# Using Radio Archives for Low-Resource Speech Recognition: Towards an Intelligent Virtual Assistant for Illiterate Users

**Moussa Doumbouya,**[1] **Lisa Einstein,**[1,2] **Chris Piech**[2]
[1] GNCode [2] Stanford University
moussa@gncode.org, lisae@stanford.edu, piech@cs.stanford.edu

## Abstract

For many of the 700 million illiterate people around the world, speech recognition technology could provide a bridge to valuable information and services. Yet, those most in need of this technology are often the most underserved by it. In many countries, illiterate people tend to speak only low-resource languages, for which the datasets necessary for speech technology development are scarce. In this paper, we investigate the effectiveness of unsupervised speech representation learning on noisy radio broadcasting archives, which are abundant even in low-resource languages. We make three core contributions. First, we release two datasets to the research community. The first, West African Radio Corpus, contains 142 hours of audio in more than 10 languages with a labeled validation subset. The second, West African Virtual Assistant Speech Recognition Corpus, consists of 10K labeled audio clips in four languages. Next, we share West African wav2vec, a speech encoder trained on the noisy radio corpus, and compare it with the baseline Facebook speech encoder trained on six times more data of higher quality. We show that West African wav2vec performs similarly to the baseline on a multilingual speech recognition task, and significantly outperforms the baseline on a West African language identification task. Finally, we share the first-ever speech recognition models for Maninka, Pular and Susu, languages spoken by a combined 10 million people in over seven countries, including six where the majority of the adult population is illiterate. Our contributions offer a path forward for ethical AI research to serve the needs of those most disadvantaged by the digital divide.

## Introduction

Smartphone access has exploded in the Global South, with the potential to increase efficiency; connection; and access to critical health, banking, and education services (MHealth 2011; Harris and Cooper 2019; Avle, Quartey, and Hutchful 2018). Yet, the benefits of mobile technology are not accessible to most of the 700 million illiterate people around the world who, beyond simple use cases such as answering a phone call, cannot access functionalities as simple as contact management or text messaging (Chipchase 2006).

Speech recognition technology could help bridge the gap between illiteracy and access to valuable information and services (Medhi et al. 2011), but the development of speech recognition technology requires large annotated datasets. Unfortunately, languages spoken by illiterate people who would most benefit from speech recognition technology tend to fall in the "low resource" category, which in contrast with "high resource" languages, have few available datasets. Transfer learning, transferring representations learned on high resource unrelated languages, has not been explored for many low resource languages (Kunze et al. 2017). Even if transfer learning can help, labeled data is still needed to develop useful models.

This data deficit persists for multiple reasons. Developing commercial products for languages spoken by smaller populations can be less profitable and thus less prioritized. Furthermore, people with power over technological goods and services tend to speak data-rich languages themselves, potentially leading them to insufficiently consider the needs of users who do not (Ogbonnaya-Ogburu et al. 2020).

We take steps toward developing a simple, yet functional intelligent virtual assistant that is capable of contact management skills in Maninka, Susu and Pular, low-resource languages in the Niger Congo family. People who speak Niger Congo languages have among the lowest literacy rates in the world, and illiteracy rates are especially pronounced for women (see Fig. 1). Maninka, Pular, and Susu are spoken by a combined 10 million people, primarily in seven African countries, including six where the majority of the adult population is illiterate (Roser and Ortiz-Ospina 2016). We address the data scarcity problem by making use of unsupervised speech representation learning and show that representations learned from radio archives, which are abundant in many regions of the world with high illiteracy rates, can be leveraged for speech recognition in low resource settings.

**In this paper, we make three core contributions that collectively build towards the creation of intelligent virtual assistants for illiterate users:**

1. We present two novel datasets (i) the West African Radio Corpus and (ii) the West African Virtual Assistant Speech Recognition Corpus. These datasets of over 150 hours of speech increase the availability of resources for speech technology development for West African Languages.
2. We investigate the effectiveness of unsupervised speech

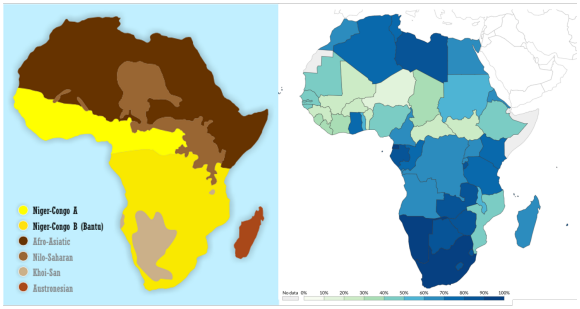



Figure 1: Speaking a language from the Niger Congo family correlates with low literacy rates. Left: Distribution of language families in Africa ("Niger-Congo" by Mark Dingemanse is licensed under CC BY 2.5). Right: Female adult literacy rates in 2015 (Our World in Data/World Bank).

representation learning from noisy radio broadcasting archives. We show that our encoder leads to a multilingual speech recognition accuracy similar to Facebook's baseline state-of-the-art encoder and outperforms the baseline on West African language identification by 13.94%.

3. We present the first-ever language identification and small vocabulary speech recognition systems for Maninka, Pular, and Susu. For all languages, we achieve usable performance (88.1% on automatic speech recognition).

The results presented are robust enough that our West African speech recognition software, in its current form, is ready to be used effectively in an intelligent virtual assistant capable of contact management skills for illiterate users.

