Creating Task-Specific Speech Recognition Datasets from Scratch for Low-Resource Languages: Assessing the Impact of Token Sequence Overlap

Adwoa Asantewaa Bremang, Dennis Asamoah Owusu, Victor Kow Quagraine, Leanne Annor-Adjaye
Ashesi University, Ghana
{ dowusu,vquagraine}@ashesi.edu.gh
{asantewaa.bremang,leanne.adjaye}@alumni.ashesi.edu.gh

Abstract

Creating a task-specific speech recognition dataset is essential for developing speech recognition applications in low-resource languages. Such applications have uses in agriculture, finance, healthcare, and others, and benefit individuals with low literacy. However, a significant challenge is the high cost of data creation. While there is some work around cost-effective dataset selection, there is little to no work on building a cost-effective dataset for a task from scratch. Our work contributes to the latter. We created a speech recognition dataset from scratch and conducted two major sets of experiments. The first aimed to observe the effect of different datasets of the same size on model performance. Our results confirmed that the same amount spent collecting data can have vastly different results. The second experiment analyzed the effect of token sequence overlap between target and training data since a natural and intuitive approach to building a dataset from scratch for a task would be having the task tokens occur in the training data. Our experiments showed that token sequence overlap was not the primary factor influencing model performance. Our work provides a counter-intuitive insight into building speech recognition datasets from scratch in low-resource settings and shows the need for further investigation.

Keywords: Token, Sequence, Automatic Speech Recognition, Dataset, Low-Resource Language

1. Introduction

The existence of large pre-trained models creates the possibility of collecting a relatively small dataset in a low-resource language to build a speech recognition model for a specific task. This can be useful in solving a wide range of problems. For instance, in some African countries, people who are illiterate or semi-literate are unable to use applications which require reading and typing in English. To solve this, applications could be created in different domains, from agriculture to finance, which allow such people to interact by speaking in their native language. Technologists could collect a dataset relevant to the application they wish to create and fine-tune a large pre-trained model using the dataset.

Typically, a technologist can only collect a relatively small dataset due to the costs involved. The first question we ask in this paper is whether different datasets of the same size could produce significantly different results. In other words, if a technologist is constrained to 4 hours of data collection, is there some 4-hour dataset that would produce a significantly better result than another 4-hour dataset? Our experiments and results show that it is indeed the case that some 4-hour dataset used in training performs better than others.

There are some situations where it is obvious that one dataset would outperform another dataset of the same size. For instance, in the extreme case, supposing one dataset consisted of repeating a single word while the other dataset comprised regular sentences, it is obvious that the latter would perform better. The question though is whether there would be a difference between two training datasets created in a reasonable manner for a target task. This was the focus of our first set of experiments.

We conducted an experiment with 45 randomly sampled training dataset. All trials were trained on the same base model. The evaluated output of these models on the same test dataset showed differences in Word Error Rate. For Twi, the Word Error Rate range was from 64.40% to 74.88%, whereas that of Lingala was from 32% to 35.89%.

One would expect that a strong predictor of which training dataset performs better would be how frequently the words needed for the task appear in the training dataset. For example, in finance, if recognizing the word *money* is important for the task, then a training dataset with more occurrences of the word *money* should do better in recognizing *money*. In general, one would expect that the more overlap there is between the token sequence used in training and the token sequence to be recognized at inference, the better the performance would be.

In this research, we studied the impact of token sequence overlap between the dataset for training and testing to see if this expectation of sequence overlap between training and test holds. Surprisingly, we found that token sequence overlap is not a strong predictor of the success of a training dataset.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3075–3082
11–16 May 2026.

For instance, a word occurring more frequently in a training dataset does not at all mean that the dataset would be better at predicting the word at test time, relative to a dataset in which the word occurs less frequently. We conducted experiments for token sequences beyond words such as: character bi-grams, trigram and 4-grams. In these experiments as well we found that token sequence overlap is not a strong predictor of which training dataset performs better. We experimented with two low resource languages: Twi and Lingala and the result still held.

