OkwuGbé: End-to-End Speech Recognition for Fon and Igbo

Anonymous ACL submission

Abstract

Language is inherent and compulsory for human communication. Whether expressed in a written or spoken way, it ensures understanding between people of the same and different regions. With the growing awareness and effort to include more low-resourced languages in NLP research, African languages have recently been a major subject of research in machine translation, and other text-based areas of NLP. However, there is still very little comparable research in speech recognition for African languages. Interestingly, some of the unique properties of African languages affecting NLP, like their diacritical and tonal complexities, have a major root in their speech, suggesting that careful speech interpretation could provide more intuition on how to deal with the linguistic complexities of African languages for text-based NLP. OkwuGbé is a step towards building speech recognition systems for African low-resourced languages. Using Fon and Igbo as our case study, we conduct a comprehensive linguistic analysis of each language and describe the creation of end-to-end, deep neural network-based speech recognition models for both languages. We present a state-of-the-art ASR model for Fon, as well as benchmark ASR model results for Igbo. Our linguistic analyses (for Fon and Igbo) provide valuable insights and guidance into the creation of speech recognition models for other African low-resourced languages, as well as guide future NLP research for Fon and Igbo. The Fon and Igbo models source code will be publicly available.

1 Introduction

$$OkwuGb\acute{e} = \underset{Igbo}{Okwu}(speech) + \underset{Fon}{Gb\acute{e}}(languages)$$

$OkwuGb\acute{e}$, the union of two words from Igbo ($Okwu$) and Fon ($Gb\acute{e}$) means the speech of languages, and signifies studying, and integrating automatic speech recognition to several African languages in an effort to unify them. African languages in the past decade received very little to no research in natural language processing (NLP) (Joshi et al., 2020a; Andrew Caines, 2019), prompting recent efforts geared towards improving the state of African languages in NLP (∀ et al., 2020b; Abbott and Martinus, 2018; Siminyu et al., 2020; ∀ et al., 2020a). However, there are few works being done on speech for these African languages, as more emphasis is being placed on their text. Due to the largely acoustic nature of African languages (mostly tonal, diacritical, etc), a careful speech analysis of African languages could provide better insight for text-based NLP involving African languages, as well as supplement the textual data needed for machine translation or language modelling. This is what inspired $OkwuGb\acute{e}$ and the focus on automatic speech recognition.

Automatic speech recognition (ASR, or speech-to-text) is a language technology where spoken words are identified, interpreted and converted to text. ASR is changing the way information is accessed, processed, and used. In recent years, ASR achieved state-of-art performances for most western and Asian languages such as English, French, Chinese, Japanese, etc, due to the availability of large quantity of quality speech resources. African languages, on the other hand are still lacking ASR applications. This is mainly due to the lack or unavailability of speech resources for most African Languages (ALs). As contributions:

- we provided specific and detailed linguistic analyses about Fon and Igbo (section 2).
- we built two end-to-end deep neural networks (E2E DNN) ASR systems for both languages. Our Fon model outperformed the unique existing Fon ASR model so far (Laleye et al., 2016).
- we designed speech models as speech encoders trained towards the Connectionist Temporal Classification (CTC) (Graves et al.,

2006) loss, and achieved promising results without huge amounts of training data.

- we demonstrated that attention mechanism (Bahdanau et al., 2016) improved the performance of acoustic models.

2 Linguistic Analyses of Fon and Igbo

In this section, we give an extensive linguistic details of both languages. Table 1 aims to summarise our analysis for the reader.

2.1 Fon

Fon (also known as Fongbe) is a native language of Benin Republic, spoken in average by more than 2.2 million people in Benin, in Nigeria, and Togo (Eberhard et al., 2020). Fon belongs to the *Niger-Congo-Gbe* languages family, and is a tonal, isolating and left-behind language according to (Joshi et al., 2020b), with a basic *Subject-Verb-Object* (SVO) word order. There are currently about 53 different dialects of the Fon language spoken throughout Benin (Lefebvre and Brousseau, 2002; Capo, 1991; Eberhard et al., 2020).

