TUNING DARI SPEECH CLASSIFICATION EMPLOYING DEEP NEURAL NETWORKS

Mursal Dawodi and Jawid Ahmad Baktash

LIA, Avignon University, Avignon, France

ABSTRACT

Recently, many researchers have focused on building and improving speech recognition systems to facilitate and enhance human-computer interaction. Today, Automatic Speech Recognition (ASR) system has become an important and commontool from games to translation systems, robots, and so on. However, there is still a need for research on speech recognition systems for low-resource languages. This article deals with the recognition of a separate word for Dari language, using Mel-frequency cepstral coefficients (MFCCs) feature extraction method and three different deep neural networks including Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Multilayer Perceptron (MLP). We evaluate our models on our built-in isolated Dari words corpus that consists of 1000 utterances for 20 short Dari terms. This study obtained the impressive result of 98.365% average accuracy.

KEYWORDS

Dari, deep neural network, speech recognition, recurrent neural network, multilayer perceptron, convolutional neural network

1. INTRODUCTION

Humans usually communicate with each other through speech. Automatic Speech Recognition (ASR) tools have played a remarkable role in human-machine interaction since the advent of more advanced technologies and smartphones. The fundamental application areas of the ASR system are robotics, translators, machine- human dialogue systems, and so on. A simple idea of using an ASR system in robotics is to instruct the robotic vehicles. Similarly, Google translate transforms speech into words and translates them to other languages. Another example of ASR application is Siri that provides the environment of question answering dialogs between humans and machines. Research on ASR for regional and many Asian languages are state of the art in recent decades. Additionally, it faces many challenges, such as different dialects and a lack of resources, for example, a dataset with adequate terminology (Sharma et al., 2008).

Dari is one of the official languages of Afghanistan and is spoken as a first or second language by a majority of about 32 million inhabitants. Recently, there are many Dari mother tongues around the world, such as European countries, the United States, Canada, Australia, and so on. Even if people living outside Afghanistan can communicate in other languages, most Dari speakers are still unable to do so. In addition, only a handful of people are familiar with modern technology devices and applications. Hence, ASR can facilitate their interaction with machines like smartphones. In contrast to universal languages such as English, research on ASR systems for most languages including Arabic (an official language of the United Nations) has only begun a few years ago.

Recent articles investigate diverse methods to establish related database and implement ASR for regional languages such as Pashto (Zada & Ullah, 2020), Urdu (Wahyuni, 2017; Qasim et al.,

DOI:10.5121/ijnlc.2023.12203

2016; Ali, H., Jianwei, A.,& Iqbal, K. (2015). Automatic Speech Recognition of Urdu Digits with Optimal Classification Approach. International Journal of Computer Applications, 118(9), 1–5. https://doi.org/10.5120/20770-3275, Hindi (Sinha et al., n.d.; Sharma et al., 2008; Ranjan, 2010; Kumar et al., 2012), and Bengali (Muhammad et al., 2009). Themost key challenge in developing an ASR system for Dari is the lack of available datasets and the great variety of dialects. For instance, the Hazara dialect is very different from the Herat residents' accent, and both vary in intonation with how Kabul natives speak. Similarly, either the corpus is not available for most NLP research purposes, or it does not have enough records or vocabulary.

To the best of our knowledge, there is no published article considering the evaluation and comparison of the performance of different models on the Dari ASR. The primary purpose of this article is to provide a more robust and less error-proven ASR system for Dari through developing Dari isolated words ASR using three different deep learning techniques including Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Long Short-term Memory (LSTM). Consequently, we compare the outcome of the differentproposed models. This research is novel in the field of Dari speech recognition that compares the efficiency offive state of the art deep learning approaches.

Authors by Dawodi et al. (2020) established an ASR system. They used Mel-frequency cepstral coefficients (MFCC) to extract features and CNN for word classification. Finally, the obtained accuracy was 88.2% on average. The current study obtained more than 10% higher accuracy compared to (Dawodi et al., 2020).

