co/spaces/chrisjay/afro-speech` respectively.

1 INTRODUCTION

Datasets are essential for the advancement of robust and beneficial deep neural networks in natural language processing (NLP) technologies Bender et al. (2021); Nekoto et al. (2020). The ImageNet (Deng et al., 2009) dataset is a prime example as it revealed the power of deep neural networks in image recognition (Krizhevsky et al., 2012; Russakovsky et al., 2015). That is to say, the more datasets there are for a given deep learning task, the better (in terms of robustness, fairness, and diversity) the model can get.

End-to-end deep learning models have pushed the state-of-the-art on speech processing tasks like automatic speech recognition (ASR) (Baevski et al., 2020; Babu et al., 2021; Radford et al., 2022), and speech synthesis (TTS). However, due to data scarcity, existing speech recognition technologies do not support African languages (Muhire, 2020; Dossou & Emezue, 2021; Afonja et al., 2021b;a). We believe that our voice defines who we are and therefore when languages are omitted from speech technologies, the identities and cultures of the speakers are gradually obscured. The AfroDigits project was created to fill this void of African speech corpora by using a community-based participatory approach (Nekoto et al., 2020) to build AfroDigits – a spoken digit dataset for all African languages. This dataset has a number of potential use cases, ranging from being used to easily introduce concepts in speech processing, to real-life applications like recognition of spoken telephone digits, street house numbers, etc.

The rest of the paper is structured as follows: we motivate AfroDigits and give an overview of our data collection efforts aimed to bridge the gap for languages in speech technology. Then we detail

*Research done while interning at Hugging Face. Correspondence to chris.emezue@gmail.com.

the AfroDigits project in section 3.1, and present the AfroDigits dataset as well as some of its useful properties in section 3.2. Finally, to demonstrate a possible use-case of the dataset we perform finetuning experiments and discuss their results in section 4.

2 RELATED WORK

We focus on related efforts in building speech corpora for speech processing tasks. Some popular large-scale open-source *monolingual* speech datasets, which have dominated research in speech processing, include LibriSpeech (Panayotov et al., 2015) (as well as its variants like LibriCSS (Chen et al., 2020), LibriLight (Kahn et al., 2019)) and TIMIT (Garofolo, John S. et al., 1993). However they do not have support for African and other non-English languages. Then came the wave of multilingual speech corpora, like Vox-Forge (vox), Babel (Gales et al., 2014), MAILABS (cai) and most notably, Mozilla's Common Voice (Ardila et al., 2019), to enable support for many more languages. However, the number of African languages supported is still meager. Out of the 2000+ African languages, only Kinyarwanda has 1000+ hours of audio on Common Voice (Muhire, 2020). Babel, the only project that contains a number of African languages, is 1) not open-source which limits its use to only those who can pay, and 2) has been shown to contain outdated styles of conversation that make it necessary to supplement models trained on it with datasets representing modern styles of communication for African languages (Dossou & Emezue, 2021).

Over the years many efforts have emerged, specially to fill the void of African speech corpora (van Niekerk et al., 2017; Dossou & Emezue, 2021; Umuganda; Meyer et al., 2022; Afonja et al., 2021a; Oyewusi et al., 2022; Babirye et al., 2022). One notable property of some of these projects has been their use of a community-based, participatory approach to data collection. One advantage of community-based data collection is that, while being cost effective, it has been shown to ensure sustainability and scalability (Adelani et al., 2022a; Emezue & Dossou, 2020; Scao et al., 2022; Joshi et al., 2020; Bender et al., 2021; Nweya et al., 2022b;a). These works have mostly focused on text-speech corpora and not digits, which is what sets our work apart.

The FSDD dataset (Jackson et al., 2018) which is most similar to ours in terms of the proposed use case, is English-based. Through AfroDigits, we contribute to the existing community-based efforts to build more African speech corpora, with a focus on digits.

3 AFRODIGITS

In this section, we expound on AfroDigits. This section begins with a description of the project, the interface and data curation process, and ends with an outline of the AfroDigits dataset.

3.1 THE PROJECT

The dataset presented in this paper was a result of the AfroDigits project. The AfroDigits project is meant to be a life-long community-driven tool for audio digit data collection. Our motivation for choosing the domain of spoken digits lies in our desire to create an Afro-centric minimalist dataset which can easily be used for speech processing tasks (e.g for making tutorials, introducing concepts or new models, training and evaluating a model), similar to the way MNIST (Lecun et al., 1998) is for the field of computer vision. The FSDD dataset (Jackson et al., 2018) which is most similar to ours in terms of the proposed use case, is English-based. The idea is that AfroDigits will 1) inspire African researchers to learn speech processing while working on their native languages, and 2) improve the discoverability of African languages to non-African researchers and practitioners. Another advantage of the digits domain is that while it's harder to find sentences (especially for African languages (Team et al., 2022; Adebara et al., 2022)), numerical digits are universal, which makes recording straightforward.

Our first challenge was the platform to use for recording the digits. We wanted a platform that required no technical expertise to use, since we were also targeting rural communities. To this end, we created the African Digits Recording platform on HuggingFace Spaces[1]. Figure 1 shows the recording platform. To make the recording entertaining, and inspired by the MNIST handwritten

[1]link to the Space will be revealed in the camera ready version

numbers dataset, we randomly displayed images of the numbers 0-9 for the participants to record. At the end of each recording session, when they had recorded numbers 0 through 9, the participants were shown a congratulatory image GIF and encouraged to do another round. Furthermore, we wrote down very simple instructions, displayed in Figure 2, at the top of the platform so that participants would quickly understand the task.

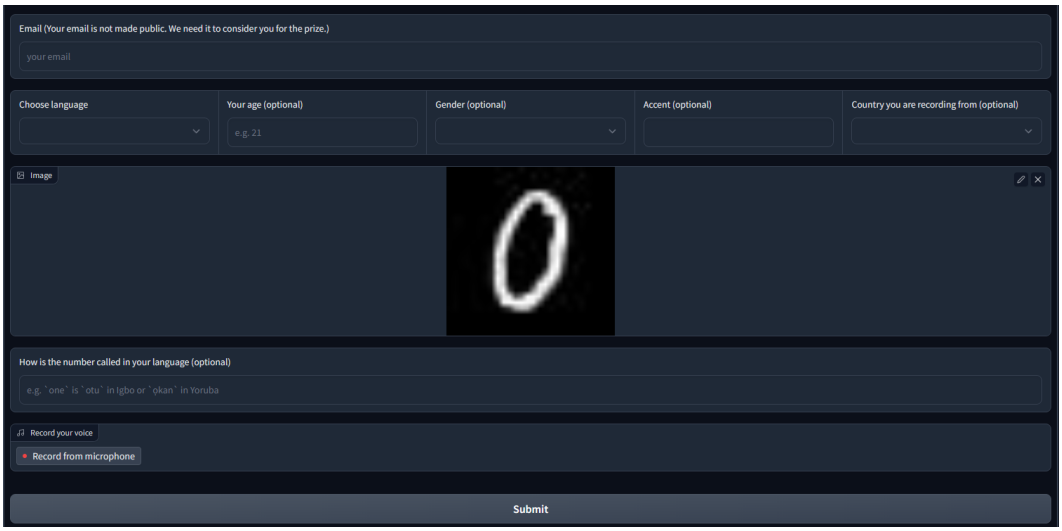

Figure 1: The AfroDigits recording platform. The participant is shown an image of the number, and recites it, while recording. The platform requires no log-in or sign-up making it very easy to use.

