Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Detection And Classification of Childhood Apraxia of Speech Using Deep Guided Convolution Neural Network

Dr. C. Thilagavathy[1], M. Saifali[2], R. Sajeena[3], A. VasanthKumar[4], S. M. Agilan
Assistant Professor[1], PG student[2,3,4,5]
Department of Information Technology[1,2,3,4,5]
CMS College of Science and Commerce (Autonomous) , Coimbatore, India[1,2,3,4,5]

ARTICLE INFO

Received: 18 Oct 2024
Revised: 10 Nov 2024
Accepted: 28 Dec 2024

ABSTRACT

Childhood Apraxia of speech is neurological motor speech disorder which is due to difficulty of brain in planning and programming the complex movement of the speech. Especially it can't be categorized on basis of muscle weakness. Thus, it becomes mandatory to design a speech recognition system towards detection of childhood apraxia of speech on recorded sounds of child and doctor conversation. However, manual speech processing technique becomes challenging to classify the childhood apraxia of speech due to its complex sampling rate. Adoption of machine learning architecture from artificial intelligence to speech recognition makes detection more feasible and accurate. Despite of several advantage of the machine learning and deep learning approaches, there exist some challenges on basis of model scalability to large vocabulary and speech variability due to accent and style. In order to mitigate those challenge, deep learning model has to be modelled. In this paper. a new deep guided convolution neural network is designed and implemented to classify Childhood Apraxia of speech. Initially preprocessing step is performed to eliminate the noise and transform signal into segmented frame. Next segmented frame is processed in fast Fourier transform to obtain the power spectrum. Obtained Power spectrum is projected to proposed model. Convolution layer of model use mel filter to MFCC features and it is organized in feature map. Extracted feature is employed to fully connected network to perform precise recognition and classification of the Childhood Apraxia of speech in order to enhance prognosis of the specified disease. Experimental analysis and performance analysis of the proposed model have been evaluated using speech dataset from Ultra Suite Repository in the Python environment. On Performance analysis of the proposed model using test data of the model through confusion matrix provides model accuracy of 98.4% which is found to be high compared other conventional architecture.

Keywords: Childhood Apraxia detection, Speech Recognition, Mel filter, Mel Frequency Cepstral Coefficient, Fast Fourier Transform

1. INTRODUCTION

Childhood Apraxia is a complex speech disfluency occurs to a child due to multiple stuttering alterations. Stuttering alteration of speech happens due to brain control in transforming signals to the body which results in complexity in pronouncing the word with brief silence to certain syllables [1]. Thus, it becomes mandatory to design a speech recognition system towards detection of childhood apraxia of speech on recorded sounds of child during doctor conversation for diagnosis as non-invasive techniques on motor speech skills[1]. However, manual speech processing technique becomes challenging to classify the childhood apraxia of speech due to its complex sampling rate and it is found as highly intensive, time-consuming and error-prone. However, Adoption of machine learning architecture from artificial intelligence to speech recognition makes detection more feasible and accurate. Despite of several advantage of the machine learning and deep learning approaches, there exist some challenges on basis of model scalability to large vocabulary and speech variability due to accent and style[2].

In order to mitigate those challenge, deep learning model has to be modelled. In this paper. a new deep guided convolution neural network is designed and implemented to classify Childhood Apraxia of speech. Initially preprocessing step is performed to eliminate the noise and transform signal into segmented frame. Next segmented frame is processed in fast Fourier transform to obtain the power spectrum. Obtained Power spectrum is projected

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

to proposed model. Convolution layer of model use mel filter to MFCC features and it is organized in feature map. Extracted feature is employed to fully connected network to perform precise recognition and classification of the Childhood Apraxia of speech in order to enhance prognosis of the specified disease[3].

The remaining part of the article is sectioned as follows; section 2 provides related work of speech recognition and classification techniques using machine learning and deep learning architectures. In Section 3, design of proposed deep learning architecture referred as deep guided convolution neural network is carried out for recognizing and classifying Apraxia on the speech. Section 4 mentions experimental analysis and performance analysis of the proposed methodology using speech signals extracted from benchmark ultra-suite repository. Finally, section 5 concludes the work with future suggestions.

