Convolution Neural Network Model for Recognition of Speech for Words used in Mathematical Expression

Vaishali A. Kherdekar,

Affiliation: Research Scholar, Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University), Pune, Maharashtra State, India. Email Id– vaishali.kherdekar@gmail.com

Dr. Sachin A. Naik,

Affiliation: Assistant Professor, Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University), Pune, Maharashtra State, India. Email Id-sachin.naik@sicsr.ac.in

Abstract: Speech recognition is translation of audio signal into human readable form. Speech recognition plays a vital role in various areas such as in signal processing, dictation system, command and control, simple data entry. Speech recognition in dictation system helps the disabled people. In this paper authors have performed the experiment for speech recognition of mathematical words which is helpful to disabled people. Now a day's the use of deep learning in various applications is challenging for the improvement of model. In this paper authors have used CNN model to improve the recognition accuracy. Authors have selected 17 mathematical words which are the most commonly used in mathematical expression. Rectified Linear unit activation function is used to train the CNN because of its fast computation. This paper evaluates the model for MFCC and Delta MFCC features for Adam and Adagrad optimizers. Result shows that Delta MFCC gives an accuracy of 83.33 % for both Adam and Adagrad optimizer. It indicates that Delta MFCC gives better result than MFCC. Result also shows that Adagrad with Delta MFCC trains the model earlier than Adam.

Keywords: Convolution Neural Network, Deep Learning, Feature Extraction, MFCC, Optimizers, Speech Recognition.

1. Introduction

FirstHCI (Human Computer Interaction) is nothing but to communicate with the machines. There are various ways to interact with the computer, speech is one of it. To interact with the machine with the help of speech plays a key role. Speech recognition is one of the applications of human computer interaction. Speech recognition is converting an audio signal into text. It is defined by various researchers separately. According to Li et al. [1] Automatic speech recognition (ASR) is "the process and the related technology for converting the speech signal into its corresponding sequence of words or other linguistic entities by means of algorithms implemented in device, a computer, or computer clusters". Jurafsky [2] define ASR technically as "building of a system for mapping acoustic signals to a string of words". Lakra et al. [3] stated that "Primary objective of ASR is to design a system that can talk just like a human being". Speech recognition is challenging because of variation in speech signal due to speaking style and rate. Due to background noise and reverberation also it is challenging. Speech recognition is used in various areas such as in education system for dictation purpose, data entry systems to input data with the help of speech, electronics appliances for controlling electronic devices. Speech recognition application [4] also used for environment control such as turn on light, TV etc. These systems are also helpful for disable people. Based on mode of the speaker Speech recognition systems are categorized as speaker reliable system, speaker which are not reliable on the system and adaptive system. Speaker dependent systems works for specific users hence more accurate and less expensive where as speaker independent systems works for any kind of speaker hence less accurate, more expensive and more complex. Adaptive system grasps the characteristics of new user and gradually improves. Speech recognition approaches are categorized into Phonetic approach, Pattern recognition and AI approach. In phonetic approach, Phonetic units produced by the human vocal organs are used to classify the speech signal. In pattern recognition approach, speech patterns are used for classification. In earlier days acoustic phonetic and pattern recognition were used. In Artificial Intelligence approach both the phonetic and pattern recognition approaches are combined. AI based approach is most widely used approach now a days. In earlier days Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) based models were used to signify the sequential structure of speech signal. In conventional system, feature extraction is very critical. In earlier days up to 1950, time domain features were used, between 1950 to 1960 frequency domain features were used, after 1960 combination of time and frequency domain features were used and since 2000 researchers focus on the deep features. Hence now a day's deep learning is most widely used. Deep learning is subset of artificial intelligence and machine learning. There are variations of deep networks present such as Deep Neural Network (DNN), Deep Belief Network (DBN), Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Deep Tensor Network (DTN). Deep features from deep learning technique are used to train the model. Deep features are the features which are obtained from hidden layers from the deep learning models. The MFCC or any appropriate features are given as input to deep learning models [5].Paragraph: use this for the first paragraph in a section.

