Investigation of the Effect of MFCC Variation on the Convolutional Neural Network-based Speech Classification

Md. Rakibul Hasan
Department of Electrical and Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
rakibul.eeekuet@gmail.com

Md. Mahbub Hasan
Department of Electrical and Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
mahbub01@eee.kuet.ac.bd

***Abstract*— This paper deals with the step-by-step implementation of a Bengali speech classification model with the variations of Mel-Frequency Cepstral Coefficient (MFCC). A deep convolutional neural network is developed for the classification purpose. Bengali vowels are the representative of the lower MFCC variational group, and words are the higher variational group. The necessary dataset derived from the vowel and word speech tokens was utilized for the analysis of the convolutional neural network. Then the model was trained and validated for both of the isolated vowels and words separately for the same number of data. The performance was measured according to four different metrics– loss, accuracy, confusion matrix, and cross-validation score. In all these cases, the performance of vowel recognition has been found superior to word recognition. The reason behind this performance variation has been studied and found that it is mostly related to the dynamic variation of the vocal tract between vowels and words at the time of speaking. MFCC is chosen as the feature of interest for classification purposes. The variation of MFCC for vowels and words have been compared and found that MFCCs of words have more variation than the vowels. As a consequence, it is concluded that the dynamic variation of the vocal tract is inversely related to the performance of recognition.**

Keywords*— *Bengali speech classification, vocal tract dynamics, Mel-Frequency Cepstral Coefficient (MFCC), convolutional neural network

I. INTRODUCTION

Day by day, speech processing-based human command dependent gadgets, such as home automation, automation of modern devices including smartphones, laptops, vehicles, etc. are becoming more and more attractive and fashionable. Speech-to-text systems also act as a helping hand for hearing impaired people. Research in these fields is emerging with particular emphasis on speech generation, speech-to-text conversion, security, cellular communications, and other core areas.

Depending on the region and culture, various spoken languages have been developed in many areas of the world. English is one of the most used languages in the world, and tremendous research has been done for the English language. As a consequence, there are various speech recognition systems available in the market, for example, Amazon Alexa, Google Assistant, Microsoft Cortana, etc. Bengali is an Indo-Aryan language primarily spoken by the Bengalis in the Indian subcontinent. It is the official language of Bangladesh and the seventh most spoken language in the world. There are 225 million native speakers in this language [1]. Therefore, the Bengali language became the point of our interest to make it more accessible to be used in modern communication devices like the other languages practiced in developed nations. Investigation to implement a reliable speech recognition system for Bengali holds primary importance to compete with comparable research accomplished in English [2, 3].

In case of speech recognition, selection of appropriate feature is an important task and there are a number of different features chosen by different researchers for different purposes. The performance of recognition depends heavily on the feature extraction phase because some important characteristics might be left out if it is not chosen properly. Authors in [4] describe state-of-the-art feature extraction techniques. Mel-Frequency Cepstral Coefficient (MFCC) is the frequently used feature and it is calculated from the short-term energy spectrum expressed on a mel-frequency scale. Linear Discriminant Analysis (LDA) is another technique which maximizes the between-class variation than the within-class variation in a data set. Fusion MFCC which is the combination of MFCC and LDA technique was practiced by Santosh Gaikwad et el. [5]. Linear Predictive Coding (LPC) Analysis is another technique that approximates the speech sample as a linear combination of past speech samples. Perceptually Based Linear Predictive (PLP) Analysis is an extension of the LPC technique which is focused on cross-speaker isolated word recognition. Formant frequency is another important feature that is referred to as the peaks of the acoustic spectrum [6].

In older days, most speech recognitions were done based on Hidden Markov Models (HMMs) classifier with Gaussian Mixture Models (GMMs). Despite having some advantages, GMMs have some serious drawbacks as it is statistically inefficient for some cases [3]. As a solution, Artificial Neural Networks (ANNs) were introduced by researchers around two decades ago. Over the last few years, advances in both the algorithms and computer hardware have been done, and these led to a more efficient and secure speech recognition system. Deep Neural Networks (DNNs) can model complex correlations in speech feature and by adding Convolutional Neural Network (CNN) the error rate can be reduced by 6%-10% [7].

A DNN based speech recognition model consists of two phases- one is training and the other is validation. In the training phase, Backpropagation algorithm is used by which the hyperparameters of the layers are gradually updated to a final value. The weights and biases of the nodes of the neural network are the hyperparameters of the network. This updating of parameters is done in such a way to reduce the

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

prediction error. This reduction of error is done step-by-step iteration over the training dataset. The number of iterations in order to update the hyperparameters in the training phase is termed as epoch. With gradual increase of the epoch number, the neural network tends to stabilize by setting final values to its hyperparameters. Stabilization of parameters denotes a state of the model where minimum error and maximum accuracy occurs. In the validation phase, the network validates for the input validation dataset by using those obtained final values of the hyperparameters [8].

The information infuses into speech through dynamic shape change of vocal-tract. So, speech dynamic actually infuses intelligence in speech. In order to speak out a particular speech, a particular dynamic shape-changing is required. Perfect dynamic shape-changing is not possible and this creates speaker to speaker variation of speech features. This variation may occur due to the variation of vocal-tract shape, culture, etc.

Speech recognition actually depends upon clustering the feature domain. The speaker related feature variation increases the intra-class variation and decreases the inter-class variation which is an undesirable effect for speech recognition purposes. These problems are being satisfactorily eliminated by DNN through step-by-step process including epoch by epoch updating of network parameters. This elimination process can be shown by the loss and accuracy of information of the neural network. As the speaker to speaker variation of speech features is a by-product of infusion of speech intellectuality, elimination of the variation is an important task for speech recognition.

Isolated words and isolated vowels are the core building blocks of any speech. So, the performance of any speech recognizer depends on the recognition of those building blocks. But the performance of their recognition varies, and no works have been done so far in order to find the root cause of that variation and its removing steps with DNN. That's why we have done it. By considering these issues found in this study, the developed speech recognition system will be an efficient one with fewer errors and higher accuracy.

II. Recongnition Model Formation

A. Convolutional Neural Network (CNN)

Convolutional layer applies a convolution operation to the input where a small matrix called kernel or filter is passed over the input matrix, and thus a transformed matrix is obtained depending on the values from the filter [9]. Subsequent feature map values are calculated according to (1).

$$G[m,n] = (f * h)[m,n] = \sum_j \sum_k h[j,k] f[m-j, n-k] \quad (1)$$

Where, f is the input matrix, and h is the filter. The resultant matrix has m rows and n columns.

A CNN consists of an input and an output layer [10], as well as multiple densely connected hidden layers as shown in Fig. 1. Typically, a convolutional layer is followed by a pooling layer, and a flatten layer is added before the densely connected layer [11]. Max polling was used in our experiment that takes the maximum value in each window, and thus it reduced the dimensionality. Flatten layer converts the 3D data of the convolutional layer to 1D data. CNNs have two components mainly- one is the hidden layers where the convolution and pooling operations are performed. Another component is the classification part where the final result is obtained based on the extracted information in the hidden layer.

In our model, classification portion consists of two densely connected layers with a sigmoid activation function at the last layer. Two dropout layers were used to randomly set a number of features to zero in order to regularize the weights.

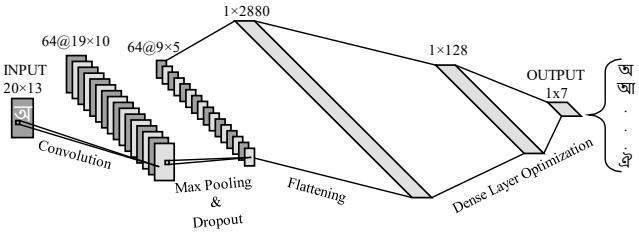

Fig. 1. Architecture of the model.

B. MFCC Feature

To determine MFCC features, the first step involves Fourier transformation on the input signal. Thus, the obtained power spectrum is compared to the Mel scale according to (2).

$$mel = 2595 \log_{10}\left(1 + \frac{x}{100}\right) \quad (2)$$

where, x is the input to the filterbank, and mel is the output of the mel filterbank [12].

In the next step, logarithm is taken on each of the above Mel frequencies. After that, Discrete Cosine Transform on the bank of the Mel log powers is required. Finally, the conversion of the log Mel spectrum back to time is called the Mel-Frequency Cepstral Coefficient (MFCC). By using this Cepstral representation, a good representation of the local spectral properties of the speech signal is extracted [13].

III. Methodology

A. Sound Capture

The very first step of this work was creating the dataset. In order to analyze the performance between the recognition of isolated vowels and isolated words, we have chosen seven isolated vowels and seven isolated words. The chosen vowels were- /অ/ [/a/], /আ/ [/ã/], / ই/[/i/] , /উ/ [/u/], /ঋ/[/ri/], /এ/ [/e/], /ঐ/ [/ai/], and the words were- বোতল, বন, কপি, দোকান, শেষ, সঠিক, উপরে. These particular words were chosen because these are more prone to vary when spoke by different speakers. Sounds were recorded by using the 'sound recorder' function of 'Xiaomi Redmi 3' smart-phone as a continuous stream of those vowels with a little silence in between them. There were about twenty different male and female speakers aged between 20-26. For some speakers, the record was doubled at a different accent in order to create variety in the dataset. There were 40 data in each class of the two sets, i.e. each of the vowel class (অ, আ, etc.) contains 40 variations, and each of the word class (বোতল, বন, etc.) also contains 40 variations.

B. Data Cleaning

The raw captured sounds were in two-channel, and it contained all the vowels in a single audio file for each speaker. In order to get the distinct feature, Audacity software was used which is a free and open-source digital audio editor and recording application software. In Audacity software, the recorded stereo channel audio is converted to the mono channel by using the function of the software from *Tracks > Stereo Track to Mono*. The distinct vowels and words have been clipped according to the waveform, and the selected

audio was exported and saved as 32-bit float datatype in different classes. The clipped spectrum of the 'অ' vowel and the 'বোতল' word is shown in Fig. 2 and Fig. 3 respectively.

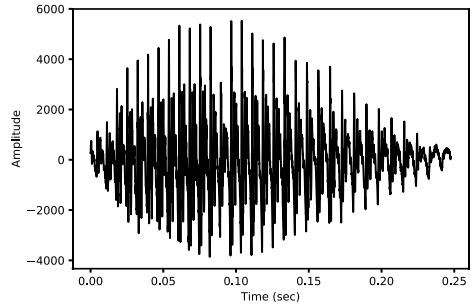

Fig. 2. Waveshape of the 'অ' vowel.

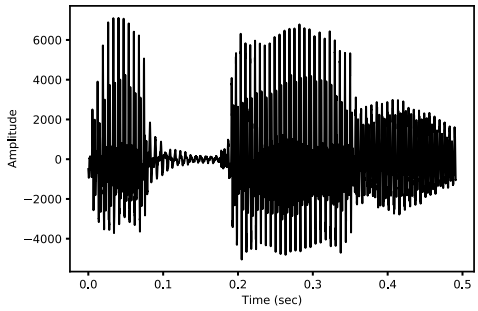

Fig. 3. Waveshape of the 'বোতল' word.

Fig. 2 and Fig. 3 depict that some discontinuity presents in the word spectrum. It's a natural phenomenon as the vocal organs including mouth, tongue, etc. are kept fixed when we speak vowels, like অ. On the contrary, those vocal organs change rapidly whenever we speak words, like বোতল.

C. *Feature Extraction*

MFCCs were extracted by using the *librosa* package of *Python* programming language. The raw audio data were down-sampled and 20 MFCC coefficients were extracted for each audio entry.

D. *Data Preprocessing*

Audio files of the vowels and words were in different time length and so the extracted MFCC vectors had also different lengths. But in order to feed the dataset to the neural network, all input data need to be of same size. For this, zero values were padded if the MFCC vector length was less than 13 and for those which are greater than 13 in length, rest MFCCs were discarded as they have less significance. The length was chosen as 13 because most of the MFCC length was around 13. Thus, each feature was in fixed 20×13 size – 20 represents number of the MFCCs, and 13 represents number of the time-samples. MFCCs were calculated once and they were saved as *numpy* Python array files in the working directory because these features might be needed many times in the trial and error process, and computing MFCCs each time is just the waste of time. That's why MFCCs were extracted form that *numpy* file in further operations.

IV. Results & Discussion

Among all the vowels and words, MFCC values of the vowel 'উ' and the word 'শেষ' are depicted in Fig. 4 and Fig. 5 respectively. It can be easily observed that the MFCC varies more rapidly for word as compared to vowel.

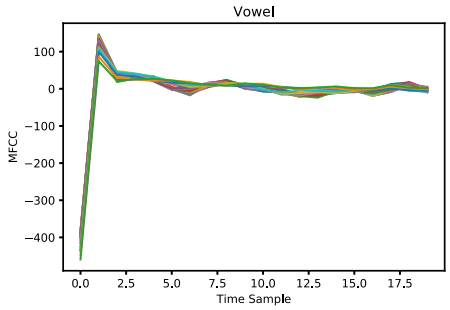

Fig. 4. MFCC over time for the 'উ' vowel.

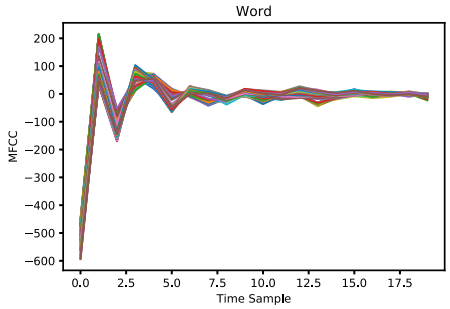

Fig. 5. MFCC over time for the 'শেষ' word.

Loss and accuracy are two different metrics to determine the performance of recognition. Fig. 6 and Fig. 7 shows the loss and accuracy comparison, respectively.

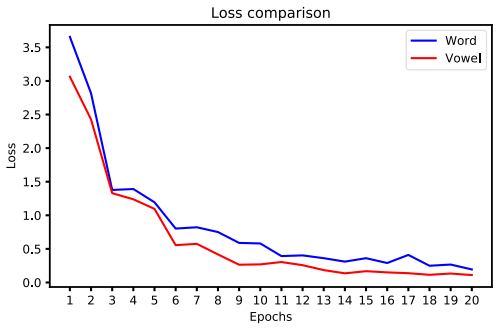

Fig. 6. Loss comparison between vowel and word recognition.

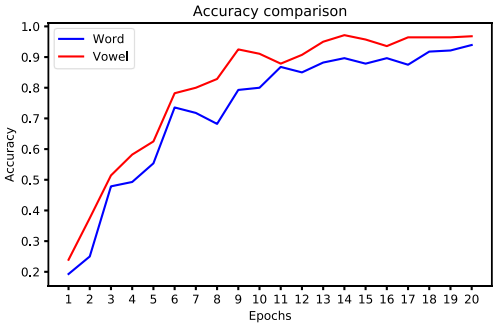

Fig. 7. Accuracy comparison between vowel and word recognition.

More accuracy and less loss are always desirable for a recognition system, and so it is clear from Fig. 6 and Fig. 7 that the recognition performance of the vowels is better than that of the words.

The above performance deviation was also verified by using confusion matrix and K-fold cross-validation. A confusion matrix is a table used to describe the performance of a classification model. This matrix maps the rate of correct prediction among all the input classes. Fig. 8 and Fig. 9 shows

the confusion matrix for vowel and word recognition, respectively. In Fig. 8, for True label ই, the value 1.00 in the predicted label ই means that in 100% test cases, the classifier predicted ই as ই, i.e. the correct prediction. On the other hand, the 0(zero) values in all other predicted labels mean that the classifier did not predict ই as any other different class.

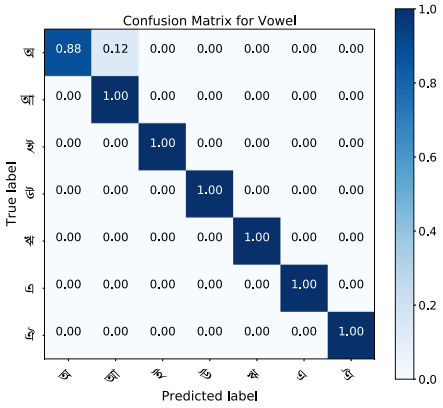

Fig. 8. Confusion matrix for vowel recognition.

Fig. 9. depicts that the word recognition was more confusing as compared to vowel recognition. For example, the word 'বন' was predicted 18% cases as দোকান, 73% as বন, and 9% as বোতল. Again, the recognition performance of the vowels was found better than the words.

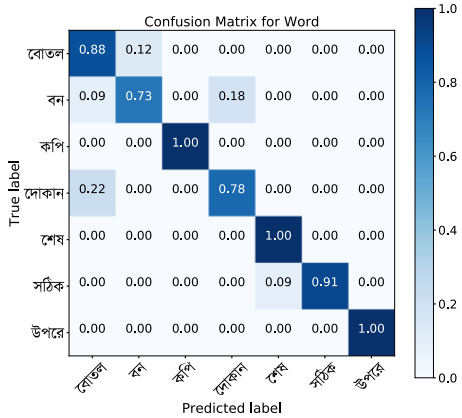

Fig. 9. Confusion matrix for word recognition.

K-fold cross-validation was also done to determine the performance of recognition which is shown in TABLE I. Number of Fold was chosen as four which means the total dataset was split into 4 equal sections. Each time any three sections were used to train the model, and the rest section was used to validate the model. This process continued until all the sections were used to validate the model for once. Finally, four validation accuracies were found for those four cases, and the overall accuracy denotes the average accuracy calculated from those four validations, which in turn proves that the performance of vowel recognition is greater than word recognition.

TABLE I. 4 – FOLD CROSS-VALIDATION RESULT

Recognition	Fold -1	Fold-2	Fold-3	Fold-4	Overall Accuracy
Vowel	95.71%	94.29%	92.86%	92.86%	93.93%
Word	90.00%	87.14%	90.00%	92.86%	90.00%

So, in all cases, the performance of vowel recognition was found greater than the word recognition, which is resulted from the variation of input classes as we have found previously, that vowels have less variation than the words. In essence, it is easily concluded that the recognition performance is inversely proportional to the variation of the input dataset.

V. CONCLUSION

Starting from speech data acquisition, performance between the vowel and word recognition is analyzed in this work. A speech token classification model is developed using CNN architecture, where the hyperparameters are selected as a brute-force approach. Seven isolated vowels and seven isolated words were fed to the model, and classification performance was measured according to four different metrics. In all those cases, vowel recognition outperforms word recognition. The reason behind such a result was also analyzed, and it was found that the dynamic change of the vocal tract plays a significant role in this case, as it changes more dynamically at the time of speaking words than the vowels. Therefore, there is more variation in the word feature as compared to vowel feature, and this deviation of feature results in deviation of performance. In essence, it can be said that intra-class variations of features affect the performance of recognition: more variation results in low performance and fewer variation results in high performance.

REFERENCES

[1] P. Barua, K. Ahmad, A. A. S. Khan and M. Sanaullah, "Neural network based recognition of speech using MFCC features," in *2014 international conference on informatics, electronics & vision (ICIEV)*, 2014.

[2] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1995.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath and others, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine,* vol. 29, pp. 82-97, 2012.

[4] T. S. Shanthi and C. Lingam, "Review of feature extraction techniques in automatic speech recognition," *International Journal of Scientific Engineering and Technology,* vol. 2, pp. 479-484, 2013.

[5] S. Gaikwad, B. Gawali, P. Yannawar and S. Mehrotra, "Feature extraction using fusion MFCC for continuous marathi speech recognition," in *2011 Annual IEEE India Conference*, 2011.

[6] G. Fant, Acoustic theory of speech production, Walter de Gruyter, 1970.

[7] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing,* vol. 22, pp. 1533-1545, 2014.

[8] F. Chollet, Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek, MITP-Verlags GmbH & Co. KG, 2018.

[9] C. K. Dewa, "Javanese vowels sound classification with convolutional neural network," in *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2016.

[10] B. Prasad and S. R. M. Prasanna, Speech, audio, image and biomedical signal processing using neural networks, vol. 83, Springer, 2007.

[11] D. Palaz, R. Collobert and others, "Analysis of cnn-based speech recognition system using raw speech as input," 2015.

[12] A. Winursito, R. Hidayat and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018.

[13] K. V. K. Kishore and P. K. Satish, "Emotion recognition in speech using MFCC and wavelet features," in *2013 3rd IEEE International Advance Computing Conference (IACC)*, 2013.