2. RELATED WORK

In this section, several machine learning and deep learning architecture employed to process the speech signals against apraxia of speech have been detailed on its architectural elements and experimental setup were as follows

2.1. Apraxia speech classification using Convolution Neural Network

In this architecture, a convolutional neural network is employed to process speech signal towards classification of speech apraxia of the child. Architectural elements of the model composed of convolution layer to extract the speech features, pooling layer to extract the optimal speech features through suitable filters. Extracted optimal features is processed in fully connected layer using SoftMax function to classify apraxia of speech along loss function to reduce the interclass variabilities of the model on cross fold validations[4].

2.2. Apraxia speech classification using Graph Convolution Network

In this architecture, a Graph convolution network is employed to process speech signal towards classification of speech apraxia of the child. Architectural elements of the model composed of dilated convolution layer to extract the speech features, attention mechanism to extract the optimal speech features through attention coefficients. Extracted optimal features is processed further in the fully connected layer using SoftMax function to classify apraxia of speech along loss function to reduce the intraclass variabilities of the model on cross fold validations[5].

3. PROPOSED MODEL

In this section, design of proposed deep learning architecture referred as deep guided convolution neural network is carried out for recognizing and classifying Apraxia on the speech for speech signals is carried out. Architecture of proposed model composed of deep guided convolution neural network is represented as follows

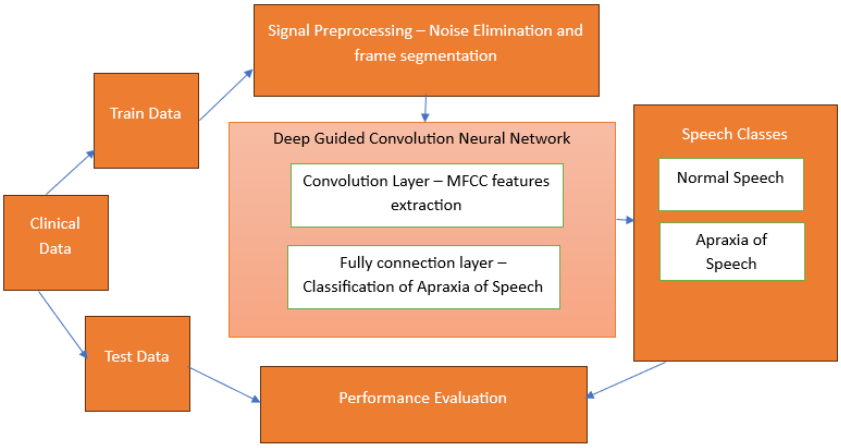

Figure 1: Architecture diagram of the proposed model

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

3.1. Signal Preprocessing

Signal Preprocessing is carried out on speech signal to eliminate the noises and normalize the speech signal. Further it is used to segment the signal into different speech frames. Furthermore, each segmented technique is employed to fast Fourier transform to generate power spectrum[6]. Further power spectrum of the signal is mentioned as

power spectrum $P_s = \frac{l(l+1)}{2\pi} c_{l....}$ Eq.1

3.2. Deep Guided Convolution Neural Network

Deep Guided Convolution Neural Network is modeled with multiple layers to process power spectrum of the signal. Functional outcomes of each layer is as follows

3.2.1. Convolution layer

Convolution layer uses the mel filter to extract the Mel Frequency Cepstral Coefficient features. Mel filter composed of the bank of filter to obtain the Mel energy of power spectrum with its mel coefficient. Log of the mel energy is applied to discrete cosine transform to extract the MFCC feature on correlating its value with mel spectrum[7]. MFCC coefficients contain rate changes of different spectrum bands. MFCC features is represented as

MFCC features $=4\int_0^\infty cos(2\pi wr)dr$....Eq.2

3.2.2. Pooling layer

Pooling layer uses max pooling function to extract the dense MFCC features and represents in form of dense feature map. Table 1 provides the hyperparameter setting of Deep Guided Convolution Neural Network architecture.

Table 1: Hyper parameter of Deep Guided Convolution Neural Network Architecture

Hyperparameter	Value
Epoch	65
Loss function	Cross Entropy
Activation Function	ReLU
Batch Size	35
Learning rate	10^{-6}

3.2.3. Fully connected layer

Fully connected layer of the model process MFCC features using SoftMax function and loss function. Support vector machine is used as SoftMax function and cross entropy is used as loss function to process the Dense MFCC features and its feature map[8].