This paper consists of sections including Section 2 represents the literature review. Proposed model is given in section 3. Dataset and feature extraction technique used are presented in section 4 and section 5 respectively. Section 6 and 7 presents CNN with training details. Results obtained through experimentation are given in section 8. Last section put light on conclusion of the paper.

2. Literature Review

ThisEarlier speech recognition system converts audio signals into text but not in the form of expression. There is lack of tools for implementing creation and modification of mathematical content using speech which are helpful for the disabled people. To design and develop a mathematical expression into e-document is tedious, time-consuming and possibility of making errors for abled person therefore for newcomers and blind people it becomes very difficult. Experts are required for it. It shows that there is need of an alternative type of interface for such people to edit mathematical content more effectively. [6,7]. In [8] Medjkoune et al, developed the model based on handwritten recognition and speech recognition for mathematical symbols. They have used 74 mathematical symbols from CIEL dataset for handwritten recognition and HAMEX (Handwritten Audio Mathematical Expression) dataset for speech recognition. By using MFCC feature extraction technique 39 features were extracted from each frame. SVM classifier was used to train the model. Result showed 50.09% recognition rate for speech recognition and 81.55% recognition rate for handwritten recognition on test data set. In [9] Jaitly et al. presented DBN network trained by using 2 dataset named VoiceSearch and Youtube. Voicesearch dataset includes 5780 hours of mobile search data and 1400 hours of Youtube data. Features were extracted using MFCC and log filter bank. To train the model 55 hours of Youtube data and 321 hours of Voicesearch data were used for per epoch. Result shows WER of 16% for VoiceSearch and 52.3% for YouTube dataset. Dong Yu et al. [10] proposed a context dependent DNN-HMM model. Authors have extended the DNN to DTNN (Deep Tensor Neural Network) which consists of tensor layers and one or more layers are double projections. Corpus used to evaluate the model was Switchboard which consists of 30 hours of data. 39 dimensional HLDA and 13 dimensional PLP features were extracted. Learning rate was set to 0.0003 for first 5 epoch and 0.000008 for next 5 epoch. Back Propagation algorithm was used to train the DTNN. It shows reduction in WER by 5% than DNN, when double projection layer is put at the uppermost layer of the DNN. WER is monitored for three designs, in first all the layers of double projection like DNN, In second last layer is put back with DP layeand in third design upper hidden layer is replaced with DP layer. In [11] Maas et al. worked on Switchboard and Fisher corpus which consists of 300 hours telephone speech. Features were extracted using MFCC and trained on DNN, DCNN and DLUNN. Python was used to implement the experiment. Result shows that increase in the no. of parameters increases the representational capacity of DNN model. Result also shows that large DNN reduces the WER on training set. NAG (Nesterov's Accelerated Gradient) and CM (Classical Momemtum) optimizers were used and authors revealed that for good result NAG is vigorous than CM. A model with optimization technique indicates better result than simple DNN. Outcome of the experiment indicates DNN is more aggressive than DCNN (Deep Convolutional Neural Networks) and DLUNN (Deep Locally Untied Neural Networks). Model size also affected on the accuracy of the mode. Mitra et al. [12] proposed HCNN (Hybrid Convolution Neural Network) model. In this model two parallel layers were used to join the acoustics and articulatory space. Aurora-4 and WSJ1 datasets were used for this study. Experiment was performed by adding various noises such as car, bubble, restaurant, street, airport and train station. Acoustic models were trained using DNN, CNN and TFCNN (Time Frequency CNN). Observations revealed that noise in the acoustic signal reduces the performance of the classifier, training the model using noisy data raises the noise robustness. HCNN based model indicates Lower word error rate than CNN/DNN based systems. Nagajyothi&Siddaiah [13] proposed Telgu language speech recognition based Airport Enquiry system using CNN. They have considered the commonly asked question at Airport enquiry to prepare a dataset. They have selected weight connectivity; local connectivity features to train the model. Model is trained and tested using CNN (Convolution Neural Network). Tanh and Relu activation function were used during training. Authors have concluded that CNN gives better performance than conventional network. Soe et al. [4] trained various CNN acoustic model. Myanmar speech dataset was used for this study. This conversational speech dataset is of 452 hours out of which 438.5 and 13.5 hours of data were used for training and validation respectively. Model was built using N-gram. MFCC features were extracted using hamming window of 25 millisecond and framerate of 10 milliseconds. 8 Models were trained using varying filter size of convolution layer. Recognizer Output Voting Error Reduction (ROVER) algorithm is used to integrate the acoustic models of CNN for the last experiment. This study results in the reduction of WER by 4.32% in integrating CNN model. Passricha and Aggarwal [14] designed CNN model which is derived from the raw speech data of TIMIT dataset. Conditional probability was computed for each class of phoneme. Acoustic model consists of the two steps first feature step in which feature were extracted using MFCC technique and it is given as input to convolution layer, second step is classifier where ANN, CRF and MLP were used. Various parameters were considered during training such as 100 to 700ms input window size, 10 to 90 samples of kernel width for first convolution layer, 1 to 11 samples of kernel width for n^{th} convolution layer, 20 to 100 filters per kernel, 2 to 6 frames of kernel width for max-pooling and 200 to 500 hidden units. Nassif et al. [15] reviewed the speech recognition systems using DNN (Deep Neural Network). Authors concluded that doing research using deep RNN specially LSTM will be useful for speech recognition.