The contributions of this paper are as follows:

1. We motivate the need to study approaches for building speech recognition datasets from scratch in low-resource settings. We show that for the same resource spent creating a dataset for a task, one could potentially have better results depending on the approach. While there is existing work on cost-effective selection of speech to label(Abraham et al., 2020; Ardila et al., 2020; de Vries et al., 2014; Azunre and Ibrahim, 2023), there is little to no work on creating a speech dataset from scratch.

2. We show that what is, perhaps, a most intuitive idea when building a task-specific dataset from scratch - maximizing token sequence overlap between what is needed for the task and the training dataset - does not yield results as one would expect.

3. We contribute a new dataset: the Twi dataset used in testing our hypothesis.

2. Related Work

2.1. Low Resource Automatic Speech Recognition(ASR) models

Wav2vec 2 is a self-supervised ASR model pretrained on billions of hours of unlabelled audio data (Baevski et al., 2020). This and similar models have transformed ASR model development for low-resource languages due to its high performance in low-resource settings; requiring a small amount of training dataset in the target language to achieve effective results.

2.2. Dataset creation

2.2.1. Supervised Dataset Creation

Creating a dataset is a fundamental step in ASR model building. Existing datasets such as Librispeech(Panayotov et al., 2015) and GigaSpeech(Chen et al., 2021) are large supervised datasets created from audiobooks, YouTube, and podcast resources. The audio-to-transcript pair is generated from extensive available data.

2.2.2. Dataset Creation in Low Resource-Settings

The process of dataset creation for low-resource languages comes with these challenges: data scarcity(Imam et al., 2025) and high cost of supervised dataset creation (Azunre and Ibrahim, 2023). The research by (Azunre and Ibrahim, 2023) contributed the first-ever dataset for Dagbani, a low-resource language in the Northern region of Ghana. They curated a total of 9 hours of supervised dataset. Their unique approach for data collection was through the Spell4Wiki app, an app designed to record audio data of Dagbani texts on Wikimedia. While their approach efficiently manages data collection, it does not provide a cost-effective method for task-specific data collection.

2.2.3. Dataset Selection in Low-Resource Settings

Recent research has been exploring unsupervised approaches for selecting task-specific datasets for labelling in speech recognition. Given a large library of audios and a task of interest, techniques are provided for optimally selecting which audios to label for best results.

(Zheng et al., 2023) utilized unsupervised models such as wav2vec-U 2.0 (Schneider et al., 2019) and HuBERT(Hsu et al., 2021) to extract discrete representations like phonemes, k-mean IDs, and word representations from a universal pool of unlabelled audio data. They identified audio data for labelling using contrastive loss, which measures the differences in perplexity between two language models(LMs). The first LM was trained on discrete representations from a general domain, while the second model was a fine-tuned version of the first LM on discrete representation of domain-specific data.

(Park et al., 2022) introduced the concept of submodular function for data labelling. The submodular function selected target-specific frames based on the principle of diminishing returns, which ensures diversity in data selection. The function scored selected audio data for labelling by comparing the ratio of contrastive loss from the model trained on the target dataset to the loss from the model trained on the general training dataset. Contrastive loss measures the similarities and dissimilarities of data points in a vector space.

While this research addresses cost-effective building of datasets in low-resource settings, it assumes a large pool of unlabelled audio for the target language which is potentially suitable for the task at hand. It also assumes that labelling existing audio is the way to build a new dataset. From our experience working with Ghanaian languages, these assumptions are not always valid. Some-