The next section discusses some related works in this field. Section 3 briefly introduces the corpus structure. Subsequently, sections 4 to 6 describe an overview of Dari speech recognition along with MFCC features extraction and deep neural network models. The result and discussion on this work are illustrated in sections 7.Finally, section 8 concludes this article and reports future work.

2. RELATED WORKS

Several investigations employed machine learning and NLP techniques in speech recognition tasks. Recently, many researchers used deep learning technics to establish ASR systems. Usually, the majority of conducted studies focused on international and non-regional languages such as English. As state-of-the-art research, Mitra et al. (2017) proposed a novel hybrid convolution neural network (HCNN). The proposed architecture consists of two parallel layers for modeling acoustic and articular spaces. The layers were then joined at the output context-dependent state level. The comparison results demonstrate that the performance of CNN / DNN is analogous to HCNN. However, the HCNN model showed a lower word error rate. Similarly, Grozdic and Jovicic (Grozdic & Jovicic, 2017) constructed a new framework based on the automatic deep decoder encoder coefficient and Teager energy storage coefficients. They claimed that Teager energy-based cepstral properties are more powerful and describe whispers better than MFCC and GMM-HMM. The new model in the whisper scenario achieved 31% higher accuracy than traditional approaches and 92.81%-word recognition rate.

During the recent decade, some studies focused on more regional languages. An ASR for the Panjabi languagewas developed by Dua et al. (2012) that used MFCC for features extraction and HTK toolkit for recognition. They prepared the dataset containing 115 Panjabi terms by the utterance of 8 Punjabi native speakers. The overall mean accuracy achieved was between 94%–96%. Bakht Zada and Rahimullah (2020) developed a Pashto isolated digit ASR to detect Pashto digits from zero to nine. They used MFCC to extract features and CNN to classify digits. The utilized dataset consisted of 50 sounds for each number. Their proposed model contained four convolutional layers, followed by ReLU and max-pooling layers. Subsequently, they obtained an average accuracy of 84.17%. G. E. Dahl et al. (2012) developed a novel context-dependent model for speech recognition using deep neural network techniques. They proved that their proposed model is superior to previouscontext-dependent methods. Similarly, O. Abdel-Hamid (2014) used CNN in the speech recognition context and proved its efficiency in decreasing the error-rate and increasing robustness. A. Graves et al. (2013) proposeda hybrid model that involved bidirectional Long Short-term Memory (LSTM) RNNs and weight noise. They evaluate their model on the TIMIT phoneme recognition benchmark. As a result, the error rate decreased by upto 17.7%.

Few recently published articles focus on ASR for the Persian language using different models. H. Sameti et al. (2009) implemented a Persian continuous speech recognition system. They used MFCC with some modificationto learn features of speech signals and model-based techniques, speech enhancement approaches like spectral subtraction, and Wiener filtering to gain desirable robustness. Likewise, H. Hasanabadi et al. (2008) used a database containing Persian isolated words and developed Persian ASR in 2008. Finally, they create a wheeled mobile robot to navigate using Persian spoken commands. They investigated simple Fast Fourier Transform (FFT) to catch attributes and MLP to classify patterns. S. Malekzadeh et al. used MFCC for extracting featuresfrom Persian speeches and MLP for detecting vowel and consonant characters. The used dataset involved 20 categories of acoustics from utterances of 10 people. The proposed model demonstrated 61% - 87% conversionaccuracy. S. Malekzadeh utilized a deep MLP deep model to recognize Persian sounds to improve voice signalprocessing. Similarly, M. Namnabat and M. Namnabat (2006) established a Persian letter and word to sound system. They utilized rule-based for the first layer and MLP for the other layers of the network model. The average accuracy of 60.7% and 83% were obtained for letter and word predictions respectively. Recently, Veisiand Mani (2020) used a hybrid model of deep belief network (DBN) for extracting features and DBLSTM with Connectionist Temporal Classification (CTC) output layer to create the acoustic model. The study indicates thatDBLSTM provides higher accuracy in Persian phoneme recognition compared to the traditional models. This paper presents the design of an ASR for Dari isolated words. The works presented in this paper is based on three different deep learning techniques.