3. Methodology

Speech recognition system using CNN accepts audio data as input. For this study, it accepts 1230 .wav file of 17 mathematical words as input. Authors have selected most commonly used words in mathematical expression. Feature learning and classification is performed by CNN. It consists of convolution, pooling, fully connected and Softmax layer. Convolution and pooling are used for feature learning where as fully connected and Softmax are used for classification. Convolution layer is used to extract the features from audio data. Pooling is used to select the information which is useful and remove the negligible information from given features. It is used to reduce the dimension of feature map vector. Figure 1 depicts the methodology used to perform this study.

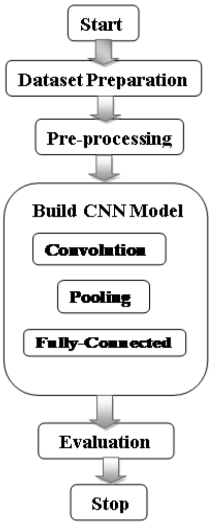


Figure 1. Proposed Model

4. Dataset

Existing dataset for audio mathematical expression is present only in French language not in English. Hence to create a dataset of audio for recognition of speech for mathematical expression is challenging for the author. For this study Authors have used Audacity tool to create dataset. Authors have recorded 17 mathematical words from 7 persons having age group between 14 to 40. The dataset is recorded at sampling rate of 22.5KHz. During recording, Authors have considered the most commonly used operators, symbols and functions used in mathematical expression. After recording audio files are pre-processed. The dataset consists of 1230 audio files in the form of .wav format. Dataset is splitted as 80% and 20% into training set and testing set respectively. Table 1 represents average number of samples used for training.

5. Feature Extraction Technique

ThisFeature extraction acts as a vital role during training a model. It is necessary to extract set of features from audio signal. A group of extracted features is given as input to classifier. In speech recognition feature vector represents the speech waveforms. There are various feature extraction techniques available to extract the features from audio signal [6] such as MFCC, delta MFCC, LPCC, PCA etc.