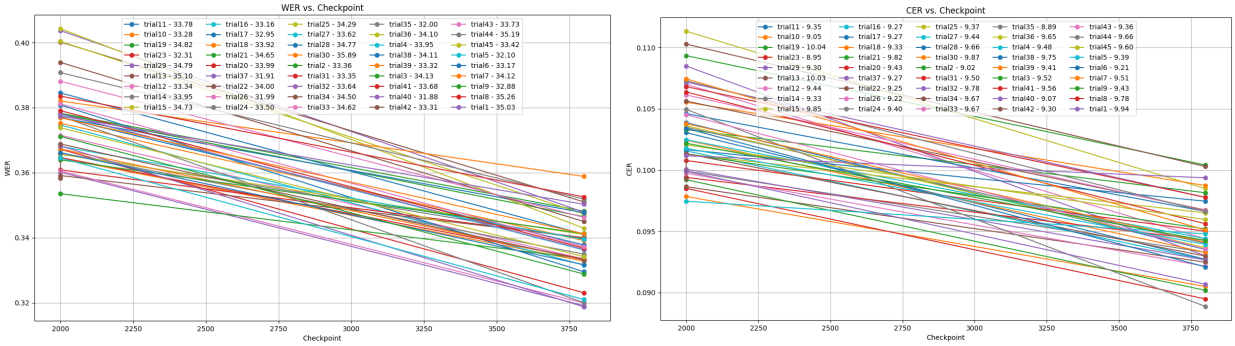


Figure 1: WER(left) and CER(right) performance for 45 trials fine-tuned on different datasets on the same domain for Lingala

times, there is the need to create an audio dataset from scratch for the task of interest. To the best of our knowlege, no work currently addresses doing this cost-effectively.

3. Our Datasets

Our research focused on two low-resource languages: Lingala and Twi. Lingala is a Bantu language spoken primarily in the Democratic Republic of the Congo (DRC) (Samarin, 1991). Twi, on the other hand, is spoken by the Akan ethnic group in Ghana (Dolphyne, 1986). We created a new dataset for Twi, called the alpha dataset. For Lingala, we used an open source supervised dataset obtained from Mendeley data (Kimanuka et al., 2023). The Lingala dataset had 32 unique speakers: 13 male and 19 female speakers(Kimanuka et al., 2024).

3.1. Alpha Dataset

The alpha dataset is a supervised speech-to-text dataset. The sentences collected for this dataset were from Twi novels and conversations. Previous datasets were either domain-specific or focused solely on the Twi Bible (Asamoah Owusu et al., 2022; George, 2025). To enable the efficient collection of Twi audio data, a mobile app was developed specifically for audio recording. This app automatically selected transcripts for users from a pool of Twi texts. Users then took audio recordings of the transcriptions with an option to playback and listen to their recorded audio. [1]

Users log in to the mobile app using their credentials, and each is assigned a random token as a unique ID to help protect their identity. The final curated alpha dataset comprises 4 hours of audio from 42 unique speakers, including 22 females and 20 males. In total, there are 134 unique Twi sentences.

[1] Link to Dataset: https://adr.ashesi.edu.gh/datasets/26

4. Experiment Set 1: Same Size, Different Results

To begin, we tested the hypothesis that different datasets of the same size could produce significantly different results when fine-tuned on a pre-trained model in the same manner. We wanted to clarify and demonstrate that when creating a speech dataset for a specific task from scratch, one could get better results for the same amount of time or money spent depending on how they go about determining which audio people should record.

There are some situations where it is obvious that one dataset would outperform another dataset of the same size.

For example, consider an extreme case where one dataset consists of a single word repeated over and over, while the other contains coherent, natural sentences; it is obvious the latter would yield better performance. However, the real question is whether meaningful differences still arise between two well-constructed datasets that are both reasonably designed for the same target task. For example, both datasets are gender balanced, have a comparable number of speakers and variation in sentences, and so on. This was the focus of the set of experiments performed here.

4.1. Experiments

4.1.1. Dataset Pre-processing

Each low-resource dataset used was prepared and processed following the steps below:

1. The train data was randomly shuffled using a seed number corresponding to the trial number. For example, trial one was randomly shuffled with seed one.
2. We selected 60% of the new randomly shuffled training dataset.
3. The selected 60% subset from the train data set was further split into a 80% train and a 20% validation set.