Its alphabet is based on the Latin alphabet, with the addition of the letters: ɔ, ɖ, ɛ, and the digraphs gb, hw, kp, ny, and xw. There are 10 vowel phonemes in Fon: 6 said to be *closed* [i, u, ĩ, ũ], and 4 said to be *opened* [ɛ, ɔ, a, ã]. There are 22 consonants (m, b, n, ɖ, p, t, d, c, j, k, g, kp, gb, f, v, s, z, x, h, xw, hw, w). Fon has two phonemic tones: *high* and *low*. *High* is realized as rising *(low–high)* after a consonant. Basic disyllabic words have all four possibilities: *high-high*, *high-low*, *low-high*, and *low-low*. In longer phonological words, like verb and noun phrases, a *high* tone tends to persist until the final syllable. If that syllable has a phonemic *low* tone, it becomes falling *(high–low)*. *Low* tones disappear between *high* tones, but their effect remains as a *downstep*. Rising tones *(low–high)* simplify to *high* after *high* (without triggering *downstep*) and to *low* before *high* (Lefebvre and Brousseau, 2002; Capo, 1991).

Fon makes extensive use of a rich system of tense or aspect markers, express many semantic features by lexical items, and the periphrastic constructions often used are of a more agglutinative nature (Capo, 1986). Fon nominals are generally preceded by a prefix consisting of a vowel (eg. the word aɖú: 'tooth'). The quality of this vowel is restricted to the subset of non-nasal vowels (Capo, 1991; Duthie and Vlaardingerbroek, 1981).

Reduplication is a morphological process in which the root or stem of a word, or part of it, is repeated. Fon, like the other Gbe languages, makes extensive use of reduplication in the formation of new words, especially in deriving nouns, adjectives, and adverbs from verbs. For instance, the verb *lã*, which means *to cut* (both in Fon and Ewe), is nominalized by reduplication, yielding *lãlã : the act of cutting*. Triplication is used to intensify the meaning of adjectives and adverbs (Capo, 1991; Duthie and Vlaardingerbroek, 1981).

2.2 Igbo

Igbo is a native language of the Igbo people, an ethnic group majorly located in the southeastern part of Nigeria, like Abia, Anambra, Ebonyi, Enugu, and Imo states, as well as in the northeast of the Delta state and in the southeast of the Rivers state. Outside Nigeria, it is spoken a little bit in Cameroon and Equatorial Guinea. Igbo belongs to the Benue-Congo group of the Niger-Congo language family and is spoken by over 27 million people (Eberhard et al., 2020). There are approximately 30 Igbo dialects, some of which are not mutually intelligible. To illustrate the complexity of Igbo, we quote (Nwaozuzu, 2008): "*...almost every community living as few as three kilometers apart has its few linguistic peculiarities. If these tiny peculiarities are isolated and considered to be able to assign linguistic dependence to each of these communities, we shall therefore be boasting of not less than one thousand languages in what we now know as the Igbo language.*"

This large number of dialects and peculiarities inspired the development of a standardized spoken and written Igbo in 1962, called the Standard Igbo (Ohiri-Aniche, 2007) (which we will refer to when we say "Igbo"). However, studies have shown that there are many sounds (mainly consonants) found in some other dialects of Igbo which are lacking in the Standard Igbo orthography. For example, Achebe et al. (2011) discovered about 50 unique speech sounds in Igbo. Morphologically, Igbo is an agglutinating language, with a compounding word formation: e.g., ugbo (vehicle) + igwe (iron) = ugboigwe (locomotive). Igbo also uses reduplication like Fon. Igbo has 28 consonants and 8 vowels, totalling 36 letters of the alphabet.

The sound system of Igbo consists of eight vowel phonemes, and 28 consonant phonemes (Ikekeonwu, 1999). There are four different types

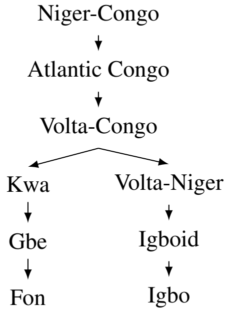
Characteristics	**Fon**	**Igbo**
Spoken where	mostly in Benin. Some part of Nigeria and Togo	mostly in southeastern Nigeria. A little bit in the Equatorial Guinea and Cameroon
Speakers (Eberhard et al., 2020)	2.2 million	22 million
Language family tree	 Niger-Congo Atlantic Congo Volta-Congo Kwa Volta-Niger Gbe Igboid Fon Igbo	
Language structure	Isolating language	Agglutinating language
Alphabet structure	32 letters: 22 consonants, 10 vowels	36 letters: 28 consonants, 8 vowels
Special alphabets besides Latin	ɔ, ɗ, ɛ, ã, gb, hw, kp, ny, and xw.	ch, gb, gh, gw, kp, kw, nw, ny, and sh
Tonal ?	Yes. 3 tones: high (/), low (\) and down step (-)	Yes. 4 tones: high tone (/), low (\), down step (−), and down drift (−)
Phoneme structure	10 vowel phonemes and 22 consonant phonemes. Nasalization is present	28 consonant phonemes and 8 vowel phonemes. Nasalization is present
Number of dialects	about 53	about 30
Reduplication ?	Yes, especially in deriving nouns, adjectives, and adverbs from verbs.	Yes, sometimes in compounding word formation: e.g., ugbo (vehicle) + igwe (iron) = ugboigwe (locomotive).
Code-switching?	No	Yes

