ISSN: 2583-9292 (ONLINE)
DOI: 10.xx/ijdsml.2023.xx

ANALYZING LANGUAGE IN MULTILINGUAL SPEECH USING DEEP NEURAL NETWORK

Mehali Vyas[1], Awadh Kishor Singh[2] and Nidhi Parmar[3]

[1]*Department of Computer Engineering, Sarvajanik College of Engineering and Technology, India*
[2,3]*Department of Computer Engineering, UPL University of Sustainable Technology, India*

Abstract

Language recognition is the process of determining the language spoken. Motivated by the impressive gain in the performance of language recognition, we adapt the deep neural networks to the problem of language recognition analysis. In the prior work we consider the application of education institute, where teachers comprehend their speech in multiple languages. We then analyze the different aspects of primary (English) and secondary (Hindi, Gujarati) languages spoken. We have prepared a basic work flow of the proposed solution. Speech features are modelled by MFCC parameterization method. In the current work we have used Long Short-Term Memory (LSTM) neural network as our deep neural network model. Results are carried out on Indic TTS dataset provided by IIT Madras. Our result shows that Deep Neural Network (DNN) gives evident results especially when the amount of training data is more. Analysis is done considering various cases 1) 1 hidden layer - 4 hidden layer LSTM network 2) 10s – 15s audio 3) small – large dataset. The analysis carried out can be helpful at administration level to measure the quality of teaching in a class which can lead to improvement in the education system.

Keywords:

Language Recognition, Deep Neural Network, LSTM Neural Network, Multilingual Speech

1. INTRODUCTION

Automatic Spoken Language Recognition (SLR) is the process of determining the language spoken. Language Recognition system has become essential technology for multilingual application [1]. The process of Spoken Language Recognition (SLP) is the front end of many applications such as human-machine interaction through speech multilingual conversational system, multilingual speech recognition, spoken language translation and spoken document retrieval [1].

Language recognition system can also be used in emergency call routing, where the fluent native operator might be critical for the response time [6]. Many of such applications face the problem of presence of multiple languages in the input. So, here very important point is ability of machine to distinguish between the languages. Spoken Language Recognition systems are further classified into two groups [2], (1) explicit language recognition system and (2) implicit language recognition system. The former system requires segmented and labelled speech corpus as input, whereas the latter system requires only raw speech as input.

Person can identify the language if he knows the phonemes, words, syllables and sentence structure. Cues about any language can be obtained from: phonotactic, phonetic, lexical and prosodic [2]. It's easy to identify the language from same family, but for language belonging to different family requires high degree of training. For example, if the person knows Hindi, languages like Gujarati, Bangali are easy to understand even with less knowledge. While it is difficult to understand the language like Kanad, Tamil and Telugu because they belong to another family.

Humans comprehend their speech in multiple languages, words from other languages might be used. It does not create much problem to listeners to identify the primary language of the speaker, tag the foreign words and identify them too if the listeners have the basic knowledge about those foreign words. But what if the listeners do not have any knowledge about those foreign words? This gives exposure to have a reliable language recognition system which is capable to identify even dialects of any language.

Spoken Language Recognition System Consist of two main stages viz. Feature Extraction and Classification [2]. Signal pre-processing operation could be carried out which includes activities like, noise removal and speech signal segmenting.

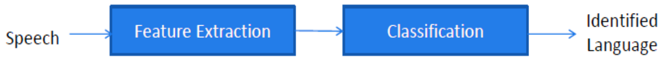

Fig.1. Stages of Language Recognition System

Spoken language recognition enables the technology for wide range of multilingual speech processing application, such one an application can be for educational institutes, where the teachers deliver their lecture in a particular language, say English. Variety of students from different regions or country may study together in a same classroom. In this case, sometimes teachers may comprehend their speech in multiple languages or use words from the local language. Following analysis can be carried out on the speech being delivered in class to improve the teaching learning process

- How many times teacher has spoken other than the primary language (English)?
- What was the other local language used by teacher?
- Is there any disturbance (noise/chaos) in class, by students?
- What was the context or subject delivered by teacher?

The further paper is organized as follows: Section 2 includes literature review, Section 3 consist of the proposed work and experimental setup, which include the details of deep neural network structure and dataset. Section 4 shows the experimental results and the analysis carried out, followed by the conclusion.