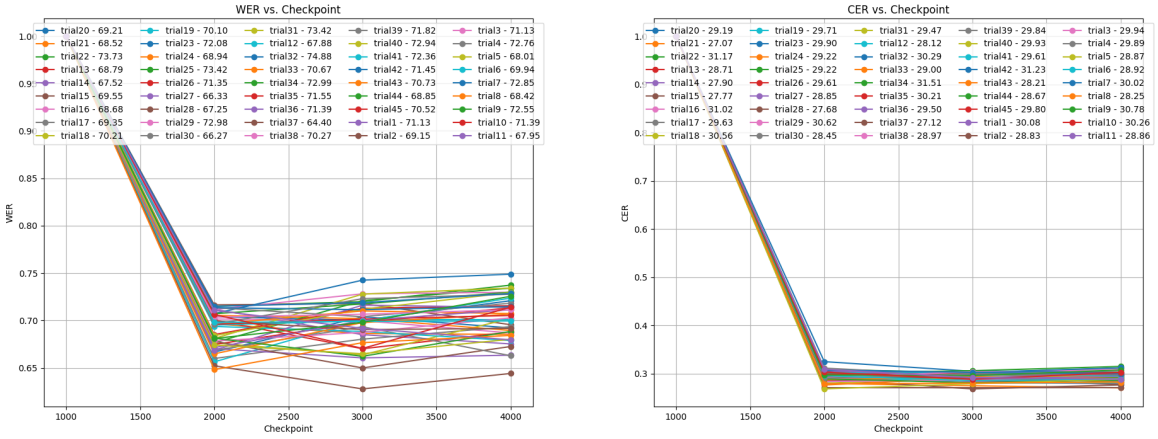


Figure 2: WER (left) and CER (right) performance for 45 trials fine-tuned on different datasets in the same domain for Twi

4. The audio information and transcripts of each subset were stored in well-labelled text files.
5. The dataset was pre-processed by removing unwanted special character and converting numbers to words.
6. Dataset pairs with either corrupted audio files and empty transcriptions were removed.
7. Audio data were sampled at 16 kHz.

This process was followed for each trial. In all, we run 45 trials for each language.

4.1.2. Model and Evaluation

The model used for the experiment was the Facebook's wav2vec large lv60 pre-trained model(Baevski et al., 2020). All subsets of labelled audio data were fine-tuned on the pre-trained wav2vec large lv60 model. A total of 90 models were evaluated, 45 for Lingala and 45 for Twi. The experiment was conducted on a Quadro RTX 6000 GPU. Each trial model was evaluated on the same test dataset. Specifically, models from trials 1-45 of Twi were evaluated on the same Twi test dataset from the Alpha dataset created. The performance of the trials were evaluated with two main metrics: the Word Error Rate(WER) and Character Error Rate(CER).

4.1.3. Results

In Figure 2 and 1, we observed a WER range from 32% to 35.89% across the 45 trials of Lingala, while, for Twi, we observed a WER range from 64.40% to 74.88% respectively. The 3.89% difference in WER for Lingala and a 10.48% difference in WER for Twi show that the same size of training dataset could produce different results even though both were collected in a reasonable manner. What is it that makes one dataset better at predicting the target since they are of the same size and the same source?

5. Experiment Set 2: Token Sequence Overlap

When creating a speech dataset from scratch for a specific task, one has the kind of sentences they want to be able to recognize. An intuitive approach to maximize performance would be to ensure that the audio which is collected contains as much as possible the sentences we want to be able to recognize. In other words, it is intuitive to seek to maximize the word sequence (token sequence) overlap between the dataset which is collected and what one seeks to predict. We performed experiments to determine if this intuition is correct.