Table 1: Summary analysis of Fon and Igbo

of tones in Igbo language (Odinye and Udechukwu, 2016). They include: High tone (/), Low tone (\), Down step (−) (Rice, 1992), Down drift (−). Down drift is only observed in Igbo sentences because one can raise or lower the pitch before a sentence is completed. Tone is an integral part of a word in Igbo. It is the interface of phonology and syntax in Igbo because it performs both lexical and grammatical functions (Nkamigbo, 2012). Igbo has three syllable types: consonant + vowel (the most common syllable type), vowel or syllabic nasal.

Code-switching, the act of "alternation of two languages during speech" (Poplack, 1979), is very common among Igbo-English bilingual speakers, making it an interesting feature for speech recognition research. Therefore, we will go deeper into it.

G.O and Mbagwu (2007); Obiamalu and Mbagwu (2010) did an extensive research on code-switching among Igbo speakers, where they classified it into three types: borrowing, quasi-borrowing and true code-switching (see Table 2). Borrowing in Igbo arises when words from English are inserted into Igbo during speech and the words go through phonological and morphological transformation (mark -> *maakigo*, table -> *tebulu*). This is usually because the speaker can not quickly find the Igbo equivalent of the word or such equivalent does not exist. This is illustrated by 1 and 2. In quasi-borrowing, the Igbo equivalents of the English words exist, but the English words are more often used by both monolinguals and bilinguals. It may or may not be assimilated into Igbo, like in borrowing. This is illustrated by 3 and 4. The third situation, called true code-switching, occurs when the speaker purposely chooses to use the English word, even though the Igbo equivalent is known and always used. This is most common among Igbo-English bilinguals. 5 and 6 are good examples.

Type	Examples (Igbo | English)	Explanation
borrowing	1.Ọ maakigo (mark) ule ahụ. | He has marked the examination 2.Ọ dị na tebulu (table) | It is on the table	The words ‘mark’ and ‘table’ had been borrowed and assimilated into Igbo because there are not readily available in Igbo.
quasi-borrowing	3.Obi zụrụ car ọhụrụ. | Obi bought a new car 4.Ọbi zụrụ ụgbọala ọhụrụ. | Obi bought a new car	Even though Igbo has words for ‘car, some bilinguals still use English words.
true code-switching	5.Fela *na e*criticize *onye ọbụla*.| Fela criticizes everybody 6.Jesus turn*ụrụ* water *ọ ghọrọ* wine. | Jesus turned water into wine	These cases are true code-switching because the Igbo words for ‘criticize’, ‘turn’, ‘water’ and ‘wine’ are readily available in Igbo, but the speaker chooses to use the English equivalents.

Table 2: Code-switching types and examples. Adapted from (G.O and Mbagwu, 2007)

3 Related Works

In this section, we review some related works according to the data resources, the model architectures and the state of ASR research for Fon and Igbo.

Previous works according to data resources: Xu et al. (2020) classified previous works on ASR, according to data resources, into rich-resource, low-resource and unsupervised settings, as shown in Table 3.

In the rich-resource setting, a large amount of paired speech and text data is available for training. This amounts up to hundreds of hours by multiple speakers. Furthermore, pronunciation lexicon is also leveraged while training for better results. These are the languages with ASR models already deployed in the industry. English is a main example of this setting. In the low-resource setting, there are only about dozen minutes of single-speaker high-quality paired data, and few hours of multi-speaker low-quality paired data. Compared to the rich-resource setting, these data resources contained fewer paired data.

In the extremely low-resource setting, which is where our work lies, there are little to no paired speech data resources, very low online presence, and sometimes no developed pronunciation lexicon rule or language model to improve ASR models. Some of these languages also contain few unpaired multi-speaker data. This is the case of many African languages.

Previous works according to model architecture: While traditional phonetic-based approaches (Hidden Markov Models) have produced considerable results in the past, we focus on end-to-end speech recognition with deep learning (Chorowski et al., 2014a,b; Hannun et al., 2014; Amodei et al., 2015; Chan et al., 2016; Chiu et al., 2018) because they have been shown to produce better results, with little dependence on handcrafted features and phoneme dictionaries.

