Self-supervised and multilingual learning applied to the Wolof, Swahili and Fongbe

Prestilien Djionang Pindoh, Paulin Melatagia Yonta

Self-supervised and multilingual learning applied to the Wolof, Swahili and Fongbe

Prestilien Djionang Pindoh[1] **and Paulin Melatagia Yonta**[1,2]

[1]Department of Computer Sciences, University of Yaoundé I, Yaoundé, Cameroon
[2]IRD, UMMISCO, F-93143, Bondy, France

*E-mail : prestilienpindoh@gmail.com, paulinyonta@gmail.com

Abstract

Under-resourced languages encounter substantial obstacles in speech recognition owing to the scarcity of resources and limited data availability, which impedes their development and widespread adoption. This paper presents a representation learning model that leverages existing frameworks based on self-supervised learning techniques—specifically, Contrastive Predictive Coding (CPC), wav2vec, and a bidirectional variant of CPC—by integrating them with multilingual learning approaches. We apply this model to three African languages: Wolof, Swahili, and Fongbe. Our evaluation of the resulting representations in a downstream task, automatic speech recognition, utilizing an architecture analogous to DeepSpeech, reveals the model's capacity to discern language-specific linguistic features. The results demonstrate promising performance, achieving Word Error Rates (WER) of 61% for Fongbe, 72% for Wolof, and 88% for Swahili. These findings underscore the potential of our approach in advancing speech recognition capabilities for under-resourced languages, particularly within the African linguistic landscape.

Keywords

Self-supervised learning ; Multilingual representation learning ; Automatic speech recognition ; Under-resourced languages.

I INTRODUCTION

Speech recognition technology has seen significant advancements in recent years, leading to numerous applications such as virtual assistants, transcription services, and voice command systems. However, these advancements have predominantly benefited languages with abundant resources, extensive datasets, and well-developed linguistic models. In contrast, many African languages, classified as under-resourced, have been largely sidelined in the development of speech recognition systems due to the limited availability of high-quality data and resources.

Developing effective speech recognition systems for under-resourced languages like Wolof, Fongbe, and Swahili is crucial for broadening access to technology and ensuring these languages are digitally preserved. The scarcity of large labeled datasets presents a unique challenge for applying traditional supervised learning methods to these languages.

In recent years, self-supervised learning, particularly the contrastive learning approach [7], has emerged as a promising paradigm for learning representations from unlabeled data. Contrastive learning is a general framework that aims to construct a feature space wherein related points are

brought into proximity while unrelated points are distanced. Several methods have been developed within this framework, including Wav2Vec [8], Contrastive Predictive Coding (CPC) [6], and Bidirectional CPC [10]. CPC, in particular, represents an unsupervised machine learning approach that seeks to derive meaningful, higher-level semantic representations from unprocessed data such as text and audio.

Nevertheless, these methods remain data-intensive for learning high-quality representations, posing a significant challenge for languages with limited resources. Kawakami et al. [10] addressed this issue by employing multilingual learning, an approach that aims to learn a shared representation of speech from data originating from diverse languages. They demonstrated the efficacy of learning representations with a large amount of multilingual data (predominantly English) and then evaluating the transferability of these representations to other under-resourced languages, including Wolof, Swahili, and Fongbe, yielding promising results.

In this work, we aim to construct a representation model specifically tailored to African languages, capable of capturing the underlying features unique to each language. To this end, we leverage CPC, wav2vec, and bidirectional CPC within the context of multilingual learning. Our study focuses on three under-resourced languages: Wolof, Fongbe, and Swahili. This approach allows us to augment the available training data and investigate how these methods capture the distinctive characteristics of each language by evaluating them in the context of automatic speech recognition tasks.

The fundamental question we address is whether combining under-resourced languages that share similar linguistic and phonetic characteristics enhances the quality of features extracted for each language individually. Our objective is to improve speech processing tasks for these languages.

The remainder of this paper is structured as follows: Section II presents an overview of the self-supervised learning approaches employed in our framework. Section III delineates the contrastive stacking model for multilingual learning with three African languages. Section IV is devoted to our experiments and the presentation of results. Finally, we conclude and discuss future directions in Section V.

