SPEECH RECOGNITION DATASETS FOR LOW-RESOURCE CONGOLESE LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large pre-trained Automatic Speech Recognition (ASR) models have begun to perform better in low-resource languages, as a result of the availability of data and transfer learning. However, a small number of languages have sufficient resources to benefit from transfer learning. This paper contributes to expanding speech recognition resources for under-represented languages. We release two new datasets to the research community namely: Lingala Read Speech Corpus consisting of 4 hours labeled audio clips and Congolese Speech Radio Corpus containing 741 hours of unlabeled audio in 4 major spoken languages in the Democratic Republic of the Congo. Additionally, we obtain state-of-the-art results for Congolese wav2vec2. We observe an average decrease of 2% in WER when a Congolese multilingual pre-trained model is used for finetuning on Lingala. Importantly, our study is the first attempt towards benchmarking speech recognition systems for Lingala and the first-ever multilingual model for 4 Congolese languages spoken by a combined 65 million people. Our data and models will be publicly available, and we hope they help advance research in ASR for low-resource languages.

1 INTRODUCTION

Low-resource languages are an important and under-represented area in the field of conversational AI. These languages suffer from a serious lack of training data for speech recognition systems, which are mostly developed for high-resource languages like English.

Automatic speech recognition (ASR), considered one of the primary building blocks of Conversational AI, use training corpora created by researchers in order to construct acoustic models capable of speech-to-text conversion. However, the development of speech recognition technology needs vast datasets with annotations. Unfortunately, the languages spoken in Sub-Saharan Africa tend to fall into the "low-resource" category, which, in contrast to "high-resource" languages, has fewer datasets accessible. Transfer learning, or the transfer of representations gained on unrelated high-resource languages, has not been studied for the majority of low-resource languages (Barnard et al., 2010; Besacier et al., 2014). Even with the assistance of transfer learning, labeled data are still required to construct viable models (Kunze et al., 2017).

It is challenging to create high-quality ASR models for such broad and varied array of languages. For the development of high-quality ASR systems, many contemporary ASR models depend on vast quantities of labeled data for each language. Such techniques are costly and not scalable, restricting the use of ASR technology to a subset of languages and populations. In addition to these issues of the availability of labeled data, Congolese languages [1] confront a number of other issues that must be addressed. Most of these languages have less digital content on the web, that pose obstacles for adopting language models. Likewise, there are prospects from an united standpoints. As numerous sources may be shared across different languages, the collection of unlabelled data for pre-training may be accomplished collaboratively. A vastly overlapping phoneme inventory may be used to construct multilingual models in which transfer learning is successfull (Javed et al., 2021).

In this contribution, we report initial development work on a broadband automatic speech recognition (ASR) system for the Lingala language and build a multilingual acoutic model for Congolese

[1] https://en.wikipedia.org/wiki/languages_of_Africa

languages. As in (Doumbouya et al., 2021), we employ unsupervised speech representation learning to solve the data scarcity issue and confirm that representations acquired from radio archives, which are rich in many areas of the globe, may be used for speech recognition in low-resource conditions. In this regard, it is essential to take notice of recent research that demonstrates the advantages of unsupervised pretraining and multilingual fine-tuning for dramatically improving ASR quality for low-resource languages (Baevski et al., 2020; Conneau et al., 2020). In particular, the Wav2Vec2.0 model has produced two significant outcomes for English ASR. First, SOTA outcomes are established through an end-to-end DNN design that borrows the popular Transformer architecture from NLP. In addition, pretraining on a large dataset decreases the labeled fine-tuning time from hundreds of hours to a few hours or even tens of minutes. It is important to investigate if these results from English ASR may be applied to Lingala ASR, considering the variety and aforementioned difficulties associated with Niger-Congo Bantu languages. Our objective is to answer the following research questions:

- Can a model similar to Wav2Vec2.0 provide SOTA performance on existing benchmarks for Lingala?
- Could extensive pretraining eliminate the requirement to gather vast quantities of labeled data?
- Could pretraining a multilingual model spanning a few Congolese languages provide beneficial results for these related languages?