2. LITERATURE REVIEW

There is different state of art methods proposed for language identification system. The following section gives overview of previously proposed methods for language identification.

Ignacio Lopez-Monero et al. [6] Have adapted deep neural network approach to identifying the language of the given spoken utterance from the short-term acoustic feature. Here they have used fully connected Feed Forward Neural Network (FFNN). Wei Wang et al. [16] In their work for language recognition have used total variability algorithm for I-vector feature extraction which is combined with deep learning theory for identification task.

Cristian Bartz et al. [19] have proposed the method of hybrid Convolution Recurrent Neural Network (CRNN) for language recognition in image domain, when spectrogram images are operated, they generate the audio snippets. They have used convolutional feature extractor; the extracted feature is used as input to Bidirectional Long Short-Term Memory (BLSTM) for prediction.

Ignacio Lopez-Monero et al. [7] Have proposed two DNN based approaches, one to use DNN as language identification classification and second to use DNN for extracting bottleneck feature.

Output coding strategy is adopted by Bin Ma et al. [1] Where multiclass language recognition problem is decomposed into many binary classification tasks. Each of which addresses a language recognition subtask by using a component classifier, the results are combined to form an output code.

Applications such as multilingual spoken dialog system, database search and retrieval system, face problem of potential presence of multiple unknown languages in the input signal [3]. For language rejection, Phonotactic background normalization method is used. The monolingual Large Vocabulary Continuous Speech Recognition (LVCSR) is used in Phonotactic background language modeling for combination. Language rejection should find possible links of language recognition to dialect identification or foreign ascent and which despite the importance for some applications is an under represented issue.

The literature survey carried out shows that DNN is widely used for speech recognition purpose [2] [6] [10]. The main advantage of DNN is the multilevel distributed representation of input in DNN and it has been proven successful in exploiting large amounts of data [15]. All these factors motivate the use of DNN for language Identification.

3. PROPOSED WORK AND EXPERIMENTAL SETUP

The main objective is to build the Deep Neural Network model which helps to analyze the language in multilingual speech. In the proposed solution for language recognition analysis, we have identified three important phases:

- Data Pre-processing: Raw input speech is pre-processed and segregated according to their duration.
- Feature Representation: For feature extraction we used the Mel-Frequency Cepstral Coeeficient (MFCC) technique.
- DNN model training: LSTM neural network model is used for language model training.

The Fig.2 show the general flow of proposed solution.

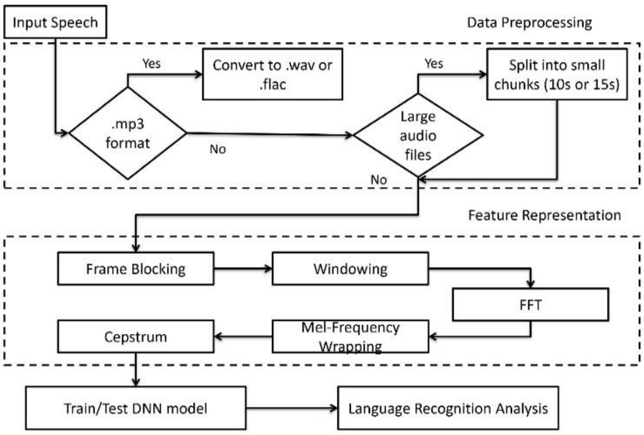

Fig.2. Flow of proposed solution

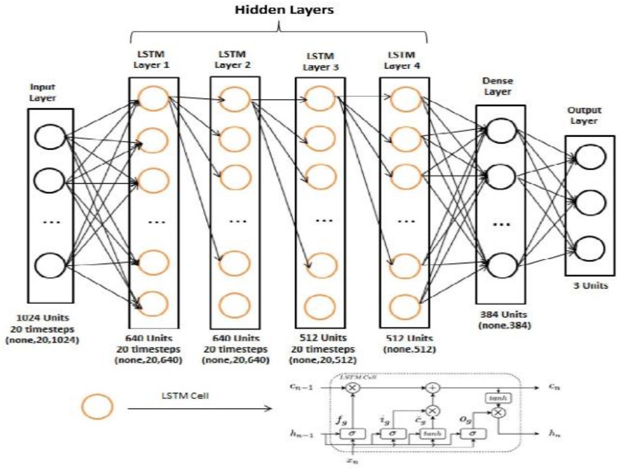

Fig.3 LSTM Network Structure

3.1 DATASET

We have obtained dataset in three language English, Hindi and Gujarati from Indic TTS provided by IIT Madras [32]. Indic TTS is the project developing text-to-speech synthesis in Indian languages [33]-[36]. The speech corpus which we have used covers major Indian languages: Hindi and Gujarati. Each language corpus comprises of utterances each of mono (in our case Hindi or Gujarati) and English recorded by both female and male native speakers. The total speech corpus available including all three languages was around 65 hours. The details and duration of audio available for each language are shown in Table.1. The speech waveform files are available in WAV format.