3. DARI WORD CORPUS

In this research, we used our previously created data set (Dawodi et al., 2020). This collection has 1000 sounds, which is related to 20 short Dari terms. The speakers expressed the words in different dialects. They were 20 Dari native speakers both male and female. In particular, there are 20 words listed in table 1. The speakers recorded audios in their home and office in a noise-free environment. The utterances were recorded by a smartphone audio recorder tool then they were transferred to the PC using a USB cable. We removed background information from speech signals and limit the length of each file to one second using Adobe auditionsoftware. Finally, we saved all audio files with the .wav extension format. Every word was saved in a separate file with its name. Subsequently, all utterance files related to a single term are located within the same folder, as an example, all audio records associated with the “Salaam” term from all speakers are stored in the “fold 20”folder. Consequently, there are 20 separate folders each relevant to a single term. In the end, we obtained 1000utterances since some of the files were corrupted.

Table 1 Labels

No.	Label	No.	Label
1	بد (Bad)	11	گل (Gol)
2	ببین (Bebeen)	12	حاضر (Hazer)
3	برادر (Brother)	13	جان (Jaan)
4	بیا (Beya)	14	خوب (Khoob)
5	بگو (Begoo)	15	خواهر (Khahar)
6	بوت (Boot)	16	کوه (Koh)
7	برو (Boro)	17	نیک (Nek)
8	چشم (Chashm)	18	پا (Paa)
9	دست (Dest)	19	قند (Qand)
10	گفت (Goft)	20	سالم (Salaam)

4. DARI ASR SYSTEM

This study is one of the states of the art research in the field of Dari ASR. The core building block of the proposed ASR system are audio signals that are inputs, MFCC to extract features from voice signals, and five separate deep learning models to recognize utterances. Besides, we divide the dataset to train and test sets with different ratios for evaluating the performance of architectures. Each method was trained and tested. Finally, the entire models are able to illustrate the detected term as output.

5. FEATURE EXTRACTION

Feature extraction is a fundamental task in analyzing data and figuring out the relationships between various objects, data, audios, and so on. Models are not able to recognize the utterances. Therefore, the speech signals need to be transformed into an understandable format. Besides, the features should be extracted from raw waveform to minimize variability in speech signals to produce input. This research uses MFCC that is the most dominant method for extracting key attributes (Wahyuni, 2017). In particular, the MFCC of a signal is a set of features that represents the entire shape of a spectral envelope. Generally, the feature set contains about 10 to 20 dominant signal properties. Therefore, this investigation prevents alteration in the variability of sound signalswhile reducing their magnitude.

High frequencies are usually suppressed when producing speech. Hence, the pre-emphasis phase rectifies theserepressed frequencies by transmitting the signal via the filter. Equation 1 describes pre-emphasis where out indicates the output, x is the variable, in stands for input, and K can have the value between 0.95-0.97.

$$o(x)=in(x)-Kin(x-1) \quad (1)$$

Afterward, each signal is divided into frames or small segments with a size of 20-40 milliseconds. Then, we applied the windowing technique to maintain continuity at the beginning and endpoints in a frame. Hence, eachframe was multiplied with a humming window to decrease the edge effect. Equation 2 defines the outcome of windowing signal, where the input signal in(x) is multiplied with the humming window w(x). The hummingwindow w(x) is described in equation 3, where N shows the total number of samples in each frame.

$$\mathrm{out(x)=in(x)}\times w(x) \ (2)$$
$$(x)=0.54-0.46\ co[2\pi nN-1],0\leq n\geq N-1 \ (3)$$

Finally, we converted the signals from the time domain to the frequency domain using the Fast Fourier transform (FFT). Equation 4 describes FFT.

$$o(x)=\Sigma in(x)eiwn\propto n=-\propto \ (4)$$

FFT has a very high frequency range which leads the speech signal to a non-linear scale. Hence, here Mel-filterbank (which is defined by equation 5) came into play to overcome the mentioned problem by estimating the mean energy per frame and capturing the algorithm of all filter bank energies. F(mel) is the Mel-filter bank.

$$(mel)=2595\times log10\ (1+f)700 \ (5)$$

Consequently, we obtained acoustic vectors by transforming the log Mel spectrum to the time domain using a discrete cosine transform (DCT). The output of DCT is the MFCC features.

6. DEEP NEURAL NETWORK MODEL FOR ASR

Diverse approaches and methods were applied to make ASR most robust and less error proven for various languages. HMM, the hybrid model of GMM-HMM and ANN-HMM are the most dominant techniques used for several decades. In recent years, the effectiveness of using artificial neural network algorithms in the field of speech recognition systems was proved. Lately, deep neural networks demonstrated improvements in the performance of ASR systems (Abdel-Hamid et al., 2014). A neural network model with more than one hidden layer is called a deep neural network (Abdel-Hamid et al., 2014). This research presents five artificial neural network models to build Dari ASR, including MLP, RNN, and CNN models. The feature maps representing MFCC features are used as the inputs for each model. In neural network architecture, the first hidden output layer is played the rule of input for the second layer, and so on. Setting the number of units in the hidden layer of the neural network is a challenge. Particularly, the number of units in the hidden layer is related to the complexity of the problem. In this structure, the weights are modified using the back propagation algorithm. In the back propagation, the second hidden layer is considered as the output layer for the first hidden layer, and allparts of the learning algorithm remain unchanged.

6.1. Convolutional Neural Network

Our deep CNN model consisted of four layers each containing 2D convolutional layers followed by max-poolinglayers except for the last layer that uses global-average-pooling. Subsequently, it has a fully connected layer inthe last as an output layer. The input of the first convolution layer is a 2D tensor of 40174 (representing MFCCfeatures) which convolve with 16 convolutional kernels. The next convolution layer is convolved with 32 filters.We doubled the number of kernels in the subsequent layers. The output of the proceeding layer is the input of the succeeding layer. Every convolutional filter generates a feature map based on input. As an example, the firstlayer generates 16 feature maps while the last layer provides 128 feature maps. Each convolution layer convolves with 2 filters. We used the linear activation function to overcome the gradient vanishing problem (Mitra et al., 2017). We used ReLU which is a well-known activation function in this study.

In the next step, max-pooling is applied to the features to reduce the sampling of the feature map by dividing the feature map into a small rectangular area, usually called the window, which does not overlap each other (Ide& Kurita, 2017). The studies proved that max pooling is more effective in comparison with other pooling types(Ide & Kurita, 2017). Hence, we used max pooling with the pool size of 2 instead of other pooling layers for the first three blocks. Max-pooling selects the maximum unit while average-pooling chooses the average unit fromthe entire units in a certain window as defined in equation 7 and 8, respectively.

$$g_{max}(x) = \max(x_i) \qquad (7)$$

$$g_{avg}(x) = \frac{1}{M} \sum_{i=1}^{M} x_i \qquad (8)$$

Usually, the performance of a model is affected by noise samples during training. Therefore, dropout is applied to address this problem and prevent a neural network from overfitting (Srivastava et al., 2014). We tried variantprobability values (in the range of 0 and 2.5) for dropout. As a result, the most effective probability value is 2.0which enhances the accuracy and minimizes the loss due to the limitation of our dataset for each term.

The output size is the same as the total number of classes which is 20 classes each related to a certain word. We examined the impact of different epochs and batch size on this model during training. Consequently, the optimalresult obtained with 80 epochs with a batch size of 84.