Chorowski et al. (2014b) introduced an end-to-end continuous ASR using a bidirectional recurrent neural network (RNN) encoder with an RNN decoder that aligns the input and the output sequences using the attention mechanism. The model achieved a word error rate (WER) of 18.57% on the TIMIT data set. Hannun et al. (2014); Amodei et al. (2015) presented a state-of-the-art ASR system using E2E DNNs. They introduced a system that does not use any hand-designed language component, nor even the concept of "phoneme". Their result was achieved, as the authors stated in their original paper, through a well-optimized RNN training system that uses multiple GPUs, as well as a set of novel data synthesis techniques and language models.

Following the promising features that E2E DNNs offer, Mamyrbayev et al. (2020) showed in their recent studies that, using them with CTC works without the need for direct inclusion of language models.

State of ASR research for Fon and Igbo: Fon, unlike Igbo, has little to no digital presence. With very few speakers, and almost no online presence, there have been understandably very few tonal analysis or ASR research for this language. The few that exist are mostly by researchers who are native speakers of the language. To the best of our knowledge, there has only been a notable effort from Laleye et al. (2016) to build an ASR for Fon, with a word error rate *(WER)* of 44.04%, keeping the diacritics whose crucial importance for both performant ASR and neural machine translation (NMT) has been proved by Orife (2018); Dossou and Emezue (2020). This will be discussed later in section 4.3.2.

Igbo, on the other hand, has had a lot of tonal and speech analysis research in the past decade, but no public research on E2E DNN ASR, to the best of our knowledge. We opine that this is largely because 1) many old works in the past on

Setting		Rich-Resource	Low-Resource	Extremely Low-Resource
Data	pronunciation lexicon	✓	✓	✗
	paired data (single-speaker, high-quality)	dozens of minutes	several minutes	✗
	paired data (multi-speaker, high-quality)	hundreds of hours	dozens of hours	several hours
	unpaired speech (single-speaker, high-quality)	✓	dozens of hours	✗
	unpaired speech (multi-speaker,low-quality)	✓	✓	dozens of hours
	unpaired text	✓	✓	very few
Related Work	ASR	[(Chorowski et al., 2014a; Chiu et al., 2018; Chan et al., 2016; Hannun et al., 2014; Li et al., 2019; Mamyrbayev et al., 2020; Chorowski et al., 2014b; Hori et al., 2019; Rosenberg et al., 2019; Schneider et al., 2019)]	[(Tjandra et al., 2017)]	[**Our Work**, Laleye et al. (2016); Xu et al. (2020); Baevski et al. (2020); Liu et al. (2020); Ren et al. (2019)]

Table 3: Data sources to build ASR models and the corresponding related works in the different settings. Adapted from (Xu et al., 2020)

Igbo focused solely on tonal analysis (Odinye and Udechukwu, 2016; Nkamigbo, 2012), and 2) there is a lack of open-source speech data to encourage further research on exploring ASR with deep learning methods, which are known to be data-hungry.

4 Speech-to-Text Corpora and Data Preprocessing

4.1 Fon Speech-to-Text Corpus

We got our speech dataset for Fon from the existing Fon speech corpus[1]. The dataset contains recordings of the texts by native speakers (including 8 women and 20 men) of Fongbe in a noiseless environment. The recordings are sampled at a frequency of *16Khz*. The 28 native speakers have spoken around 1500 phrases (daily conversations domain). These recordings were made with the LigAikuma[2] android application. The minimum length of a speech sample is *2 seconds* and the maximum is *5 seconds*, giving us an average of *4 seconds* content length. Overall, there are around *10 hours* of speech data that have been collected. The dataset has been split into training, validation and test data sets. The training set contains *8 hours* of speech *(8235 speech samples)*, the validation data set contains *1500 speech samples* and finally the test data set contains *669 speech samples*. The text corpus made of the *1500* sentences used to build the speech data set has been scraped from BéninLangues [3].

4.2 Igbo Speech-to-Text Corpus

It was very hard to find the data set of Igbo audio samples and their transcripts. We realized that there is a great lacuna: even though there's been much research on Igbo phonology, there has really not been any (public) efforts to gather any speech-to-text data set for it.

The data set for our experiments on Igbo was got through a license from the Linguistic Data Consortium (LDC2019S16: IARPA Babel Igbo Language Pack) (Nikki et al., 2019). It contains approximately 207 hours of Igbo conversational and scripted telephone speech collected in 2014 and 2015 along with corresponding transcripts. The data set (hereafter called IgboDataset) is made up of telephone calls representing the Owerri, Onitsha, and Ngwa dialects spoken in Nigeria, sampled at 8kHz, and 48kHz. The recordings were very noisy as they were made of phone calls in various environments including streets, homes, offices, public places, and inside vehicles.