The rest of this paper reads as follows. First, we formalize the problem of speech recognition for virtual assistants for illiterate users. Then we provide background information and summarize prior work. We then introduce the novel datasets. Next, we introduce the methodology for our intelligent virtual assistant and present our experimental results. We conclude with a discussion of results and future work.

## Contact Management Virtual Assistant

To demonstrate how speech recognition could enable the productive use of technology by illiterate people, we propose a simple yet functional virtual assistant capable of contact management skills in French, Maninka, Susu, and Pular. Fig. 2 illustrates the states and transitions of the virtual agent, the performance of which greatly depends on its ability to accurately recognize the user's utterances.

**Automatic Speech Recognition (ASR).** As demonstrated in Table 1 and Fig. 2, the recognition of a small utterance vocabulary covering wake words, contact management commands (search, add, update, delete), names, and digits is sufficient to make the assistant functional.

There are no existing speech recognition systems or data sets for Maninka, Pular, or Susu. Therefore, we first collected and curated the West African Virtual Assistant dataset, which contains the utterances described in Fig. 2

| Role | Utterance | VA State |
|---|---|---|
| User: | "Guru!" | |
| VA: | "Yes, what would you like to do?" | 1 |
| User: | "Add a contact" | |
| VA: | "What is the name of the contact?" | 2 |
| User: | "Fatoumata" | |
| VA: | "What is the phone number of Fatoumata?" | 4 |
| User: | "698332529" | |
| VA: | "Are you sure to add Fatoumata - 698332529 ?" | 5 |
| User: | "Yes" | |
| VA: | "OK. Done" | 6 |

Table 1: Example dialog between a user and the virtual assistant (VA). The last column shows the new state of the VA as a result of its interaction with the user.

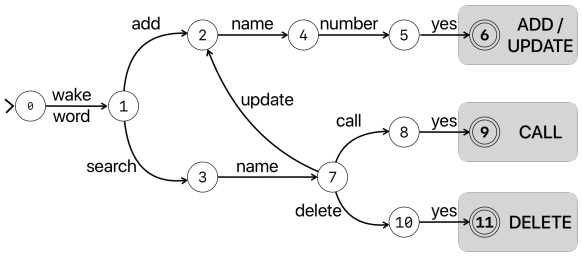



Figure 2: Simplified diagram of the states of the virtual assistant. In non-final states, it recognizes a subset of utterances, e.g., "add"/"search" in state 1. Upon transition to state 6, 9, or 11, it invokes the contact management application.

in French, Maninka, Pular, and Susu, before creating the speech recognition models.

Because of the small size of our dataset, we used wav2vec, the state of the art unsupervised speech representation learning method from Facebook (Schneider et al. 2019). We compared the baseline wav2vec model to its counterpart trained on the West African Radio Corpus we collected and conducted West African language identification experiments to validate the learned speech features.

**Language Identification (Language ID).** It is not completely clear what happens when speech representations learned from high resource languages (e.g., English for the baseline wav2vec) are used to solve speech recognition tasks on unrelated low resource languages such as Maninka, Pular, and Susu. To shed light on the semantics encoded by wav2vec features, we compare the difference in performance on a West African language identification task by the baseline wav2vec with its counterpart trained on the West African Radio Corpus and conduct a qualitative analysis of the acoustic features on which they focus.

## Prior Work

**User Interfaces for Illiterate People.** Many researchers agree on the importance of facilitating technology access in populations with low literacy rates by using speech recognition and synthesis, local language software, translation, accessibility, and illiterate-friendly software (Patra, Pal, and

Nedevschi 2009; Ho et al. 2009; Ahmed, Zaber, and Guha 2013). Medhi et al. confirmed illiterate users' inability to use textual interfaces and showed that non-text interfaces significantly outperform their textual counterparts in comparative studies (Medhi et al. 2011). Graphical content and spoken dialog systems have shown promise in allowing illiterate users to perform tasks with their phones or interact with e-government portals (Taoufik, Kabaili, and Kettani 2007; Medhi et al. 2011; Friscira, Knoche, and Huang 2012). Studies so far have relied on "Wizard of Oz" voice recognition, where the speech-to-text function is simulated with humans remotely responding to spoken instructions, rather than true ASR as we demonstrate here.

**Existing Speech Datasets for African languages.** Some work has been done to collect speech datasets in low-resource African languages. Documentation exists for three African languages from the Bantu family: Basaa, Myene, and Embosi (Adda et al. 2016). Datasets also exist for other African languages such as Amharic, Swahili, Wolof (Abate, Menzel, and Tafila 2005; Gelas, Besacier, and Pellegrino 2012; Gauthier et al. 2016). Several research efforts have focused on South African languages (van Niekerk et al. 2017; Nthite and Tsoeu 2020; Badenhorst et al. 2011). We were unable to identify any datasets that include the three Niger-Congo languages we focus on here.

**Exploiting "found" data, including radio broadcasting.** Cooper et al. explored the use of found data such as ASR data, radio news broadcast, and audiobooks for text-to-speech synthesis for low resource languages. However, the radio broadcast used is high quality and English language, as opposed to noisy and in low-resource languages (Cooper 2019). Radio Talk, a large-scale corpus of talk radio transcripts, similarly focuses on English language speakers in the United States (Beeferman, Brannon, and Roy 2019). Some research has focused on speech synthesis from found data in Indian languages (Baljekar 2018; Mendels et al. 2015). None of these found data projects include noisy radio data for low-resource languages, a data source that is abundant in many countries with low literacy rates since speech is the method by which citizens must consume information.