To encourage participation during the launch of the platform, we created the African Digits Recording Sprint which lasted for one month. Through widespread advertisement, especially within communities, such as Masakhane, with native speakers of African languages, we ensured active participation during the sprint. We further included prizes that were given to the top ten recording contributors. In order to obtain additional meaningful metadata besides the audio, we included optional fields for users to indicate their age, gender, accent, and country of residence. Additionally, we did not require the name, address, or any other personal information of the participants, following standard practice in audio data collection (Ardila et al., 2019).

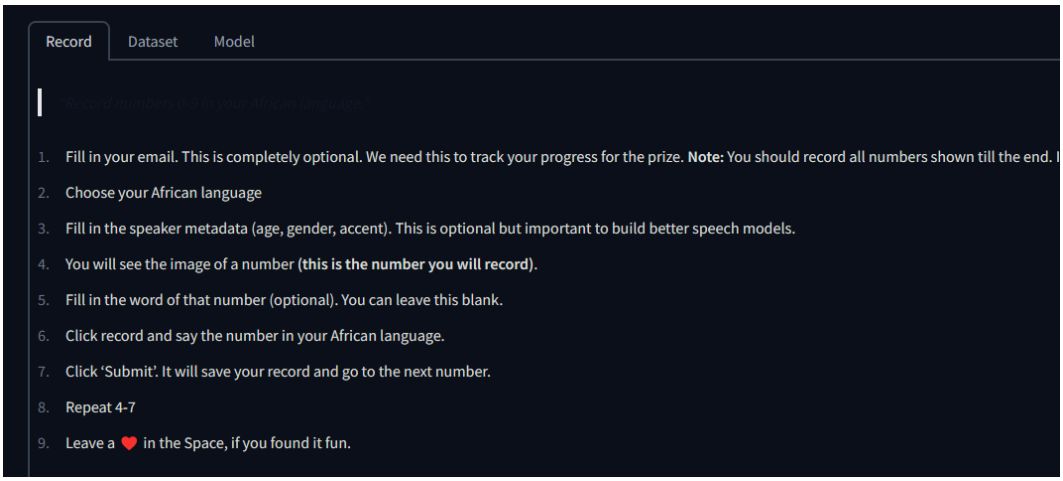

Figure 2: The set of instructions which the participant sees on entering the platform.

3.2 THE DATASET

Table 1 shows the current statistics of AfroDigits, which currently has 2,185 audio samples covering 38 African languages. AfroDigits is freely available for download. Following existing data governance principles (Benjamin et al., 2019; Jernite et al., 2022), the dataset is gated, meaning that one needs to provide details like name, email address and affiliation before getting access to the dataset. The whole dataset is housed in a `data` directory, which consists of sub-directories, each of which is named with randomly generated audio ids and contains an `audio.wav` file and a `metadata.jsonl` file where the metadata (audio id, language name, language id, digit, text of the digit, audio frequency, age, gender, and country of residence of the participant) for the specific audio file can be retrieved. All audios are mono-channel with a sampling rate of 48kHz. From Table 1, we see that Oshiwambo (kua) language has the highest number of recordings contributed (1,721). Using the HuggingFace Datasets[2] and Transformers[3] (Wolf et al., 2019) libraries, one can integrate the dataset directly into their training pipeline.

Table 1: Current data statistics of AfroDigits. The table is sorted by the language ISO-639-3 code in alphabetical order.

Language	Code	# Clips	Duration (seconds)
aasáx	aas	1	2.22
abua	abn	10	17.1
abon	abo	1	2.34
adamorobe sign language	ads	1	19.2
arabic, tunisian spoken	aeb	10	23.94
afrikaans	afr	11	25.02
qimant	ahg	2	5.27
amharic	amh	10	25.26
arabic, sudanese spoken	apd	5	12.96
arabic, moroccan spoken	ary	10	20.4
arabic, egyptian spoken	arz	1	2.64
bambara	bam	1	2.88
basaa	bas	10	33.6
andaandi	dgl	2	5.04
ezaa	eza	11	76.38
fon	fon	3	9.6
oromo, borana-arsi-guji	gax	40	83.1
hausa	hau	1	1.74
igbo	ibo	138	355.95
kinyarwanda	kin	21	84.96
oshiwambo	kua	1721	3376.34
dholuo	luo	1	1.92
luwo	lwo	10	47.34
massalat	mdg	1	10.8
ndebele	nde	12	51.3
ndonga	ndo	1	1.38
!xóõ	nmn	1	2.58
rundi	run	35	142.79
shona	sna	30	70.89
somali	som	12	11.54
swahili	swa	11	11.28
turkana	tuv	1	1.8
tswapong	two	10	9.81
makhuwa	vmw	10	40.38
wolof	wol	10	9.81
maay	ymm	1	3.42
yoruba	yor	28	27.48
zulu	zul	1	2.68

4 EXPERIMENTAL SETTING

To demonstrate the use-case of the AfroDigits dataset, we run finetuning experiments using pre-trained speech models. In this section we discuss the focus languages for our experiments, as well as the models utilized.

[2] `https://huggingface.co/docs/datasets/index`
[3] `https://huggingface.co/docs/transformers/index`

4.1 FOCUS LANGUAGES

For our experiments, we focused on the six African languages with the most significant number of audio samples in AfroDigits – Igbo (ibo), Yoruba (yor), Rundi (run), Oshiwambo (kua), Shona (sna), and Oromo (gax). The distribution of the digits and gender for each language is shown in Figures 5 - 16 in the Appendix section. Figure 3 shows the gender distribution across the number of recorded clips for the focus languages. We see representation of both male and female voices in yor, ibo and kua. Together with the pie chart on the right, there is a comparably similar representation of male and female voices in all our audio samples. All this is in line with previous work (Ardila et al., 2019; Bender et al., 2021; Adelani et al., 2021) opining that community-based crowd-sourcing fosters a wide representation of participants and improves diversity in data collection, thereby making the dataset more representative and less biased to a particular gender, race or location.