5.1. Experimentation

5.1.1. Experiment

Token sequence is referred to as a unique word, pairs of characters (bi-gram), a sequence of three characters (tri-gram) or a sequence of four characters (4-gram). We conducted an experiment to analyse the effect of token sequence overlap between the training and test datasets on the performance of an ASR model. We evaluated the trial models through the following process:

1. We selected the **best** and the **worst-performing** trials for each language.
2. We determined all the **unique words** in the test set. *Lingala had a total of 1108 and Twi had a total of 129*.
3. We determined the occurrence of the test word in the training datasets of the best and worst

trials, as shown in the table 1, 2.

4. We repeated the process above for character bi-gram, trigram, and 4-gram sequence.

Test Word	TW	F37	F32	R37	R32
owo	3	2	2	50%	0%
ase	3	96	100	100%	100%
fie	8	38	41	87.5%	75%
awie	3	3	1	33.3%	0%
oman	2	27	37	100%	100%

Test Word Test word.
TW number of times Test word occurs in test set .
F37 number of times Test word occurs in training dataset of **trial 37**.
F32 number of times Test word occurs in training dataset of **trial 32**.
R37 percentage of correct predictions of Test word in trial 37.
R32 percentage of correct predictions of Test word in trial 32.

Table 1: The table illustrates how frequently a Twi test word occurs in the training set of trial 37 and 32, and the percentage correctly predicted in each trial.

Test Word	TW	F35	F30	R35	R30
kodiongo	15	1	2	46.7%	20%
bango	27	129	123	96.2%	96.2%
barumbu	26	4	4	23.1%	23.1%
tango	18	60	61	44.4%	38.9%
bato	24	85	86	83.3%	95.8%

Test Word Test word.
TW number of times Test word occurs in test set.
F35 number of times Test word occurs in training dataset of **trial 35**.
F30 number of times Test word occurs in training dataset of **trial 30**.
R35 percentage of correct predictions of Test word in trial 35.
R30 percentage of correct predictions of test word trial 30.

Table 2: The table illustrates how frequently a Lingala test word occurs in the training set of trial 35 and 30, and the percentage correctly predicted in each trial.

5.1.2. Analysis of Token Sequence Patterns

Table 1 and Table 2 show the different patterns in token sequence overlap between the test set and the corresponding training dataset, as well as the predictions of the test dataset across the selected trials. For example, the word "fie" in Twi had a higher word occurrence in the training dataset of trial 32 relative to trial 37, but trial 32 struggled to correctly predict the word at inference. In contrast, the word "owo" had the same token sequence occurrence in the training dataset of both trials, but it was predicted correctly more in trial 32. Similar patterns were observed in Lingala[2]. The word "kodiongo" had a higher token sequence occurrence in the training dataset of trial 30 relative to trial 35, yet it struggled to correctly predict the word in trial 30.

Twi	
Sequence Type	**Number of observations**
Words	9
Bi-gram	45
Tri-gram	92
4-gram	58
Lingala	
words	92
bi-gram	102
tri-gram	376
4-gram	556

Table 3: A situation where a sequence occurs more in a trial and fares better in predicting the sequence.

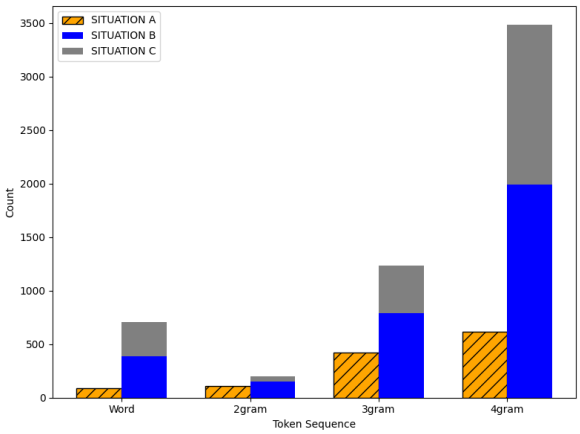
Twi	
Sequence Type	**Number of Observations**
Words	25
Bi-gram	83
Tri-gram	176
4-gram	102
Lingala	
words	387
bi-gram	156
tri-gram	837
4-gram	2051

Table 4: A situation where a sequence occurs more in a trial, but that trial fares worse in predicting the sequence.