**Unsupervised speech representation learning.** Unsupervised speech representation learning approaches such as Mockingjay and wav2vec aim to learn speech representation on unlabeled data to increase accuracy on downstream tasks such as phoneme classification, speaker recognition, sentiment analysis, speech recognition, and phoneme recognition while using fewer training data points (Liu et al. 2020; Schneider et al. 2019).

In this work, we compare the baseline "wav2vec large" model, trained on LibriSpeech - a large (960 hours) corpus of English speech read from audiobooks (Panayotov et al. 2015) - to its counterpart we trained on a small (142 hours) dataset of noisy radio broadcasting archives in West African languages for the downstream tasks of language identification and speech recognition on West African languages.

**Transferring speech representations across languages.** It has been shown that speech representations learned from a high resource language may transfer well to tasks on unrelated low resource languages (Rivière et al. 2020). In this work, we compare such representations with representations learned on noisy radio broadcasting archives in low resource languages related to the target languages. We present quantitative results based on performances on downstream tasks, and a qualitative analysis of the encoded acoustic units.

## West African Speech Datasets

In this paper, we present two datasets, the West African Speech Recognition Corpus, useful for creating the speech recognition module of the virtual assistant described in the introduction section, and the West African Radio Corpus intended for unsupervised speech representation learning for downstream tasks targeting West African languages.

### West African Virtual Assistant Speech Recognition Corpus

The West African Speech Recognition Corpus contains 10,083 recorded utterances from 49 speakers (16 female and 33 male) ranging from 5 to 76 years old on a variety of devices. Most speakers are multi-lingual and were recorded pronouncing all utterances in their native language in addition to other languages they spoke. First names were recorded only once per speaker, as they are language independent.

Following the virtual assistant model illustrated in Fig. 2, the ASR corpus consists of the following utterances in French, Maninka, Susu, and Pular: a wake word, 7 voice commands ("add a person", "search a person", "call that", "update that", "delete that", "yes", "no"), 10 digits, "mom" and "dad". The corpus also contains 25 popular Guinean first names useful for associating names with contacts in a small vocabulary speech recognition context. In total, the corpus contains 105 distinct utterance classes.

82% of the recording sessions were performed simultaneously on 3 devices (one laptop, and two smartphones). This enables the creation of acoustic models that are invariant to device-specific characteristics and the study of sensitivity with respect to those characteristics.

Each audio clip is annotated with the following fields: Recording session, speaker, device, language, utterance category (e.g., add a contact), utterance (e.g., add a contact in Susu language), and the speaker's age, gender, and native language. Speakers have been anonymized to protect privacy. Table 2 (top) summarizes the content of the West African Virtual Assistant Speech Recognition Corpus.

### West African Radio Corpus

The West African Radio Corpus consists of 17,091 audio clips of length 30 seconds sampled from archives collected from six Guinean radio stations. The broadcasts consist of news and various radio shows in languages including French, Guerze, Koniaka, Kissi, Kono, Maninka, Mano, Pular, Susu, and Toma. Some radio shows include phone calls, background and foreground music, and various noise types.

**Dataset 1**
**West African Virtual Assistant Speech Recognition Corpus**

| Utterance Category | French | Maninka | Pular | Susu |
|---|---|---|---|---|
| Wake word | 66 | 95 | 67 | 109 |
| Add | 66 | 95 | 67 | 111 |
| Search | 66 | 95 | 67 | 111 |
| Update | 66 | 95 | 67 | 111 |
| Delete | 66 | 95 | 67 | 111 |
| Call | 66 | 95 | 64 | 111 |
| Yes | 66 | 95 | 67 | 111 |
| No | 66 | 95 | 67 | 111 |
| Digits (10) | 660 | 946 | 670 | 1,110 |
| Mom | 36 | 53 | 43 | 51 |
| Dad | 36 | 53 | 43 | 51 |
| **Total/Language** | 1,260 | 1,812 | 1,289 | 2,098 |
| Names (25) | | 3,624 | | |
| **Total** | | 10,083 | | |

**Dataset 2**
**West African Radio Corpus**

| | Clips | Duration |
|---|---|---|
| Unlabeled Set | 17,091 | 142.4 hours |
| Validation Set | 300 | 2.5 hours |
| **Total** | 17,391 | 144.9 hours |

Table 2: Description of collected datasets. **Dataset 1:** Record counts by utterance category in the West African Virtual Assistant Speech recognition corpus. We aggregated record counts for digits (10 per language) and names (25 common Guinean names). **Dataset 2:** West African Radio Corpus includes noisy audio in over 10 local languages and French collected from six Guinean radio stations.

Although an effort was made to filter out archive files that mostly contained music, the filtering was not exhaustive. Therefore, this dataset should be considered uncurated. Segments of length 30 seconds were randomly sampled from each raw archive file. The number of sampled segments was proportional to the length of the original archive, and amounts to approximately 20% of its length.

The corpus also contains a validation set of 300 audio clips independently sampled from the same raw radio archives, but not included in the main corpus. The validation clips are annotated with a variety of tags including languages spoken, the presence of a single or multiple speakers, the presence of verbal nods, telephone speech, foreground noise, and background noise among other characteristics.

