SautiLearn: Improving Online Learning Experience with Accent Translation

Tejumade Afonja
AI Saturdays Lagos
tmafonja1@gmail.com

Clinton Mbataku
AI Saturdays Lagos
mbatakuc@gmail.com

Olumide Okubadejo
Audionamix
olumideokubadejo@gmail.com

Ademola Malomo
AI Saturdays Lagos
demlabz@gmail.com

Munachiso S. Nwadike
New York University
msn307@nyu.edu

Lawrence Francis
AI Saturdays Lagos
lawrencedikeu@gmail.com

Oluwafemi Azeez
AI Saturdays Lagos
azeezfemi17937@gmail.com

Abstract

Our work addresses the problem of accent translation, for the modification of spoken audio into Nigerian accents. In this work, we propose a unique speech dataset, SautiDB, consisting of 919 voices from a mixture of different Nigerian accents (Yoruba, Igbo, Hausa, Efik-Ibibio, Igala, Edo) collected from a distributed network of volunteer speakers. We show that by using phonetic posteriorgrams in a sequence-to-sequence model, we can achieve convincing performance in accent conversion. We intend to use our resulting model to convert the accents of lectures provided in American English to a mixture of Nigerian accents that are more easily understood by native speakers of the Nigerian language. The end result is will be a proposed tool, SautiLearn.

1 Introduction

Many advances in areas such as education, technology, transportation, and economics have contributed to the increase in verbal communication between people from different cities, regions, countries, and other parts of the world [1]. In many cases, even when people speak the same language, they have difficulty understanding each other because different speakers use different accents. In recent years, for example, a plethora of online courses have emerged, most of which are offered in English [2, 3]. These online courses are sometimes free and supposedly accessible to learners around the world. However, learners often have to adapt to the teaching style of the instructor, who may deliver the course in an accent that is unfamiliar to them, which can place a significant cognitive burden on learners and put them at a disadvantage in learning the course material in a timely manner compared to peers who are familiar with the instructor's accent. Therefore, it would be desirable to find a way to present online video content to learners in an accent style that is familiar to them. Motivated by this question, our work contributes the following:

1. We present SautiDB, a dataset collection platform that collects speech recordings of various Nigerian accents through the power of crowdsourcing. Together with SautiDB, we release the SautiDB-919 dataset, which contains 919 speech samples collected via crowdsourcing through our SautiDB platform. Our dataset covers a wide range of ethnicities and Nigerian accent variants.

Preprint. Under review.

2. We demonstrate the usecase for SautiLearn, a tool that can help students learn in accents they are familiar with, if they wish. In this work, we focused on converting audio speech from one accent to another using a sequence-to-sequence neural network model that uses speaker-independent linguistic features such as the phonetic posteriorgram as input, following the work of [4].

2 Related Works

Accent Conversion or translation has been the subject of much research in recent years with the goal to learn speaker-independent representations that capture the subtleties of a larger subset of individuals who share a common native language. A central theme in work such as [5, 6, 7] is the use of phonetic posteriorgrams in accent conversion. Phonetic posteriorgrams (PPG) consist of unique phoneme-level embeddings that represent the presence or absence of linguistic phonemes over time and because phonemes can be represented with English letters, they provide human-interpretable features that can complement traditional audio modalities such as mel spectrograms or MFFCs.

In other work, [8] show that speech can be decomposed into the different components of timbre, pitch, and rhythm. Therefore, a neural network can learn to identify individual components of speech in order to morph audio from one speaker to sound similar to that of another speaker. MelGan-VC [9] and similar approaches [10, 11] convert human speech to audio using generative adversarial networks (GAN), where the source domain is the speech of a source speaker and the target domain is the speech of the target speaker. The goal is then to learn a function from the source to the target domain.

However, it is important to distinguish between accent and voice conversion. While both tasks are speech conversion, voice conversion aims to imitate the speaking style of a single speaker. Accent conversion, on the other hand, aims to imitate the broader style of an established accent. The need for speaker-independent representations to guide the neural network learning process leads us naturally to a framework in which PPG is valuable.

When collecting a dataset of spoken audio for speech processing applications, it is often desirable to select a sentence that has the property that phonemes occur at the same frequency as in English. For example, the Harvard Sentences [12] contains 72 lists of 10 phonetically balanced sentences that adhere to this property. Another sentence set carefully selected for its phonetic properties was used in the CMU ARCTIC dataset [13]. This dataset contains 1132 uttered sentences taken from non-proprietary texts from the Gutenberg Project. However, these sentences were uttered in standard native English accents. The L2- ARCTIC corpus [14] sought to build on CMU ARCTIC by providing utterances spoken by non-native (L2) speakers of English who have a pronounced accent peculiar to their first language. The languages selected were Hindi, Korean, Mandarin, Spanish and Arabic.

Our goal was to create, in the spirit of L2- ARCTIC, a phonetically-balanced dataset of non-native speakers with a pronounced Nigerian accent. To this end, we developed a mechanism for collecting local speech datasets.

3 Dataset

3.1 Collection of Speech Data

To collect the dataset, we designed and built a simple web application that anyone can access through a web browser. Our web application [15], called SautiDB (Sauti means sound in Swahili), was built using Angular and Firebase. We used the built-in audio web API to collect speech samples. The speech samples were collected at 48KHz sample rate and we left all other parameters as default.

3.2 Choice of Sentence Prompt

To create the corpus, we used the 1132 sentences of the CMU ARCTIC prompts [13]. There were several reasons that influenced our decision: First, as described in [14], the ARCTIC prompts are phonetically balanced (100%, 79.6%, and 13.7% coverage for phonemes, diphones, and triphones, respectively); the sentences are also non-proprietary and can produce about an hour of edited speech. Second, the Arctic corpus itself has proven useful for various tasks such as speech synthesis [16] and speech conversion tasks [4]. Perhaps most importantly, this dataset provides us with a parallel

sentence that has been shown in previous work to be an important factor in accent conversion tasks [17].

3.3 Marketing the webapp / Social Media Campaign

We created some social media flyers for the two-week data collection campaign. Figure 1 shows our marketing flyers. We reached out to our AI Saturdays Lagos community members and also shared the webapp within our various networks. Our strategy resulted in 1500 voice samples over the course of the two-week campaign.

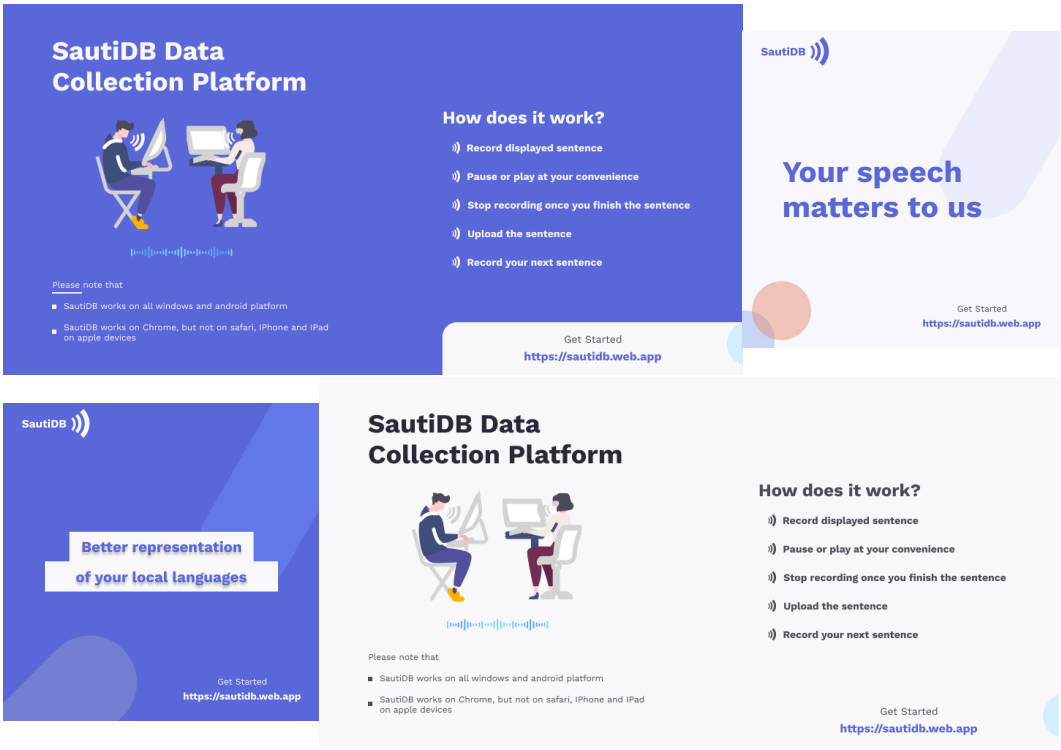

Figure 1: Social Media Flyers

3.4 Data Visualization

There are about 525 languages spoken in Nigeria. While Hausa, Igbo and Yoruba are the most widely spoken languages in Nigeria, other languages such as Fulfulde, Efik-Ibibio, Tiv, Kauri, Edo, Igala, Nupe and Izon are also spoken by millions of Nigerians as L1 or L2. The linguistic diversity of Nigeria was strongly considered in data collection to ensure that our data is comprehensive and unbiased. Other factors considered in data collection were age at which English was introduced, most fluent mother tongue, and country of residence. Figure 2 and 3 shows breaks down the collated dataset accross the various considered factors. From these figures, we can observe that our dataset skewed towards the Yoruba ethnicity and we did not capture the Hausa speakers in the course of our campaign.

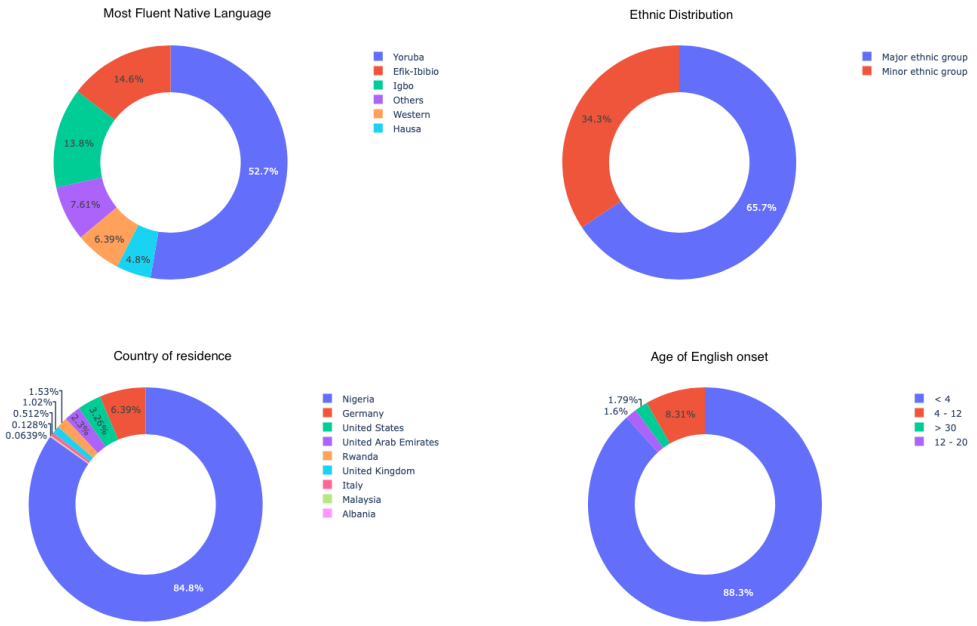

Figure 2: Distribution of SautiDB dataset across various factors considered

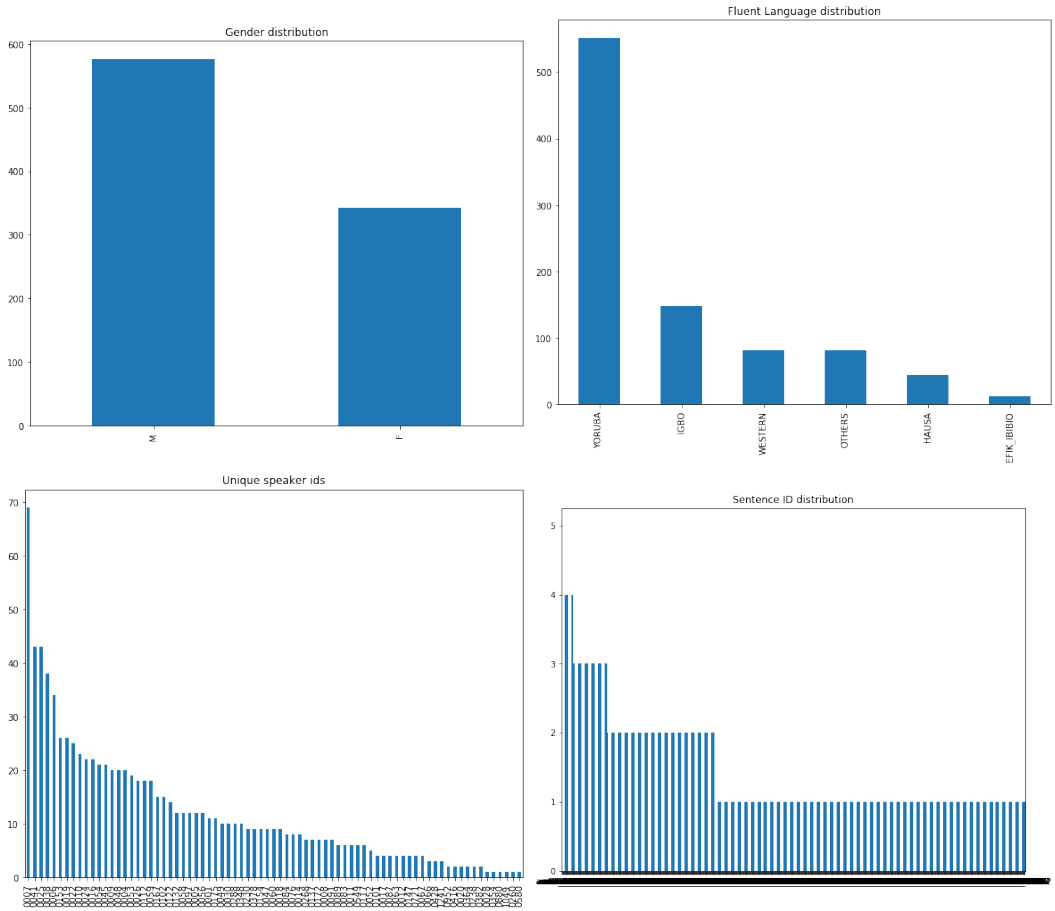

Figure 3: Histogram representation of various factors considered

3.5 Cleaning and Postprocessing

Since we collected our speech data in an uncontrolled manner, we post-processing our dataset in stages.

3.5.1 Filtering Audio

Two members of the project (one expert and another project member) listened to all 1615 samples and rejected some samples based on the following:

- Repetition of words, stuttering or self-editing during reading
- Distortion on speech (from codec or other elements of the recording chain), sometimes the distortion is just on the front/attack portion of each vocalization
- Transient distortion (clicks/pops/), whistling sounds, non-stationary noise in background (like Danfo driving by)
- Breathing or turbulent noise on microphone and proximity effect on microphone
- Audio gating effect on the microphone/recording-software
- Singing or loud discussion in the background.
- Muffled or unintelligible speech, extremely quiet audio, cut off words or incomplete sentences.
- Noise immediately the microphone open

3.5.2 Add Missing Labels

Since we omitted gender and sentence IDs from the data collection, we set up another labeling job on Label Studio[1]. We deployed this application on an AWS server so that our labelers could access it via a public url[2]. Two volunteers (male and female) from the AI Saturdays Lagos[3] community were paid to label the remaining 970 samples.

3.5.3 Labeling Task Analysis

After the labeling task, we ran a program to check the agreement between the two labelers on some important fields, sentence_id and gender. More samples were mislabeled by the two speakers for gender (28) than for sentence ID (18). We corrected these manually[4].

We removed speech samples with invented sentences and non-Nigerian speakers. After our post-processing, 919 audios remained.

3.5.4 Post-processing of Dataset

We post-processed the dataset to normalize the audio to -0.1db and remove the leading and trailing silence[5]. We also renamed the audio files and then uploaded the dataset to Zenodo [18]. Figure 4 shows a pictorial representation of the data collection pipeline described in this section.

[1] https://labelstud.io/guide/

[2] http://54.71.199.167:8889/tasks?tab=1

[3] https://aisaturdayslagos.github.io/

[4] https://github.com/AISaturdaysLagos/sautidb_postprocessing_scripts/blob/main/sautidb_dataset_exploration.ipynb

[5] https://github.com/AISaturdaysLagos/sautidb_postprocessing_scripts

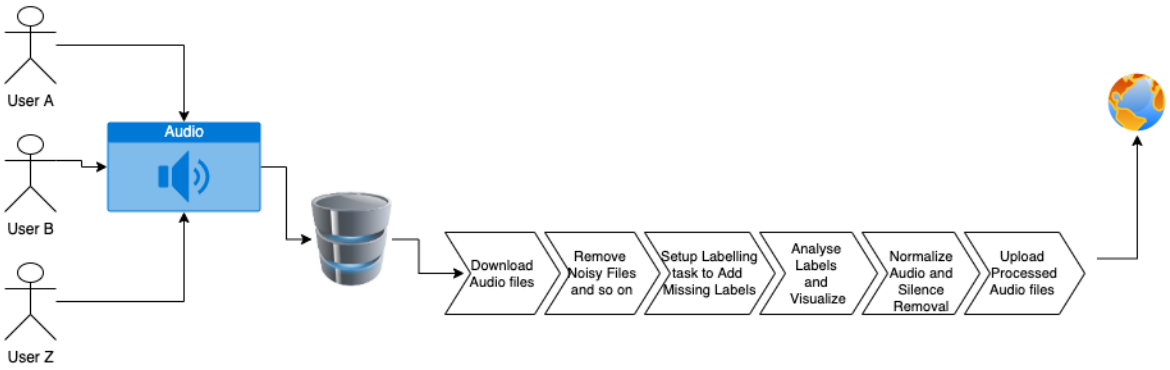

Figure 4: SautiDB data processing pipeline

4 Modeling

To validate our approach, we considered two models from the accent conversion literature. The MelGAN-VC and Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams (FAC-via-PPGs).

4.1 MelGAN-VC

MelGAN-VC is a generative model proposed by [9] for the task of voice conversion. The generator takes as input the mel-spectrogram of a source voice and generates the mel-spectrogram of the target voice. An interesting experiment using MelGAN- VC in [9] was voice conversion from male to female. This was achieved by training on the male and female voices from one dataset, CMU-ARCTIC. Since this model works well for voice conversion, we decided to adapt it to the accent conversion/translation task, recognising that the expected input and output are the same in both cases, namely mel-spectrograms. We hypothesise that this adaptation can be achieved by training the model with a dataset targeting accent conversion (i.e., a dataset containing speech in two different accents), and the generator would be conditioned on one accent (the input) to generate speech in the other accent (the output). To verify this hypothesis, we first reproduced the original voice conversion task, then collated a dataset, and finally trained the model to investigate the qualitative result.

4.1.1 Reproducing Voice Conversion with MelGAN-VC

Using the official MelGAN-VC code and data containing male and female speech from the CMU-ARCTIC database, we trained the model for 50 epochs. The data contains 1132 speech utterances from each part of the database named "CMU US BDL ARCTIC 0.95" and "CMU US CLB ARCTIC 0.95", containing male and female US English speakers, respectively.

Good results[6] were obtained when converting voices from the same dataset, but the model failed when tested with data not included in the database. We judged the results based on quality, determined by our opinion of how good it sounds. As far as we know, this is the metric used to evaluate the performance of speech-generative models. However, we did not assign an opinion score.

With this result, we attributed the generally good performance of MelGAN-VC to overfitting. This suggests that for good performance in any task of converting from a source domain to a target domain, the model should be provided with both the source data and the target data and then trained for the task. In the end, it would learn to generate something similar to the target.

4.1.2 Using MelGAN-VC for Accent Conversion

To extend this model to the task of accent conversion, it should be trained with a dataset that allows the generator to be conditioned on data from one accent to generate data in another accent. For the dataset, we used a combination of data from the CMU ARCTIC database and the L2 ARCTIC

[6] Audio Samples: http://bit.ly/SautiLearn

database. Since our dataset was smaller and consisted of multiple accents and speakers, we thought it might be a bit difficult to evaluate the performance of the model on our dataset. Therefore, we decided to use the L2 ARCTIC, a database of speech from non-native English speakers. The goal is to have a generator that generates non-native English speech when conditioned on US English speech. For this task, we used the "CMU US BDL ARCTIC 0.95" part of CMU ARCTIC and the Male Arabic speaker speech from L2 ARCTIC. Training was performed with the same code but with a different dataset (in this case, a combination of CMU ARCTIC and L2 ARCTIC).

When evaluated, the result showed no evidence of accent conversion. In our evaluation of the result, the model instead attempted to perform voice conversion. We therefore concluded that MelGAN-VC cannot be used for the accent conversion task, at least not in the current parameter settings. If accent conversion is to be tackled, we should be able to extract some notion of accent from our data. Therefore, we considered the model described in Section 4.2.

4.2 FAC-via-PPG

In this paper, the authors trained an acoustic model on a native English speech corpus to extract speaker-independent phonetic posteriorgrams (PPGs) and then trained a speech synthesizer to map the PPGs of a non-native English speaker to the corresponding spectral features, which are in turn converted to audio waveform using a high-quality neural vocoder. The proposed system produces speech that sounds clear, natural and similar to that of the non-native speaker and significantly reduces the perceived foreign accent of the non-native speaker's utterances. This means that the speech content from a native English speaker is used to synthesize a non-native speaker's speech, but with a reduced accent, i.e. the non-native speaker still sounds like themselves but with a less pronounced accent. This is somewhat different from our aim. In our case, we would like the synthesized non-native speaker to retain their accent, but to be as close as possible to the native speaker's voice, so as not to distort the learning experience. When we evaluated the pre-trained models with a native American professor's lecture downloaded from Youtube [7], we found that the model performed as expected. We therefore hypothesized that we could adapt this model to work with our dataset, i.e., instead of using the provided pre-trained ppg-to-mel model trained with speakers from the L2 Arctic corpus, we trained the model with the SautiDB corpus. We have described our experimental approach below.

Inorder to validate the model, we did the following:

1. Downloaded the model checkpoints folder which contains the tacotron2 and waveglow checkpoints for the two native speakers (ykwk and zhaa)
2. Download a lecture video with an American professor accent on Youtube. The lecture is about 45min long
3. Use Audacity tool[8] to convert video to audio, then we chunk the 45mins audio into 5secs audio speeches
4. Run the ppg-to-mel (tacotron2) model on the raw wavfiles of the source (american professor) to get the predicted mel spectrograms
5. Run the mel-to-wav (waveglow) model on the predicted mel spectrograms to reconstruct this back to audio wave form
6. Use the Audacity tool to combine the different chunks.

From this experiment, we can confirm that the model was able to reconstruct native audio speech in the non-native speaker's voice, which is quite impressive[9]. We then repeated this experiment with our dataset. The paper mentions that we only need about 1 hour of speech to train these models (from a single speaker). Although our corpus (clean set) is 53 minutes long, it contains multiple speakers and thus might not perform as expected. Nevertheless, we proceeded with it and present the results of the ppg-to-mel model. At the time of writing, we have not yet completed training the mel-to-audio model and thus have not yet published results for this model. Our experiment is therefore inconclusive at the moment, we intend to report the result as soon as we have it.

[7] https://www.youtube.com/watch?v=DwqNsSTjgbc

[8] https://www.audacityteam.org/

[9] Audio Samples: http://bit.ly/SautiLearn

4.3 PPG-to-Mel Model

We trained a ppg-to-mel model, as described in the paper's Github repository[10], with the default hyperparameters. Table 1 shows the details of the dataset partitioning, as well as the gender representation in each subset. We trained this model on Azure, Nvidia Tesla 80 12GiB server and the training was done for 4 days. We visualize the training process and show some results in Figure 5, 6, 7 and 8. The predicted mel-spectrogram suggests that the model has learned to reconstruct the target mel-spectrogram with some notable differences in some high-frequency bands. We hope to validate this hypothesis with our trained mel-to-wav model.

Subset	**# Female**	**# Male**	**Total**
Train	275	443	718
Valid	64	121	185
Test	4	12	16
Total	343	576	919

Table 1: Train-Test-Val Split of SautiDB Corpus

[10]https://github.com/guanlongzhao/fac/-via/-ppg

Step:17K

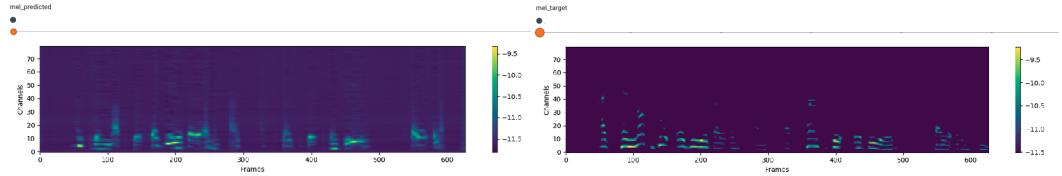

Figure 5: Left: Predicted Mel-spectrogram after Postnet at 17K step. Right: Target Mel-spectrogram at 17K step

Step:30.3K

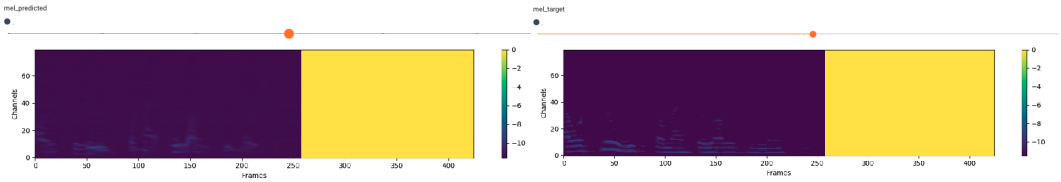

Figure 6: Left: Predicted Mel-spectrogram after Postnet at 30K step. Right: Target Mel-spectrogram at 30K step

Step:63.4K

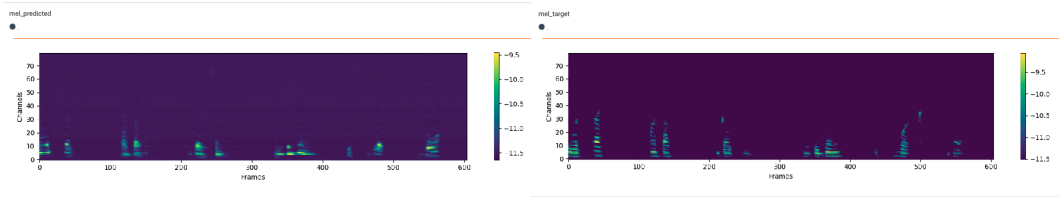

Figure 7: Left: Predicted Mel-spectrogram after Postnet at 63.4K step. Right: Target Mel-spectrogram at 63.4K step

Step:108.8K

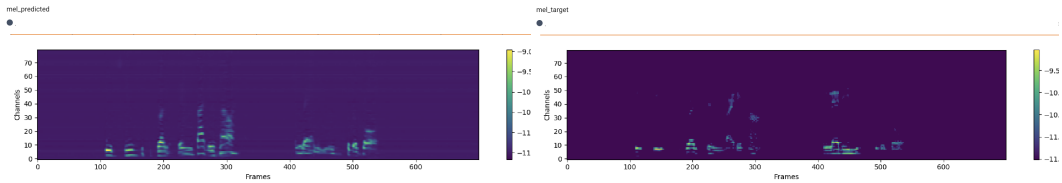

Figure 8: Left: Predicted Mel-spectrogram after Postnet at 108.8K step. Right: Target Mel-spectrogram at 108.K step

5 Larger Vision

Consistency with our motivations for this project is found in the 4th and 10th Sustainable Development Goals (SDG). The 4th SGD concerns Quality Education, while the 10th SGD concerns reducing inequality. For students who prefer to learn in a familiar accent, our tool could be useful to enrich their learning experience, which inevitably improves the quality of their education. The quality of education is one of the factors that influence an individual's income, which also contributes to unequal opportunities.

6 Ethical Review

A major concern, among others, is the potential for abuse, manipulation, or ridicule to which our technology may be directly or indirectly exposed. We have formulate our accent translation problem in a way that ensure only the change of speech style and not speech content. This formalization ensures that the inherent use of the developed technology is only for educational purposes and not for misinformation. We have also released our dataset under creative commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) to ensure that our dataset is only used for research purposes. It is critical to note that the goal of SautiLearn is to reduce the barrier to learning often created by accent variations and to ensure that online learning resources are accessible to all.

7 Conclusion

We presented SautiDB, an English database of non-native Nigerian accents to support the development of machine learning models for accent conversion or translation and classification tasks. We evaluated a sequence-to-sequence model on our dataset and the results show that the model is able to synthesize the target speaker, despite our small and multi-speaker dataset. In the future, we would like to complete the experiment described in section 4.2 and evaluate its effectiveness on our local dataset. We would also like to collect more data using our SautiDB platform, especially extending its application to the northern part of Nigeria, since our current corpus does not consist of users from this region.

8 Acknowledgement

This work was supported by AI4D-IndabaX awards IDRC Grant Number: 109187-002. We would like to express our deep gratitiude to Iroro Orife for guidance during the second half of this project, his contribution to the post-processing phase of the SautiDB-919 dataset was invaluable. We would also like to thank SautiDB app users for their contribution to this research.

References

[1] Leo Parker Dirac, Fabian Moerchen, and Edo Liberty. Accent translation, December 25 2018. US Patent 10,163,451.

[2] Nehal Mamgain, Arjun Sharma, and Puneet Goyal. Learner's perspective on video-viewing features offered by mooc providers: Coursera and edx. In *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)*, pages 331–336. IEEE, 2014.

[3] Adriana A Vinke, Joke Snippe, and Wim Jochems. English-medium content courses in non-english higher education: a study of lecturer experiences and teaching behaviours. *Teaching in Higher Education*, 3(3):383–394, 1998.

[4] Guanlong Zhao, Shaojin Ding, and Ricardo Gutierrez-Osuna. Foreign accent conversion by synthesizing speech from phonetic posteriorgrams. In *INTERSPEECH*, pages 2843–2847, 2019.

[5] Guanlong Zhao, Sinem Sonsaat, John Levis, Evgeny Chukharev-Hudilainen, and Ricardo Gutierrez-Osuna. Accent conversion using phonetic posteriorgrams. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5314–5318. IEEE, 2018.

[6] Guanlong Zhao and Ricardo Gutierrez-Osuna. Using phonetic posteriorgram based frame pairing for segmental accent conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1649–1660, 2019.

[7] Yi Zhou, Xiaohai Tian, Haihua Xu, Rohan Kumar Das, and Haizhou Li. Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6790–6794. IEEE, 2019.

[8] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR, 2020.

[9] Marco Pasini. Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms. *arXiv preprint arXiv:1910.03713*, 2019.

[10] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6820–6824. IEEE, 2019.

[11] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273. IEEE, 2018.

[12] EH Rothauser. Ieee recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246, 1969.

[13] John Kominek and Alan W Black. The cmu arctic speech databases. In *Fifth ISCA workshop on speech synthesis*, 2004.

[14] Guanlong Zhao, Sinem Sonsaat, Alif O Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-arctic: A non-native english speech corpus. *Perception Sensing Instrumentation Lab*, 2018.

[15] Ademola Malomo, Clinton Mbataku, Olumide Okubadejo, and Tejumade Afonja. Sautidb data collection platform, 2020.

[16] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, and Keiichi Tokuda. The hmm-based speech synthesis system (hts) version 2.0. In *SSW*, pages 294–299. Citeseer, 2007.

[17] Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. Accentdb: A database of non-native english accents to assist neural speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5353–5360, Marseille, France, May 2020. European Language Resources Association.

[18] Tejumade Afonja, Clinton Mbataku, Ademola Malomo, Olumide Okubadejo, Lawrence Francis, Munachiso Nwadike, and Iroro Orife. Sautidb: Nigerian accent dataset collection, February 2021.