II RELATED WORK

Self-supervised learning has emerged as a revolutionary paradigm in machine learning, enabling algorithms to extract meaningful representations from unlabeled data [9]. This approach is particularly valuable in contexts where labeled data is scarce or expensive to obtain, such as in the domain of under-resourced languages. Self-supervised learning leverages the inherent structure of the data to generate its own supervisory signals, distinguishing it from traditional unsupervised learning methods that primarily seek to uncover hidden patterns or structures within the data itself.

Contrastive learning aims to learn discriminative representations by distinguishing between pairs of similar (positive) examples and pairs of dissimilar (negative) examples [7]. A positive example typically constitutes a slightly different view (e.g., a minor transformation) of the same original example. Negative examples are different examples drawn from the same dataset. The objective is to learn robust representations that minimize the distance between positive examples in the representation space while maximizing the distance between negative examples.

Conversely, generative learning approaches train models to generate new data that resembles

the training dataset. These models attempt to capture the underlying distribution of the training data, enabling the sampling of new, synthetic data points. Generative models find widespread applications in text generation, image synthesis, and audio production..

A speech signal encodes more information than text, such as speaker identity and prosodic features, which makes it harder to generate. However, to generate all details of the input, the model must encode all information in the speech signal. Hence, a model that learns to perfectly reconstruct its input may not necessarily have learned to isolate the features of interest and will encode redundant information for a given downstream task[11]. Given these considerations, we have opted for a contrastive approach in our work. This choice is motivated by the contrastive method's capacity to capture discriminating features in the data without the need for complete signal reconstruction.

The InfoNCE metric is a commonly used measure in contrastive learning to assess the quality of learned representations. It is based on the principle of maximizing the normalized mutual information between positive pairs and negative pairs.

The mathematical formula for the InfoNCE metric is as follows:

$$\text{InfoNCE} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(x_i, x_i^+))}{\exp(\text{sim}(x_i, x_i^+)) + \sum_{j=1}^{K}\exp(\text{sim}(x_i, x_i^{-j}))} \tag{1}$$

This cost function measures the probability that the positive example (x^+) is closer to the original example (x) than any negative example (x^-). The rationale behind this formulation is that the model should maximize the probability of selecting the positive among the negatives.

Aaron et al. [6] proposed contrastive predictive coding (CPC) (Figure 1), a self-supervised machine learning method for extracting high-level semantic representations from raw data, such as text or sound. In the case of sound, this refers to the audio waveform of the signal, meaning the numerical features of the audio. It comprises two networks: an encoder that generates latent representations and an autoregressive network that generates contextual representations. The cost function used here is a mutual information (MI) lower bound called InfoNCE. This approach has produced impressive results for speaker identification and phoneme classification tasks. However, it was not used for speech recognition.

In general, CPC [6] work as follows: given x a signal sliced into frames, the encoder network g_{enc} encodes the signal x at each time step t, yielding the latent representations z_t. Then the second autoregressive network g_{ar} produces the contextual representations c_t, taking into account the previous representations up to time step t, and predicts the future z_{t+k} from c_t, while maximizing the MI (Mutual Information) between c_t and the predicted z_{t+k}. Both representations, z_t and c_t, can be used as input for the automatic speech recognition model. In our case, we specifically use the contextual representations, c_t.

The Contrastive Predictive Coding (CPC) uses a lower bound of mutual information called InfoNCE, formulated as follows:

Let $Z = \{z_1, \ldots, z_N\}$ be a set containing one positive sample from the probability distribution $p(z_{t+k}|c_t)$ and $N-1$ negative samples from a "noise" distribution $p(z)$, the approximate lower bound is written as:

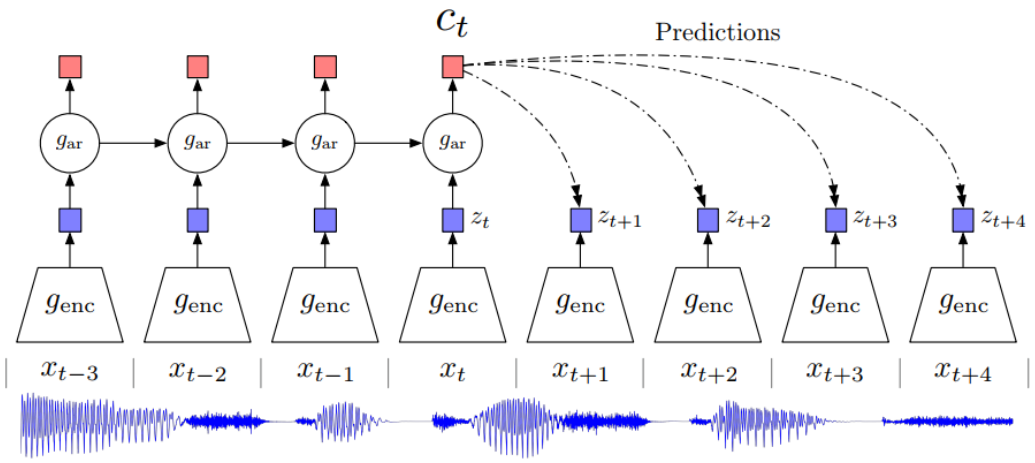


Figure 1: Overview of CPC, the proposed representation learning approach. [6]

$$L_N = E_z[\log \frac{f_k(c_t, z_{t+k})}{\frac{1}{N}\sum_{\tilde{z}\in Z} f_k(c_t, \tilde{z})}] \tag{2}$$

Where $f_k(c_t, z_{t+k})$ is a scoring function that measures the similarity between the representations of c and z [6] given by:

$f_k(c_t, \tilde{z}) = \exp(c_t^T W_k \tilde{z})$

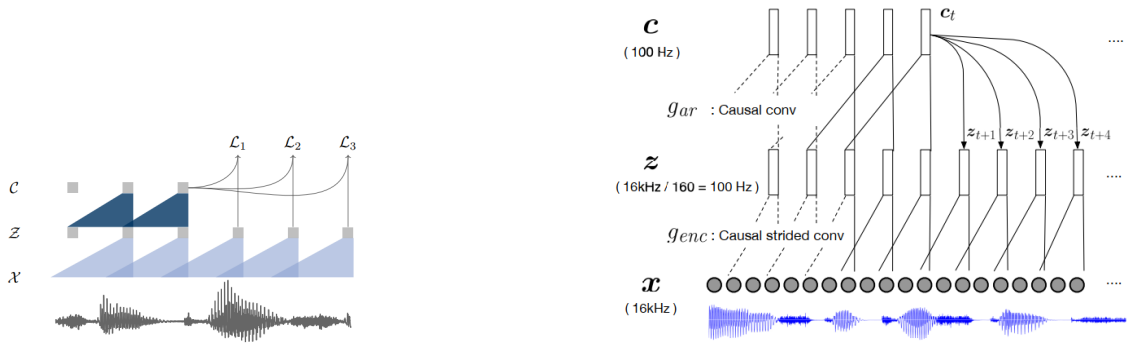
where c_t is the context at time t, $\tilde{z}$ is a negative sample from $p(z)$, and W_k is a specific weight matrix at time shift k.

The loss function to maximize is the sum of InfoNCE lower bounds (L_N) for each time step t and each time shift k, i.e., $\sum_t \sum_k L_N$

Schneider et al [8] proposed Wav2Vec (Figure 2(a)), which is almost identical to CPC, but they use NCE (Noice Contrastive Estimation) as loss function instead of InfoNCE (a lower bound estimation of MI), and the encoder and context networks are made up of layers of causal convolutions, unlike CPC which uses convolutional networks on the encoder and a layer of GRU (Gated Recurrent Units) in the autoregressive network. Causal convolutional layers are used to ensure that predictions at time t depend only on inputs up to time t, and not on future inputs. This allows for properly modeling the sequential nature of audio signals. The model is optimized to solve a next-time step prediction task. Apart from this, they took the experiments further in the automatic speech recognition task. However, these experiments were conducted only in English and were not used in the context of low-resource languages.

Bidirectional Contrastive Predictive Coding (Figure 2(b)) was used for multilingual learning of speech representation, with the main aim of assessing the robustness of representations to domain changes and the transferability of representations to other poorly endowed languages. For the BCPC which combines forward and backward directions, the InfoNCE loss is calculated as the sum of InfoNCE in both directions:

$$L_N^{BCPC} = L_N^{fwd} + L_N^{bwd}$$



(a) (b)

Figure 2: **(a).** Illustration of pre-training from audio data X which is encoded with two convolutional neural networks that are stacked on top of each other [8]. **(b).** Bidirectional CPC Illustration [10]

where L_N^{fwd} and L_N^{bwd} are the forward and backward InfoNCE loss, respectively. They pre-trained their model on 8,000 hours of multilingual audio data, mainly in English (95%). They evaluated the transferability of learned representations to Wolof, Fongbe, and Swahili, but these languages were not included in the data used for pre-training. The Word Error Rates (WER) obtained for Wolof, Swahili, Fongbe and Amharic were 55%, 70%, 57%, and 65% respectively.

These advancements in self-supervised learning techniques have paved the way for improved speech recognition capabilities, particularly in the challenging domain of under-resourced languages. Our work builds upon these foundations, adapting and extending these methods to address the unique challenges posed by African languages with limited available data.

III STACKING CONSTRASTIVE MODELS

To address the challenge of limited data availability in acquiring high-quality representations, we have devised and implemented a multilingual representation learning approach (Figure 3). Our methodology leverages the self-supervised learning techniques proposed by Aaron et al. [6], Schneider et al. [8], and Kawakami et al. [10]. Given that the original CPC and wav2vec models were not initially designed for multilingual contexts, we have adapted the multilingual representation learning approach proposed by Kawakami et al. [10] and applied it to all three methods before their utilization in subsequent tasks.

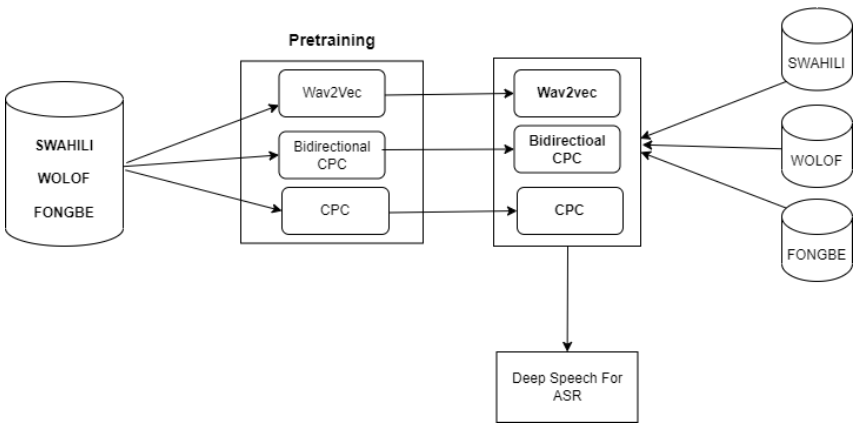


Figure 3: Architecture of the model

Our study focuses on speech recognition datasets in three African languages, collected as part of the ALFFA [1] (African Languages in the Field: Speech Fundamentals and Automation) project: Fongbe (A. A Laleye et al. [5]), Swahili (Gelas et al. [1]), and Wolof (Gauthier et al. [4]). These languages are characterized by unique phonological properties, including pitch harmony and distinct phonetic inventories. It is noteworthy that these African languages are resource-constrained, with less than 20 hours of transcribed speech data available for each.

Once we obtained these datasets, we followed the following steps for representation learning as shown in Figure 3:

- The data was preprocessed and normalized to reduce noise and mitigate volume differences between the data. Preprocessing involves applying filters to reduce background noise and normalize volume across recordings. This step ensures consistency in input data for the learning model.
- We used the data set to train the representation learning algorithms, namely CPC, wav2vec, and bidirectional CPC. We thus obtained three pre-trained models in all languages.
- To predict representations for a single language, we fed the data from that language into one of the pre-trained models. The predicted representation vectors were then used to train the automatic speech recognition (ASR) model.

Contrastive learning, using a mixture of audio from several languages, presents an interesting opportunity for multilingual speech representation learning. This approach is based on the principle that, according to the contrastive learning methodology, positive sampling is independent of a specific language.

By combining data from diverse languages during the contrastive learning process (Figure 3), the model is exposed to a rich variety of acoustic and linguistic structures. This exposure facilitates the emergence of general and shared representations. Consequently, the model can learn to extract speech-relevant features from different languages while simultaneously capturing the similarities and differences between them.

In conclusion, the application of contrastive learning to a mixture of multilingual data offers a promising approach for developing speech representations that are adaptable to different languages. This method maximizes the utilization of available data for resource-constrained languages, potentially leading to more robust and versatile speech recognition systems.

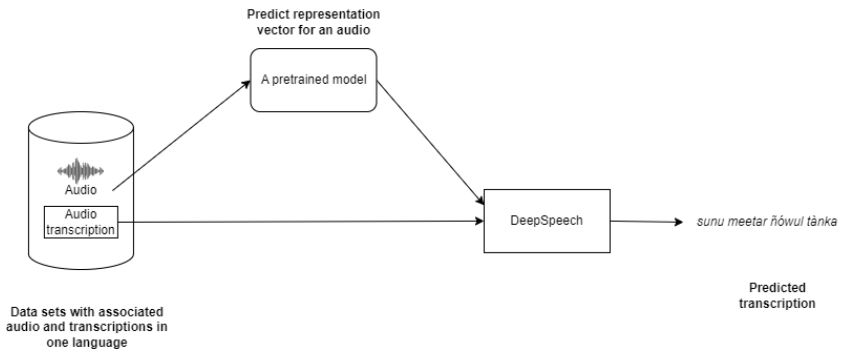


Figure 4: Architecture of the model

Once the representations are learned, we use the obtained models to predict the representation vectors of the data (Figure 4), which will be used as inputs to the speech recognition algorithm

[1] http://alffa.imag.fr

DeepSpeech2 Small (a reduced version of DeepSpeech2 [3]). DeepSpeech2 is an automatic speech recognition model based on deep learning. It uses a recurrent neural network (RNN) with long short-term memory (LSTM) to perform the task of transcribing speech into text. It uses the spectrograms (numerical representation vector) as model inputs but in our case we have replaced the spectrograms by the representation vectors predicted by the pre-trained models.

During the training phase of the ASR model, the parameters of the representation models were kept constant, while the parameters of the speech recognition models were supervised and trained using one dataset at a time [10]. The models were evaluated using the standard Word Error Rate (WER) metric on held-out test data. The WER is calculated as the number of words incorrectly predicted divided by the total number of words. It measures the accuracy of the model's transcription, where a lower value indicates better performance of the automatic speech recognition model. To evaluate the impact of multilingual learning, we also trained monolingual representation learning models to formulate hypotheses and compare the results.

IV EXPERIMENTS AND RESULTS

4.1 Dataset details

ALFFA (African Languages in the Field: Speech Fundamentals and Automation) is a research project aimed at collecting high-quality linguistic data for understudied African languages using speech recognition to enable speakers of these languages to access information in their language. The ALFFA project has gathered data for several African languages, including Wolof, Swahili, and Fongbe, available on GitHub[2]. These languages are considered low-resource, hence the inherent interest in studying them. Moreover, from a more objective standpoint, these languages were chosen because there existed significant and usable public data sources for each.

4.1.1 Wolof

The Wolof language, primarily spoken in Senegal, is represented in a dataset featuring 21 hours of recorded speech from 18 distinct speakers. Wolof distinguishes itself with an extensive vowel system and a unique possession indicator, "ñu." Noteworthy features include its nominal class system and the use of specific prefixes based on these classes, contributing to Wolof's linguistic distinctiveness. This dataset is meticulously curated, drawing from diverse sources such as proverbs, narratives by Kesteloot and Dieng (1989), transcriptions of healer debates, a song titled "Baay de Ouza," and two dictionaries: "Dictionnary wolof-french" and "Dictionnary french-wolof." Additionally, data from the Bible, Wikipedia, and the Universal Declaration of Human Rights enrich the corpus (Table 1).

	Male	Female	Utterances	Duration
Training	8	6	13998	16h49
Development	1	1	2000	2h12
Testing	1	1	2000	2h20
Total	**10**	**8**	**17998**	**21h21**

Table 1: Description of Wolof dataset

[2]ALFFA: https://github.com/besacier/ALFFA_PUBLIC

4.1.2 *Fongbe*

Fongbe, also known as Fon, is predominantly spoken in Benin, where it holds the status of a national language. It boasts a tonal system influencing word meaning through varying pitch. The Fongbe dataset consists of 9 hours of recorded speech from 29 different speakers (Table 2). The textual corpus content includes primarily biblical texts, a variety of texts related to daily life, the Universal Declaration of Human Rights, texts on education, songs, and folktales.

	Speakers	Utterances	Duration
Training	25	8234	7h 35
Testing	4	2163	1h45
Total	**29**	**10397**	**9h20**

Table 2: Description of Fongbe dataset

4.1.3 *Swahili*

Swahili, spoken primarily in East Africa, serves as the national and official language in both Kenya and Tanzania. It extends its influence to countries like Uganda, Rwanda, Burundi, and parts of the Democratic Republic of the Congo. As a Bantu language, Swahili is characterized by an agglutinative structure, incorporating prefixes and suffixes for grammatical nuances. The Swahili dataset consists of 12 hours (Table 3) of recorded speech from various speakers, representing diverse socio-economic and educational backgrounds. Recordings originate from web-based news broadcasts and Swahili text extracted from informational websites. Manual transcriptions are available for each recording.

	Utterances	Duration
Training	10180	10h
Testing	1991	2h
Total	**12171**	**12h**

Table 3: Description Swahili dataset

For implementation, we used an unofficial code from GitHub [3] for the CPC model that we modified for our application case and the official implementation [4] of the Wav2Vec model for our experiments. Since no bidirectional CPC model implementation was available, we modified the code of the Wav2Vec model to make it bidirectional[5] (as mentioned in the article by Kawakami et al. [10]). Similarly for DeepSpeech, we used an implementation available on the Keras framework site [6] that we adapted to use the numerical vectors predicted by our pre-trained models rather than spectrograms.

4.2 Descriptions and parameters of the representation and ASR learning algorithms

The parameters of the algorithms we used to learn the representations are shown in Table 4.

[3] https://github.com/jefflai108/Contrastive-Predictive-Coding-PyTorch
[4] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
[5] https://shorturl.at/b7V2v
[6] https://keras.io/examples/audio/ctc_asr/

For speech recognition, we used DeepSpeech2 Small. The features of this model are as follows: the model features two 2d-convolutions with kernel sizes (11, 41) and (11, 21) and stride sizes (2, 2) and (1, 2), as well as a one-way recurrent neural network (GRU) above the output of the convolution layers. A linear transformation and a softmax function are applied to predict frame-level character probabilities. Training is performed using a batch size of 8 and a learning rate of 0.0001.

We use WER, which is a popular metric for evaluating the performance of automatic speech recognition algorithms because it takes into account errors such as word substitution, insertion, and deletion, which are common types of errors in speech recognition. A lower WER value indicates better performance, as it means fewer errors in the output generated by the speech recognition system, and its value is expressed as a percentage.

The Word Error Rate (WER) is calculated as follows:

$$\text{WER} = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{Total number of reference words}}$$

4.3 Monolingual learning

Table 5 presents the results of our experiments utilizing a single language for representation learning, subsequently employed to train the automatic speech recognition model. The metric used is the Word Error Rate (WER), expressed as a percentage.

	CPC	BCPC	Wav2vec	Kawakami et al.
WOLOF	**80**	85.7	80.1	**55**
FONGBE	83	81	**68**	**57**
SWAHILI	96	95	**90**	**70**

Table 5: WER for monolingual learning

The results demonstrate the relatively limited performance of the CPC, BCPC, and wav2vec self-supervised approaches when applied to the low-resource languages Wolof, Fongbe, and Swahili. For Wolof, CPC and wav2vec achieve comparable word error rates of 80%, while BCPC exhibits slightly diminished performance at 85.7%. In the case of Fongbe, wav2vec distinguishes itself with a WER of 68%, outperforming both BCPC (81%) and CPC (83%). However, all three approaches encounter significant challenges with Swahili, yielding WERs exceeding 90% (CPC 96%, BCPC 95%, wav2vec 90%).

These findings underscore the formidable challenges posed by the scarcity of data for these under-resourced languages. Simultaneously, they hint at performance disparities potentially linked to the unique linguistic characteristics of each language. The superior performance of wav2vec on Fongbe, for instance, may be attributed to its ability to better capture the tonal features characteristic of this language.

4.4 Multilingual learning

Table 6 presents the results obtained using representations learned across all datasets for the downstream task, trained with the previously described parameters:

	CPC	BCPC	Wav2vec	Kawakami et al
WOLOF	78.9	75.2	**72.6**	**55**
FONGBE	82	77.5	**61**	**57**
SWAHILI	95	93	**88**	**70**

Table 6: WER for multilingual learning

The multilingual learning approach, wherein representations are learned jointly across several languages, yields substantial performance gains. This improvement suggests an enrichment of representations through the sharing of information across languages, enabling the model to better capture both common and language-specific linguistic features.

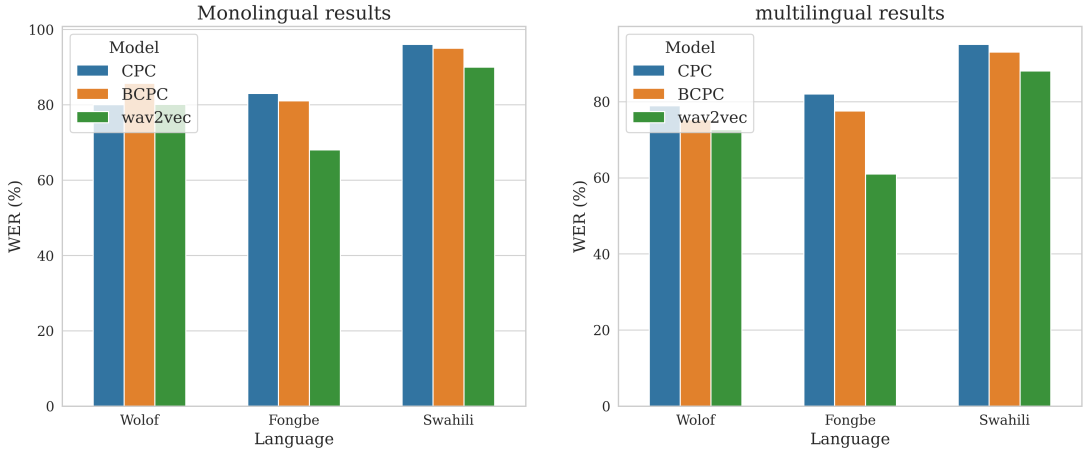


Figure 5: The contribution of multilingual learning to automatic speech recognition in a low-resource context.

Notably, wav2vec demonstrates remarkable improvement in the multilingual setting, achieving WERs of 72.6% for Wolof, 61% for Fongbe, and 88% for Swahili. This significant enhancement over its monolingual performance (80.1%, 68%, and 90% respectively) underscores the potential of multilingual learning in low-resource scenarios.

The bidirectional CPC (BCPC) also shows considerable improvement in the multilingual context, particularly for Wolof and Fongbe. This suggests that the bidirectional nature of BCPC allows it to capture more comprehensive linguistic information when exposed to multiple languages simultaneously.

Interestingly, while all approaches benefit from multilingual learning, the magnitude of improvement varies across languages. Fongbe, a tonal language, appears well-modeled by multilingual wav2vec (61% WER), suggesting that this approach may be especially effective for languages with complex tonal systems. However, the performance gap between Swahili and the other two languages highlights that not all languages benefit equally from the same multilingual framework.

Despite these improvements, it is important to note that there remain substantial performance gaps between these results and state-of-the-art systems for well-resourced languages (typically <5% WER). This disparity highlights the ongoing challenges in automatic speech recognition for under-resourced languages and underscores the need for continued research and development.

In comparing our results to those reported by Kawakami et al., we observe that while our monolingual models generally underperform their reported figures (e.g., 55% WER for Wolof compared to our best monolingual result of 80%), our multilingual learning results are more competitive. This suggests that our approach to multilingual learning enables better capture of common linguistic features and effectively compensates for the lack of monolingual data. However, disparities persist in some languages such as Swahili.

These results collectively demonstrate the potential of multilingual contrastive learning in improving automatic speech recognition for under-resourced languages. They also highlight the importance of considering language-specific characteristics in the development of speech recognition systems for diverse linguistic contexts.

V CONCLUSION

This study demonstrates that pre-training on limited yet diverse datasets, consolidated within multilingual models, can yield robust representations that effectively capture the common and unique acoustic characteristics of individual languages. Our approach shows promise in building a single model for multiple languages, potentially streamlining the development process for under-resourced languages.

While our work has made significant strides, there remain avenues for future exploration. Expanding the training data corpus across a broader spectrum of African languages could lead to even more diverse and tailored speech representations. Additionally, enhancing speech recognition models with language-specific knowledge has the potential to yield more accurate transcriptions, better reflecting the unique structures of each language.

In conclusion, this research demonstrates that self-supervised and multilingual learning approaches are promising solutions for overcoming the challenges of limited data in speech recognition for under-resourced languages. With continued efforts to expand training data and refine model architectures, we believe that these techniques will play a critical role in making speech recognition systems more inclusive and accessible to speakers of all languages.

REFERENCES

Publications

[1] H. Gelas, L. Besacier, and F. Pellegrino. “Developments of Swahili resources for an automatic speech recognition system.” In: *SLTU*. 2012, pages 94–101.

[2] Y. Bengio, A. Courville, and P. Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pages 1798–1828.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin". In: *International conference on machine learning*. PMLR. 2016, pages 173–182.

[4] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui. "Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof". In: *10th Language Resources and Evaluation Conference (LREC 2016)*. 2016.

[5] F. A. LAleye, L. Besacier, E. C. Ezin, and C. Motamed. "First automatic fongbe continuous speech recognition system: Development of acoustic models and language models". In: *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. 2016, pages 477–482.

[6] A. v. d. Oord, Y. Li, and O. Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[7] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. "A theoretical analysis of contrastive unsupervised representation learning". In: *arXiv preprint arXiv:1902.09229* (2019).

[8] S. Schneider, A. Baevski, R. Collobert, and M. Auli. "wav2vec: Unsupervised pre-training for speech recognition". In: *arXiv preprint arXiv:1904.05862* (2019).

[9] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. "A survey on contrastive self-supervised learning". In: *Technologies* 9.1 (2020), page 2.

[10] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. v. d. Oord. "Learning robust and multilingual speech representations". In: *arXiv preprint arXiv:2001.11128* (2020).

[11] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, et al. "Self-supervised speech representation learning: A review". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pages 1179–1210.

	CPC	BCPC	Wav2vec
Number of negative samplings	10	10	10
Timestep	12	12	12
Audio window length (frames)	20480	150000	150000
Optimizer	Adam	Adam	Adam
Initial learning rate	0.0004	0.0001 with gradient clipping, maximum norm of 5.0	0.0001
Batch size	8	128	8
Loss	InfoNCE	InfoNCE (sum of backward and forward)	InfoNCE
Encoder size	512	512	512
Encoder Layers	convolutional layers with kernel size (10, 8, 4, 4, 4) and strides (5, 4, 2, 2, 2).	5 causal convolutional layers with kernel size (10, 8, 4, 4, 4, 1, 1) and strides (5, 4, 2, 2, 2, 1, 1).	5 causal convolutional layers with kernel size (10, 8, 4, 4, 4) and strides (5, 4, 2, 2, 2).
Decoder size	256	(fwd: 256, bwd: 256): concatenation (c = [fwd, bwd] => 512)	512
Decoder Layer	An unidirectional GRU layer	13 causal convolutional layers with kernel size 1, 2, ..., 13 and stride 1.	9 causal convolutional layers with kernel size 3 and stride 1.

Table 4: Parameters of the models