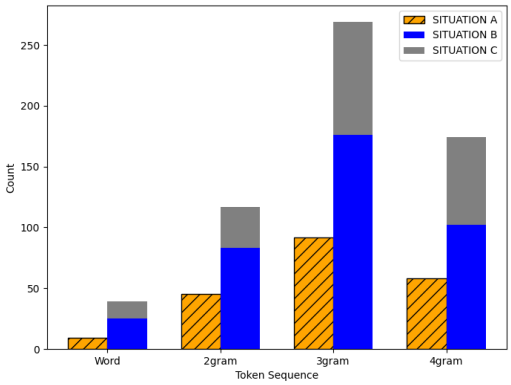
Twi	
Sequence Type	**Number of Observations**
words	14
bi-gram	34
tri-gram	93
4-gram	72
Lingala	
words	312
bi-gram	50
tri-gram	496
4-gram	1590

Table 5: A situation where a sequence occurs more in a trial, but that trial performs the same as other trial predictions.

These observations prompted further experiments to gain a better understanding of token sequence overlap across four main situations:



(a) For Lingala, instances where a test token sequence appears more frequently in the training data and the model makes more correct predictions (indicated by the striped orange bars for Situation A) occur less often than the other scenarios (Situations B and C, shown in blue and grey).



(b) For Twi, instances where a test token sequence appears more frequently in the training data and the model makes more correct predictions (indicated by the striped orange bars for Situation A) occur less often than the other scenarios (Situations B and C, shown in blue and grey).

Twi	
Sequence	**Number of Observations**
words	21
bi-gram	7
tri-gram	74
4-gram	149
Lingala	
words	148
bi-gram	51
tri-gram	213
4-gram	607

Table 6: A situation where a sequence occurs the same in both trials, but prediction success differs.

1. **SITUATION A**: A situation where a sequence occurs more frequently in a trial training dataset, and that trial performs better in predicting the sequence[3]. For example, the word "*awie*" appeared 3 times in training dataset of trial 37 and 1 time in trial 32. At inference, *awie* was predicted correctly 1 time by trial 37 and 0 times in trial 32. This situation describes what one would expect intuitively.

2. **SITUATION B**: A situation where a sequence occurs more in a trial training dataset, but that trial fares worse in predicting the sequence[4]. For example the word "*fie*" appeared 38 times in the training dataset of trial 37 and 41 times in trial 32. At inference, *fie* was predicted correctly 7 times by trial 37 and 6 times in trial 32.

3. **SITUATION C**: A situation where a sequence occurs more in a trial training dataset, but that trial performs the same as other trial predictions[5]. For example, the word "*bango*" appeared 123 times in the training dataset of trial 30 and 129 times in trial 35. At inference, *bango* was predicted correctly 26 times by the trial 35 and 26 times by trial 30.

4. **SITUATION D**: A situation where a sequence occurs the same in both trials training dataset, but the prediction success is different for each trial[6]. For example the word "*owo*" appeared 2 times in the training dataset of trial 37 and 2 times in trial 32. At inference, *owo* was predicted correctly 1 time by trial 37 and had 0 correct predictions in trial 32.

5.2. Results

The results across the 4 situations are presented in tables 3,4,5,6.

To answer the question of whether the token sequence overlap occurs more in the best trial, we visualized this bar plot:

1. A bar plot of the total occurrence of situation A[3] against the sum of the total occurrence of situation B[4] and the total occurrence of situation C[5].

The outputs from the graphs 3a, 3b justify the fact that token sequence occurrence in the training dataset does not necessarily predict model performance on that token sequence. The graphs show that there are more situations where trials with fewer sequence occurrences in the training dataset (situation B and C) correctly predicted the token sequences relative to the situations where the trial had more sequence occurrences in its training dataset (situation A).