# Method

## West African wav2vec (WAwav2vec)

To maintain comparability with wav2vec, WAwav2vec was obtained by training wav2vec as implemented in the fairseq framework (Ott et al. 2019) on the West African Radio Corpus. We used the "wav2vec large" model variant described in (Schneider et al. 2019) and applied the same hyperparameters, but we trained on 2 Nvidia GTX 1080 Ti GPUs instead of 16 GPUs as did (Schneider et al. 2019). We trained for 200k iterations (170 epochs) and selected the best checkpoint based on the cross-validation loss. The audio clips from the West African Radio Corpus were converted to mono channel waveforms with a sampling rate of 16 kHz and normalized sound levels. The baseline wav2vec and the WAwav2vec were used as feature extractors in all our experiments. We experimented with both their context (C) and latent (Z) features. We used quantitative and qualitative observations on the downstream tasks and analysis to make conclusions about the effectiveness of unsupervised speech representation learning and transfer learning in two settings: The first, where representations are learned from high-quality large-scale datasets in a high resource language not directly related to the target languages, and the second, where representations are learned from noisy radio broadcasting archives in languages related to target languages.

| Model | wav2vec | mel spectrogram |
|---|---|---|
| Language ID | 1,651 | 499 |
| ASR | 186,393 | 180,249 |

Table 3: Parameter counts of the CNNs for language identification and speech recognition using wav2vec features and mel spectrograms (respectively 512 and 128 dimensional).

## Neural Net for Virtual Assistant

We used the convolutional neural network architecture illustrated in Fig. 3 for both the language identification and the speech recognition experiments. Of its variants we explored, the following performed the best. The network comprises a 1x1 convolution followed by 4 feature extraction blocks. Each feature extraction block contains a 3x1 convolution, the ELU activation function (Clevert, Unterthiner, and Hochreiter 2015), a Dropout layer (Srivastava et al. 2014) and an average pooling layer with kernel size 2 and stride 2. The output of the last 3 feature extraction blocks are max pooled across the temporal dimension and then concatenated to make a fixed-length feature vector that is fed to the fully connected layer. This design allows extracting acoustic features at multiple scales and makes the neural network applicable to any sequence length. In order to mitigate overfitting issues, we apply Dropout in each of the convolutions feature extractors and before the fully connected layer.

The language identification model uses 3, 1, 3, 3 and 3 convolution channels, resulting in a 9 dimensional feature vector used for a 3 class classification. The ASR model uses 16, 32, 64, 128 and 256 convolutional chanels, resulting in a 448 dimensional representation used for a 105 class classification. In both experiments, we used the Adam optimiser (Kingma and Ba 2014) with learning rate $10^{-3}$ to minimize a cross entropy loss function. We also compared learned wav2vec features with spectrograms based on 128 mel filter banks. Table 3 summarizes the number of parameters of each of our models. We implemented and trained our speech

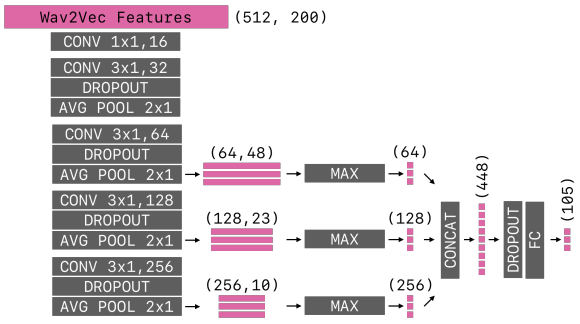



Figure 3: Architecture of the speech recognition CNN. The language identification CNN has a similar architecture.

.

recognition and language identification networks using PyTorch (Paszke et al. 2019).

## Results

In addition to establishing the baseline accuracies for speech recognition on the West African Virtual Assistant Speech Recognition Corpus and language identification on the validation set of the West African Radio Corpus, our experiments aimed at answering the following questions:

- Is it possible to learn useful speech representations from the West African Radio Corpus?
- How do such representations compare with the features of the baseline wav2vec encoder, trained on a high-quality large-scale English dataset, for downstream tasks on West African languages?
- How does the West African wav2vec qualitatively compare with the baseline wav2vec encoder?

### Language Identification

We used the annotated validation set of the West African Radio Corpus, which is disjoint from its unlabeled portion on which WAwav2vec is trained, to train the language identification neural network for the task of classifying audio clips in Maninka, Pular, and Susu.

We selected clips where the spoken languages include exactly one of Maninka, Susu, or Pular. For balance, we selected 28 clips per language for a total of 84 clips. Because of the small data size, we performed 10-fold cross-validation with randomly sampled training (60%) and validation (40%) portions. The mean test accuracies and their standard errors are reported in Table 4, showing that the West African wav2vec features outperform the baseline wav2vec, which outperforms mel spectrograms.

Fig. 4 shows the 84 audio clips used in the language identification experiments. The aggregated concatenated 9-D convolutional features of the best model for each of the 10 cross-validation training sessions were concatenated to make 90-D feature vectors. As bolstered by the qualitative results in the *Acoustic Unit Segmentation* section, the t-SNE (Maaten and Hinton 2008) projection of those feature vectors suggests that the WAwav2vec encoder is more sensitive than the baseline wav2vec to the specificities of the Maninka, Susu, and Pular languages.

| Features | Test Accuracy | | | |
|---|---|---|---|---|
| | Overall | MA | PU | SU |
| mel spectrogram | $60.00 \pm 2.80$ | 60.79 | 72.93 | 40.55 |
| wav2vec-z | $65.15 \pm 2.20$ | 58.05 | 78.97 | 60.08 |
| WAwav2vec-z | $\mathbf{79.09 \pm 1.32}$ | **88.43** | 83.76 | 63.44 |
| wav2vec-c | $60.61 \pm 0.61$ | 70.31 | 72.43 | 34.61 |
| WAwav2vec-c | $78.48 \pm 1.51$ | 80.09 | **88.20** | **67.02** |

Table 4: Language ID accuracies using mel spectrograms, and the latent (z) and context (c) features of the baseline wav2vec, and the West African wav2vec (WAwav2vec)

.