To ensure a good data quality, we cleaned the raw speeches by filtering based on: (a) length of text transcriptions, (b) the most frequent sample rate, and (c) the level of noise in the the recording. We obtained a final dataset of 2.5 hours, which we split into train, dev and test sizes of 4000, 100 and 100 audio samples respectively. The diacritics were originally removed from the transcripts.

[1] https://paperswithcode.com/dataset/fongbe-speech-recognition

[2] https://lig-aikuma.imag.fr/

[3] https://beninlangues.com/fongbe

4.3 Data Preprocessing

4.3.1 Speech Preprocessing

Speech signals are made up of amplitudes and frequencies. To get more information from our speech samples, we decided to map them into the frequency domain using the Discrete Fourier Transformation (DFT) (Boashash, 2003; Bracewell, 2000), and converted each speech into a narrow-band spectrogram[4].

We used 512 as length of the FFT window, 512 as the hop-length (number of samples between successive frames) and a hanning windows size is set to the length of FFT window.

For handling the audio data, we used the torchaudio utility from PyTorch (Paszke et al., 2019). We used Spectrogram Augmentation (SpecAugment) (Park et al., 2019) as a form of data augmentation: we cut out random blocks of consecutive time and frequency dimensions. Mel-Spectograms were generated from each speech samples with some fine-tuned hyper-parameters:

- sample_rate: sample rate of audio signal, set to 16000 hertz (16 kHz) for Fon and 8000 hertz (8 kHz) for Igbo.
- n_mels: number of mel filterbanks, set to 128 for Fon and 64 for Igbo.

4.3.2 Text Preprocessing

Scholars like Orife (2018) and Dossou and Emezue (2020) have shown in their studies that keeping the diacritics reduces lexical disambiguation and provides more morphological information to neural machine translation models. Additionally, diacritics relay the pronunciation tone and the sound generated, leading to an improved understanding of the sentences and their contexts.

Fon (a)	**Igbo (b)**
to (ambiguous and uncertain)	akwa (ambiguous and uncertain)
tó (*ear*) tò (*sea*) tô (*country*)	ákwà (*cloth*) ākwá (*egg*) ákwá (*to cry*)

Table 4: Examples of tonal inflection with Fon and Igbo

For Fon, a good example of the importance can be seen in Table 4 (a) where we demonstrate that removing diacritics from a word could lead to ambiguity and result in the confusion of the word. Therefore, we preprocessed the textual data without removing the diacritics.

For Igbo, unfortunately the IgboDataset was originally stripped of its diacritics. Therefore, we were not able to encode any diacritical information.

5 Models Architectures and Experiments

Whereas in data scenarios as the ones we presented in this paper, it would have been wiser to typically start from a pre-trained model, we chose to explore RNN-based models from scratch. The main reasons for this choice are: (a) the lack of computational resources, (b) the goal of having a simpler model architecture, which achieves reasonable performance, (c) the end goal of having a simple but efficient library for lower communities that can not afford expensive computational resources to build ASR systems for their languages.

5.1 Model Architecture

Related works have shown that we can increase model capacity, in order to efficiently learn from large speech datasets, by adding more hidden layers rather than making each layer larger. Graves et al. (2013) explored increasing the number of consecutive bidirectional recurrent layers, and Amodei et al. (2015) proposed the Deep Speech2, which among a number of optimization techniques, extensively applied batch normalization (Ioffe and Szegedy, 2015) to the deep RNNs.

Furthermore, Chorowski et al. (2014a) showed that the use of Bahdanau (additive) attention mechanism (Bahdanau et al., 2016) could reduce the phoneme or word error rate (WER) of the ASR model. This is possible because the attention mechanism forces the decoder to make monotonic alignment and hence improve the predictions.

Our model architecture, shown in Figure 1, draws inspiration from these research findings. While our model at its core is similar to Deep Speech 2, our key improvements are:

- the exploration of the combination of Bidirectional Long Short Term Memory (BiLSTMs) and Bidirectional Gated Recurrent Units (BiGRUs) for low-resource ASR.
- the integration of the Bahdanau attention mechanism, which effect has been demonstrated on Fon.

