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 Pindoh Djionang[1] **and Paulin Yonta Melatagia**[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 face significant challenges in speech recognition due to limited resources and data availability, hampering their development and usage. In this paper, we present a speech recognition model built upon existing frameworks based on self-supervised learning (Contrastive Predictive Coding (CPC), wav2vec and bidirectional version of CPC) by combining them with multilingual learning. This model is experimented on Wolof, Swahili, and Fongbe which are African languages. The results of our evaluation of representations on the automatic speech recognition task, using a similar architecture to DeepSpeech, highlight the model's capability to discriminate language-specific linguistic features, achieving a Word Error Rate (WER) of 61% for Fongbe, 72% for Wolof and 88% for Swahili

Keywords

Self-supervised learning ; Multilingual representation learning ; Automatic speech recognition ; Low endowed languages.

I INTRODUCTION

The technology of voice recognition has experienced significant advancements in recent years, offering numerous applications such as virtual assistants, transcription services, and voice command systems. However, these advancements have mainly benefited languages with abundant resources and extensive datasets, relegating under-resourced languages to the sidelines. These low-resourced languages, particularly African languages, face major challenges in developing accurate and efficient voice recognition systems due to limited data availability and the absence of dedicated linguistic models.

In recent years, self-supervised learning, in particular the contrastive learning approach [7] has emerged as a promising approach for learning representations from unlabeled data. Contrastive learning is a general framework that attempts to learn a feature space in order to bring together points that are related and discard points that are unrelated. Several methods exist today, such as Wav2Vec [8], CPC [6] and BCPC [10]. CPC represents an unsupervised machine learning approach that aims to derive meaningful, higher-level semantic representations from unprocessed data like text and audio. However, these methods also remain data-intensive for learning high-quality representations, which is a challenge for languages with limited resources. Kawakami et al. [10] use multilingual learning, which is an approach to machine learning that aims to

learn a shared representation of speech from data coming from different languages. They learn representations with a large amount of multilingual data (With a high quantity of English) and then evaluate the transferability of these representations to other (sparsely endowed) languages, including Wolof, Swahili, and Fongbe, which yielded interesting results.

In this work, we aim to build a representation model specific to African languages capable of capturing the underlying features of each language. To this end, we use CPC, wav2vec, and bidirectional CPC in the context of multilingual learning. We use three low-resource languages: Wolof, Fongbe and Swahili. This allows us to have more data for training and to see how these methods manage to capture the unique characteristics of each language by evaluating them in the task. The underlying problem is to assess whether combining low-resource languages that share similar linguistic and phonetic characteristics improves the quality of features extracted for each language individually. This aims to enhance speech processing tasks.

The rest of this paper is structured as follows: we will first present the self-supervised learning approaches in our framework in section II, then we present the contrastive stacking model for multilingual learning with three African languages in section III. Section IV is devoted to the experiments and the presentation of the results. We conclude this document in Section V.

II RELATED WORK

Self-supervised learning is a machine learning method that allows an algorithm to learn from data without labels [9]. Self-supervised learning uses the structure of the data to generate its labels. This is different from unsupervised learning, which seeks to discover hidden patterns or structures in the data itself. This approach has seen growing interest in recent years thanks to the advent of new techniques, such as contrastive and generative learning.

Contrastive learning is an approach aimed at learning 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 bring positive examples closer together in representation space while pushing negative examples apart. On the other hand, generative learning is a machine learning approach in which the model is trained to generate new data that resembles the training dataset. The generative model attempts to model the distribution of the training data so that new data can be sampled from this distribution. Generative models are often used for the generation of text, images, sounds, etc.

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]. So we opted for a contrastive approach that would allow us to capture discriminating features in the data.

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^-) (Figure **??**). 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.