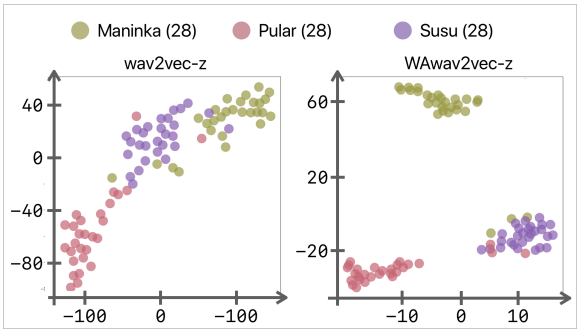



Figure 4: t-SNE projection of the language ID CNN's intermediate features for 84 audio clips encoded using the baseline wav2vec (left), and the West African wav2vec (right).

### Multilingual Speech Recognition

Next, we compared WAwav2vec to the baseline wav2vec encoder for the downstream task of speech recognition on the West African Virtual Assistant Speech Recognition Corpus containing 105 distinct utterance classes across 4 languages. Table 5 summarizes the speech recognition accuracies, from which we conclude that the features of the West African wav2vec are on par with the baseline wav2vec for the task of multilingual speech recognition.

### Acoustic Unit Segmentation

The previous experimental results showed that while features of the baseline wav2vec were overall marginally better than those of WAwav2vec for multilingual speech recognition, the features of WAwav2vec outperformed the baseline on the task of West African language identification. In this section, we attempt to qualitatively analyse the nature of the salient acoustic units encoded by both speech encoders.

We identified important acoustic segments that influence the language classification decision by computing the gradients of the input features with respect to the output of the language identification neural network, similarly to (Simonyan, Vedaldi, and Zisserman 2013), but with speech instead of images. We computed an attention signal by first normalizing (using softmax) the magnitude of the gradients, then summing them across the 512 input features, and finally

| Features | Test Accuracy | | | | | | |
|---|---|---|---|---|---|---|---|
| | Overall | Names | French | Maninka | Pular | Susu | Native Language |
| mel spectrogram | $74.05 \pm 0.74$ | $75.88 \pm 0.74$ | $67.79 \pm 1.98$ | $73.37 \pm 0.56$ | $71.60 \pm 1.78$ | $78.59 \pm 1.51$ | $71.75 \pm 0.97$ |
| wav2vec-z | $88.36 \pm 0.45$ | $89.41 \pm 0.75$ | $85.44 \pm 1.32$ | $86.80 \pm 0.96$ | $\mathbf{91.59 \pm 0.79}$ | $88.27 \pm 0.95$ | $87.73 \pm 0.39$ |
| WAwav2vec-z | $87.64 \pm 0.63$ | $89.59 \pm 1.21$ | $83.27 \pm 0.84$ | $85.92 \pm 0.81$ | $87.72 \pm 1.24$ | $89.74 \pm 0.53$ | $86.49 \pm 0.40$ |
| wav2vec-c | $\mathbf{88.79 \pm 0.46}$ | $\mathbf{89.78 \pm 0.48}$ | $\mathbf{86.92 \pm 1.38}$ | $87.51 \pm 0.96$ | $88.75 \pm 0.72$ | $\mathbf{89.88 \pm 1.11}$ | $87.54 \pm 0.40$ |
| WAwav2vec-c | $88.01 \pm 0.43$ | $87.74 \pm 1.10$ | $84.50 \pm 1.03$ | $\mathbf{88.53 \pm 0.00}$ | $89.19 \pm 1.27$ | $89.59 \pm 0.40$ | $\mathbf{87.99 \pm 0.53}$ |

Table 5: Multilingual ASR Accuracies on the West African Virtual Assistant Speech Recognition Corpus: Overall (Overall), for Guinean Names (Names), for utterances in specific languages (French, Maninka, Pular, Susu), and utterances spoken in the native language of the speaker (Native Language). We compare models using mel spectrograms, the latent (z) and context (c) features of the baseline wav2vec, and those of WAwav2vec.

normalizing again over the input sequence. Fig. 5a shows the attention signal aligned with the audio waveform.

Important acoustic units were segmented by considering contiguous time steps where the attention signal was 2 standard deviations above its mean. Fig. 5c shows the histogram of the duration $d = (c-1) \times p + f$ of the segmented acoustic units computed as a function of the number of wav2vec time steps ($c$), the period of the wav2vec latent features $p = 10ms$, and their receptive field with respect to the raw input waveform $f = 30ms$.

In order to get a sense of the semantic relevance of the segmented acoustic units, we computed their representation by averaging the wav2vec features over their span, and then projected those 512 dimensional features to 2D using t-SNE. Fig. 5b shows those projections for the baseline wav2vec and WAwav2vec. We observe that the acoustic units seem to separate by language in the West African wav2vec, but not in the baseline.

The duration of the segmented acoustic units and their language-specificity hint that they might be phoneme-like. In future work, we would like to investigate this approach to phoneme discovery, which to the best of our knowledge has not been explored before.

Fig. 5 reveals neighborhoods of language-specific acoustic units of duration between 40 and 200 milliseconds particularly prominent in the features of WAwav2vec, suggesting the semantic relevance of its encoding of Maninka, Pular, and Susu.