Our model has two main neural network modules: N-blocks of Residual Convolutional Neural Networks (rCNNs) (He et al., 2015, 2016) and M-blocks each of BiLSTMs and BiGRUs. Each rCNN block is made of two CNN layers, two dropout lay-

[4] http://www.glottopedia.org/index.php/Spectrogram

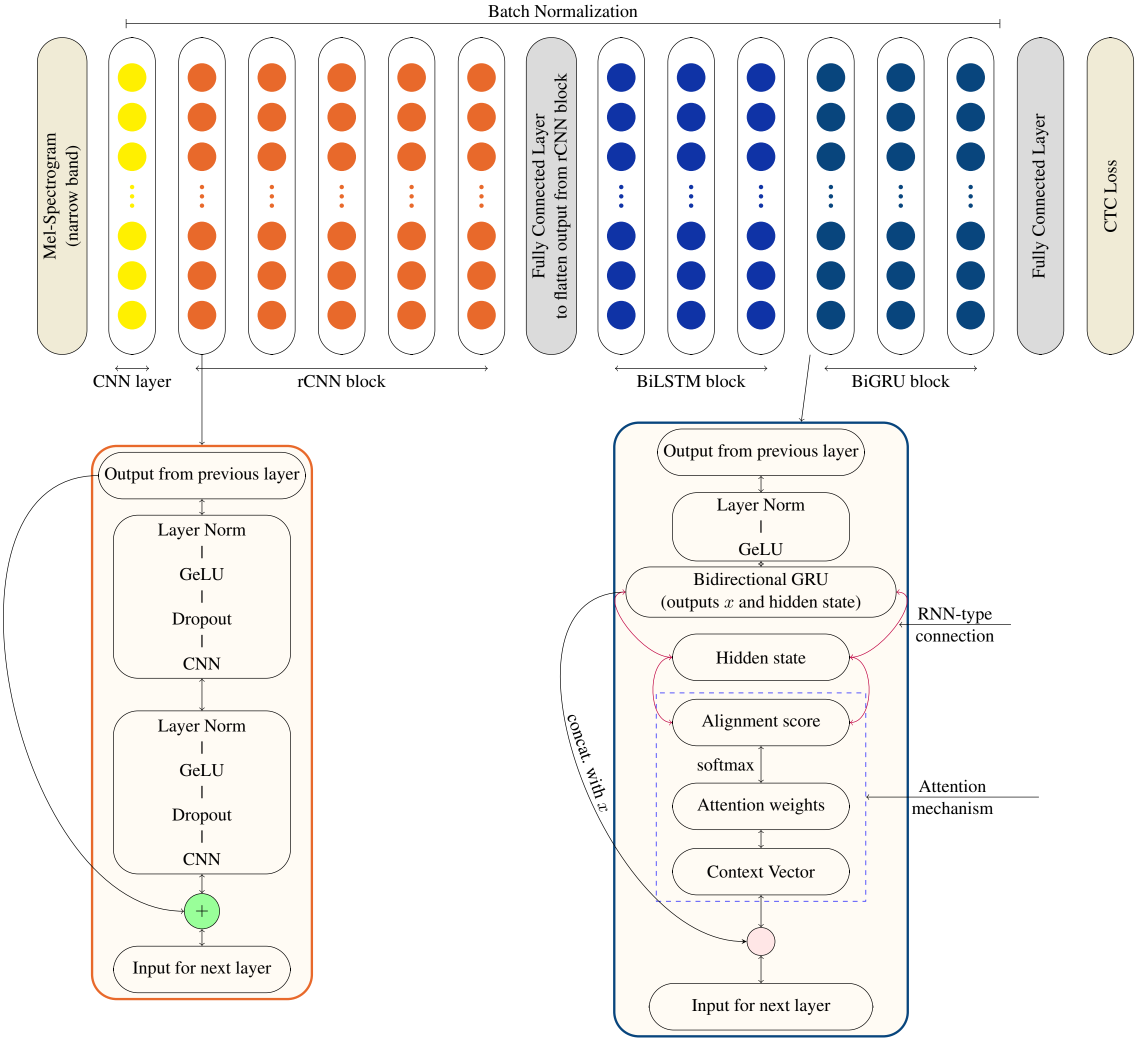


Figure 1: Architecture of the best model for Fon and Igbo, with an expansion of each component of the rCNN block and BiGRU block.

ers and two normalization layers (Ba et al., 2016) for the CNN inputs. We leveraged the power of convolutional neural networks (CNNs) to extract abstract features by converting speeches into spectrograms. RNNs process the abstract features produced by the rCNNs, step by step, making a prediction for each frame while using context from previous frames. We use BiRNN's (Schuster and Paliwal, 1997) because we want the context of not only the frame before each timestep, but the following as well. This helps the model make better predictions.

5.2 Experiments

For Igbo, to support our claim of the importance of data quality, we trained two ASR models: one with all speech samples, and the another solely with the cleaned samples.