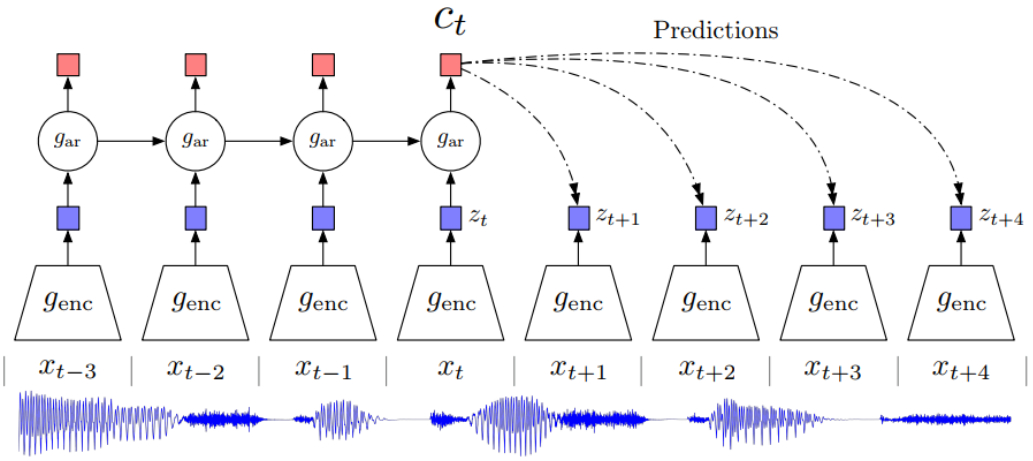
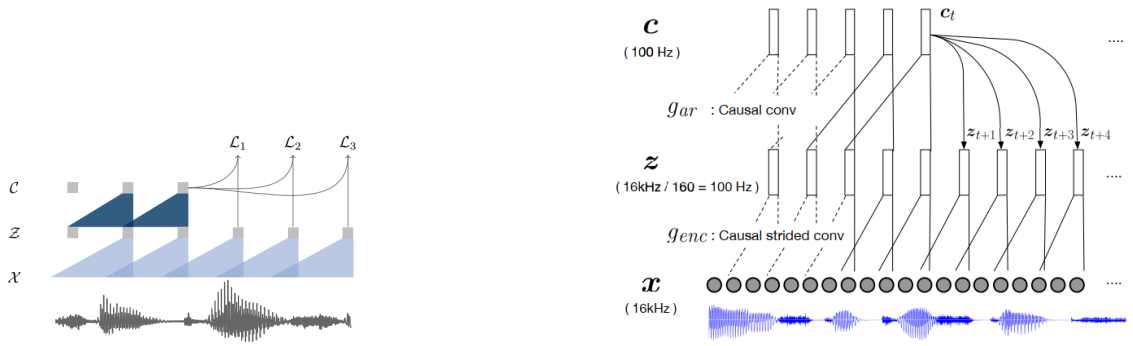


Figure 1: Overview of CPC, the proposed representation learning approach. [6]

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:



(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]

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

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:

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

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.

III STACKING CONSTRASTIVE MODELS

To overcome the issue of limited data for obtaining high-quality representations, we implemented multilingual representation learning (Figure 3). For this purpose, we used the self-supervised learning methods proposed by Aaron et al. [6], Schneider et al. [8], and Kawakami et al. [10]. Since the CPC and wav2vec models were not originally trained in a multilingual context, we adopted the multilingual representation learning approach proposed by Kawakami et al. [10], which we applied to all three methods before using them for subsequent tasks.

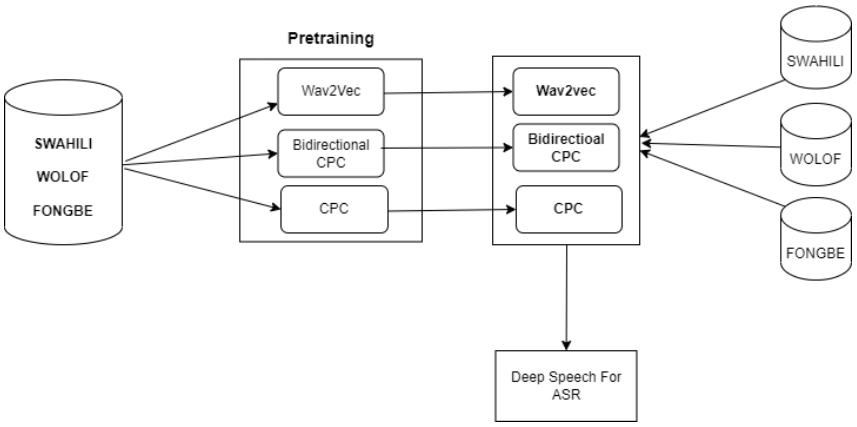


Figure 3: Architecture of the model

We used speech recognition datasets in three African languages, collected as part of the ALFFA project [1]: 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 such as pitch harmony and distinct phonetic inventories. It is worth noting that these African languages have limited resources available, with less than 20 hours of transcribed speech data for each of them.