- **Activation Function**

Activation function employs the ReLu activation function to generate linear feature vector and minimize the errors[9].

- **Loss function**

Loss function employs cross entropy function to eliminate the intraclass variability of the feature vector[10].

- **SoftMax function – Support vector Machine Classifier**

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Support vector machine processes the linear MFCC feature vector. It transforms linear MFCC feature vector into support vector. Hyperplane is established to support vector and class to vector is determined using decision boundaries to determine the disfluencies[11].

Decision Boundary Function of Support vector $D_{f=6}(\sum_{i=0}^{n} I)M(f)$...Eq.3

Decision boundary function provides the speech classes to the support vector as normal and apraxia of speech class.

Algorithm 1: Deep Guided Convolution Neural Network

Input: Speech Signal
Output: Class Label={Apraxia, Normal}
Process
Signal Preprocessing ()
Noise Filtering (Signal)
Signal Normalization (Noise filter signal)
Power spectrum= Segment (Frames with Windowing)
Deep Guided Convolution Neural Network ()
Convolution layer ()
Mel filter (power spectrum)
MFCC features
Pooling layer _Max(MFCC features)
Dense MFCC features
Fully Connected layer
Activation function(Dense MFCC features)
Linear MFCC features
Loss Function_Cross entropy (Linear MFCC features)
Eliminate the overfitting issues
SoftMax functions_SVM(Linear MFCC features)
Class of speech = {Apraxia, Normal}

4. EXPERIMENTAL RESULTS

Experimental analysis of the proposed deep guided convolution neural network has carried out using speech signal extracted from benchmark ultra-suite repository in python environment [12]. Performance analysis of the model is carried out using test data through confusion matrix to obtain the parameter value of true positive, true negative, false positive and false negative.

4.2. Performance metrics

The model performance such as precision, recall and f measure are computed using parameter of the confusion matrix. Confusion matrix classifies power spectrum of the speech signal into normal and apraxia of speech with respect to the matrix elements of the MFCC features.

- **Precision**

It is computed as correctly predicted MFCC feature to speech class among extracted MFCC feature. In other words, it is defined as ratio of true positive to combination of true positive and false positive of prediction outcomes.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

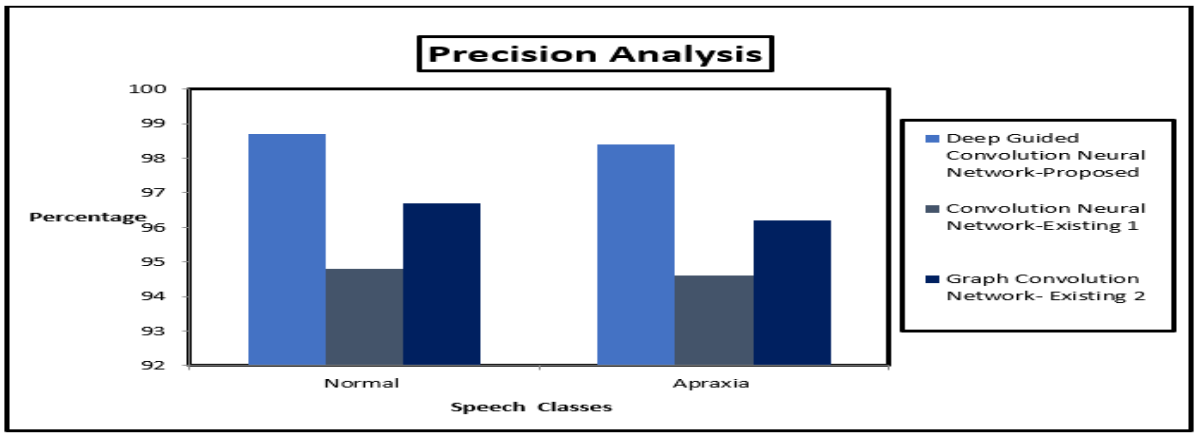

Figure 2: Precision Analysis

It is represented as

$$\text{Precision} = \frac{TP}{TP+FP} \text{ ...Eq.3}$$

Figure 2 provides precision analysis of the speech recognition and classification technique using Deep guided convolution neural network. It performs better while compared to existing architectures such as convolution neural network and graph convolution network[13].