Table.1. Details of Dataset (In Hours)

	Hindi		Gujarati		Total
	Mono	English	Mono	English	
Female	5.18	7.22	10.33	10	32.73
Male	5.16	7.22	10.92	10.13	33.43
Total	10.34	14.44	21.25	20.13	66.16

3.2 LSTM ARCHITECTURE

LSTM neural network structure with 4 hidden layers is used for DNN model building. The structure of LSTM used is shown in Fig.3. MFCC features are given as input to the model. The shape of input is (20,313), where 20 MFCC features are calculated on 313 frames. As in input layer we have used 1024 units, the input data can be represented as [none, 20,1024]. The parameters for input layer can be calculated as,

$$\text{Parameters} = 4 * (1024 * (1024 + 313) + 1024) = 54,80,448$$

Here the 4 hidden layers will help to make a better use of parameters by distributing them over the space through multiple layers [34]

First two hidden layers have 640 LSTM units while other two hidden layers have 512 LSTM units. This is followed by a dense layer with 384 units and an output layer with 3 units. We have used ReLu activation function on hidden layers and SoftMax activation function on output layer. Dropout with 30% is applied on hidden layer in order to reduce the dependency on training set. We have fixed the learning rate (lr) and mini batch size to 0.001 and 64 samples. The output layer is configured with cross entropy cost function.

4. EXPERIMENTAL RESULT AND ANALYSIS

We have proposed our idea with LSTM, but to evaluate the same we have tested our algorithm with three different scenarios.

4.1 CASE 1: DIFFERENT NETWORK STRUCTURE

We tested different network structure because there is no readily available structure for such purpose, which is otherwise available for image classification. So, we have to make a model to ensure that we get good accuracy with less complexity. We have used two different network structures:

- LSTM network structure with 1 hidden layer
- LSTM network structure with 4 hidden layers

The result comparison of both is shown in Table.2.

Table.2. Result Comparison for 1 and 4 Hidden Layer Network

	Accuracy	Loss
1 Hidden Layer Network	81.1 %	0.502
4 Hidden Layer Network	86.7 %	0.6581

Training and validation accuracy chart for 1 hidden layer network in Fig.4 shows that validation accuracy is getting into the stagnant state where it is rarely increasing at some epochs. Same for validation loss (Fig.5) it is hardly decreasing after some epochs. Here as in Fig.5 the loss is equivalent to 1 show that the training model is not sufficient to process the input layer parameters. In Fig.6 and Fig.7 show the chart for 4 hidden layer network.

