SautiDB-Naija: A Nigerian L2 English Speech Corpus

Tejumade Afonja, Iroro Orife, Lawrence Francis, Clinton Mbataku, Olumide Okubadejo, Ademola Malomo, Munachiso Nwadike, Oladimeji Mudele, Kenechi Dukor, Oluwafemi Azeez, Duru Goodness

DEEP LEARNING INDABA

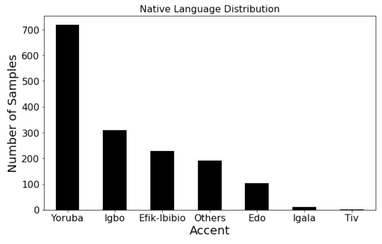
Abstract

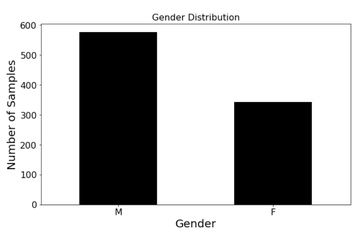
In this paper, we introduce SautiDB-Naija: a speech corpus of non-native speakers of English intended for research in accent translation, voice conversion, and accent classification. This initial release of our corpus includes over 900 recordings of non-native speakers of English whose first language (L1) is amongst the most common in Nigeria, namely Yoruba, Igbo, Edo,, Efik-Ibibio and Igala. To the best of our knowledge, this is the first documented effort to curate a corpus of Nigerian accents for machine learning research to date. We demonstrate that neural networks are capable of learning linguistic features that distinguish between different accent classes by training a discriminative classifier on our corpus. Our results demonstrate the potential of SautiDB-Naija as a valuable resource for future computational linguistic research.

Introduction

Advances in education, technology and transportation have made the world a much smaller place --- with people from different cities, regions and countries commonly speaking a hypercentral or global language[1]. As a result, there is the need to build voice-enabled tools that are adaptable to different accents. To this end, we make the following contributions:

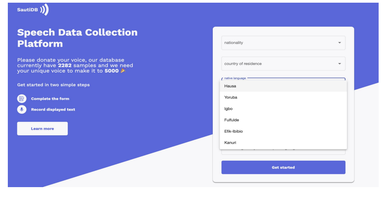
1. **SautiDB:** a web-app for crowdsourcing English language speech recordings from a distributed network of volunteer Nigerian speakers. The name *sauti* is the Swahili word for sound.
2. **SautiDB-Naija**: a non-native English speech corpus consisting of 919 speech recordings from an assemblage of first-language (L1) speakers of over 5 Nigerian Languages, Yoruba, Igbo, Edo, Efik-Ibibio, and Igala. The number of samples per accent considered are presented in the figure below:

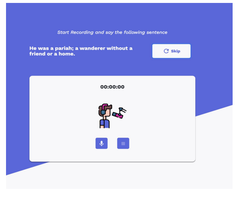


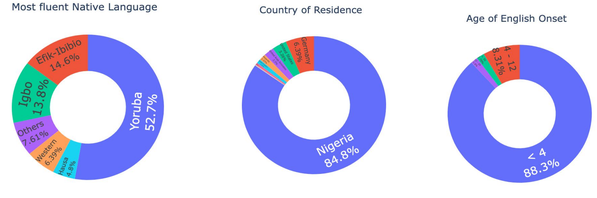


Corpus Creation and Statistics

We built a simple web application with text prompts of short sentences from 1132 phonetically balanced sentences from the CMU Arctic Database [2].

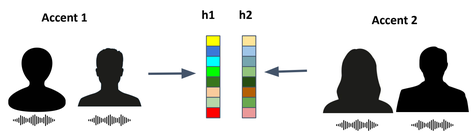




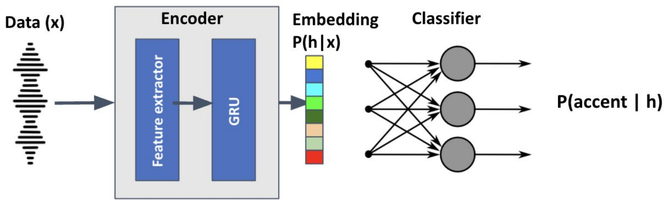


Experiments and Results

We attempt to quantify the fidelity of the accent information in our SautiDB-Naija corpus by learning the accent embeddings through an accent classification supervised learning task.



Model Architecture



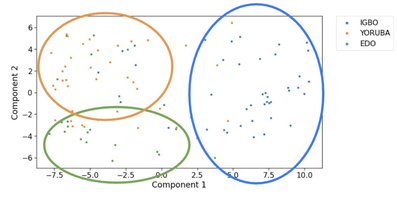
For the feature extractor, mel-spectrogram and a fine tuned Wav2vec [3] model representation were tested.
Batch normalization (BN) is applied to the feature extractor output to reduce overfitting.

Results

Our experiments are performed with Yoruba, Igbo and Edo accented speech.

Our results show that our data contain informative differentiating accents. We obtained our best model using wav2vec feature extractor with batch normalization.
The two-dimensional projection plot of that the embedding space clusters the audios by speaker accents.

Model	Accuracy	F1 Score
Base model	0.2667	0.2413
Mel+GRU	0.4140	0.3809
wav2vec+GRU	0.5333	0.4881
wav2vec+GRU+BN	**0.6952**	**0.6457**



Conclusion

We presented SautiDB-Naija, a non-native English speech database of short sentences consisting of 5 Nigerian languages to support accent classification, conversion or translation tasks. Our experiments point towards possible future use cases for SautiDB-Naija. Future work will focus on expanding and diversifying the corpus while commencing research on L2 accent translation tasks

Acknowledgement

This work was supported by AI4D-IndabaX awards IDRC Grant Number: 109187-002. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AI4D-IndabaX. We appreciate and thank everyone who took our survey and contributed their voices to our database.

References

[1] De Swaan, A., 2013. Words of the world: The global language system. John Wiley & Sons

[2] Kominek, J., Black, A.W. and Ver, V., 2003. CMU ARCTIC databases for speech synthesis.

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