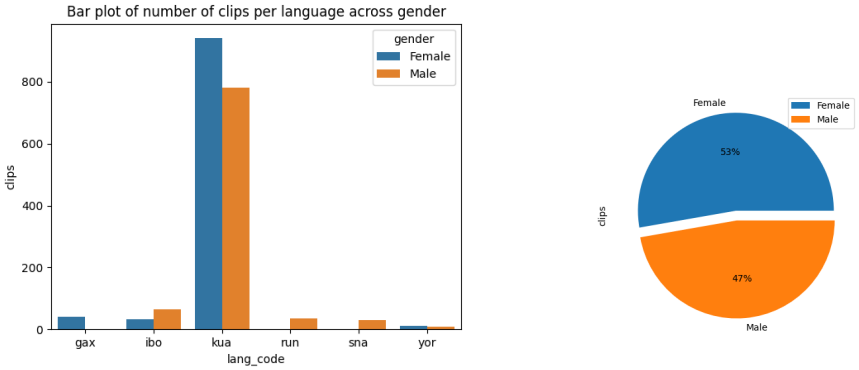

Figure 3: Left: Barplot of the number of clips for each of our focus languages segmented by the gender. Right: Gender distribution of audio samples for our focus languages

Table 2 shows information about each of our focus languages. XLS-R pretraining contained audio data from three of the six focus languages (namely yor,run,and sna) while Wav2Vec2-Large has none. The African audio data used in pretraining XLS-R came mostly from Babel Gales et al. (2014), which is not a free dataset. AfroDigits, being open-source and free for all, offers a major contribution to creating open-source and free speech data for African languages.

Table 2: Information about the languages used in our experiments, and their number of training and evaluation clips. For each pretrained model, we denote (✓) if the model was pretrained on audio data from that language and (✗) otherwise. `mixed` refers to the setting where we mix all the audio data from all languages.

Code	#Train / #Eval	XLS-R	Wav2Vec2-Large
ibo	96 / 42	✗	✗
yor	19 / 9	✓	✗
run	24 / 11	✓	✗
kua	1204 / 517	✗	✗
sna	21 / 9	✓	✗
gax	28 / 12	✗	✗
`mixed`	1392 / 600	-	-

4.2 MODELS

Pretrained speech models are powerful large neural network-based models that have been trained on a gigantic speech corpora. They are trained to learn and capture meaningful abstract features from speech (Schneider et al., 2019; Hsu et al., 2021; Radford et al., 2022). The knowledge learned can then be transferred to downstream tasks (Bengio, 2012; Wang & Zheng, 2015), and they are

particularly useful in low-resource settings (Zoph et al., 2016; Radford et al., 2022; Adelani et al., 2022a). This motivates our choice of using pretrained speech models on the downstream task of spoken digit classification. For our finetuning experiments, we utilized two large pretrained speech models: Wav2Vec2.0-Large (Baevski et al., 2020) and XLS-R (Babu et al., 2021).

Wav2Vec2.0-Large: The Wav2Vec2.0-Large model was pre-trained through a self-supervised learning of representations from raw audio data by masking spans of its discretized latent speech representations, similar to masked language modeling (Devlin et al., 2018), and using contrastive learning, where the true latent is to be distinguished from the false ones (van den Oord et al., 2018; Baevski et al., 2020; Rivière et al., 2020; Schneider et al., 2019). The authors note that jointly learning discrete speech units with contextualized representations helps Wav2Vec2.0 outperform the original Wav2Vec model (Schneider et al., 2019) in downstream recognition tasks. The Wav2Vec2.0 model was pretrained on an English-only LibriSpeech corpus (Panayotov et al., 2015)

XLS-R: In automatic speech recognition, researchers (Rivière et al., 2020; Schultz & Waibel, 2001; Stolcke et al., 2006; Huang et al., 2013; Hsu et al., 2021) have shown that it is beneficial to finetune with models that were pretrained on multilingual audio data – especially if the multilingual corpora contains some of the languages (or language family) of your downstream language. Motivated by this, we decided to include this model in our experiments. While the Wav2Vec2.0 model was pretrained on an English-only LibriSpeech corpus (Panayotov et al., 2015), XLS-R – built on the Wav2Vec2.0 backbone – was pretrained on a combined multilingual dataset of 128 languages, including 17 African languages. Table 2 shows which of our target African languages are represented in the XLS-R model.

XLS-R-Mix: Some studies have underlined the relevance of mixing training datasets in NLP, particularly for low-resource languages. In machine translation, Adelani et al. (2022b) showed that as low as 2000 translation sentences were sufficient to effectively finetune a large pre-trained model and obtain a significantly good performance. In the speech domain, Babu et al. (2021); Chan et al. (2021) demonstrated that combining audio data from several languages and domains improves transfer learning capabilities in settings where the training data is very small or noisy. Inspired by this, we set out to answer the following: since the individual train samples for each language are very small (see Table 1), can we have some improvement, for each language, if we finetune one model on the mix of audio samples from all the languages? For this, we finetuned XLS-R on a combined dataset from all the languages using the same hyperparameters above. The resulting model is called *XLS-R-Mix*.

4.3 Training Settings

Relatively Equal Model Parameters: The XLS-R model consists of $315,703,690$ model parameters while the Wav2Vec2.0-Large model has $315,693,962$ – a requirement enforced in order to ensure that neither model has an edge over the other based on their size.

Handling Class Imbalance: The AfroDigits dataset currently has very small audio samples for our focus languages, with an unequal balance of digits for each language (see Figures 5 - 16). In order to prevent the model from overfitting on the classes with many samples, we implemented weighted sampling (Monaco, 2013). With weighted sampling in the data loading process, different from the normal sampling which favors the majority classes (Caswell et al., 2020; Dunn, 2020; Fan et al., 2020), the labels are chosen with a probability inversely proportional to their size in the training set. This means that at each training step and for each language, the labels with few samples are more likely to be chosen for training the model. This is similar to studies (Arivazhagan et al., 2019; Team et al., ai) that have leveraged upsampling for under-represented languages during pre-training large language models in machine translation.

Training Setup: All audio samples were resampled to 16kHz for the finetuning experiments. We froze the encoders of each model and finetuned for 100 epochs. We used the Adam optimizer Kingma & Ba (2015), with a learning rate of $3e-5$ for both models. We did not do any search for optimal hyperparameters but instead used the recommended settings from the authors. We ran our finetuning experiments with five different seeds, then we took the average over the different

runs as well as the standard deviation. Training for each language took less than 30 minutes with a GPU, indicating the feasibility of the AfroDigits dataset as a 'Hello World' dataset for speech processing, just like MNIST is for computer vision. Finetuning the dataset on large speech models like Wav2Vec2.0-Large and XLS-R, however, needs a larger GPU resource. Where needed, we used an NVIDIA A100-SXM GPU.