For this study authors have used MFCC and Delta MFCC feature extraction technique. Total of 20 MFCC coefficients are extracted to perform the experiment. First order derivative is considered in delta MFCC. Figure 2 depicts the process of evaluation of the MFCC features. The Fourier transformation of time domain audio signal into frequency domain is called as spectrum. By using fast fourier transformation samples from each frame are converted into frequency domain i.e. spectrum. Mel scales for frequency f is find out by using equation

$$Mel(f) = 2595\ \log_{10}\left(\frac{f}{700} + 1\right) \qquad \ldots\ldots\ldots(1)$$

Log magnitude of mel is called as mel spectrum. DCT (Discrete Cosine Transform) applied to mel spectrum and mel frequent cepstral coefficients (MFCC features) are computed. Computation of MFCC features consists of various steps such as pre-processing, framing, windowing, calculation of discrete fourier transform, mel frequency and inverse document frequency.

Table 1. Average number of samples used for training

Sr. No.	Label Name	Number of Recordings	Average Number of Samples
1	Alpha	64	9696
2	Beta	64	8985
3	Cube	68	7619
4	Delta	56	10002
5	Divided by	48	15493
6	Equal to	60	14171
7	Gama	56	8580
8	Greater than	56	14925
9	Integration	56	16782
10	Less than	56	12077
11	Minus	64	12436
12	Pie	48	6986
13	Plus	69	9397
14	Square	65	11677
15	Square Root	53	15369
16	Summation	56	15332
17	Theta	48	8880

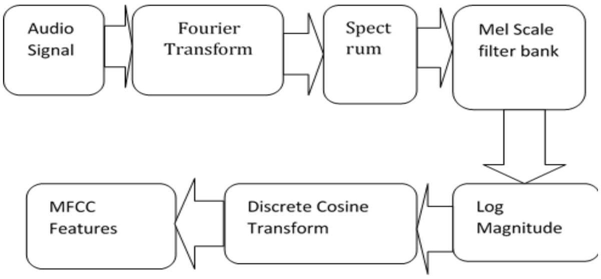


Figure 2. Process of MFCC feature extraction

6. Convolution Neural Network

Convolution Neural network [4,14,16,17] is the most popular variation of deep learning used in speech recognition. According to Passrichaand Aggarwal [14] Mitra V. et al.[12] Convolution neural network is supervised learning algorithm. CNN's have achieved great performance because of it's advance features such as sharing of weight, convolution and pooling operation. CNN's are commonly used for learning high level features and it's pooling operation is used for dimensionality reduction.

CNN is kind of deep learning algorithms. It is feed forward, supervised deep learning model. It consists of one or more convolution layer, pooling and one or more fully connected layer. It consists of sparse interactions, parameter sharing, and equivariant representation as basic concepts[14]. Convolution layer and pooling layer are the building blocks of CNN. Locality, weight sharing and pooling are the main three aspect of CNN [16]. All these aspect act as vital role in the performance improvement of speech recognition. Number of network weight is reduced by locality. Weight sharing is used to decrease the overfitting which gains the robustness of the model. The features extracted in locality and weight sharing are bring all together in pooling.

Convolution means combining, it combines the input vector with filter or weight vector using dot product and gives convolved feature as output. The main concept of convolution is to use the filter for extracting the feature map from input vector. In convolution layer filters works as feature extractor to extract time-frequency spectral variations. Multiscale features can be extracted using different filter size. In convolution operation features at particular layer is calculated as fv(i)

$$fv(i) = \sigma(W(i) * h(i-1) + b(i)) \qquad \ldots\ldots\ldots\ldots(2)$$

Where,

h(i-1) = feature vector in previous layer

W(i) = filter,

b = bias

σ = activation function.

Pooling layer act as a backbone of CNN. Pooling method decreases the dimensionality of feature vector. It also reduces the computational cost. Dense layer also called as Fully connected layer. It provides meaningful and low dimensional feature map.

During training of CNN following parameters plays important role for the improvement of model.