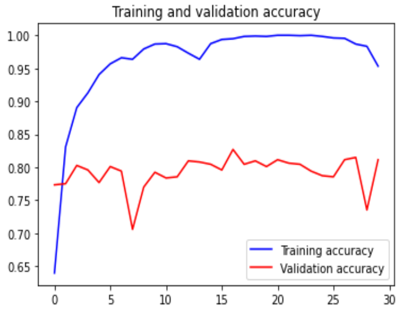

Fig.4. Training vs Validation Accuracy for 1 Hidden Layer Network

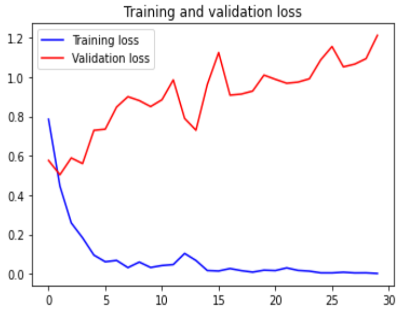

Fig.5. Training vs Validation Loss for 1 hidden layer network

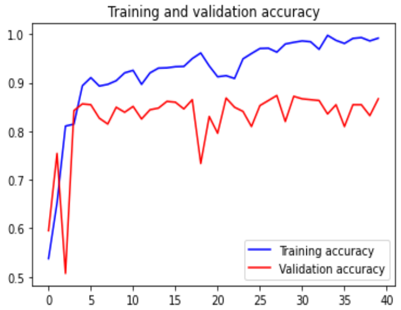

Fig.6. Training vs Validation Accuracy for 4 hidden layer network

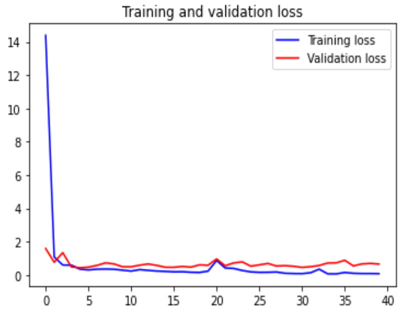

Fig.7. Training vs Validation Loss for 4 hidden layer network

4.2 CASE 2: DURATION OF AUDIO

We have analyzed the performance of network when the length of audio is very short or large. We have used different scenario to see the effect of duration of audio on network.

We have evaluated two different durations of audio:

- 10 seconds of audio
- 15 seconds of audio

Result comparison of both is shown in Table.3.

Table.3. Result Comparison for 10 and 15 seconds of Audio.

	Accuracy	Loss
10 seconds audio	86.7 %	0.6581
15 seconds audio	83.04%	0.7980

The total training audio samples available for 15 seconds duration was about 1 hour 34 minutes and the same for 10 seconds of duration was about 6 hours. Despite of this difference in available training data, model trained with 15 seconds of audio gives good result with accuracy of 83.04%. Figures 8 and 9 show the graphs of training vs validation and loss for 15 seconds of audio.

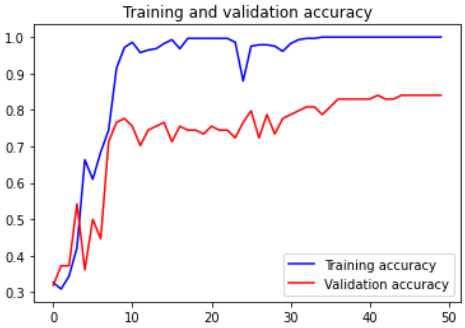

Fig.8 Training vs Validation Accuracy for 15 seconds of audio

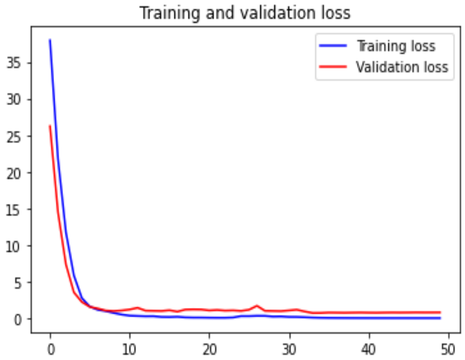

Fig.9. Training vs Validation Loss for 15 seconds audio

4.3 CASE 3: SIZE OF DATA

As we are experimenting with deep neural network and in literature it is mentioned more data could give better result, so in order to test that we have used two different sizes of training data:

- Large dataset with 6 hours of audio
- Small dataset with 1 hr 40 min of audio

The result comparison is shown in Table.4.

Table.4. Result Comparison for 10 and 15 seconds of Audio.

	Accuracy	Loss
6 Hours	86.7%	0.6581
1 hours 40 minutes	32.6%	1.1.03

The accuracy gained by model trained with small data is very low that is, 32.6%. The training and validation accuracy chart (Fig.8) shows that the accuracy ranges from 0.30 to 0.38 only. This is because the insufficient data points tend to the wrong feature scaling.

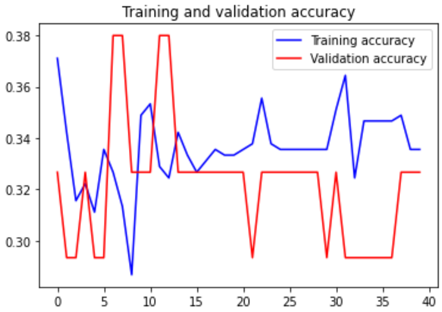

Fig.10. Training vs Validation Accuracy for small dataset

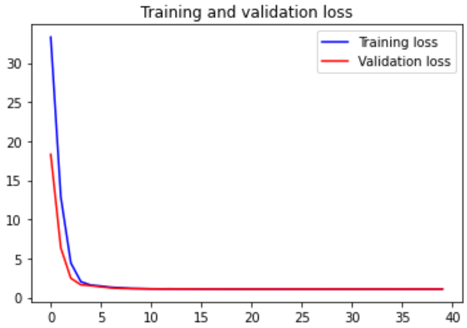

Fig.11. Training vs Validation Loss for small dataset

The analysis of model was carried out by providing unseen data to the model trained with 6 hours of data on 4 hidden layer network. From this analysis we observed the following,