## Discussion

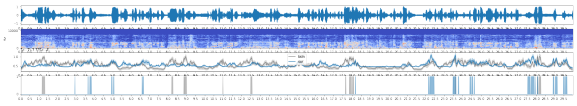
**We developed the first-ever speech recognition models for Maninka, Pular and Susu.** To the best of our knowledge, the multilingual speech recognition models we trained are the first-ever to recognize speech in Maninka, Pular, and Susu. We also showed how this model can power a voice interface for contact management.

**We enabled a multilingual intelligent virtual assistant for three languages spoken by 10 million people in regions with low literacy rates.** The state diagram shown in Fig. 2 demonstrates that the virtual assistant is simple yet functional and usable for the task of contact management, provided a working speech recognition module capable of recognizing the utterances described in Table 2. We built a speech recognition model capable of classifying those utterances with more than $88\%$ accuracy. We expect good generalization performance given the diversity of devices used for data collection, and the low variance of accuracy across the validation folds. The virtual assistant has a distinct wake word for each language. Therefore, after activation, it only needs to recognize utterances in the language corresponding to the used wake word. Additionally, as Fig. 2 shows, at each state there is only a subset of the utterance vocabulary that the assistant needs to recognize. Consequently, in practice the virtual assistant's speech recognition accuracy will be above the accuracy reported in our experiments.

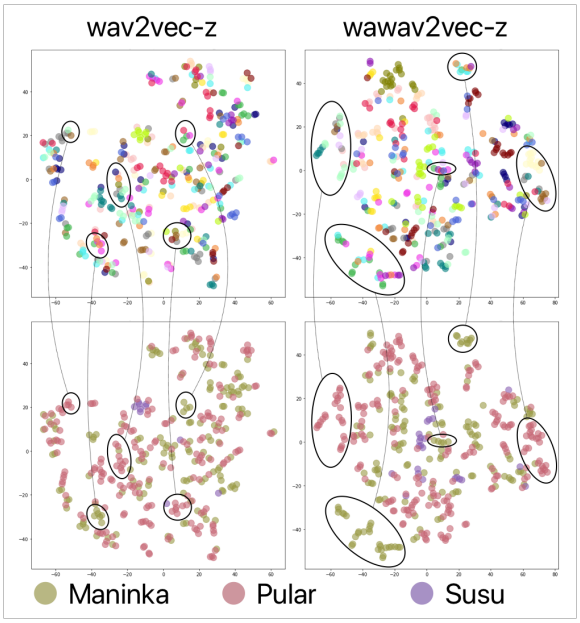
**Noisy radio archives are useful for unsupervized speech representation learning in low resource languages.** WAwav2vec features lead to significant improvements over mel spectrograms in both ASR accuracy ($88.01\%$ vs $74.05\%$) and language ID accuracy ($79.09\%$ vs $60.00\%$).

**WAwav2vec is on par with wav2vec on multilingual speech recognition.** Speech features learned from the West African Radio corpus lead to $88.01\%$ speech recognition accuracy, which is on par with the accuracy obtained with the baseline wav2vec, $88.79\%$. This result may be surprising given that the radio corpus is of lower quality (noise, multi-speakers, telephone, background and foreground music, etc.), and smaller size (142 vs 960 hours) compared to LibriSpeech, the training dataset of the baseline wav2vec. However, this result may be justified because the languages spoken in the West African Radio Corpus are more closely related to the target languages compared to English.

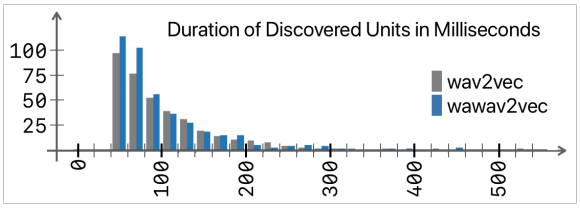
**WAwav2vec outperforms wav2vec on West African Language Identification.** On the task of language identification, WAwav2vec features outperformed the baseline by a large margin, achieving $79.09\%$ accuracy compared to the baseline accuracy of $65.15\%$. Our qualitative analysis indicated that the language classifier's decision was influenced by acoustic units of duration 40 to 200 milliseconds. Data visualization suggested that the acoustic units segmented from WAwav2vec features were more language-specific than the ones segmented from the baseline wav2vec features.



(a) Raw audio, attention signal, and segmented acoustic units.




(b) t-SNE Projection of segmented acoustic units. Left: baseline wav2vec. Right: West African wav2vec




(c) Histograms of the duration of the acoustic units in milliseconds.

Figure 5: Visualization of segmented acoustic units. (a) First two rows: raw wave form and spectrogram of a classified audio clip. Last two rows: attention signals and segmented units computed using wav2vec and WAwav2vec. (b) t-SNE projection of segmented acoustic units colored by audio clip (top row) and language (bottom row) using wav2vec features (left) and WAwav2vec features (right) with highlighted clusters of acoustic units of the same language extracted from different audio clips. (c) Histogram of the duration of the segmented acoustic units in milliseconds.

**English speech features can be useful for speech recognition in West African languages.** Using the baseline wav2vec resulted in $88.79\%$ speech recognition accuracy, compared to $74.05\%$ with mel spectrograms.

**There are non-obvious trade-offs for unsupervised speech representation learning.** WAwav2vec performs as well as the baseline wav2vec on the task of multilingual speech recognition, and outperforms the baseline wav2vec on West African language identification. This indicates the need for a more rigorous investigation of the trade-offs between relevance, size and quality of datasets used for unsupervised speech representation learning.