In our journey to answer these questions, we offer the following contributions:

- **We assemble two novel datasets:** The first is the Congolese Speech Radio Corpus, while the second is the Lingala Read Speech Corpus. In the first dataset, we gather 741 hours of raw audio data for pretraining across four languages from the Niger-Congo Bantu language family, making it one of the biggest and most varied collections of Niger-congo B language data. The second dataset is a preliminary attempt to produce linguistic resources for speech recognition systems in Lingala, 4.35 hours of labeled data including a text corpus, a speech corpus, a lexical dictionary, and a language model.
- **We provide a baseline for the collected datasets:** Starting with the Supervised Learning (SL) method, we perform contrastive studies against Self-Supervised Learning (SSL) paradigms, particularly the Wav2Vec 2.0 model, and then we conduct extensive experimentations in order to arrive at a training and decoding rule that is effective for the Lingala language.
- To the best of our knowledge this is the first work carried out towards creating baselines for Lingala and the only created multilingual acoustic model combining four Congolese languages (i.e Kikongo, Tshiluba, Lingala and Swahili).
- **To encourage further research** in the field of underrepresented languages such as Niger-Congo Bantu ASR, we will make all of our artifacts available. This comprises (a) sources of pretraining data as well as scripts for their collection and preprocessing, (b) pretraining, finetuning, and decoding scripts, (c) language models, and (d) our top ASR models.

2 Related Work

Previous work in Natural Language Processing has made it possible to gather data, construct monolingual and multilingual ASR models, and push ASR research further for Niger-Congo Bantu languages.

The current state of speech datasets for African languages. Efforts have been made to compile speech datasets in low-resource African languages. The Niger-Congo Bantu language family have a total of six documented dataset collection: Basaa, Myene, and Embosi (Godard et al., 2017), Maninka, Susu, and Pular (Doumbouya et al., 2021). Amharic, Swahili, and Wolof are only few of the various African languages for which open datasets are available (Abate et al., 2005; Gelas et al., 2012; Gauthier et al., 2016a). additionally, East African languages(Claire Babirye et al., 2022), as well as South African languages(Van Niekerk et al. 2017; Nthite and Tsoeu 2020; Badenhorst et al. 2011), have been the focus of several efforts. The open-sourcing of some of these African

Table 1: Numbers of speakers, phonological features and orthographic conventions

Languages	ISO	Population	Tones	Diacratical	Example
Lingala	ln	40M	high,low	Yes	lɛlɔ 'today'
Kongo/Kikongo	kon	6.9M	high,low	Yes	mbó 'to hit'
Congolese Swahili	swc	11M	high,low	No	mtoto 'child'
Luba-Kasai/Tshiluba	lua	7M	high,mid,low	Yes	kutúla 'finished'

datasets has been a driving force in ASR research, as it encourages experimentation, training, and the development of better models. To this end, we were able to identify a dataset (Conneau et al., 2022) that included only one of the four Niger-Congo B languages we focused on. However, the dataset was noisy and lacked curation. Which prevented further testing.

Utilizing "found" data, such as radio archives (Doumbouya et al., 2021) investigated the use of found data, including ASR data and radio broadcasting, for three Niger-Congo Bantu languages. However, the data mostly included West African languages and did not include the central African of Niger-Congo Bantu languages. (Cooper, 2019) studied the use of radio news and audiobooks for low-resource language text-to-speech synthesis. However, the radio broadcast is of good quality and in English, as opposed to being noisy and having limited resources. Some research has focused on speech processing using "found" data in Indian languages (Mendels et al., 2015; Baljekar et al., 2018; Gupta et al., 2021; Javed et al., 2021). None of these found data projects includes noisy radio data for low-resource Niger-Congo languages in the Central African Regions.