5 Results & Discussion

In Table 3 we report the weighted F1 values on the held-out evaluation averaged over the 5 runs. Figure 4 shows the evolution of the finetuning performance on the held-out test set. We discuss our findings in the sub-sections below.

Table 3: Weighted F1 scores of each target language's evaluation set averaged over 5 runs. We see that in most cases, XLS-R performs better than Wav2Vec2-Large. XLS-R-Mix outperforms all other models in all languages.

Method	ibo	yor	run	kua	sna	gax
Wav2Vec2.0-Large	0.64 ± 0.29	0.13 ± 0.12	0.09 ± 0.08	0.40 ± 0.47	0.03 ± 0.06	0.37 ± 0.07
XLS-R	0.85 ± 0.03	0.16 ± 0.05	0.18 ± 0.07	0.98 ± 0.00	0.01 ± 0.03	0.56 ± 0.04
XLS-R-Mix	**0.86** ± 0.02	**0.27** ± 0.07	**0.55** ± 0.10	**0.98** ± 0.00	**0.65** ± 0.08	**0.57** ± 0.01

Bridging the gap for African Speech Datasets We observe, first of all, that despite being large pretrained models and finetuned for 100 epochs, the performance on some languages like yor, run, and sna is very low. This supports the claim by many research (Nekoto et al., 2020; Team et al., 2022; Fan et al., 2020; Kreutzer et al., 2021; Babirye et al., 2022; Oyewusi et al., 2022; Nweya et al., 2022b;a) that there is need to build more African datasets to improve the generalization of large pretrained models to low-resource African languages, and therefore the relevance of AfroDigits.

Effect of having African speech data in the model pretraining: We observe from Table 3 that the XLS-R model, which was pretrained on a larger set of African languages than Wav2Vec2-Large, performed better across all languages (except sna) than Wav2Vec2-Large which is Anglo-centric. Furthermore, Figure 4 shows the evolution of each of the model's performance (F1 metric) while it was being finetuned on each language. Each evaluation point is actually an average of the 5 rounds, with the confidence interval. Using kua as an example, we clearly see that the XLS-R model was able to quickly attain a high performance very early on in the finetuning, unlike Wav2Vec2.0-Large.

We also see that both models had difficulty with languages like sna, yor, and run: for XLS-R, we see a slight improvement for these languages only after the 50th epoch of finetuning for XLS-R, while for Wav2Vec2.0-Large, their performance oscillates between a rather low F1 score of 0.0 and 0.2. Both models perform very poorly on sna.

Effect of mixing audio samples during finetuning: XLS-R-Mix, which is XLS-R finetuned on a mix of training audio samples from the six focus languages, outperforms all the other models as shown in Table 3. More interestingly, even in sna, run, and yor, where the previous models perform very poorly, we see a significant boost in the performance of XLS-R-Mix. While the effect of a multilingual speech corpora has been shown in pretraining models Conneau et al. (2020); Chan et al. (2021); Radford et al. (2022), we present a useful insight on the effect of the mixing (especially for low-resource African languages) while finetuning on spoken digit classification.

6 Limitations of AfroDigits

The primary constraint observed in the initial release of the AfroDigits dataset is its small size, particularly for certain languages where only one sample is available. It is noteworthy that this project is a continuing effort and the platform used for recording voices is accessible to the general public. As such, it is anticipated that the number of recorded samples for certain languages in the dataset will expand in the future.

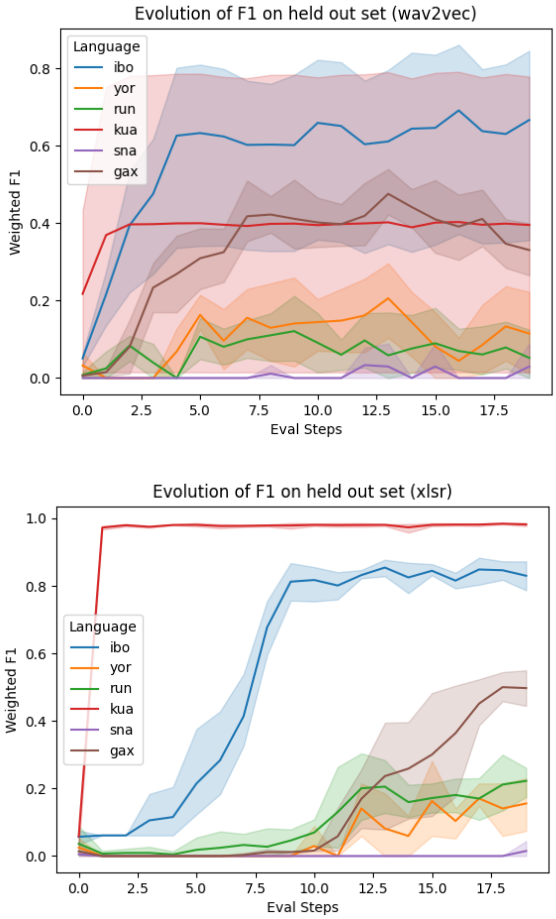

Figure 4: ↑F1 scores for each evaluation epoch during finetuning on held-out evaluation set for each of our focus African languages. Top is for Wav2Vec2.0-Large and Bottom is for XLS-R

7 Conclusion

In this work, we present AfroDigits: a minimalist, community crowd-sourced dataset of recorded digits in African languages, which can scale to any African language through community effort. AfroDigits, the first African digits dataset of its kind, was created with the aim of filling the void in African speech corpora and is released as a freely accessible public dataset. We further present the current contents and statistics of AfroDigits and show spoken digits classification experiments on six African languages using the speech corpus.

References

The mailabs speech dataset. https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/. [Accessed 01-Feb-2023].

Free Speech... Recognition (Linux, Windows and Mac) - voxforge.org — voxforge.org. http://www.voxforge.org/home. [Accessed 01-Feb-2023].

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv: Arxiv-2210.11744*, 2022.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, jul 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, D. Klakow, Peter Nabende, Ernie Chang, Tajuddeen R. Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris C. Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, T. Ajayi, Yvonne Wambui Gitau, Jade Z. Abbott, Mohamed Ahmed, Millicent A. Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Kabore, Godson Kalipe, Derguene Mbaye, A. Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, A. Awokoya, Happy Buzaaba, Blessing K. Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *naacl*, 2022b. doi: 10.48550/arXiv.2205.02022.

Tejumade Afonja, Clinton Mbataku, Ademola Malomo, Olumide Okubadejo, Lawrence Francis, Munachiso Nwadike, and Iroro Orife. Sautidb: Nigerian accent dataset collection, February 2021a. URL https://doi.org/10.5281/zenodo.4561842.

Tejumade Afonja, Oladimeji Mudele, Iroro Orife, Kenechi Dukor, Lawrence Francis, Duru Goodness, Oluwafemi Azeez, Ademola Malomo, and Clinton Mbataku. Learning nigerian accent embeddings from speech: preliminary results based on sautidb-naija corpus. *arXiv preprint arXiv: Arxiv-2112.06199*, 2021b.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *LREC*, 2019.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv: Arxiv-1907.05019*, 2019.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, and Davis David. Building text and speech datasets for low resourced languages: A case of languages in east africa. In *3rd Workshop on African Natural Language Processing*, 2022. URL https://openreview.net/forum?id=SO-U99z4U-q.

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. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv: Arxiv-2111.09296*, 2021.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver (eds.), *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pp. 17–36, Bellevue, Washington, USA, 02 Jul 2012. PMLR. URL https://proceedings.mlr.press/v27/bengio12a.html.

Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. Towards standardization of data licenses: The montreal data license. *CoRR*, abs/1903.12262, 2019. URL http://arxiv.org/abs/1903.12262.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6588–6608, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.579. URL https://aclanthology.org/2020.coling-main.579.

William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv: Arxiv-2104.02133*, 2021.

Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li. Continuous speech separation: dataset and analysis. *arXiv preprint arXiv: Arxiv-2001.11482*, 2020.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv: Arxiv-2006.13979*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: Arxiv-1810.04805*, 2018.

Bonaventure F. P. Dossou and Chris C. Emezue. Okwugbé: End-to-end speech recognition for fon and igbo. *arXiv preprint arXiv: Arxiv-2103.07762*, 2021.

Jonathan Dunn. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*, 54:999–1018, 2020.

Chris C. Emezue and Bonaventure F. P. Dossou. Lanfrica: A participatory approach to documenting machine translation research on african languages. *arXiv preprint arXiv: Arxiv-2008.07302*, 2020.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *arXiv preprint arXiv: Arxiv-2010.11125*, 2020.

M.J.F. Gales, K.M. Knill, A. Ragni, and S.P. Rath. Speech recognition and keyword spotting for low-resource languages : Babel project research at cued, May 2014. URL https://eprints.whiterose.ac.uk/152840/. © 2014 ISCA. Reproduced in accordance with the publisher's self-archiving policy.

Garofolo, John S., Lamel, Lori F., Fisher, William M., Pallett, David S., Dahlgren, Nancy L., Zue, Victor, and Fiscus, Jonathan G. Timit acoustic-phonetic continuous speech corpus, 1993. URL https://catalog.ldc.upenn.edu/LDC93S1.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, T. Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *INTERSPEECH*, 2021. doi: 10.21437/interspeech.2021-236.

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7304–7308, 2013. doi: 10.1109/ICASSP.2013.6639081.

Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. Jakobovski/free-spoken-digit-dataset: v1.0.8, August 2018. URL https://doi.org/10.5281/zenodo.1342401.

Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 2206–2222, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534637. URL https://doi.org/10.1145/3531146.3534637.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, jul 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. Libri-light: A benchmark for asr with limited or no supervision. *arXiv preprint arXiv: Arxiv-1912.07875*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv: Arxiv-2103.12028*, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998. doi: 10.1109/5.726791.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, A. Oktem, Daniel Whitenack Julian Weber, Salomon Kabongo KABENAMUALU, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris C. Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, S. Olanrewaju, Jesujoba Oluwadara Alabi, and Shamsuddeen Hassan Muhammad. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. *INTERSPEECH*, 2022. doi: 10.48550/arXiv.2207.03546.

Jane Monaco. *Weighted Sample*, pp. 2043–2044. Springer New York, New York, NY, 2013. ISBN 978-1-4419-1005-9. doi: 10.1007/978-1-4419-1005-9_1082. URL https://doi.org/10.1007/978-1-4419-1005-9_1082.

Remy Muhire. How rwanda is making voice tech more open, Sep 2020. URL https://foundation.mozilla.org/en/blog/how-rwanda-making-voice-tech-more-open/.

Wilhelmina Nekoto, V. Marivate, T. Matsila, Timi E. Fasubaa, T. Kolawole, T. Fagbohungbe, S. Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo KABENAMUALU, Salomey Osei, Sackey Freshia, Andre Niyongabo Rubungo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, L. Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Z. Abbott, Iroro Orife, I. Ezeani, Idris Abdulkabir Dangana, H. Kamper, Hady ElSahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan Van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris C. Emezue, Bonaventure F. P. Dossou, Blessing K. Sibanda, B. Bassey, A. Olabiyi, A. Ramkilowan, A. Oktem, Adewale Akinfaderin, and A. Bashir. Participatory research for low-resourced machine translation: A case study in african languages. *FINDINGS*, 2020. doi: 10.18653/v1/2020.findings-emnlp.195.

Gerald Okey Nweya, Solomon Oluwole Akinola, Emeka Felix Onwuegbuzia, Samuel Obinna Ejinwa, Anita Adiboshi, Daniel Success Nwokwu, Ihunna Peter, and Amarachi Akudo Osuagwu. Replication data for igbo natural language processing tasks ii, 2022a. URL https://doi.org/10.7910/DVN/YB9FWK.

Gerald Okey Nweya, Akinola Solomon Oluwole, Emeka Felix Onwuegbuzia, Samuel Obinna Ejinwa, Anita Adiboshi, Daniel Success Nwokwu, Ihunna Peter, and Amarachi Akudo Osuagwu. Replication Data for Igbo Natural Language Processing Tasks I, 2022b. URL https://doi.org/10.7910/DVN/RXBNCZ.

Wuraola Fisayo Oyewusi, Sharon Ibejih, Soromfe Uzomah, Elizabeth Mawutin Joseph, Jon Cynthia, Folakunmi Ojemuyiwa, Benedicta Johnson-Onuigwe, Omolola Taiwo, Akintunde Akinpelumi, Olabisi Adesina, Ayodele Noutouglo, Adeola Adeleke Adeoba, Andrew Akoh, Chukwuemeka Nwachukwu, Opeyemi Agbabiaje, Itunu Falade, Olukemi Erhunmwunsee, Oluwatobiloba Dada, Olúwatóbi David OSIBELUWO, Ehis Akene, Udim Akpan, Moira Amadi-Emina, Jaiyeola Marquis, Michael Senapon Bojerenu, Gbolahan Olumade, Oluwagbemi Lesi, Timothy Ezeh, Oluwadamilola Oguntoyinbo, Tosan Mogbeyiteren, Felicia Oresanya, Samuel Chika, and Sodiq Akinjobi. TCNSpeech: A community-curated speech corpus for sermons. In *3rd Workshop on African Natural Language Processing*, 2022. URL https://openreview.net/forum?id=r_-PYcf4LZc.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *PREPRINT*, 2022.

M. Rivière, Armand Joulin, Pierre-Emmanuel Mazar'e, and Emmanuel Dupoux. Unsupervised pretraining transfers well across languages. *Icassp 2020 - 2020 Ieee International Conference On Acoustics, Speech And Signal Processing (icassp)*, 2020. doi: 10.1109/ICASSP40776.2020.9054548.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh Haji-Hosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish

Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model. *BigScience Workshop, arXiv preprint arXiv: Arxiv-2211.05100*, 2022.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv: Arxiv-1904.05862*, 2019.

Tanja Schultz and Alex Waibel. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1):31–51, 2001. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(00)00094-7. URL https://www.sciencedirect.com/science/article/pii/S0167639300000947. MIST.

A. Stolcke, F. Grezl, Mei-Yuh Hwang, Xin Lei, N. Morgan, and D. Vergyri. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 1:I–I, 2006. doi: 10.1109/ICASSP.2006.1660022.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv: Arxiv-2207.04672*, 2022.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *META*, ai.

Digital Umuganda. Kinyarwanda dataset. URL https://digitalumuganda.com/dataset/.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv: Arxiv-1807.03748*, 2018.

Daniel van Niekerk, Charl van Heerden, Marelie Davel, Neil Kleynhans, Oddur Kjartansson, Martin Jansche, and Linne Ha. Rapid development of tts corpora for four south african languages. In *Proc. Interspeech 2017*, pp. 2178–2182, 2017. URL http://dx.doi.org/10.21437/Interspeech.2017-1139.

Dong Wang and Thomas Fang Zheng. Transfer learning for speech and language processing. *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1225–1237, 2015. doi: 10.1109/APSIPA.2015.7415532.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv: Arxiv-1910.03771*, 2019.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. *Conference On Empirical Methods In Natural Language Processing*, 2016. doi: 10.18653/v1/D16-1163.

A APPENDIX

Analysis of gender and digits distribution in the recorded audio samples: In the figures below, we plot their distribution, for both gender and the digits, across the Igbo, Oshiwambo and Rundi languages.

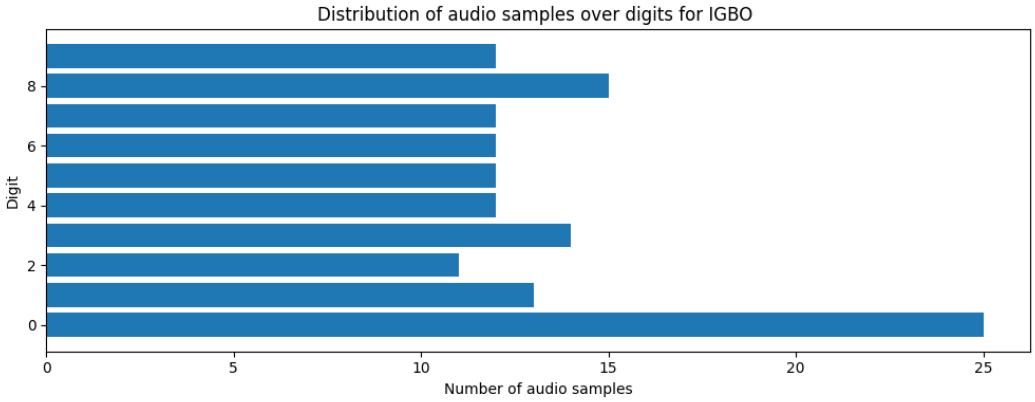

Figure 5: Distribution of recorded audio samples across the digits (ibo)

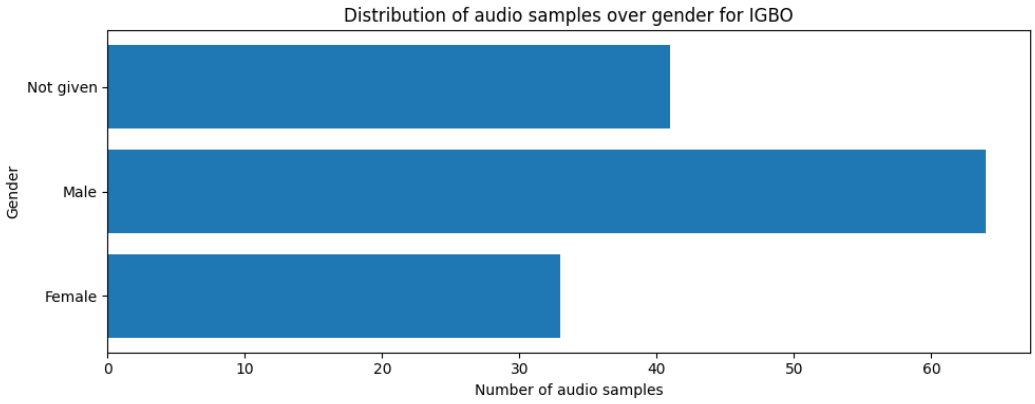

Figure 6: Distribution of recorded audio samples across the gender (ibo)

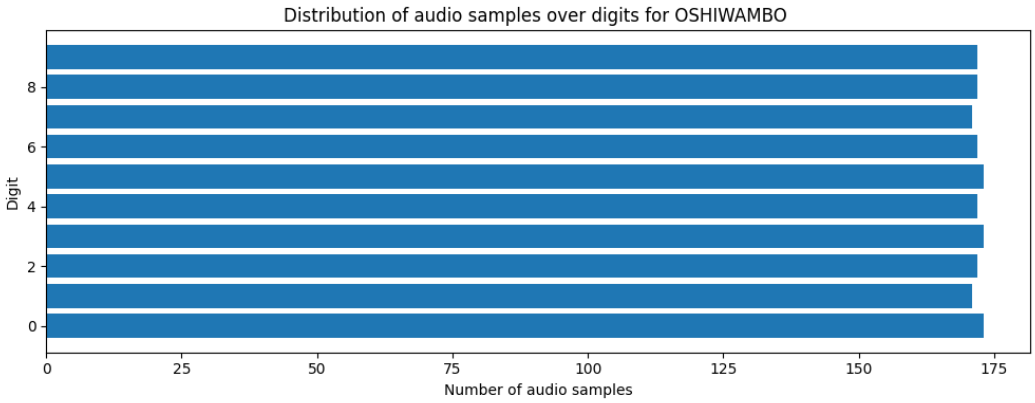

Figure 7: Distribution of recorded audio samples across the digits (kua)

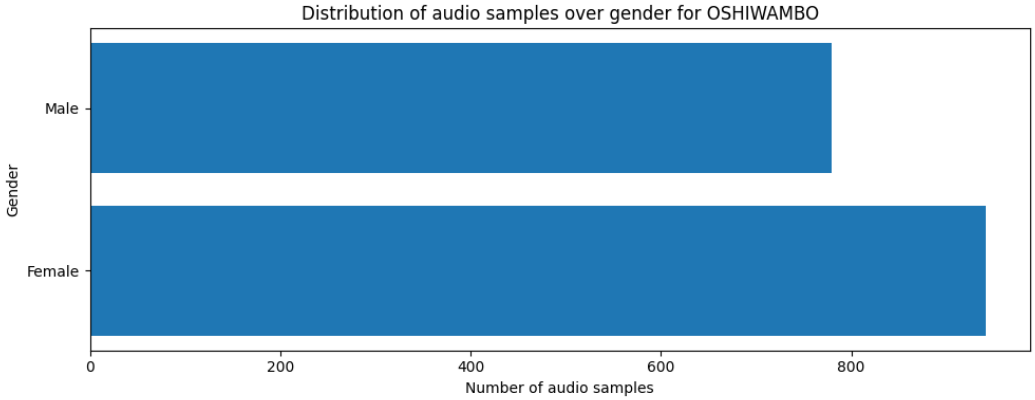

Figure 8: Distribution of recorded audio samples across the gender (kua)

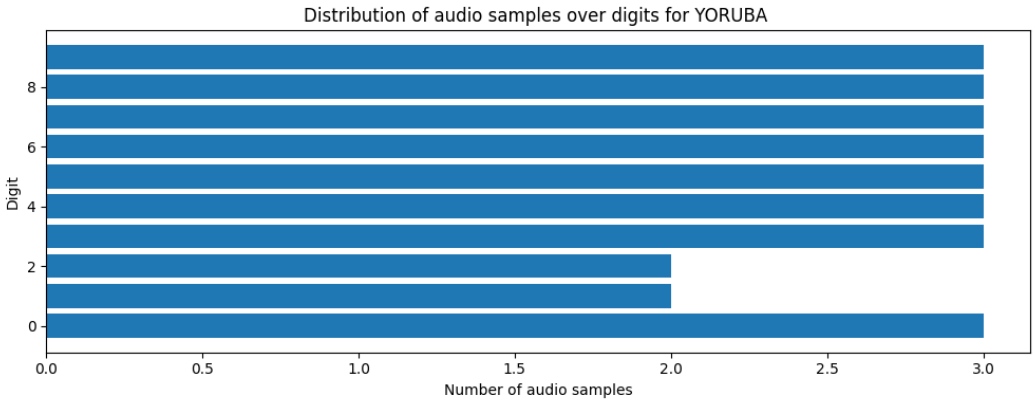

Figure 9: Distribution of recorded audio samples across the digits (yor)

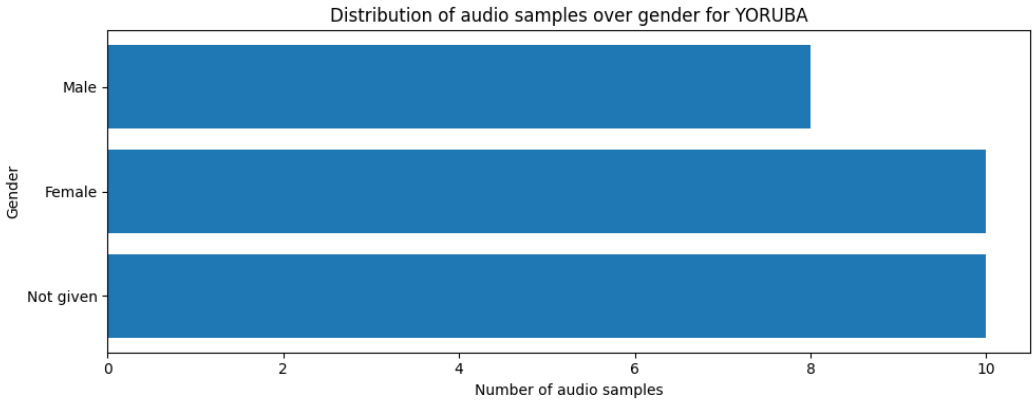

Figure 10: Distribution of recorded audio samples across the gender (yor)

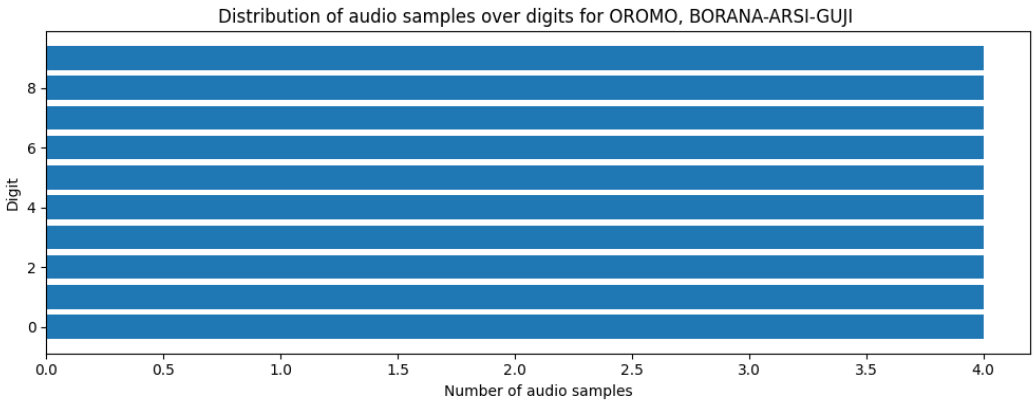

Figure 11: Distribution of recorded audio samples across the digits (gax)

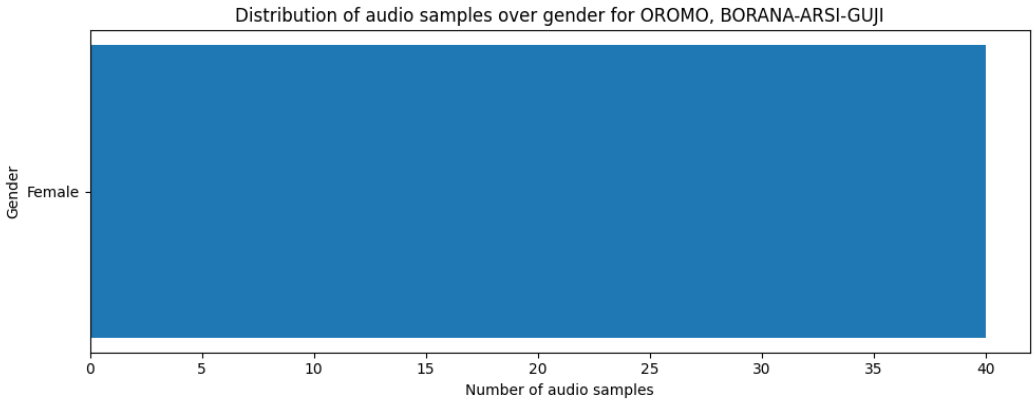

Figure 12: Distribution of recorded audio samples across the gender (gax)

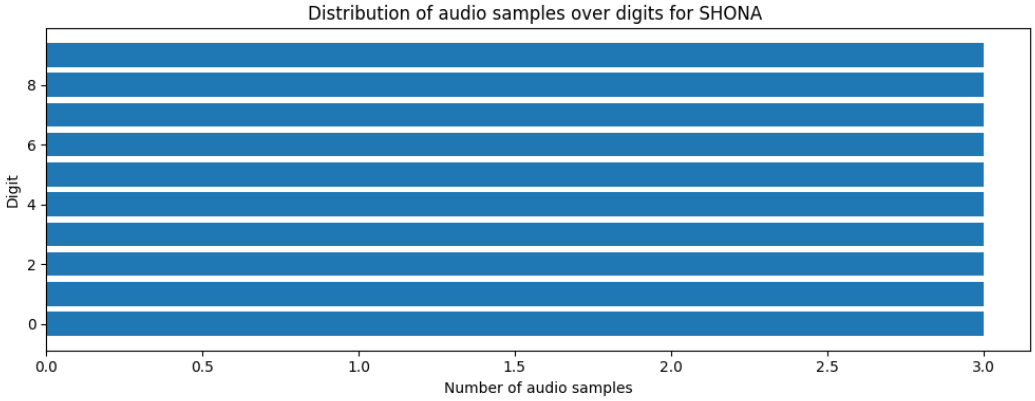

Figure 13: Distribution of recorded audio samples across the digits (sna)

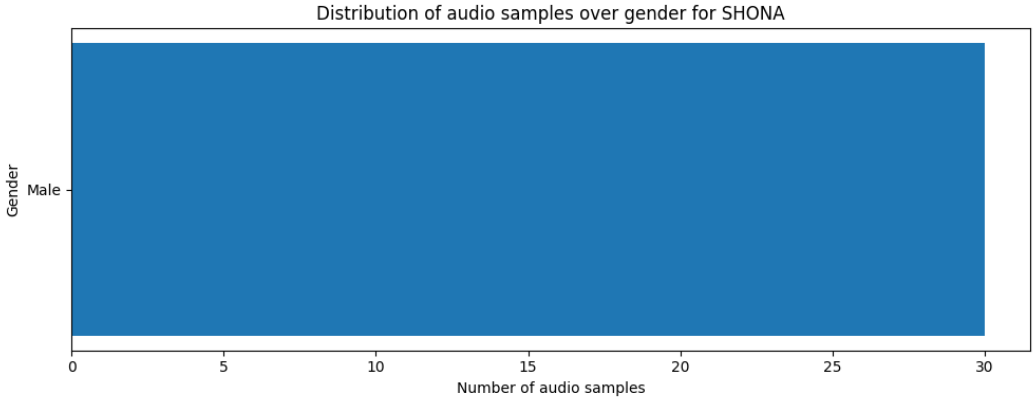

Figure 14: Distribution of recorded audio samples across the gender (sna)

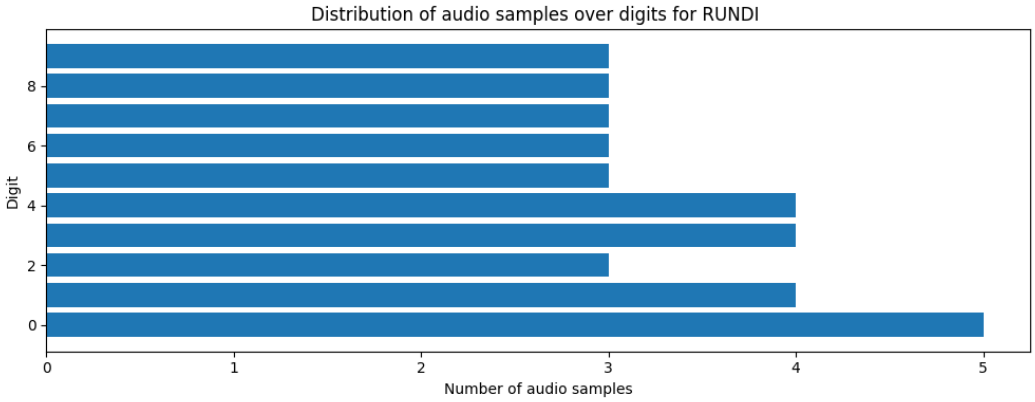

Figure 15: Distribution of recorded audio samples across the digits (run)

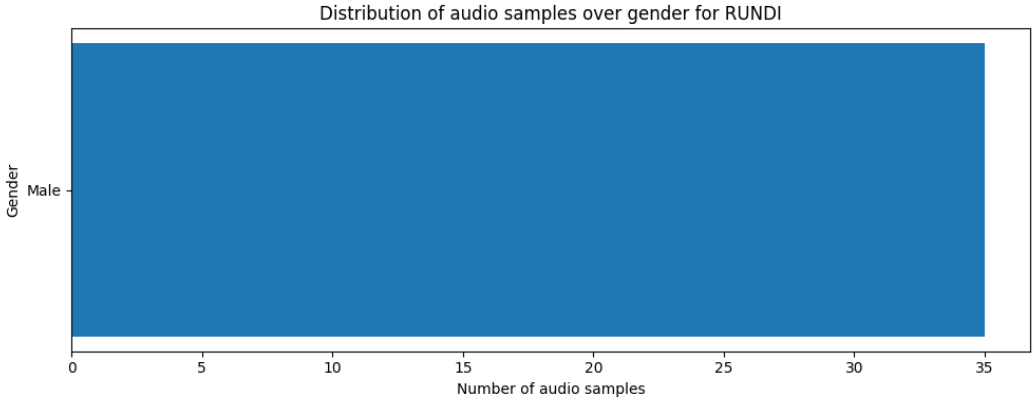

Figure 16: Distribution of recorded audio samples across the gender (run)