Activation Function B. Optimizer and C. Loss function

Activation function helps to decide whether the neuron get fire or not. There are various types of activation function present such as Sigmoid, tanh, ReLU, LeakyReLU etc. ReLU activation function is defined as y=max(0,x). Most commonly used activation function for CNN is ReLU and tanh. Optimizer is used to update the weight parameters. Adam, Adagrad, Adadelta, RMSProp, Nadam are the commonly used optimizers. Loss function is used to compute error i.e. the difference between actual output and predicted output. Here we use softmax cross entropy loss fuction which is used for multiclass classification problem. It converts the output of last layer into probabilities by using the formula

$$S(Yi) = \frac{e^{Yi}}{\sum e^{Yi}} \qquad \ldots\ldots\ldots\ldots\ldots(3)$$

this function takes the exponent of each output so that all probabilities should add up to one.

7. Training

To train a model using neural network or deep learning, it is necessary to consider the parameters such as sample, learning rate, batch size, number of epochs etc. Sample is a number of rows of data that is given as input to train the model. Training dataset consists of samples which is also called as input vector or feature vector. Batch is group of number of samples [19]. Before updating the neural network parameters training dataset is grouped into batches. Batches are classified into three types. When a batch is created using all the training samples then it is called as gradient descent batch. If a batch is created using only one sample then it is called as stochastic gradient descent batch. When a batch size consists of samples greater than one and less than the size of training dataset then it is called as mini batch gradient descent. Epoch is the number of times learning algorithm is executed during training.

Authors have considered the learning rate of 0.001 with a batch size of 10 to perform the experiment. Initially authors have conducted the experiment for 100 iteration, after that carried out the analysis for 200, 300 and 500 iterations. As the no. of iterations proceeds, accuracy of the training also increases. To update the weight parameters Adam and Adagrad optimizer are used. Authors have executed the study by considering MFCC and delta MFCC features. Because of its fast computation we picked up ReLU (Recurring Linear Unit) activation function to find out the models output in terms of accuracy. Softmax Cross Entropy loss function is selected to find out error rate.

In machine learning algorithm, performance of the training and test model is measured in terms of accuracy, loss, precision, recall etc. In multiclass classification accuracy and loss metric is used to measure the performance. To evaluate this study authors have used accuracy metric which is computed using equation 4.

$$Accuracy = \frac{(TP+TN)}{P+N} \qquad \ldots\ldots\ldots\ldots(4)$$

where, TP is predicted and observed labels are positive, TN is predicted and observed labels are negative, P and N are the total no. of positive and negative labels.

8. Results and discussion

Table 2 shows the accuracy for speech recognition of mathematical words for MFCC and Delta MFCC features using Adam optimizer whereas Table 3 represents the accuracy of Adagrad optimizer. Both the table also shows the time required in seconds to train the model.

Table 2. Training Accuracy of Adam Optimizer with MFCC and ΔMFCC Feature Extraction

No. of Iterations	MFCC		Delta MFCC	
	Accuracy in %	Time in Sec.	Accuracy in %	Time in Sec.
100	53.33	162.35	79.99	182.71
200	63.33	306.8	83.33	335.82
400	71.66	798.48	78.33	735.68
500	74.99	2655	83.33	976.08

Table 3. TrainingAccuracy of Adagrad Optimizer with MFCC and ΔMFCC Feature Extraction

No. of Iterations	MFCC		Delta MFCC	
	Accuracy in %	Time in Sec.	Accuracy in %	Time in Sec.
100	78.33	259.22	78.33	239.03
200	81.66	398.3	81.66	273.06
400	79.99	645.52	83.33	543.016
500	79.99	828.51	83.33	605.74

Figure 3 indicates loss of Adam optimizer during training and Figure 4 shows loss of Adam optimizer during testing. It indicates how loss goes on decreasing as the no. of iterations goes on increasing during training and testing. Figure 5 represents training accuracy of Adam optimizer and Figure 6 exhibits the training accuracy for Adagrad optimizer for MFCC features. Training accuracy for delta MFCC features is represented in Figure 7 for Adam optimizer and Figure 8 represents the training accuracy for Adagrad optimizer.

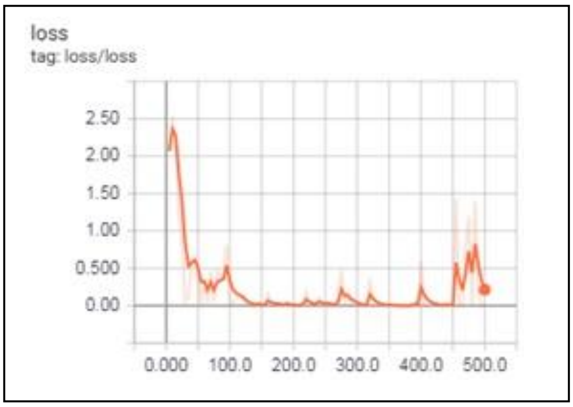


Figure 3. Adam optimizer during training

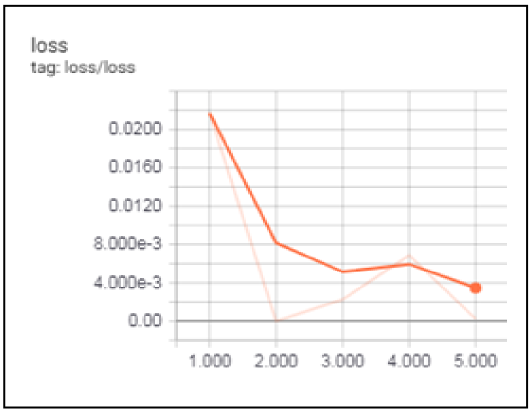
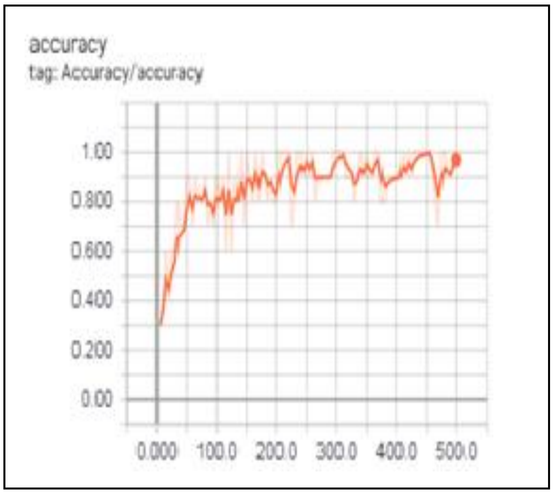


Figure 4. Adam optimizer during testing



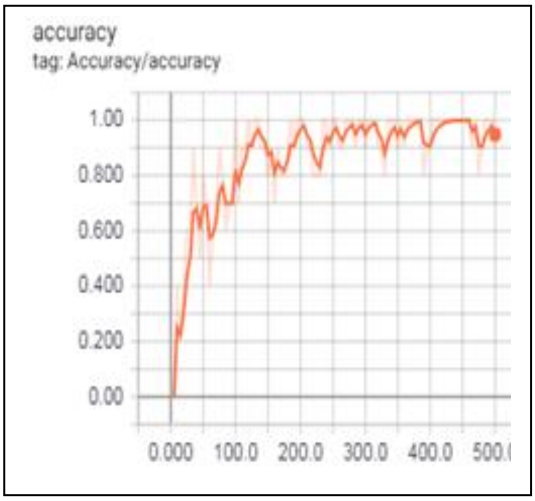


Figure 5. Training Adam using MFCC **Figure 6.** Training Adagrad using MFCC

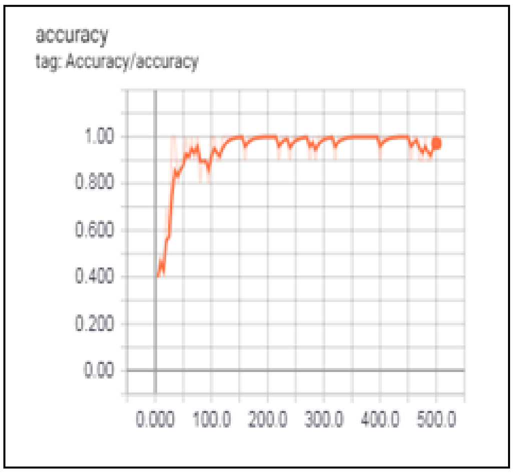


Figure 7. Training Adam using delta MFCC

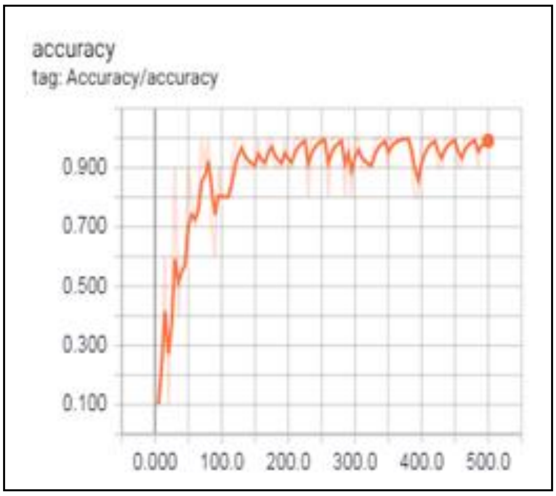


Figure 8. Training Adagrad using delta MFCC

9.Conclusion

Speech recognition is very important in human computer interaction but to implement it is very crucial because of lots of challenges present in it. Accuracy of the speech recognition systems depends on various factors such as technique used to develop the model, dataset size, speaker dependency, microphone quality, noise present during recording and many more. Anusaya&Katti, [20] stated that to improve the accuracy of speech recognition it is necessary to consider the points such as definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, database and performance evaluation techniques.

There is lack of soft computing techniques in previous study for speech recognition of audio mathematical expression so accuracy is not up to the mark. Hence there is need to focus on deep learning for improvement of the model. To design a model for audio data using deep learning is challenging task. In this paper, model is trained using convolution neural network to recognize speech for most commonly words used in mathematical expression. It is evaluated by considering MFCC and Delta MFCC feature extraction technique with Adam and Adagrad optimizer. Result shows that Delta MFCC gives an accuracy of 83.33% for both Adam and Adagrad optimizer and MFCC gives 74.99% and 79.99% for Adam and Adagrad optimizer respectively. It indicates that Delta MFCC gives better result than MFCC. Result also shows that Adagrad with Delta MFCC train the model earlier than Adam. We hope it will help more researchers to work on audio data of mathematics using convolution neural network. In future author want to extend the experiment to improve the recognition rate by using different classification techniques for more complicated mathematical expression.

References

[1] Li, J., Deng, L., Gong, Y., &Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), 745-777.

[2] Jurafsky D. (2000). Speech & language processing. Pearson Education India; 2000.

[3] Lakra, S., Prasad, T. V., Sharma, D. K., Atrey, S. H., & Sharma, A. K. (2012). Application of fuzzy mathematics to speech-to-text conversion by elimination of paralinguistic content. arXiv preprint arXiv:1209.4535.

[4] Soe, T., Maung, S. S., &Oo, N. N. (2018, July). Applying Multi-scale Features in Deep Convolutional Neural Networks for Myanmar Speech Recognition. In 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (pp. 94-97). IEEE.

[5] Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. Applied Acoustics, 158, 107020.

[6] Isaac M., Pfluegel E., Hunter G., Denholm-Price J., Attanayake D., &Coter G. (2016, October). Improving Automatic Speech Recognition for Mobile Learning of Mathematics Through Incremental Parsing. In Intelligent Environments (Workshops). (pp. 217-226).

