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

Bonaventure F. P. Dossou*
Jacobs University Bremen
f.dossou@jacobs-university.de

Chris C. Emezue*
Technical University of Munich
chris.emezue@tum.de

Abstract

Language is a fundamental component of human communication. African low-resourced 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. OkwuGbé is a step towards building speech recognition systems for African low-resourced languages. Using Fon and Igbo as our case study, we build two end-to-end deep neural network-based speech recognition models. We present a state-of-the-art automatic speech recognition (ASR) model for Fon, and a benchmark ASR model result for Igbo. Our findings serve both as a guide for future NLP research for Fon and Igbo in particular, and the creation of speech recognition models for other African low-resourced languages in general. The Fon and Igbo models source code have been made publicly available. Moreover, ***Okwugbe***, a python library has been created to make easier the process of ASR model building and training.

1 Motivation and State of ASR research for Fon and Igbo

$$OkwuGbé = \underset{Igbo}{Okwu}(speech) + \underset{Fon}{Gbé}(languages)$$

$OkwuGbé$, the union of two words from Igbo ($Okwu$) and Fon ($Gbé$) literally means the speech of languages, and signifies studying, and integrating speech technologies to several African languages in an effort to unify them. In this paper, we introduce ASR systems for two low-resourced languages: Fon and Igbo. We show that using end-to-end deep neural networks (E2E DNN) with Connectionist Temporal Classification (CTC) (Graves et al., 2006) allows us to achieve promising results. We also demonstrate that leveraging an attention mechanism (Bahdanau et al., 2016) improves the performance of acoustic models.

Fon (Capo, 1991; Capo, 1986), unlike Igbo (MustGo et al, 2015), has little to no digital presence. With very few speakers, and almost no online presence, there have been understandably very few tonal analyses or ASR research for this language. Up to now, the unique effort for Fon ASR is described in (Laleye et al., 2016), and it achieved a word error rate (WER) of 44.04%.

Igbo (Nkamigbo, 2012), 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 of the lack of open-source speech data to encourage further research on exploring ASR with deep learning methods, which are known to be data-hungry.

2 Speech-to-Text Corpora and Data Preprocessing

2.1 Fon and Igbo Speech-to-Text Corpora

We used around *10 hours* of speech from the existing Fon speech corpus[1] built upon the tedious task of text recordings, pronounced by 28 native speakers in a noiseless environment, who spoke around 1500

*These authors contributed equally to this work.
[1]https://paperswithcode.com/dataset/fongbe-speech-recognition

phrases[2] (daily conversations domain). The dataset was splitted into training *(8235 speech samples)*, validation (*1500 speech samples*) and test data (*669 speech samples*) sets.

For Igbo, 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). The data set (hereafter called IgboDataset) contains approximately 207 hours of Igbo conversational and scripted telephone speech collected in 2014 and 2015. The audio samples were very noisy, and we implemented cleaning strategies (like filtering based on length of words, upsampling, exploring different mel-spectrograms units and number of Fast Fourier Transform (FFT) bins). Our cleaning strategies gave us a reduced data set of 2.5 hours, which we split into train, dev and test sizes of 4000, 100 and 100 audio samples respectively.

2.2 Data Preprocessing

We processed the speech samples by using the FFT (Boashash, 2003; Bracewell, 2000), converting each speech into a narrow-band spectrogram[3].

The importance of diacritics has been showed for Fon, in Table 1 (a), where we can see that removing diacritics from a word could lead to ambiguity and result in the confusion (Orife, 2018; Dossou and Emezue, 2020; Dossou and Sabry, 2021). Therefore, we preprocessed the textual data without removing the diacritics.

Conversely, for Igbo, even though Table 1 (b) showcases the importance of the diacritics, unfortunately the IgboDataset was originally stripped of its diacritics. Therefore, we were not able to encode any diacritical information.

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 1: Fon and Igbo Examples showing the the importance of diacritics

3 Model, Experiments and Results

We present our best model[4] and findings using 5 blocks of rCNNs (He et al., 2015; He et al., 2016) with, 3 blocks solely of BiGRUs (Hendrycks and Gimpel, 2016), 3 blocks solely of BiLSTMs, 3 blocks each of BiLSTMs and BiGRUs, and 3 blocks each of BiLSTMs and BiGRUs + Attention mechanism (Bahdanau et al., 2016). Our main performance evaluation metrics were the Character Error Rate (CER) and the WER, based on the Levenshtein distance (Levenshtein, 1966).

Models (rCNN)	Fon		Igbo (without cleaning)	
	CER	WER	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 2: CER (%), and WER (%) of different models on Fon and Igbo (original and cleaned) test datasets.

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

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

[4]Okwugbé Model: https://drive.google.com/file/d/1sEfVtRwIxRn1g2r3POaNE_WwqWxPOZPm/view

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 3: Decoded Predictions and Targets of the best Fon ASR Model

Our work shows promising results considering the small training sizes, and we have presented a state-of-the-art ASR model for Fon. We show that implementing attention mechanism reduced for both languages, the WER by 5-6%, and Fon ASR model outperformed the current Fon ASR model of (Laleye et al., 2016).

An important observation we show in Table 2 is the effect of our cleaning procedure on the model's ability to learn: for our large IgboData with background noise, uneven audio length, low sampling rate, etc, the model found it very difficult to learn the speech representations. Taking time to sieve through the data (in Section 2.1) mitigated this issue by helping the model learn the abstract features better, albeit on a small training set. Our preliminary results serve as a benchmark for ASR on Igbo. The source code for the model can be accessed here.

In Table 2, 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 3 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).

Table 3 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. The native speakers included in this study have testified to the fact that those mismatched words or characters are often practically not distinguishable in speaking. The model source code is open-sourced here.

3.1 Improving the Model

To improve the proposed models, we are exploring approaches like leveraging language models, deeper model structures, transformers and crowd-sourcing/compiling speech-to-text data set for Igbo and Fon. For Igbo language, the next stage involves incorporating diacritical information in the ASR model. We have begun by gathering new speech dataset which include the diacritics.

3.1.1 Wav2Vec

One of our attempts towards improving the ASR system for Fon was taking part in the HuggingFace XLSR-Wav2Vec2 Finetuning Week where we finetuned a pretrained Wav2Vec model (Conneau et al., 2020) on the Fon language using the same dataset and splits used in our project. The full Colab implementation can be found here. This setup achieved a WER of **14.97**.

4 Acknowledgements

We are grateful to Professor Graham Neubig of Carnegie Mellon University for coming to our aid by providing us with an Amazon EC2 instance for training our model when we were very low on computational resources. We also thank Dr Frejus Layele for giving us access to the Fon data set, and Dr Iroro Orife, for his guidance on designing the ASR model and cleaning the IgboData.

References

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.

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.

Bonaventure F. P. Dossou and Mohammed Sabry. 2021. Afrivec: Word embedding models for african languages. case study of fon and nobiin.

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.

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).

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.

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

MustGo MustGo et al. 2015. Igbo Language - Structure, Writing & Alphabet - MustGo.

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.

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