- On an average, 18 out of 20 unseen audios were predicted correctly for each language (English, Hindi, Gujarati)
- The unseen audio was predicted correctly by model trained with 15 seconds of audio, which were wrongly predicted by the model trained with 10 seconds of audio.
- It is observed that if the sentence includes 20-35% of words which are pronounced in different language, than there are more chances that the sentence will be classified incorrectly.

5. CONCLUSIONS

We aimed to propose an efficient approach for Spoken Language Recognition and analyze language in multilingual speech using deep neural network. We have demonstrated the experiment results for the same and the highest accuracy we achieved is 86.7 %. From the experimental results we conclude that, audio files with comparatively more duration work better. If the uttered sentence consists of 25-35% of different language words, chances of wrong prediction increase. We also conclude that, model trained with large number of audio data gives acceptable accuracy. The analysis carried out can be helpful at administration level to measure the quality of teaching in a class which can lead to improvement in the education system.

REFERENCES

[1] Bin Ma, Haizhou Li and Rong Tong, "Spoken Language Recognition using Ensemble Classifiers", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 7, pp. 2053-2062, 2007.

[2] Himanish Shekhar Das and Pinky Roy, "A Deep Dive into Deep Learning Techniques for Solving Spoken Language Identification Problems", *Intelligent Speech Signal Processing, Elsevier*, pp. 81-100, 2019.

[3] Jiri Navratil, "Spoken Language Recognition-A Step Toward Multilinguality in Speech Processing", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 6, pp. 678-685, 2001.

[4] Fred Richardson, Douglas Reynolds and Najim Dehak. "Deep Neural Network Approaches to Speaker and Language Recognition", *IEEE Signal Processing Letters*, Vol. 22, No. 10, pp. 1671-1675, 2015.

[5] Shivesh Ranjan, Chengzhu Yu, Chunlei Zhang, Finnian Kelly and John H.L. Hansen, "Language Recognition using very Limited Training Data", *Proceeding of IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 5830-5834, 2016.

[6] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez and Pedro Moreno, "Automatic Language Identification using Deep Neural Networks", *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 5337-5341, 2014.

[7] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, David Martinez, Oldrich Plchot, Joaquin Gonzalez-Rodriguez and J. Pedro Moreno, "On the use of Deep Feedforward Neural Networks for Automatic Language Identification", *Computer Speech and Language*, Vol. 40, pp. 46-59, 2016.

[8] Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin and Vidhyasaharan Sethu, "Language Recognition: A Tutorial", *IEEE Circuits and Systems Magazine*, pp. 82-108, 2011.

[9] K.V. Mounika, H.R. Lakshmi, Suryakanth V. Ganga Shetty and Anil Kumar Vuppala, "An Investigation of Deep Neural Network Architectures for Language Recognition in Indian Languages", *Proceedings of IEEE International Conference on Interspeech*, pp. 2930-2933, 2016.

[10] Haizhou Li, Bin Ma and Kong Aik Lee, "Spoken Language Recognition: from Fundamentals to Practice", *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1136-1159, 2013.

[11] Bihong Zhang, Lei Xie, Yougen Yuan, Huaiping Ming, Dongyan Huang and Mingli Song. "Deep Neural Network Derived Bottleneck Features for Accurate Audio Classification", *Proceedings of IEEE International Conference on Multimedia and Expo Workshops*, pp. 1-6, 2015.

[12] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition", *International Journal for Advance Research in Engineering and Technology*, Vol. 1, No. 6, pp. 1-5, 2013.

[13] Najim Dehak, Reda Dehak, Pierre Dumouchel, Pierre Ouellet and Patrick J. Kenny, "Front-End Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788-798, 2011.

[14] Mitchell Mclaren, Luciana Ferrer and Aaron Lawson, "Exploring the Role of Phonetic Bottleneck Features for Speaker and Language Recognition", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5575-5579, 2016.