Once we obtained these datasets, we followed the following steps for representation learning as shown in Figure 4:

- 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.

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

By combining data from different languages during contrastive learning (Figure 3), the model is exposed to a variety of acoustic and linguistic structures, which promotes the emergence of general and shared representations. As a result, the model can learn to extract speech-relevant features from different languages, while capturing the similarities and differences between them.

In conclusion, the use of contrastive learning with a mixture of multilingual data offers a promising approach to develop speech representations that are adapted to different languages, while maximizing the use of available data for low-resourced languages.

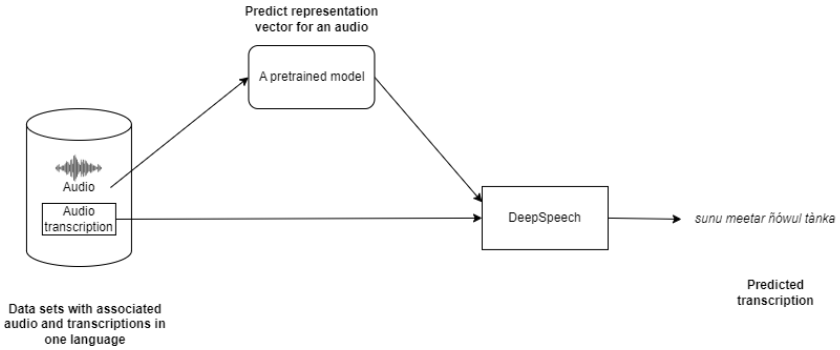


Figure 4: Architecture of the model

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 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

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

languages were chosen because there existed significant and usable public data sources for each. The Wolof language, primarily spoken in Senegal, is represented in a dataset featuring 21 hours of recorded speech from 18 distinct speakers.

4.1.1 Wolof

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

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

We used an unofficial implementation 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.

For speech recognition, 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

The results of the experiments we conducted using a single language to learn the representations and then using them to train the automatic speech recognition model are presented in Table 5. The metric used is the Word Error Rate (WER) and is expressed as a percentage.

[3] https://github.com/jefflai108/Contrastive-Predictive-Coding-PyTorch

[4] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

[5] https://drive.google.com/drive/folders/1IWhFaQzTPADUJo0j8tnSzorvcFaUloji?usp=sharing

[6] https://keras.io/examples/audio/ctc_asr/

	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 show relatively limited performance of the CPC, BCPC and wav2vec self-supervised approaches on the low-resource languages Wolof, Fongbe and Swahili. For Wolof, CPC and wav2vec achieve a word error rate (WER) of 80%, while BCPC performs slightly less well at 85.7%. For Fongbe, wav2vec stands out with 68% WER, ahead of BCPC (81%) and CPC (83%). However, all three approaches struggle on Swahili, with WERs above 90% (CPC 96%, BCPC 95%, wav2vec 90%). These results underline the challenges posed by the lack of data for these poorly endowed languages, while hinting at differences in performance potentially linked to the linguistic specificities of each.

4.4 Multilingual learning

The results obtained using the learned representations on all of these datasets for the downstream task (trained with the previous parameters) are as the table 6:

	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

Multilingual learning, where representations are learned jointly across several languages, brings substantial gains. This suggests an enrichment of representations through the sharing of information across languages, enabling common and specific linguistic features to be better captured. However, there are still performance gaps between languages, probably due to their distinct linguistic properties. Fongbe, a tonal language, seems to be particularly well modeled by wav2vec multilingual (61% WER).

First of all, we note that self-supervised learning approaches are able to obtain relevant representations for the automatic speech recognition task, even in low-resource languages such as Wolof, Fongbe, and Swahili.

The multilingual approach, where representations are learned jointly on several languages, brings substantial performance gains over monolingual learning. This suggests that sharing information across languages enriches learned representations, and better captures both common and language-specific linguistic features.