- **Recall**

It is computed as incorrectly predicted MFCC feature to speech class among extracted MFCC feature. In other words, it is defined as ratio of true positive to combination of true positive and false negative of classification outcomes. It is represented as

$$\text{Recall} = \frac{TP}{TP+FN} \text{..Eq.4}$$

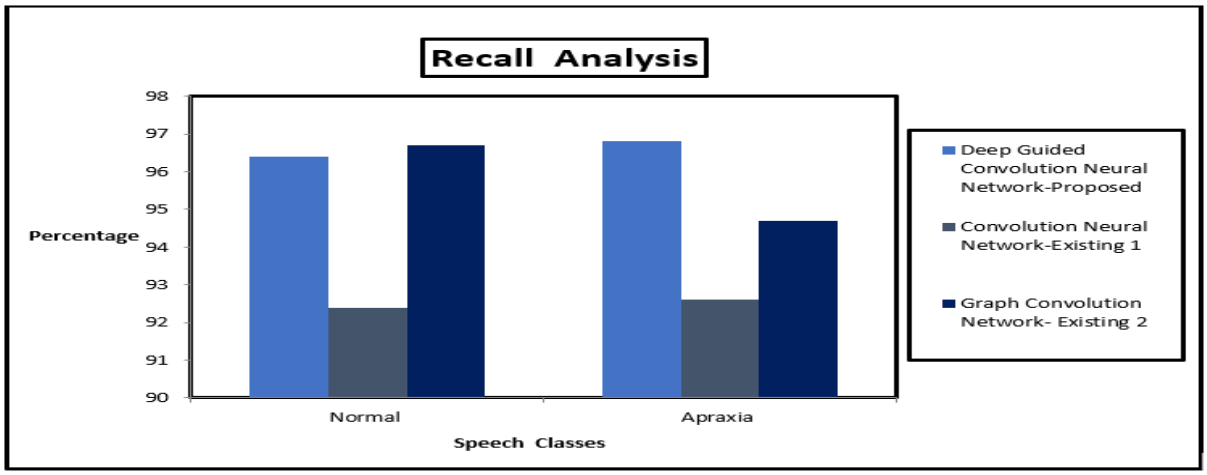

Figure 5: Recall Analysis

Figure 3 provides recall analysis of the speech recognition and classification technique using Deep guided convolution neural network. It performs better while compared to existing architectures such as convolution neural network and graph convolution network[14].

- **Accuracy**

It is defined as ratio of True positive to combination of true positive and false negative on classifying speech data on basis of the MFCC features It is represented as

$$\text{Accuracy} = \frac{TP}{2TP+FN} \text{..Eq.7}$$

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

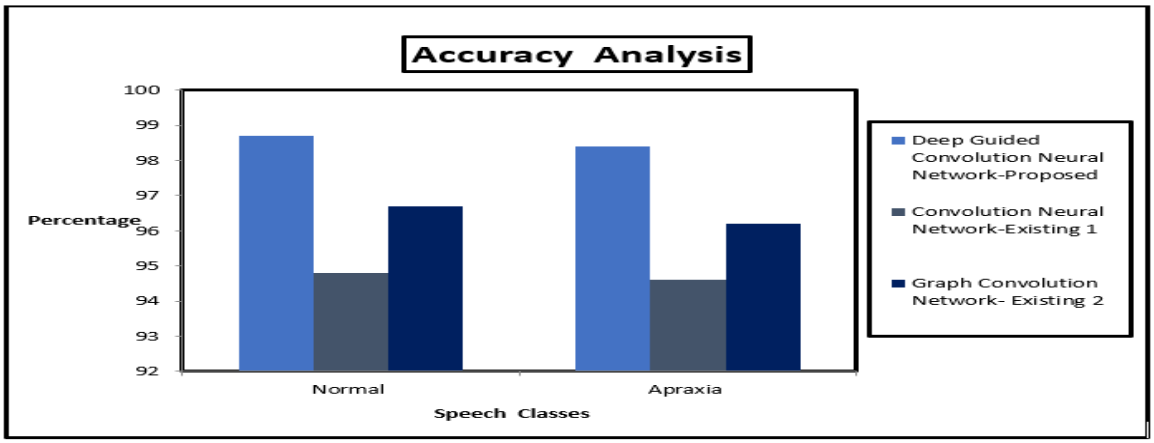

Figure 6 : Accuracy Analysis

Figure 4 provides accuracy analysis of the speech recognition and classification technique using Deep guided convolution neural network. It performs better while compared to existing architectures such as convolution neural network and graph convolution network[15].