Throughout our experiments, we explored different model architectures with various number of convolutional and bidirectional recurrent layers. The best ASR model, shown on Figure 1 has 5 blocks of rCNNs, 3 blocks each of BiLSTMs and BiGRUs, with attention incorporated into each component of the BiGRU block. Also, we used a form of Batch Normalization throughout the model.

We got the best evaluation results with the following hyper-parameters:

- max learning_rate: 5e-4 (for Fon), 3e-4 (for Igbo)
- batch_size: 20 (for Fon), 20 (for Igbo).
- (N, M): (5, 3) for Fon and Igbo.
- embedding_size: 512
- epochs: 500 (for Fon) and 1000 (for Igbo), with early stopping after 100 epochs.
- activation_function: GeLU (Hendrycks and Gimpel, 2016)
- optimizer: AdamW (Loshchilov and Hutter,

2019) (Fon), Nesterov accelerated descent (Nesterov, 1983) (Igbo)

We used two metrics to evaluate the models: the Character Error Rate (CER) and the WER. WER uses the Levenshtein distance (Levenshtein, 1966) to compare reference text and hypothesis text in word-level.

6 Results

Models (rCNN)	Fon CER	 WER	Igbo (without cleaning) CER	 WER
+BiGRUs	22.0831	59.66	-	-
+BiLSTMs	24.2783	61.46	-	-
+BiLSTMs+BiGRUs	**16.9581**	47.05	56.00	64.00
+BiLSTMs+BiGRUs+Attn	18.7976	**42.50**	**50.12** (92.67)	**55.03** (97.99)
(Laleye et al., 2016)	-	44.09	-	-

Table 5: CER (%), and WER (%) of different models on Fon and Igbo (original and cleaned) test datasets.

Fon Decoded Predictions	**Fon Decoded Targets**
tɔ ce xwe yɔyɔ din tɔn ɔ ci gblagadaa	tɔ ce xwe yɔyɔ din tɔn ɔ ci gblagadaa
eo mi sa aakpan nu mi	eo mi sa akpan nu mi
fitɛ a gosin xwe yi gbe	fitɛ a go sin xwe yin gbe
e kpo kpɛɗé	e kpo kpɛɗe
akwɛ cɛ gbadé jí ɗaximɛ	akwɛ jɛ gbadé ji ɗaximɛ

Table 6: Decoded Predictions and Targets of the best Fon ASR Model

We show that implementing attention mechanism reduced the WER by 5-6% for both languages. Our Fon ASR model outperformed the current Fon ASR model with diacritics of Laleye et al. (2016). For Igbo, our results serve as a benchmark.

Table 6 shows some decoded predictions and targets from the Fon ASR model, which are very similar. Common mistakes (colored), happen most often at a character level where a character is either omitted, added or replaced by another one; mistakes often practically not distinguishable in speaking.

In Table 5, one may observe the large difference between the CER and WER on Fon language, unlike Igbo. We strongly believe that this is due to the fact that the character set for Fon contains all the possible diacritics for each letter of the Fon alphabet, making it extremely large (compared to the set of Igbo characters which had no diacritical information). To further support our claim, a close observation of the targets and predictions in Table 6 reveals that the errors are mostly due to omission or mismatch of diacritics for the characters ('e' predicted instead of 'é ' in row 4 or a space added between 'go' and 'sin' in row 3).

7 Future Work

Our work shows promising results considering the small training sizes, and we have presented a state-of-the-art ASR model for Fon. Our future attempt towards improving the current Fon system will be leveraging transfer learning techniques on transformers-based like (Conneau et al., 2020), whose core is its Transformer encoder, which has been trained on huge corpora of French Speech data.

For Igbo language, as we showed that data quality is crucial for the performance, the next stage involves incorporating diacritical information in the ASR model. We have begun by gathering new speech dataset which include the diacritics, soon available on CommonVoices.

References

Jade Z. Abbott and Laura Martinus. 2018. Towards neural machine translation for african languages. *CoRR*, abs/1811.05467.

Ike Achebe, Clara Ikekeonwu, Cecilia Eme, Nolue Emenanjo, and Nganga Wanjiku. 2011. A composite synchronic alphabet of igbo dialects (csaid). *IADP, New York*.

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. 2015. Deep speech 2: End-to-end speech recognition in english and mandarin.

Andrew Caines. 2019. The Geographic Diversity of NLP Conferences.

Jimmy Ba, J. Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2020. Effectiveness of self-supervised pre-training for speech recognition.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

B. Boashash. 2003. *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*. Oxford: Elsevier Science.

R. N Bracewell. 2000. *The Fourier Transform and Its Applications*. Boston: McGraw-Hill.

Hounkpati B. C. Capo. 1986. *Renaissance du gbe, une langue de l'Afrique occidentale: étude critique sur les langues ajatado, l'ewe, le fon, le gen, laja, le gun, etc.* Université du Bénin, Institut national des sciences de l'éducation.