[15] Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Pedro J. Moreno and Joaquin Gonzalez-Rodriguez, "Frame-by-Frame Language Identification in Short Utterances using Deep Neural Networks", *Neural Networks*, Vol. 64, pp. 49-58, 2015.

[16] Wei Wang, Wenjie Song, Chen Chen, Zhaoxin Zhang and Yi Xin, "I-Vector Features and Deep Neural Network Modeling for Language Recognition", *Proceedings of International Conference on Identification, Information and Knowledge in the Internet of Things*, pp. 36-43, 2018.

[17] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg and Oriol Nieto, "Librosa: Audio and Music Signal Analysis in Python", *Proceedings of IEEE International Conference on Python in Science*, pp. 1-6, 2015.

[18] Alex Graves, Abdel-Rahman Mohamed and Geoffrey Hinton, "Speech Recognition with Deep Recurrent Neural Networks", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1-6, 2013.

[19] Christian Bartz, Tom Herold, Haojin Yang and Christoph Meinel, "Language Identification Using Deep Convolution Recurrent Neural Network", *Proceedings of International Conference on Neural Information Processing*, pp. 256-265, 2017.

[20] A. Lozano-Dez, R. Zazo Candil, J. Gonzlez Domnguez, D.T. Toledano and J. Gonzlez-Rodrguez, "An End-to-End Approach to Language Identification in Short Utterances using Convolutional Neural Networks", *Proceedings of International Conference on International Speech and Communication Association*, pp. 562-569, 2015.

[21] Language Data Consortium, "NIST Evaluation", Available at https://www.ldc.upenn.edu/collaborations/evaluations/nist, Accessed in 2020.

[22] Altexsoft, "Machine Learning Tools", Available at https://www.altexsoft.com/blog/datascience/the-best-machine-learning-tools-experts-top-picks/, Accessed in 2020.

[23] Sanket Doshi, "Music Feature Extraction in Python", Available at https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d, Accessed in 2020.

[24] Librosa, "Tutorial", Available at https://librosa.org/doc/latest/tutorial.html, Accessed in 2020.

[25] Linguistic Data Consortium, "CALLFRIEND American English-Southern Dialect Second Edition", Available at https://catalog.ldc.upenn.edu/LDC2020S08, Accessed in 2020.

[26] Analytics Vidhya, "Fundamentals of Deep Learning - Introduction to Recurrent Neural Networks", Available at https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/, Accessed in 2020.

[27] Artifexian, "Phonotactics", Available at https://www.youtube.com/watch?v=1Up5hSm7LYI, Accessed in 2020.

[28] Autism Life, "Prosody-Jargon of the Day", Available at https://www.youtube.com/watch?v=4uu3jhtxqXY, Accessed in 2020.

[29] The Linguistics Channel, "An Introduction to Morphology", Available at https://www.youtube.com/watch?v=syjbhT45J14, Accessed in 2020.

[30] John McGonagle, Jose Alonso García and Saruque Mollick, "Feedforward Neural Networks", Available at https://brilliant.org/wiki/feedforward-neural-networks/, Accessed in 2020.

[31] Jason Brownlee, "How to Develop a Bidirectional LSTM For Sequence Classification in Python with Keras", Available at https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/, Accessed in 2020.

[32] "Indic TTS", Available at https://www.iitm.ac.in/donlab/tts/index.php, Accessed in 2021.

[33] ResearchGate, "Structure of LSTM Unit", Available at https://www.researchgate.net/figure/The-structure-of-the-LSTM-unit_fig2_331421650, Accessed in 2021.

[34] Hasim Sak, Andrew Senior and Francoise Beaufays, "Long Short-term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling", *Proceedings of International Conference on Google Research*, pp. 338-342, 2014.

[35] Uddyalok Chakraborty, D. Thilagavathy, Suresh Kumar Sharma and Awadh Kishore Singh, "Hybrid Deep Learning with Alexnet Feature Extraction and Unet Classification for Early Detection in Leaf Diseases", *ICTACT Journal on Soft Computing* Vol. 14, No. 3, pp. 3255-3262, 2024.

[36] M.S. Banu, V. Elanangai, U. Chakraborty, D. Sharmiladevi and P. Rajeswari, "Skin Cancer Disease Classification using Tasnet V2 in Image Classification", *Proceedings of International Conference on Disruptive Technologies*, pp.1174-1179, 2024.