[7] Attanayake, D., Denholm-Price, J., Hunter, G., Pfluegel, E., &Wigmore, A. (2015). Speech interfaces for mathematics: opportunities and limitations for visually impaired learners. In IMA international conference on barriers and enablers to learning maths: Enhancing learning and teaching for all learners.

[8] Medjkoune, S., Mouchère, H., Petitrenaud, S., & Viard-Gaudin, C. (2011, September). Handwritten and audio information fusion for mathematical symbol recognition. In 2011 International Conference on Document Analysis and Recognition (pp. 379-383). IEEE.

[9] Jaitly, N., Nguyen, P., Senior, A., &Vanhoucke, V. (2012). Application of pretrained deep neural networks to large vocabulary speech recognition.

[10] Yu, D., Deng, L., &Seide, F. (2012). Large vocabulary speech recognition using deep tensor neural networks. In Thirteenth annual conference of the international speech communication association.

[11] Maas, A. L., Qi, P., Xie, Z., Hannun, A. Y., Lengerich, C. T., Jurafsky, D., & Ng, A. Y. (2017). Building DNN acoustic models for large vocabulary speech recognition. Computer Speech & Language, 41, 195-213.

[12] Mitra V, Sivaraman G, Nam H, Espy-Wilson C, Saltzman E, Tiede M. (2017 May). Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. Speech Communication. 89:103-12.

[13] Nagajyothi D, Siddaiah P. (2018). Speech Recognition Using Convolutional Neural Networks. International Journal of Engineering and Technology.

[14] Passricha, V., & Aggarwal, R. K. (2018). Convolutional neural networks for raw speech recognition. In From Natural to Artificial Intelligence-Algorithms and Applications. IntechOpen.

[15] Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. (2019 Feb). Speech recognition using deep neural networks: A systematic review. IEEE access. 7:19143-65.

[16] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), 1533-1545.

[17] Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in neural information processing systems, 22, 1096-1104.

[18] Sainath, T. N., Vinyals, O., Senior, A., &Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4580-4584). IEEE.

[19] Json B. Difference Between a Batch and an Epoch in a Neural Network. 2019 Retrieved from https://machinelearningmastery.com

[20] Anusuya, M. A., &Katti, S. K. (2010). Speech recognition by machine, a review. arXiv preprint arXiv:1001.2267.

[21] Collobert, R., Puhrsch, C., &Synnaeve, G. (2016). Wav2letter: an end-to-end convnet-based speech recognition system. arXiv preprint arXiv:1609.03193.

[22] Deng, L., Hinton, G., & Kingsbury, B. (2013, May). New types of deep neural network learning for speech recognition and related applications: An overview. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8599-8603). IEEE.

[23] Huang, X., & Deng, L. (2010). An Overview of Modern Speech Recognition. Handbook of natural language processing, 2, 339-366..

[24] Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9(4), 393-404.

[25] Lawrence R. (2008). Fundamentals of speech recognition. Pearson Education India.

[26] Lu, L., Zhang, X., Cho, K., &Renals, S. (2015). A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition. In Sixteenth Annual Conference of the International Speech Communication Association.

[27] Ogawa, A., & Hori, T. (2017). Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. Speech Communication, 89, 70-83.

[28] Sadaghiani, M. H., Shafiabady, N., & Isa, D. (2015). A Novel Approach for Allocating Mathematical Expressions to Visual Speech Signals. SAGE Open, 5(4), 2158244015611937.

[29] Sarkar, A., Dasgupta, S., Naskar, S. K., & Bandyopadhyay, S. (2018, April). Says who? deep learning models for joint speech recognition, segmentation and diarization. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5229-5233). IEEE.

[30] Vishnupriya, R., & Devi, T. (2014, March). Speech recognition tools for mobile phone-a comparative study. In 2014 International Conference on Intelligent Computing Applications (pp. 426-430). IEEE.