**We publicly released useful resources for West African speech technology development.** To advance speech technology for West African languages we released the West African Radio Corpus [1], the West African Virtual Assistant Speech Recognition Corpus [2], and a prototype of our multilingual intelligent virtual assistant along with our trained models and code to reproduce our experiments.[3]

## Limitations and Future Work

The virtual assistant only recognizes a limited vocabulary for contact management. Future work could expand its vocabulary to provide capabilities covering areas such as micro-finance, agriculture, or education. We also hope to expand its capabilities to more languages from the Niger-Congo family and beyond, so that literacy or ability to speak a foreign language are not prerequisites for accessing the benefits of technology. The abundance of radio data should make it straightforward to extend the encoder to other languages. Potentially, by training on even more languages in a language family (e.g., Mande or Bantu languages) the model will perform better. In our results, the West African wav2vec found "acoustic units" which were highly correlated with the different languages in our dataset. This hints at the potential use of speech encoders to learn the linguistic features of each language. We have only scratched the surface regarding the potential to use unsupervised speech representation learning to better articulate what makes each language unique.

## Conclusion

We introduced a simple, yet functional virtual assistant capable of contact management for illiterate speakers of Maninka, Pular, and Susu, collected the dataset required to develop its speech recognition module, and established baseline speech recognition accuracy.

In order to address the low-resource challenge, we explored unsupervised speech representation learning in two contexts: First, where representations are learned from high resource languages and transferred to unrelated low resource languages. Second, where representations are learned from low quality "found" data, radio archives abundant in many low-resource languages, in the target low resource languages. We gathered quantitative comparative results on those two forms of learning, and developed an effective qualitative analysis method of the learned representations.

We created the first-ever speech recognition models for 3 West African languages: Maninka, Pular and Susu. We released all our developed software, trained models, and collected datasets to the research community.

[1] https://openslr.org/105
[2] https://openslr.org/106
[3] https://github.com/mdoumbouya/nicolingua

## Ethics Statement

**Social Justice & Race** It is well known that digital technologies can have different consequences for people of different races (Hankerson et al. 2016). Technological systems can fail to provide the same quality of services for diverse users, treating some groups as if they do not exist (Madaio et al. 2020). Speakers of West African low-resource languages are likely to be ignored given that they are grossly underrepresented in research labs, companies and universities that have historically developed speech-recognition technologies. Our work serves to lessen that digital divide, with intellectual contributions, our personal backgrounds, and the access to technology we seek to provide to historically marginalized communities.

**Researchers** This research was conducted by researchers raised in Guinea, Kenya, Malaysia, and the United States. Team members have extensive experience living and working in Guinea, where a majority of this research was done, in collaboration with family members, close friends, and the local community.

**Participants** All humans who participated in data creation in various languages were adults who volunteered and are aware of and interested in the impact of this work.

**Data** The data contains the ages, genders and native languages of participants, but names have been erased for anonymity.

**Copyright** Radio data is being made public with permission from the copyright holders.

**Finance** The authors are not employed by any company working to monetize research in the region.

## Acknowledgements

We thank Voix de Fria, Radio Rurale de Fria, Radio Rurale Regionale de N'Zerekore, Radio Baobab N'Zerekore, City FM Conakry, and GuiGui FM Conakry for donating radio archives and the 49 speakers who contributed their voices and time to the speech recognition corpus. We thank Taibou Camara for annotating and collecting data, with support from Moussa Thomas Doumbouya, Sere Moussa Doumbouya and Djene Camara. Koumba Salifou Soumah, Mamadou Alimou Balde and Mamadou Sow provided valuable input on the vocabulary of the virtual assistant and Youssouf Sagno, Demba Doumbouya, Salou Nabe, Djoume Sangare and Ibrahima Doumbouya supported various aspects of the project including mediating with radio stations. We thank FASeF for providing office space and relatively stable grid power essential to running our deep learning experiments. Koulako Camara provided critical resources and guidance that made this project possible.

## References

Abate, S. T.; Menzel, W.; and Tafila, B. 2005. An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition. In *INTERSPEECH-2005*.

Adda, G.; Stüker, S.; Adda-Decker, M.; Ambouroue, O.; Besacier, L.; Blachon, D.; Bonneau-Maynard, H.; Godard, P.; Hamlaoui, F.; Idiatov, D.; Kouarata, G. N.; Lamel, L.; Makasso, E. M.; Rialland, A.; Van De Velde, M.; Yvon, F.; and Zerbian, S. 2016. Breaking the Unwritten Language Barrier: The BULB Project. *Procedia Computer Science* 81(May): 8–14. ISSN 18770509. doi:10.1016/j.procs.2016.04.023.

Ahmed, S. I.; Zaber, M.; and Guha, S. 2013. Usage of the Memory of Mobile Phones by Illiterate People. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, ACM DEV '13. New York, NY, USA: Association for Computing Machinery. ISBN 9781450318563. doi:10.1145/2442882.2442930. URL https://doi.org/10.1145/2442882.2442930.

Avle, S.; Quartey, E.; and Hutchful, D. 2018. Research on Mobile Phone Data in the Global South: Opportunities and Challenges. *The Oxford Handbook of Networked Communication* .

Badenhorst, J.; van Heerden, C.; Davel, M.; and Barnard, E. 2011. Collecting and evaluating speech recognition corpora for 11 South African languages. *Language Resources and Evaluation* 45(3): 289–309. ISSN 1574020X. doi:10.1007/s10579-011-9152-1.

Baljekar, P. 2018. *Speech synthesis from found data*. Ph.D. thesis, Google London.

Beeferman, D.; Brannon, W.; and Roy, D. 2019. RadioTalk: a large-scale corpus of talk radio transcripts. *arXiv preprint arXiv:1907.07073* .

Chipchase, J. 2006. How do you manage your contacts if you can't read or write? *interactions* 13(6): 16–17.

Clevert, D.-A.; Unterthiner, T.; and Hochreiter, S. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* .

Cooper, E. L. 2019. *Text-to-speech synthesis using found data for low-resource languages*. Ph.D. thesis, Columbia University.

Friscira, E.; Knoche, H.; and Huang, J. 2012. Getting in Touch with Text: Designing a Mobile Phone Application for Illiterate Users to Harness SMS. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, ACM DEV '12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450312622. doi:10.1145/2160601.2160608. URL https://doi-org.stanford.idm.oclc.org/10.1145/2160601.2160608.

Gauthier, E.; Besacier, L.; Voisin, S.; Melese, M.; and Elingui, U. P. 2016. Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof. *LREC* .

Gelas, H.; Besacier, L.; and Pellegrino, F. 2012. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*. Cape-Town, Afrique Du Sud. URL http://hal.inria.fr/hal-00954048.

Hankerson, D.; Marshall, A. R.; Booker, J.; El Mimouni, H.; Walker, I.; and Rode, J. A. 2016. Does Technology Have Race? In *Proceedings of the 2016 CHI Conference*

*Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, 473–486. New York, NY, USA: Association for Computing Machinery. ISBN 9781450340823. doi:10.1145/2851581.2892578. URL https://doi-org.stanford.idm.oclc.org/10.1145/2851581.2892578.

Harris, A.; and Cooper, M. 2019. Mobile phones: Impacts, challenges, and predictions. *Human Behavior and Emerging Technologies* 1(1): 15–17.

Ho, M. R.; Smyth, T. N.; Kam, M.; and Dearden, A. 2009. Human-computer interaction for development: The past, present, and future. *Information Technologies & International Development* 5(4): pp–1.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kunze, J.; Kirsch, L.; Kurenkov, I.; Krug, A.; Johannsmeier, J.; and Stober, S. 2017. Transfer learning for speech recognition on a budget. *arXiv preprint arXiv:1706.00290* .

Liu, A. T.; Yang, S.; Chi, P.; Hsu, P.; and Lee, H. 2020. Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6419–6423.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080. doi:10.1145/3313831.3376445. URL https://doi.org/10.1145/3313831.3376445.

Medhi, I.; Patnaik, S.; Brunskill, E.; Gautama, S. N.; Thies, W.; and Toyama, K. 2011. Designing Mobile Interfaces for Novice and Low-Literacy Users. *ACM Trans. Comput.-Hum. Interact.* 18(1). ISSN 1073-0516. doi:10.1145/1959022.1959024. URL https://doi.org/10.1145/1959022.1959024.

Mendels, G.; Cooper, E.; Soto, V.; Hirschberg, J.; Gales, M. J.; Knill, K. M.; Ragni, A.; and Wang, H. 2015. Improving speech recognition and keyword search for low resource languages using web data. In *INTERSPEECH 2015: 16th Annual Conference of the International Speech Communication Association*, 829–833. International Speech Communication Association (ISCA).

MHealth, W. 2011. New Horizons for Health through Mobile Technologies; Global Observatory for eHealth. *World Health Organization: Geneva, Switzerland* 3.

Nthite, T.; and Tsoeu, M. 2020. End-to-End Text-To-Speech synthesis for under resourced South African languages. In *2020 International SAUPEC/RobMech/PRASA Conference*, 1–6.

Ogbonnaya-Ogburu, I. F.; Smith, A. D.; To, A.; and Toyama, K. 2020. Critical Race Theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080. doi:10.1145/3313831.3376392. URL https://doi.org/10.1145/3313831.3376392.

Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; and Auli, M. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 8026–8037.

Patra, R.; Pal, J.; and Nedevschi, S. 2009. ICTD state of the union: Where have we reached and where are we headed. In *2009 International Conference on Information and Communication Technologies and Development (ICTD)*, 357–366. IEEE.

Rivière, M.; Joulin, A.; Mazaré, P.-E.; and Dupoux, E. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7414–7418. IEEE.

Roser, M.; and Ortiz-Ospina, E. 2016. Literacy. *Our World in Data* Https://ourworldindata.org/literacy.

Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* .

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* .

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1): 1929–1958.

Taoufik, I.; Kabaili, H.; and Kettani, D. 2007. Designing an E-Government Portal Accessible to Illiterate Citizens. In *Proceedings of the 1st International Conference on Theory and Practice of Electronic Governance*, ICEGOV '07, 327–336. New York, NY, USA: Association for Computing Machinery. ISBN 9781595938220. doi:10.1145/1328057.1328125. URL https://doi-org.stanford.idm.oclc.org/10.1145/1328057.1328125.

van Niekerk, D.; van Heerden, C.; Davel, M.; Kleynhans, N.; Kjartansson, O.; Jansche, M.; and Ha, L. 2017. Rapid development of TTS corpora for four South African languages. In *Proc. Interspeech 2017*, 2178–2182. Stockholm, Sweden. URL http://dx.doi.org/10.21437/Interspeech.2017-1139.