Hounkpati B. C. Capo. 1991. *A comparative phonology of Gbe*. Foris Publications.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2018. State-of-the-art speech recognition with sequence-to-sequence models.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014a. End-to-end continuous speech recognition using attention-based recurrent nn: First results.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014b. End-to-end continuous speech recognition using attention-based recurrent nn: First results.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Bonaventure F. P. Dossou and Chris C. Emezue. 2020. Ffr v1.1: Fon-french neural machine translation.

Alan S. Duthie and R. K. Vlaardingerbroek. 1981. *Bibliography of GBE: (Ewe, Gen, Aja, Xwala, Fon, Gun, etc.): publications "on" and "in" the language*. Basler Afrika Bibliographien.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2020. Ethnologue: Languages of the world. twenty-third edition.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020a. Participatory research for low-resourced machine translation: A case study in african languages. *Findings of EMNLP*.

∀, Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020b. Masakhane – machine translation for africa.

Obiamalu G.O and D.U. Mbagwu. 2007. Code-switching: Insights from code-switched english/igbo expressions. pages 51–53. Awka Journal of Linguistics and Languages Vol 3.

A. Graves, Abdel rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus).

T. Hori, R. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. Le Roux. 2019. Cycle-consistency training for end-to-end speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6271–6275.

Clara Ikekeonwu. 1999. Igbo", handbook of the international phonetic association.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020a. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020b. The state and fate of linguistic diversity and inclusion in the nlp world.

F. A. A. Laleye, L. Besacier, E. C. Ezin, and C. Motamed. 2016. 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)*, pages 477–482.

Claire Lefebvre and Anne-Marie Brousseau. 2002. *A grammar of Fongbe*. Mouton de Gruyter.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6706–6713.

Alexander H. Liu, Tao Tu, Hung yi Lee, and Lin shan Lee. 2020. Towards unsupervised speech recognition and synthesis with quantized speech representation learning.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Orken Mamyrbayev, Keylan Alimhan, Bagashar Zhumazhanov, Tolganay Turdalykyzy, and Farida Gusmanova. 2020. End-to-end speech recognition in agglutinative languages. In *Intelligent Information and Database Systems*, pages 391–401, Cham. Springer International Publishing.

Y. Nesterov. 1983. A method for unconstrained convex minimization problem with the rate of convergence o($1/k^2$).

Adams Nikki, Bills Aric, Conners Thomas, David Anne, Dubinski Eyal, Fiscus Jonathan G., Gann Ketty, Harper Mary, Kaiser-Schatzlein Alice, Kazi Michael, Malyska Nicolas, Melot Jennifer, Onaka Akiko, Paget Shelley, Ray Jessica, Richardson Fred, Rytting Anton, and Sinney Shen. 2019. Iarpa babel igbo language pack iarpa-babel306b-v2.0c ldc2019s16. web download.

Linda Chinelo Nkamigbo. 2012. A phonetic analysis of igbo tone. ISCA Archive, The Third International Symposium on Tonal Aspects of Languages.

G.I. Nwaozuzu. 2008. *Dialects of the Igbo Language*. University of Nigeria Press.

G. Obiamalu and Davidson U. Mbagwu. 2010. Motivations for code-switching among igboenglish bilinguals: A linguistic and sociopsychological survey. *OGIRISI: a New Journal of African Studies*, 5:27–39.

Sunny Odinye and Gladys Udechukwu. 2016. Igbo and chinese tonal systems: a comparative analysis. *Ogirisi: A new Journal of African Studies*, Volume 1:48.

Chinyere Ohiri-Aniche. 2007. Stemming the tide of centrifugal forces in igbo orthography. *Dialectical Anthropology*, 31(4):423–436.

Iroro Orife. 2018. Attentive sequence-to-sequence learning for diacritic restoration of yorùbá language text.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

S. Poplack. 1979. "sometimes i'll start a sentence in spanish y termino en español": Toward a typology of code-switching.

Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Almost unsupervised text to speech and automatic speech recognition. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5410–5419. PMLR.

Keren Rice. 1992. *Language*, 68(1):149–156.

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. 2019. Speech recognition with augmented synthesized speech.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Mike Schuster and Kuldip Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.

Kathleen Siminyu, Sackey Freshia, Jade Abbott, and Vukosi Marivate. 2020. Ai4d – african language dataset challenge.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2017. Listening while speaking: Speech chain by deep learning.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition.