Applied Computer Science, vol. 16, no. 3, pp. 56–68
doi:10.23743/acs-2020-21

Submitted: 2020-06-18
Revised: 2020-09-03
Accepted: 2020-09-21

Spoken Language Recognition, Computer Vision, Image Recognition, CNN

Nancy WOODS [0000-0001-8396-3467]*, *Gideon BABATUNDE**

A ROBUST ENSEMBLE MODEL FOR SPOKEN LANGUAGE RECOGNITION

Abstract

The identity of a language being spoken has been tackled over the years via statistical models on audio samples. A drawback of these approaches is the unavailability of phonetically transcribed data for all languages. This work proposes an approach based on image classification that utilized image representations of audio samples. Our model used Neural Networks and deep learning algorithms to analyse and classify three languages. The input to our network is a Spectrogram that was processed through the networks to extract local visual and temporal features for language prediction. From the model, we achieved 95.56 % accuracy on the test samples from the 3 languages.

1. INTRODUCTION

Speech is an important means of human communication. Recently, speech serves a means of interaction between machines and humans as seen in voice control and commands, map navigation/guide, robotics, intelligent assistants like "Siri", "Alexa","Bixby" e.t.c. Thoughts and ideas are exchanged through speech and statistics shows that there exist over 7,111 unique languages of the world (Eberhard, Simons & Fennig, 2020).

There are several attributes contained in speech utterance which can be extracted via machines, and over the years, efforts have been made by researchers to create methods for extracting the fundamental information which a speech utterance conveys. This led to the development of various extraction modules such as 'Speech to Text', Speaker Recognition, Topic Identification, Spoken Language Recognition and many ways to understand the semantic meaning which speech utterance coveys via machines. (Li, Ma & Lee, 2013).

* University of Ibadan, Faculty of Science, Department of Computer Science, Oyo State Ibadan, Nigeria, chyn.woods@gmail.com, gideonbabs2@yahoo.com

Spoken Language Recognition (SLR), also known as Automatic Language Identification, is the process by which the identity of a language in a given speech sample is detected. Spoken language recognition task is perceived as a pre-processing step for Speech technologies such as Automatic Speech Recognition. SLR can also be applied as a standalone task. (Zissman, 1993; Muthusamy, Cole & Oshika, 1992)

In the area of Spoken Language Recognition, studies have shown that it is part of human intelligence to distinguish between languages, and this is a natural ability we are born with (Zhao et al., 2008). It was also discovered that with minimal exposure to a language, humans can detect the identity of the language being spoken in a conversation with reference to languages they have heard or languages they know. Although these judgments may be less precise when hard decisions need to be made for an identification task, they show that human listeners can apply auditory perception with linguistic knowledge at different levels to distinguish between broad language groups (Li, Ma & Lee, 2013). Most SLR systems are based on high level features such as Frequency, Phonotactics, Prosodic and Acoustic-Phonetic Modelling. Such systems have an inherent problem: tokenizing the features accurately.

In this study we developed a model for Automatic Spoken Language Recognition Systems (ASLRs) from a Computer-vision perspective, using deep learning algorithms, similar to that proposed by (Bartz, Herold, Yang & Meinel, 2017). We proposed an ensemble of Convolutional Neural Network (CNN)-Recurrent Neural Network (RNN) algorithm. The system adopts existing algorithms, with a variant in network architecture on the deep learning techniques. Furthermore, training and testing of the system was carried out on 3 Spoken Nigerian Languages (English, Yoruba and Igbo). The research does not consider full development of a SLRs, nor the other aspects of Speech information extraction such as Automatic Speech Recognition (Speech to text transcription), Language Translation, Machine hearing/ Language understanding, linguistic analysis etc. Our approach is restricted to the computer vision perspective and not the advanced signal processing techniques or statistical modelling approach.

2. REVIEW OF RELATED LITERATURE

Language recognition systems are usually categorized by the features they use, such as the acoustic–phonetic approach, the phonotactic approach, the prosodic approach, and the lexical approach. More recently, newer features for identification have surfaced which do not fall into any of these categories thanks to deep learning. These are lower level features like ‘Spectrogram images extracted from sounds. The mainstream research on spoken language recognition adopts techniques utilizing these higher level features (Torres-Carrasquillo et al., 2002).

Acoustic Phonetics refers to the wide range of sounds that the human speech apparatus is capable of producing. Speech sounds as concrete acoustic events are referred to as phones. Whereas speech sounds as entities in a linguistic system are termed as phonemes (Kirchhoff, 2006). The phonotactic constraints dictate the permissible phone sequences. Each language has its unique set of lexical–phonological rules that govern the combinations of different phonemes. Phonemes can be shared considerably across languages, but the statistics of their sequential patterns differ very much from one language to another. Prosody refers to supra-segmental features in running speech, such as stress, duration, rhythm, and intonation (Ashby & Maidment, 2005). The set of interrelated prosodic features are all important characteristics of spoken languages. Prosody appears to be useful for distinguishing between broad language classes (e.g., tonal versus non-tonal languages). However, human listening experiments reported in (Navratil, 2001), and (Ramus & Mehler, 1999) show that prosodic cues are less informative than the phonotactic one. In the past few decades, researchers have explored many speech and language knowledge sources, including articulatory parameters, acoustic features (Sugiyama, 1991), prosody (Adami & Hermansky, 2003), phonotactic (Zissman, 1996), and lexical knowledge (Adami & Hermansky, 2003).

(Safitri, Zahra & Adriani, 2016) carried out a study involving the identification of spoken data in three local Indonesian languages: Minangkabau, Sundanese and Javanese. In their study, two phonotactic methods were used, namely Phone Recognition followed by Language Modelling (PRLM) and Parallel Phone Recognition followed by Language Modelling (PPRLM). PRLM method showed the highest accuracy using the phone recognizer trained for English and Russian with the average of 77.42% and 75.94% respectively. From their study, observation was made that accuracy of Spoken LID system with PRLM and PPRLM methods are affected more by the performance of phone recognizer that is used.

In the study carried out by (Boussard, Deveau & Pyron, 2017), several machine learning techniques for classifying spoken language were explored. They applied algorithms which utilized various spectral features derived from English and Mandarin Chinese phone call audio to predict the language of conversation. They assert that to a large extent, a language is not distinguished by the presence of certain sound waves, but rather by the patterns they form and the sequence in which they are produced. The information was incorporated explicitly via Shifted Delta Cepstrum (SDC) features and using Gaussian Mixture Method (GMM) and neural network models, they were able to effectively capture this crucial information, leading to improved predictive power. The modelling assumptions of the GMM ultimately turned out to be vital since they had only limited data.

According to (Abdel-Hamid et al., 2014) exploration of Deep Neural Networks (DNN) revealed that the hybrid deep neural network (DNN)-hidden Markov model (HMM) showed significant improved speech recognition performance over the conventional Gaussian mixture model (GMM)-HMM. This improvement is attributed to the inherent ability of DNN's to model complex correlations

in speech features. Their study attempted to further reduce error rate in the underlying model by using Convolutional Neural Networks (CNNs).

These studies, among several others, have shown that high accuracies in language recognition can be achieved, depending on the adopted model for development of the SLRs. To this effect, we undertook our study exploring a different approach that was motivated by the recent successes in the area of deep learning and Computer Vision.

3. METHODOLOGY

The field of Computer vision deals with utilization of techniques to help machines understand the content of digital images. It involves the extraction of information from images to infer something about a real world problem. The application of computer vision techniques to this problem domain, meant audio samples had to be represented with images for further processing. We applied Convolutional Neural Network-Recurrent Neural Network (CNN-RNN) in an ensemble, to the development of our model for SLRs thus, the approach is image processing based. More specifically, these Deep Neural Networks were adapted to the problem of identifying the language of a given spoken utterance from short-term spectral features.

3.1. Dataset

Our model was trained and tested using the three languages: Yoruba (Ibadan dialect), Igbo, and English. Audio recordings of conversations in all the three languages were acquired from various sources such as Radio streams, Video Streams and other available online corpus for research purposes (Kaggle, Librovox). These audio samples served as the dataset. The total data set consists of over 2100 wav and mp3 files, with an average of 700 samples per language.

Tab. 1. Size of the Dataset

Language	Length of Files (secs)	Average Time (secs)
English	$2 < X < 5$	3
Igbo	$4 < X < 10$	5
Yoruba	$2 < X < 8$	5

From these 2100 files, we separated 80% of the sample files for training and the rest for validation and testing. The recordings had varying length of approximately 10 seconds. Some of the audio files contained background noise, intermittent laughter, music and other unwanted properties while some samples were noise free.

3.2. Pre Processing

As part of our pre-processing step, all the mp3 format files were first converted to a lossless wav format at a 22050 kHz/16bit sampling rate. We investigated audio specific techniques to denoise and remove unwanted properties of our audio samples and resorted to manually denoising each file thereby utilizing an audio engineering tool called FL Studio (Fig. 1). We removed unnecessary silences between sentences, intermittent jingles or unwanted background sounds and we removed noise stemming from recording apparatus (mic and general audio recording setup), while maintaining characterizing features of the Samples. This process produced clean versions of the audio samples as wav files.

Fig. 1. Wav Sample Pre-processing

In other to adopt an image processing approach to the Language recognition problem rather than the conventional audio processing approach, an image representation of the audio sample was utilized. We discovered that spectrograms (Fig. 2) are the standard ways to represent audio for deep learning systems from our investigations and also according to recent studies (Park et al., 2019; Amodei et al., 2015). Spectrograms are 2D visual representations of audio frequency spectra over time. The image depicts the intensity of sound around certain frequencies as time varies. A major point about speech is that the sounds generated by humans are filtered by the shape of the vocal tract including tongue, teeth, etc. This shape determines what comes out, and it gives an accurate representation of the phoneme being produced. More specifically, the shape of the vocal tract manifests itself in the envelope of short term power spectrum, and the job of the Mel-scepstrum is to accurately represent this envelope. Mel-Spectrogram images of our training files were created, and these mel-spectrogram images serve as the input to our DNN-based model.

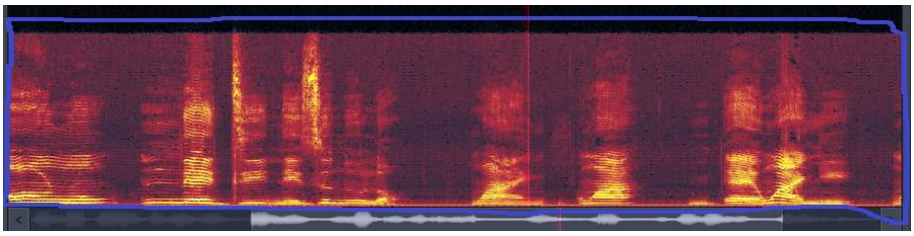

Fig. 2. Sample of Spectrogram Image of an Audio file

The mel-spectrogram can be viewed as a sequence of column vectors that consist of 256 (or 128, if only <5.5 KHz frequencies are used) numbers. We considered only frequencies of less than 5.5 KHz.

3.3. Network Architecture (CNN-RNN Ensemble)

As illustrated in Figure 3, the overall architecture of our model comprised of 3 stages. In the first stage, a convolutional feature extractor was utilized. The convolutional feature extractor transformed our input (Spectrogram image of recordings) into a feature map through several series of processing. This CNN algorithm was used in the model because they can transform high level information in images with great capacity thereby improving predictive power of our model. The output of the convolutional neural network (feature map) was then fed into a variant of the Recurrent Neural Network (RNN) architecture known as the Bi-directional Long-short term memory (having two LSTM layers). This Bi-LSTM was introduced in the model due to their ability to store information of both past and future sequences at the same time, such that at every point in time large, information is available to the model thereby improving chances of obtaining higher classification accuracy. The data sequence (RNN's Output) is further passed to a fully connected layer in order to solidify training efficiency of our model. Finally, a Softmax layer was introduced for classification purpose. The output of the model is the detected language.

20% dropout was factored in our model in order to reduce overfitting in the network. The Activation function, Rectified Linear Units activation function (ReLU) was used in this model to increase non linearity of the CNN. We also utilized batch normalization in the model to increase the stability of the neural network. This model used Adam optimization algorithm instead of the classical stochastic gradient descent procedure to update the network weights.

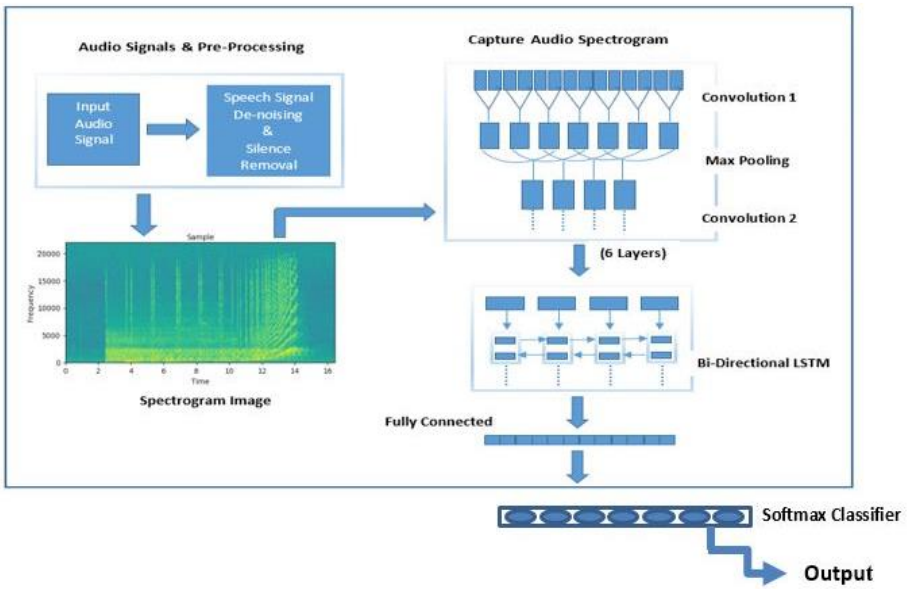

Fig. 3. Model for Spoken Language Recognition

3.3.1. Convolutional Neural Network

CNN's have the ability to capture and transform high level images with great capacity and our model leveraged on this property to explore its performance on the input features (Spectrogram images). As shown in Figure 4, the network had 6 convolutional layers thus a deep convolutional network. It consisted 6 blocks of 2D convolution, ReLU nonlinearity, 2D max pooling between each layer and batch normalization. 3x3 filters was used for all the convolutional layers with a stride 1. Pooling size was always 3x3 with a stride 2. The network used "Same" padding throughout. Learning rates were set to be higher for the first convolutional layers and lower for the top convolutional layers. We trained the CNN on Keras framework with Python with 32 images in a batch. This significantly increased the training speed.

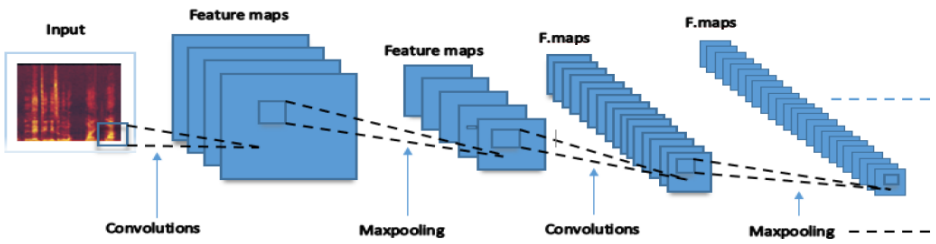

Fig. 4. Convolutional Neural Network Architecture

3.3.2. Recurrent Neural Network

The RNN accounts for the sequential characteristic of the audio data; therefore we applied a Bi-directional long short term memory (BLSTM). As shown in Figure 5, the output of the CNN were sets of several channels (*feature maps*). These feature maps were then reshaped to the RNN input dimension (which takes 3dimensional input). Two layers of LSTM in opposite directions captured and stored information on sequences from the past and future set. These units then combine and were fed to a fully connected layer. The Bi-directional LSTM makes large amount of information available in the network.

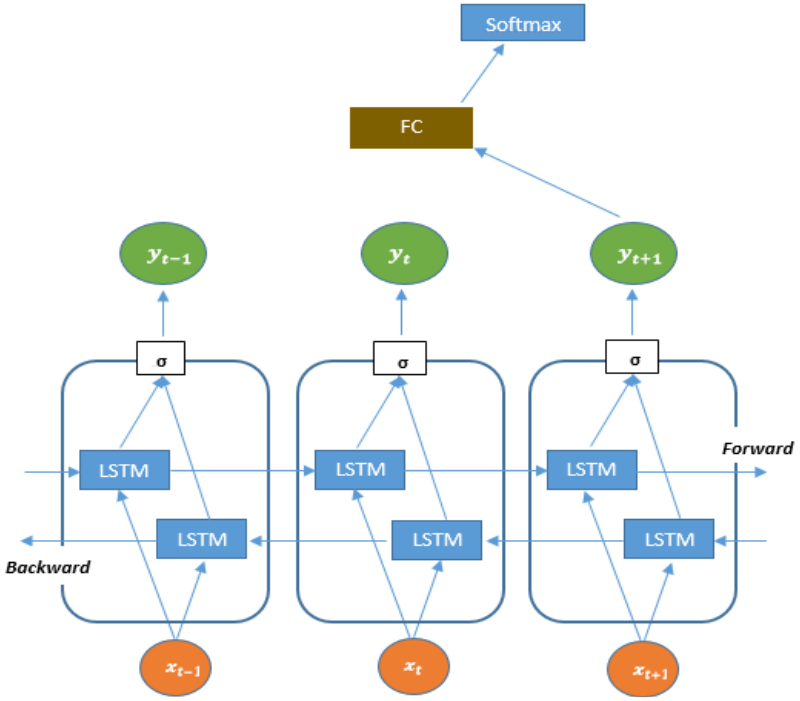

Fig. 5. Bi-directional Long Short Term Memory–RNN

Finally we used only 1 fully connected layer between the RNN (Bi-directional LSTM) and the Softmax layer, and apply 20% dropout on that. The fully connected layer had 64 neurons and was trained using a Softmax loss. The output of the Softmax layer is the predicted language. We utilized Keras framework for Model implementation.

4. RESULT

Our system was trained on 1,437 samples and tested with 360 samples all comprising the 3 languages. Although we had over 2100 files in our dataset, some samples were discarded during the processing stage, because we wanted a uniform dimension (128 x 128) for all our spectrogram images for less than 3secs of spoken utterance (~3s). The CNN algorithm which was used required uniformity for all training set, as such, all samples that did not meet this shape were discarded. For over 100 Epochs, the accuracy of the model was given at 95.56%

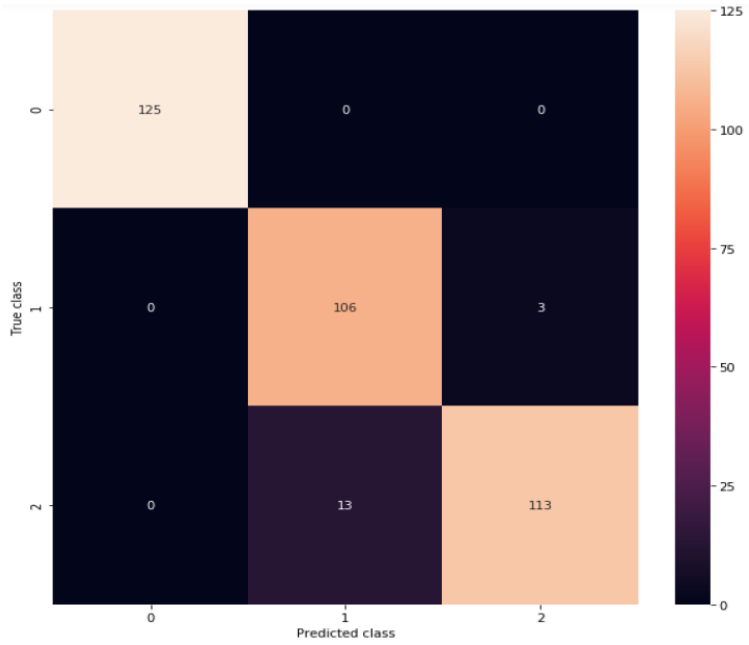

Fig. 6. Confusion Matrix

Shown in Figure 6 is the confusion matrix which depicts the classification correctness of the model. It shows the predicted class and the true class and shows where the classification lies for every input. We worked on 3 Languages (English, Yoruba, and Igbo) therefore an audio sample having an utterance or conversation in any of the languages should be detected by the model by correctly classifying it in one of the 3 classes. The English language was labelled with index 0, Yoruba was labelled with index 1 and Igbo language was labelled with index 2. We tested 125 English samples with the model, and it classified all 125 samples correctly (as belonging to English class-index 0). From observation, there was no mis-classification, hence the predicted class and the true class match. For Yoruba language, 109 audio samples were tested on the model, and from that, 106 were

accurately classified while 3 samples were misclassified as Igbo. And finally, for the Igbo language, 126 test samples were tested on the model, 113 were classified correctly, while 13 samples were misclassified as Yoruba language.

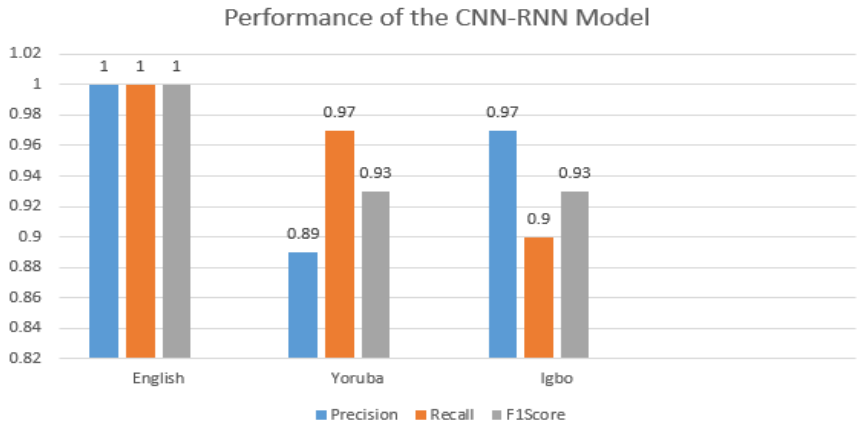

Fig. 7. Model evaluation using Precision, recall, and F1 Score as metrics

Figure 7 depicts the performance of the model in terms of precision, accuracy, and the F1 score. The precision values show the extent to which the model captures the true classes of a sample out of the total. It is a ratio of correctly predicted positive observations to the total predicted positive observations. And from the chart, we see that English has the highest precision value followed by that of the Igbo language and then the precision of the Yoruba classification is fair in relative terms. The recall value depicts that given the total test samples, how many elements were captured.it is the ratio of correctly predicted positive observations to the all observations in actual class.

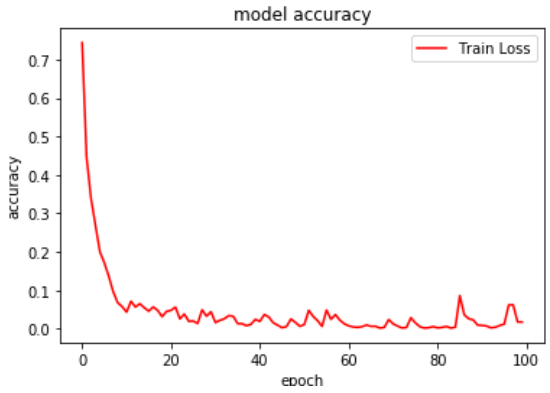

Fig. 8. Training Loss of the Model

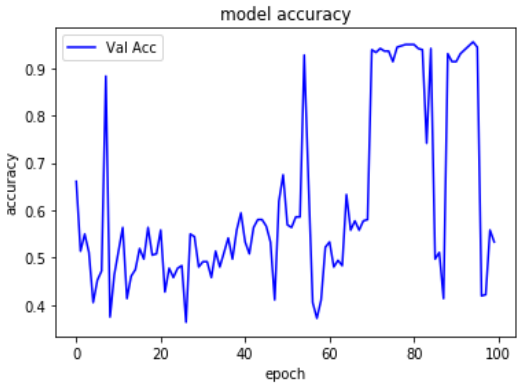

Fig. 9. Validation Accuracy of the Model

From Figure 8, it is observed that over 100 epochs of training the loss (which is the degree of error) is actually minimized. This shows that the level of accuracy of the model after training is very near optimum, which is a desired property in every model. From figure 9, it is observed that validation accuracy fluctuates but becomes stable towards last few epochs. This is also a desired property which shows that the model can be functional.

5. CONCLUSION

The high performance of our CNN-RNN deep learning algorithm based model for SLR shows that approaching the Language recognition problem domain from an Image classification perspective yields comparable optimal performance to the mainstream phonological computational and statistical modelling approach. This observation also shows that using intermediate features such as a Spectrogram is adequate enough to obtain improved performance for SLRs, thereby eliminating the need for large corpus bearing phonetically transcribed data.

Based on the performance of the system on short speech excerpts, we infer that the system can classify even short speech utterances of less than 3s (~3s) thus, it is a long enough interval to classify a spoken language correctly with this model. Noises in speech samples affect SLR performance considerably as observed in our experiments. The observation made was that the English samples were properly classified due to the very minimal noise contained in the training sample. The English dataset had far less noise because it was created specifically for training purposes (Kaggle, Librovox), while the Yoruba and Igbo language testset had some misclassifications stemming from the noise contained in the Training set. Although these samples went through series of denoising, they were not as clean as that of the English because they contained white-noise which is almost impossible to completely remove. The white-noise was from the recording apparatus in the radio station.

We observed that the architecture of the deep learning model used for developing SLR's is pertinent to the performance of a model, therefore we recommend the exploration of deeper architectures for future works. Google's Inception v3 is one of such networks which has much more layers in its architecture and is considered deeper than ours. We believe a deeper network should be able to extract more general features thus leading to increase in accuracy although they come with an increase of computational cost, as the Inception-v3 model uses up to six times more parameters, than a regular CNN. We also suggest utilizing completely noise free samples as dataset for training models.

REFERENCES

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *22*(10), 1533–1545. https://doi.org/10.1109/taslp.2014.2339736

Adami, A., & Hermansky, H. (2003). Segmentation of speech for speaker and language recognition. *EUROSPEECH-2003* (pp. 841–844). Geneva. Retrieved from https://www.academia.edu/32317887/Segmentation_of_speech_for_speaker_and_language_recognition

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., ... Narang, S. (2015). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *CoRR, abs/1512.02595*. Retrieved from https://arxiv.org/abs/1512.02595v1

Ashby, M., & Maidment, J. (2005). *Introducing phonetic science*. Cambridge University Press.

Bartz, C., Herold, T., Yang, H., & Meinel, C. (2017). Language Identification Using Deep Convolutional Recurrent Neural Networks. In D. Liu, S. Xie, Y. Li, D. Zhao, & E. El-Alfy (Eds.), *Neural Information Processing ICONIP 2017. Lecture Notes in Computer Science* (vol. 10639). Springer. https://doi.org/10.1007/978-3-319-70136-3_93

Boussard, J., Deveau, A., & Pyron, J. (2017). *Methods for Spoken Language Identification*. Retrieved from http://cs229.stanford.edu/proj2017/final-reports/5239784.pdf

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2020). *Ethnologue: Languages of the World*. Retrieved from http://www.ethnologue.com

Kirchhoff, K. (2006). Language characteristics. In T. Schultz, & K. Kirchhoff (Eds.), *Multilingual Speech Processing* (pp. 5–33). Elsevier.

Li, H., Ma, B., & Lee, K. A. (2013). Spoken Language Recognition: From Fundamentals to Practice. *Proceedings of the IEEE, 101*(5), 1136–1159. https://doi.org/10.1109/JPROC.2012.2237151

Muthusamy, Y. K., Cole, R., & Oshika, B. (1992). The OGI multi-language telephone speech corpus. *Int. Conf. Spoken Lang. Process*, 895-898. Retrieved from https://pdfs.semanticscholar.org/aad7/274fdd57191e89f9df2880a50ec14581d671.pdf

Navratil, J. (2001). Spoken language recognition A step toward multilinguality in speech processing. *IEEE Trans. Speech Audio Process*, *9*(6), 678–685. https://doi.org/10.1109/89.943345

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proc. Interspeech 2019* (pp. 2613–2617). https://doi.org/10.21437/interspeech.2019-2680

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech re-synthesis. *Journal of Acoustical Society of America*, *105*(1), 512–521. https://doi.org/10.1121/1.424522

Safitri, N. E., Zahra, A., & Adriani, M. (2016). Spoken Language Identification with Phonotactics Methods on Minangkabau, Sundanese, and Javanese Languages. *Procedia Computer Science 81* (pp. 182–187). Elsevier. https://doi.org/10.1016/j.procs.2016.04.047

Sugiyama, M. (1991). Automatic language recognition using acoustic features. *International Conference on Acoustics, Speech, and Signal Processing* (pp. 813–816). Toronto. https://doi.org/10.1109/icassp.1991.150461

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., & Deller, J. (2002). Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In *ICSLP-2002* (pp. 89–92). Denver. https://doi.org/10.1109/icassp.2002.5743828

Zhao, J., Shu, H., Zhang, L., Wang, X., Gong, Q., & Li, P. (2008). Cortical competition during language discrimination. *NeuroImage*, *43*(3), 624–633. https://doi.org/10.1016/j.neuroimage.2008.07.025

Zissman, M. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, *4*(1), 31–44. https://doi.org/10.1109/icassp.1993.319323

Zissman, M. A. (1993). Automatic language identification using Gaussian mixture and hidden Markov models. *IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 2, pp. 399–402). IEEE. https://doi.org/10.1109/tsa.1996.481450