Self-supervised speech representation learning Self-supervised speech representation learning approaches such as Wav2vec 2.0 aim to learn speech representation from unlabeled data in order to improve accuracy on downstream tasks such as phoneme classification, speaker recognition, sentiment analysis, speech recognition, and phoneme recognition while using fewer training data points (Baevski et al., 2020; Ritchie et al., 2022) This paper compares the multilingual Wav2vec 2.0 (XLS-R) large model pretrained on 128 languages - a large (436 thousounds of hours) corpus (Babu et al., 2021) - to its counterpart trained on a small (741.5 hours) dataset of noisy radio broadcasting archives in Central African languages for speech recognition on Niger-Congo B languages.

3 Congolese languages

There are more than 200 indigenous languages spoken in the Democratic Republic of the Congo (Eberhard et al., 2021). The top four languages in terms of global use are Lingala, Kongo or Kikongo, Luba-Kasai or Tshiluba and Congolese Swahili. These are Bantu languages particularly a branch of Niger-Congo B language family. They employ prefixes, suffixes, and infixes to show grammatical connections, and have a sophisticated system of noun classes and verb conjugation, and also have tonal system, agglutinative morphology, Subject Verb Object word order, complex syllable structure, and vowel harmony (Sene-Mongaba, 2015). All the three languages in this study are written using the standard 26 letters of the Latin alphabet, with the exception of lingala which possess two additional special characters: the open vowels ɛ and ɔ. Three of the four languages may use diacritical marks to indicate certain linguistic features, such as tone, vowel length, or emphasis. A summary of number of speakers, some phonological features and orthographic conventions used for the 4 languages is given in Table 1.

4 Data collection

In this research, we present two datasets: the Lingala Read Speech Corpus, which can be used to create the large vocabulary supervised speech recognition module for different applications, and the Congolese Radio Corpus, which is designed for unsupervised speech representation learning for downstream tasks involving 4 Congolese languages.

4.1 Lingala Read Speech Corpus

Description The corpus contains a total of more than 4.3 hours of voice data. In order to allow future experimentation, we offer consistent train, development, and test divides. There is no speaker

Table 2: General characteristics of the Lingala Read Speech corpus

Subset	Duration (hours)	Utterances
Train	4	2465
Dev	0.2	204
Test	0.1	180
Total	**4.3**	**2849**

Table 3: The token counts for the two sets of text sources used to create the language models task. LM_{small} relate to texts originating from the transcript of the corpus and LM_{large} are texts from additional online resources.

Language model	Sentences	No. of unique tokens
LM_{small}	2849	3980
LM_{large}	22 000	22261

overlap between the subgroups. All recordings utilise Waveform Audio File Format (WAVE), and each file is encoded with a sample rate of 16 kHz and a single track (mono).

Collection of the dataset To create the Lingala Read Speech corpus, we utilised the Lig-Aikuma app (Gauthier et al., 2016b) for speech recording. The software's elicitation mode was used to capture their readings of text scripts that had been tokenized at the sentence level. Numerous scholars have utilized the Lig-Aikuma in their own similar works (Sikasote & Anastasopoulos, 2021; Blachon et al., 2016; Gauthier et al., 2016c).

Speakers Lingala Read Speech recording speakers were from the Kinshasa area. In all, there are 2849 recordings recorded by 32 distinct speakers (13 male and 19 female). The ability to speak and read Lingala was a requirement for all of the speakers. Noting that the recordings in this corpus were not created under perfect controlled conditions. The speakers voices were recorded in a variety of conditions and with a varied range of accents. Consequently, it is to be anticipated that some utterances will include some ambient noise. This enables us to train and evaluate ASR systems in environments that more closely mimic the real world than a studio setting.

Preprocessing All utterances were validated against the transcripts to ensure the accuracy of the data. According to the utterances, the text was modified such that all numbers, dates, and times were converted into their textual equivalents. We also used speaker ID to sort the recorded audio files into groups based on the speaker. We also renamed all the audio files by appending the speaker ID to the utterance IDs.Table 2 shows the distribution of the dataset.