There are, however, differences in performance between languages. Swahili seems particularly difficult to model for all approaches. This could be due to the linguistic specificities of this Bantu language, such as its affix-rich agglutinative structure. Conversely, Fongbe and Wolof benefit greatly from multilingual learning, particularly with wav2vec, which achieves WERs of 61% and 72.6% respectively. The tonal nature of Fongbe and Wolof could explain these better performances.

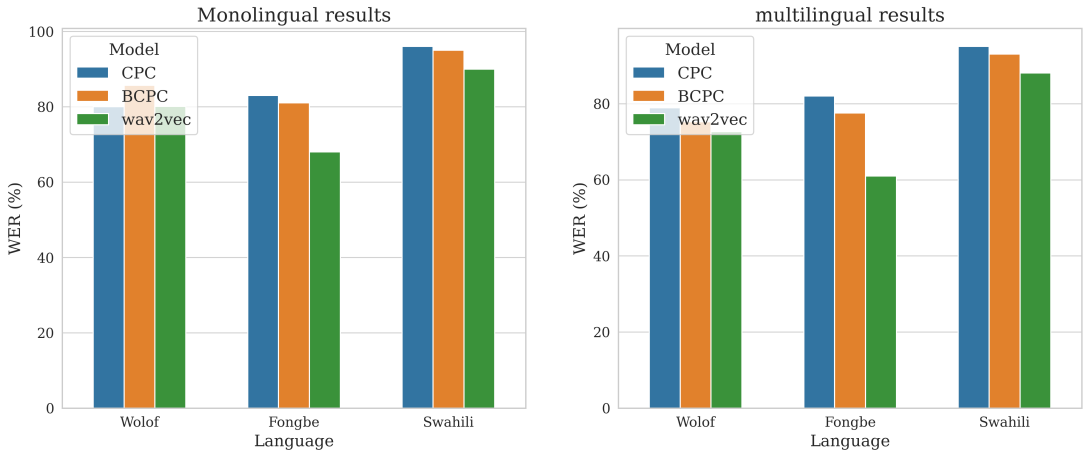


Figure 5: The contribution of multilingual learning to automatic speech recognition in a low-resource context.

Concerning the various self-supervised approaches, wav2vec and BCPC seem superior, especially on Fongbe. Their purely convolutional architectures may be better suited to capturing the acoustic features of this tonal language. CPC, with its recurrent components, could have more difficulty.

The performance of our monolingual learning models generally falls below those reported by Kawakami et al., likely due to the limited quantity of training data at our disposal. For instance, their model achieved a word error rate (WER) of 55% for Wolof, compared to 80% for our best monolingual models. Similarly, for Fongbe, their approach yielded a 57% WER, compared to 68% for our best wav2vec model. However, our results in multilingual learning are more competitive, with a WER of 75.2% for Wolof compared to 55% in Kawakami et al., and 61% for Fongbe compared to 57% in their study, suggesting that multilingual transfer enables better capture of common linguistic features and compensates for the lack of monolingual data, although disparities persist for some languages such as Swahili.

Finally, it should be pointed out that even the best results (61% WER for Fongbe with multilingual wav2vec) are still a long way from state-of-the-art systems for well-endowed languages (typically <5% WER). This shows that improvements are still needed, notably by increasing the training data, incorporating specific language models, and exploring new self- and multilingual learning techniques.

These results also show that the pretraining of the small amounts of available data, pooled in multilingual models, can allow to have a robust model that captures the acoustic characteristics of each language.

V CONCLUSION

This paper shows that pre-training on small amounts of available data, pooled in multilingual models, can allow having a robust model that captures the acoustic characteristics of each language. Thus, one could build a single model for several languages instead of building a different model for each language.

Although our work has made significant progress in self-supervised learning of speech representations for under-resourced African languages, there are still many perspectives to explore in order to further improve the results. First, it would be interesting to expand the training data

corpus for representation learning models by collecting more speech data for a broader set of African languages. This would make it possible to obtain even more diverse speech representations tailored to a wider range of under-resourced languages. Next, the speech recognition models, especially DeepSpeech, could be enhanced by developing language models specific to each language. By incorporating language-specific knowledge, it would be possible to achieve more accurate and higher-quality transcriptions.

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