Table 3: Performance Evaluation of speech Recognition Technique Techniques

Disease Classes	Technique	Accuracy	Precision	Recall
Apraxia	Deep Guided Convolution Neural Network – Proposed model	98.4	96.4	98.8
	Convolution Neural Network-Existing Model 1	94.7	92.8	94.9
	Graph Convolution Network-Existing Model 2	96.7	94.2	96.8
Normal	Deep Guided Convolution Neural Network – Proposed model	98.4	96.4	98.8
	Convolution Neural Network-Existing Model 1	94.7	92.8	94.9
	Graph Convolution Network-Existing Model 2	96.7	94.2	96.8

CONCLUSION

In this paper, deep guided convolution neural network is designed and implemented to classify Childhood Apraxia of speech. Deep guided convolution neural network composed of multiple layer process preprocessed speech signal in form of power spectrum on extracting MFCC features using mel filter. Those extracted feature is processed in the fully connected network to perform precise recognition and classification of the Childhood Apraxia of speech in order to enhance prognosis of the specified disease. Experimental analysis and performance analysis of the proposed model proves that proposed model obtains accuracy of 98.4% as it is found to be high compared other conventional architecture to the speech dataset from Ultra Suite Repository in the Python environment.

REFERENCES

[1] K. Knollman-Porter, “Acquired apraxia of speech: a review,” Topics in stroke rehabilitation, vol. 15, no. 5, pp. 484–493, 2008.
[2] J. Ogar, H. Slama, N. Dronkers, S. Amici, and M. Luisa Gorno-Tempini, “Apraxia of speech: an overview,” Neurocase, vol. 11, no. 6, pp. 427–432, 2005.

[3] R. T. Wertz, L. L. LaPointe, and J. C. Rosenbek, Apraxia of speech in adults: The disorder and its management. Singular Publishing Group, 1991.
[4] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-toend text-to-speech synthesis," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2021, pp. 5679–5683,
[5] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, "Automatic speech recognition for mixed dialect utterances by mixing dialect language models," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 23, no. 2, pp. 373–382, Feb. 2015
[6] D. S. Sisodia, S. Nikhil, G. S. Kiran, and P. Sathvik, "Ensemble learners for identification of spoken languages using mel frequency cepstral coefficients," in Proc. 2nd Int. Conf. Data, Eng. Appl. (IDEA), Feb. 2020, pp. 1–5
[7]I.Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," J. King Saud Univ. Comput. Inf. Sci., vol. 33, no. 5, pp. 497–507, Jun. 2021,
[8] D. C. Cires¸an, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "High-performance neural networks for visual object classification," arXiv preprint arXiv:1102.0183, 2011.
[9] E. M. Mohammed, M. S. Sayed, A. M. Moselhy, and A. A. Abdelnaiem, "LPC and MFCC performance evaluation with artificial neural network for spoken language identification," Int. J. Signal Process., Image Process. Pattern Recognit., vol. 6, no. 3, p. 55, 2013.
[10] D. S. Sisodia, S. Nikhil, G. S. Kiran, and P. Sathvik, "Ensemble learners for identification of spoken languages using mel frequency cepstral coefficients," in Proc. 2nd Int. Conf. Data, Eng. Appl. (IDEA), Feb. 2020, pp. 1–5
[11] B. Jan, H. Farman, M. Khan, M. Imran, I. U. Islam, A. Ahmad, S. Ali, and G. Jeon, "Deep learning in big data analytics: a comparative study," Computers & Electrical Engineering, vol. 75, pp. 275–287, 2019.
[12] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench, "Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions," arXiv preprint arXiv:1907.00835, 2019.
[13] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," in Proc. Interspeech, Oct. 2020, pp. 374–378.
[14] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," 2019, arXiv:1909.05645.
[15] B. Maji, M. Swain, and M. Mustaqeem, "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with convcaps and bi-GRU features," Electronics, vol. 11, no. 9, p. 1328, Apr. 2022.