Text sources The recorded phrases and sentences are drawn from a wide variety of Lingala language sources, including Lingala literature. Table 3 summarises the textual contributions to the Lingala Read Speech corpus and the distribution of text used for language modeling tasks. Each phrase ranges in length from one to twenty words.

Availability The corpus will be accessed at our github project repository and will be shared with the research community under a creative commons license.

4.2 Congolese Speech Radio Corpus

Description There are a total of 741 hours of audio clips in the Congolese Radio Corpus, with the longest one being 25 seconds long. These recordings were taken from the archives of four different broadcast stations in the Democratic Republic of the Congo. News and Radio programmes are aired in many Niger-Congo B languages, including Lingala, Tshiluba, Kikongo, and Congolese Swahili. There may be a variety of sounds and music playing in the background and in the front of certain radio contents. Crawling the websites of the Radio archives (R.A) to get the URLs of audio recording in 4 Congolese languages simplified the gathering procedure. We not only curated material from the radio's official website, but also from YouTube clips that the stations had released

Table 4: The number of hours of Congolese Speech Radio Corpus per language post preprocessing.

Language	Radio archives(hours)	Youtube (hours)	Total(hours)
Lingala	187.6	86.4	274.1
Kikongo	174.4	-	174.4
Tshiluba	112.8	-	112.8
Congolese Swahili	191.1	-	191.1

underneath a Creative Commons License. Although this drastically reduced the quantity of data, we could gather, particularly in the low resource Niger-Congo language family, it was necessary to guarantee that we may freely share the URLs with the scientific world, assuring repeatability of our study.

Preprocessing The data was curated in the same way as in [8] after the URLs were bought to light. To begin, we grabbed the audio clips from the websites of the various radio broadcasts; for those files hosted on YouTube, we utilised the Youtube-dl library[2]. Secondly, the information was not always mono channel, and the sampling frequency ranged from 8 kilohertz to 44 kilohertz since the data was picked from a wide variety of sources. We used the FFmpeg library[3] to upsample/downsample the data that was captured at a frequency less than/ higher than 16 khz and then decrease the number of audio channels to 1. Thirdly, we further improved the data by excluding extended periods of quietness from the audio files using the Py-webrtcvad library[4], a python interface to hte widely used WebRTC VAD(Voice Activity Detection) module built by Google. The VAD algorithm screens out noise and lets us choose a harshness parameter (number between 0 and 3), which determines how strictly to apply the filter on speech (0 is the least aggressive about filtering out non-speech, 3 is the most aggressive). Similar to the findings in (Javed et al., 2021), we found that changing this value to 2 provided the best results for our data. Also, we utilised Waveform Amplitude Distribution Analysis (WADA-SNR) (Kim & Stern, 2008) to exclude audio samples with a signal-to-noise ratio (SNR) below 15 dB from our dataset. Using a sample of the audio files, a threshold was found to be optimal. Finally, we followed industry standards by chunking audio recordings to a maximum of 25 seconds in length. Table 6 provides a summary of the data acquired in this way.

Availability The data is 200 GigaBytes in size before compression. The audio file URLs, as well as the scripts used to collect the data and clean it up, can be accessed on the following link[5]

5 EXPERIMENTS

In this section, we describe the experiments to ascertain the usefulness of the two novel speech corpus for ASR applications in supervised, weakly-supervised and self-supervised learning configurations.

5.1 SUPERVISED BASELINE MODELS

The following experiment includes two supervised baseline models; These are traditional GMM-HMM based followed by a Time-Delay neural network (T-DNN) model (Peddinti et al., 2015) and the end-to-end architectural framework (Chang et al., 2020).