6. Conclusion

In this paper, we show that one of the most intuitive approaches to building task-specific speech recognition datasets from scratch - maximizing token sequence overlap between the train dataset and target task - does not always produce the best results. This calls for the development of principled approaches to building task specific speech datasets from scratch for low-resource languages. Improving upon dataset building approaches could lead to the creation of models or systems with performance on the level of more expensive datasets at a fraction of the resources used. We also contribute a new Twi speech recognition dataset.

7. Ethics statement

An IRB approval was obtained for the creation of the Twi dataset. Participants signed an informed consent form agreeing to contribute to the dataset creation, knowing the dataset would be disseminated in a way that protected their privacy and minimized the possibility of identification. Participants received compensation for their contributions.

8. Limitations

Although we analyse the model's performance on words and different character n-grams, we did not perform experiments to explore the effects of phonemes in the different languages to better understand their role in the models performance. To the best of our knowledge, there is no phonemizer for Lingala to enable this task. Additionally, resource constraints in the form of limited compute and dataset size prevented us from experimenting to determine whether or not the pattern holds for larger datasets.

9. Bibliographical References

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819–2826, Marseille, France. European Language Resources Association.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus.

Paul Azunre and Naafi Dasana Ibrahim. 2023. Breaking the low-resource barrier for dagbani ASR: From data collection to modeling. In *4th Workshop on African Natural Language Processing*.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

G. Chen, S. Chai, G. Wang, J. Du, W. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, M. Jin J. Zhang, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Y. Zhao, and Z. Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

N. J. de Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. de Wet, E. Barnard, and A. de Waal. 2014. A smartphone-based asr data collection tool for under-resourced languages. *Speech Communication*, 56:119–131.

F. A. Dolphyne. 1986. The languages of the akan peoples. *Research Review*, 2(1):1–22.

Mark Hasegawa-Johnson and Camille Goudeseune. 2017. g2ps: Data and code for grapheme-to-phoneme transducers. https://github.com/uiuc-sst/g2ps. GitHub repository, original commit in 2017. Accessed 8 October 2025.

W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

S. H. Imam, T. D. Belay, K. Y. Husse, I. S. Ahmad, I. Abdulmumin, H. A. Umar, M. Y. Bello, J. Nakatumba-Nabende, S. M. Yimam, and S. H. Muhammad. 2025. Automatic speech recognition (asr) for african low-resource languages: A systematic literature review.

U. Kimanuka, C. wa Maina, and Büyük Osman. 2024. Speech recognition datasets for low-resource congolese languages. *Data in Brief*, 52:109796.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, page 5206–5210. IEEE.

C. Park, R. Ahmad, and T. Hain. 2022. Unsupervised data selection for speech recognition with contrastive loss ratios. In *ICASSP 2022 -*

2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), page 8587–8591. IEEE.

W. J. Samarin. 1991. The origins of kituba and lingala. *Journal of African Languages and Linguistics*, page 47–78.

S. Schneider, A. Baevski, R. Collobert, and M. Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition.

Z. Zheng, Z. Ma, Y. Wang, and X. Chen. 2023. Unsupervised active learning: Optimizing labeling cost-effectiveness for automatic speech recognition.

10. Language Resource References

D. Asamoah Owusu, A. Korsah, B. Quartey, S. Nwolley Jnr., D. Sampah, D. Adjepon-Yamoah, and L. Omane Boateng. 2022. Financial-inclusion speech dataset. https://github.com/Ashesi-Org/Financial-Inclusion-Speech-Dataset. GitHub repository, CC-BY-4.0, accessed 2025-10-16.

Kojo George. 2025. asante-twi-tts dataset. https://huggingface.co/datasets/kojo-george/asante-twi-tts. Cc-by-sa-4.0, accessed 2025-10-16.

Kimanuka, Ussen and Maina, Ciira wa and Büyük, Osman. 2023. *Speech Recognition Datasets for Congolese Languages*. Mendeley Data.