The T-DNN architecture consists of 4 hidden layers with layerwise temporal contexts of [-2,2],-1,2,-3,3 and -7,2, utilising a P-norm input dimension of 3000 and a group size of 10. The architecture follows the implementation of kaldi recipe (Povey et al., 2011; Kimanuka & Büyük, 2018). The Deep Speech2-like architecture is an end-to-end sequence-to-sequence model with two primary neural networks modules: N-blocks of Residual Convolutional Neural Networks (rCNNs) and a combined M-blocks of Bidirectional Long Short Term Memory (BiLSTMs) and Bidirectional Gated Recurrent Units (Bi-GRUs). The rCNNs function as feature extractors by translating speech into spectrograms, while the combined BiLSTMS and BiGRUs serve as encoder and decoder. We adhere to the same

[2]https://github.com/tpikonen/youtube-dl

[3]https://ffmpeg.org/

[4]https://github.com/wiseman/py-webrtcvad

[5]https://github.com/ussenuk/Congolese_Wav2Vec_datasets_experiment

Table 5: Results of experiments using fully supervised baseline models and weakly-supervised model on the Lingala Read Speech Corpus.

Model	**WER(%)**	**CER(%)**
Traditional ASR		
GMM-HMM + LM_{small}	39.9	15.9
T-DNN + LM_{small}	33.2	13.3
End-to-End ASR		
rCNN + BiLSTMs + BiGRUs	100	95
Whisper fine-tuned$(*)$	20.9	5.8

design as (Meyer, 2019). The final hidden state of the last Bi-GRU block is utilised to calculate the attention weights and the context vector, which are concatenated with the BI-GRU output to serve as the final outputs. These two baseline experiments use the 4.3 hours supervised Lingala Read Speech corpus in their implementation. Table 5 displays the outcomes of the baseline systems.

5.2 WEAKLY-SUPERVISED MODEL

The weakly-supervised Whisper model (Radford et al., 2022), is an encoder-decoder Transformer with an end-to-end auto-regressive architecture. This model has been pretrained using a total of 680 000 hours of supervised data from a variety of languages and tasks. The encoder uses a pair of convolution function of GELU, the first of which has a filter width of 3 and an activation function of GELU, and the second of which has a stride of 2. With the same width and number of transformer blocks, the encoder and decoder are architecturally identical. The Whisper model was pretrained in a weakly supervised manner using varaying sizes. For this experiment we used the medium sized model and performed supervised finetuning using the 4.3 hours supervised Lingala Read Speech corpus. Since Whisper models are pretrained for multitask operation, it is important to specify the task before finetuning operation, hence we selected the ”Transcribe task” prior to finetuning. Results are displayed in Table 5.

5.3 SELF-SUPERVISED MODEL

Using the unlabelled audio data given in the Congolese Speech Radio corpus, we implemented the pretraining of a self-supervised ASR model for Congolese languages. Note that we strictly adhere to the process outlined in (Baevski et al., 2020) for developing a self-supervised ASR system.

For this self-supervised experiment, we selected Wav2Vec2.0 architecture which is an end-to-end model consisting primarily of a feature encoder for converting raw audio into a sequence of T latent representations, a Transformer for learning contextualised representations for each of the T units, and a quantizer for discretizing the representations learned by the feature encoder. The process of training this model involves a pretext task of masking some of the T input representation before passing them to the context network. The quantizer's inputs however, are not masked, since they constitute the target of the Self-learning procedure. The objective is achieved by minimizing a contrastive loss between the context network's output and the quantized representation.

Pretraining a Wav2Vec2 model from scratch (Congolese Wav2vec / CdWav2Vec) We pretrain only on the BASE model similar to that of [9], with 7 convolutional layers, 512 channels per model, strides of (5,2,2,2,2,2), and kernel widths of (10,3,3,3,3,2,2). This BASE model is made up of 12 transformer blocks, 768 model dimensions, 3072 FFN dimension, and 8 attention heads. In the quantization module of this design, we use G=2 (codebooks) with V=320 elements per codebook. We have used the pretrained checkpoint of the equivalent (BASE) English Wav2Vec 2.0 model to kick off our pretraining process. We next use our carefully chosen data on the Congolese Speech Radio archive dataset to further pretrain the model. Shorter audio clips of 15.6 seconds in length (256k samples) are used in the BASE model. There are no changes set to any of the hyperperameters from the standard configuration of the Wav2Vec 2.0 source code.[6]

[6]https://github.com/pytorch/fairseq/tree/master/examples/wav2vec#pretrained-models

$(*)$ The Whisper model contained the non-curated Lingala subset of Fleurs dataset not considered in this experiment

Table 6: Results of experiments using self-supervised pretraining on the Congolese Speech Radio corpus and finetuning on the Lingala Read Speech corpus.

Model	WER(%)	CER(%)
XLSR-53	23.3	6.8
XLS-R-0.3B	25.8	7.0
CdWav2Vec	21.4	6.81

Finetuning a Wav2Vec2 model During fine-tuning we update all of the network parameters besides the convolutional features encoder parameters. In our finetuning experiment we considered two scenarios: (1) Finetuning the congolese Wav2vec (CdWav2Vec) model pretrained on the 4 Congolese languages(Transfer learning); (2) Finetuning multilingual XLSR-53 and XLS-R (Babu et al., 2021) model, which is also a Wav2Vec 2.0 based model trained on freely accessible speech audio data in 53 and 128 languages respectively. We explored experiments on the XLS-R-0.3B model with 600 million parameters. Both the fine-tuning experiments used the supervised Lingala Read Speech corpus implemented using either the HuggingFace transformer library (Wolf et al., 2020) or the fairseq toolkit (Ott et al., 2019).The hyperparameters used for both pretraining and finetuning are similar to the BASE model in (Baevski et al., 2020).

Transfer learning From Congolese languages to Lingala From the Congolese Wav2Vec pretrained model, we tried implementation of cross-lingual transfer learning in our quest to find a model with improved performance. to do this,we finetuned the multilingual Congolese model pretrained on the 4 congolese languages on our supervised Lingala Read Speech corpus using Adam optimizer, with a learning rate of 1e-4 and a three-stage learning rate schedule for 80 000 steps. we have included the WER results in Table 6. By the transfer learning from multilingual pretrained CdWav2Vec model to Lingala Read speech corpus we were able to obtain 3% relative WER improvement over the multilingual XLS-R models.

6 RESULTS

The results of our experimental evaluation are summarised in Table 7. The model with the best performance was the CdWav2Vec multilingual model obtained by pretraining on the Congolese speech radio corpus and then fine-tuning on the Lingala Read Speech corpus. We were able to get a WER of 21.4% using this model. We also find that the CdWav2Vec model consistently outperforms the supervised baselines and the multilingual self-supervised XLS-R model. The improvement in performance of the CdWav2Vec model compared to the baseline demonstrates the superior cross-lingual transfer ability of multilingually pre-trained models over supervised (trained from scratch) models in low resource scenarios, in this case from three Congolese languages to Lingala.These results indicate that multilingual pre-trained models perform better in cross-lingual transfers than supervised monolingual models.

6.1 IMPACT ON FINETUNING

Using checkpoints from monolingual and multilingual pretraining, we fine-tune 4.3 hours of Lingala data. We see that even with a small amount of supervised data, Lingala benefits from multilingual pretraining. Table 8 demonstrates further that fine-tuning monolingual and multilingual pre-trained models resulted in a 1% WER difference when decoding with or without a language model.

6.2 IMPACT OF CROSS-LINGUAL REPRESENTATIONS

It has been explainned earlier that the first stage of the CdWav2vec model entails computing representations of frames of audio from a learned codebook. All four languages are represented in the model using the same codebook vectors. In the same attempt as (Gupta et al., 2021),we try to answer whether these codebooks are independent or shared across languages and the degree to which the related languages share codebook entries increases with linguistic distance. To answer the question, we created a plot, which displays the quantized speech representations for all the languages in our study. To do so, 200 samples are chosen at random from each language, for a grand total of 5 hours of data. For the purpose of creating codebook vectors, we first run the audio through the feature

Table 7: Results summary reporting the best performance of each model.

Model	WER(%)	CER(%)
Traditional Supervised ASR		
GMM-HMM	39.9	15.9
T-DNN	33.4	13.4
End-to-End Supervised ASR		
rCNN + BiLSTMs + BiGRUs	100	95
multilingual self-supervised models		
XLSR-53	23.3	6.8
XLS-R-0.3B	25.8	7.0
CdWav2Vec	21.4	6.8
CdWav2Vec+ LM_{small}	16.7	6.2
CdWav2Vec+ LM_{large}	18.4	6.3

Table 8: Effect of multilingual and monolingual pretraining on WER

Pretraining	Finetuning	Decoding	WER	CER
monolingual	Lingala	Viterbi	22.1	6.8
multilingual	Lingala	Viterbi	21.4	6.8
monolingual	Lingala	KenLM	18.4	6.3
multilingual	Lingala	KenLM	17.4	6.3

encoder and then through the quantizer. For each sequence, the quantizer module's output vectors. For each language, vectors are normalised to form V x G vectors of size, followed by K-means clustering and principal component analysis to reduce the dimensions of these vectors (PCA). As seen in Figure 1, groupings of languages that are most phonetically similar to one another tend to cluster together. This gives more evidence that multilingual pretraining may promote representation learning that is transferable across linguistically related languages.

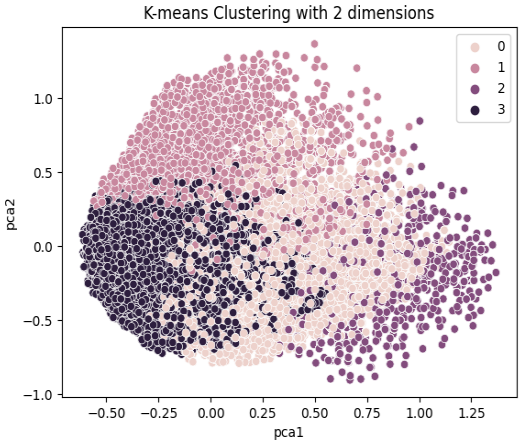

Figure 1: Quantized speech representation where 0:Kikongo, 1: Lingala, 2: Congolese Swahili and 3: Tshiluba)

7 Conclusion

Building speech recognition systems require the use of very large datasets to train the models. We created two novel datasets in this paper: a 4.3-hour labelled Lingala Read Speech Corpus for Lingala and a 741-hour unlabeled Congolese Speech Radio Corpus for four major Congolese languages. We provide an extensive evaluation using both supervised and self-supervised learning settings. Our experimental results show the potential of multilingual pretraining approaches like wav2vec 2.0 for low-resource Congolese languages. We created the first benchmark for speech recognition systems

for the Lingala language and the first-ever multilingual model for 4 major Congolese languages (CdWav2Vec). We will also publicly release all the trained models, and the collected datasets to promote further speech technology research development for currently marginalised communities.

ACKNOWLEDGMENTS

We thank Radio Okapi for donating radio archives and the 32 speakers who contributed their voices and time to the speech recognition corpus. We thank Yannis Assael for providing critical guidance that made this research possible. We also express our gratitude to the Africa PhD Fellowship Program for the resources provided during the research.

REFERENCES

Solomon Teferra Abate, Wolfgang Menzel, and Bairu Tafila. An Amharic speech corpus for large vocabulary continuous speech recognition. In *Proc. Interspeech 2005*, pp. 1601–1604, 2005. doi: 10.21437/Interspeech.2005-467.

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, 2021. URL https://arxiv.org/abs/2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020. URL https://arxiv.org/abs/2006.11477.

Pallavi Baljekar, SaiKrishna Rallabandi, and Alan W Black. An Investigation of Convolution Attention Based Models for Multilingual Speech Synthesis of Indian Languages. In *Proc. Interspeech 2018*, pp. 2474–2478, 2018. doi: 10.21437/Interspeech.2018-1869.

Etienne Barnard, Johan Schalkwyk, Charl van Heerden, and Pedro J. Moreno. Voice search for development. In *Proc. Interspeech 2010*, pp. 282–285, 2010. doi: 10.21437/Interspeech.2010-111.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100, 2014. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2013.07.008. URL https://www.sciencedirect.com/science/article/pii/S0167639313000988.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, 81:61–66, 2016. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2016.04.030. URL https://www.sciencedirect.com/science/article/pii/S1877050916300448. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe. End-to-end multi-speaker speech recognition with transformer, 2020. URL https://arxiv.org/abs/2002.03921.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, Daan Van Esch, Vera Axelrod, Simran Khanuja, Jonathan H. Clark, Orhan Firat, Michael Auli, Sebastian Ruder, Jason Riesa, and Melvin Johnson. Xtreme-s: Evaluating cross-lingual speech representations, 2022. URL https://arxiv.org/abs/2203.10752.

Erica Cooper. Text-to-speech synthesis using found data for low-resource languages. 2019.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users. *CoRR*, abs/2104.13083, 2021. URL https://arxiv.org/abs/2104.13083.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the world. twenty-third edition., 2021. URL http://www.ethnologue.com.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. Collecting resources in sub-Saharan African languages for automatic speech recognition: a case study of Wolof. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3863–3867, Portorož, Slovenia, May 2016a. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1611.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. Lig-aikuma: A mobile app to collect parallel speech for under-resourced language studies. In *Interspeech*, 2016b.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. LIG-AIKUMA: a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies. In *Interspeech 2016 (short demo paper)*, San-Francisco, United States, September 2016c. URL https://hal.science/hal-01350062.

Hadrien Gelas, Laurent Besacier, and François Pellegrino. Developments of Swahili resources for an automatic speech recognition system. In *Proc. 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2012)*, pp. 94–101, 2012.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Maynard, Markus Müller, Annie Rialland, Sebastian Stüker, François Yvon, and Marcely Zanon Boito. A very low resource language speech corpus for computational language documentation experiments. *CoRR*, abs/1710.03501, 2017. URL http://arxiv.org/abs/1710.03501.

Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chimmwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. CLSRIL-23: cross lingual speech representations for indic languages. *CoRR*, abs/2107.07402, 2021. URL https://arxiv.org/abs/2107.07402.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. Towards building ASR systems for the next billion users. *CoRR*, abs/2111.03945, 2021. URL https://arxiv.org/abs/2111.03945.

Chanwoo Kim and Richard M. Stern. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Interspeech*, 2008.

Ussen Abre Kimanuka and Osman Büyük. Turkish speech recognition based on deep neural networks. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 2018.

Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. Transfer learning for speech recognition on a budget. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 168–177, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2620. URL https://aclanthology.org/W17-2620.

Gideon Mendels, Erica Cooper, Víctor Soto, Julia Hirschberg, Mark John Francis Gales, Kate Knill, Anton Ragni, and Haipeng Wang. Improving speech recognition and keyword search for low resource languages using web data. In *Interspeech*, 2015.

Josh Meyer. Multi-task and transfer learning in low-resource speech recognition. 2019.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4009. URL https://aclanthology.org/N19-4009.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proc. Interspeech 2015*, pp. 3214–3218, 2015. doi: 10.21437/Interspeech.2015-647.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlícek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. The kaldi speech recognition toolkit. 2011.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

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

Bienvenu Sene-Mongaba. The making of lingala corpus: An under-resourced language and the internet. *Procedia - Social and Behavioral Sciences*, 198:442–450, 2015. ISSN 1877-0428. doi: https://doi.org/10.1016/j.sbspro.2015.07.464. URL https://www.sciencedirect.com/science/article/pii/S1877042815044651. Current Work in Corpus Linguistics: Working with Traditionally- conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015).

Claytone Sikasote and Antonios Anastasopoulos. Bembaspeech: A speech recognition corpus for the bemba language. *CoRR*, abs/2102.04889, 2021. URL https://arxiv.org/abs/2102.04889.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

A Appendix

You may include other additional sections here.