6.2. Multilayer Layer Perceptron

We evaluated different numbers of hidden layers using random configuration, but the best performance was obtained with only two hidden layers. Similar to the CNN model, each hidden layer is followed by the ReLU activation function. Every single layer except the output layer has 300 neurons. We examined a random numberof nodes between 25 to 350 and the best result was achieved with 300 output Perceptron. Afterward, a dropout of 2.0 is applied to decrease the likelihood of over fitting on training data by randomly excluding nodes from each cycle. The input shape for this model is a one-dimensional feature vector that contains 40 features as each sample contains 40 MFCC features. The output layer contains 20 nodes (each representing a class) is followed by the Soft Max activation function like in CNN model. The goal behind selecting Soft Max as the activation function for the last layer is that it sums up the output to 1. Finally, the model was trained with a batch size of 32 and a total number of 108,620 parameters.

6.3. Recurrent Neural Network

The structure of the implemented RNN model is also a sequential model as exposed in Figure 6. However, it consists of one Long short-term memory (LSTM) architecture block with the size of 64 which is followed by adropout with a rate of 0.2. Studies have shown that in many cases RNN works better with Tahn and sigmoid functions than ReLU (Good fellow et al., 2016). Hence, the LSTM model uses the Tanh activation function to activate the cell mode and the sigmoid activation function for the node output. The next layer is the Flatten layer to collapse the spatial dimensions of the input into the channel dimension. The output of the flatten layer passesto the output layer. The output layer consists of 20 nodes. Finally, the output layer is followed by the Soft ax activation function similar to the MLP and CNN models. The best performance was obtained using 100 epochswith a batch size of 64. The main reason for using different numbers of epochs and batch sizes is that we trainedeach model with several different epochs and batch sizes, then the most appropriate value that brings the modelto the highest accuracy was selected.

7. RESULTS AND DISCUSSION

We implemented a Dari ASR system to recognize 20 isolated words. At the initial step, we used our previously build dataset (Dawodi et al., 2020) that consist of 50 utterances for each term. MFCC algorithm was used to extract features from audio files. MLP, CNN, and LSTM were comparatively used with different numbers of parameters for speech classification. 10-fold cross-validation was used to evaluate the performance of each model. In all distinct models, we used categorical cross-entropy as loss function, accuracy for the metrics, and Adam as the optimizer.

MFCC feature extraction technique and artificial neural network techniques were implemented in python 3.6.9using a computer in windows 10 environment with Intel core i7(TM) 2.80 GHz 2.90 GHz processor. Librosa version 0.8.0 was used for feature extraction. Similarly, Tensor Flow version 2.3.1 and Keras version 2.4.3 were used for implementing MLP, CNN, and LSTM recognition models.

Table 2 illustrates average accuracy and loss during training and testing for each model. The highest average testing accuracy is obtained by CNN with 98.184% and 98.037% average testing accuracies, respectively. In the spite of the fact that LSTM is the most suitable model for many natural language processing applications due to its nonlinear nature, in this study LSTM obtained comparable average testing accuracy to MLP with a slight difference. The primary source of this result is that the dataset is small, and the sounds are simple thereforethe performance of MLP and LSTM is almost the same. However, the MLP obtained the highest average testing loss value amongst all five models.

Tables 3–5 demonstrate the sum of 10 confusion matrices associated with a certain model. As table 2 depicts the LSTM model obtained the lowest average training loss and the highest training accuracy. However, the average accuracy is decreased, and the loss is increased during testing. Similarly, CNN obtained 99.63% accuracy in fold 8 while this value raised down to 95.63% in fold 6.

Employing the MLP model, the training and testing phases in one-fold were completed in almost 11 seconds. However, this period increases to about 2 minutes and 32 seconds while training and testing LSTM architecturein 100 iterations. This execution time gets much longer (approximately 14 minutes and 12 seconds for 80 cycles)by applying the CNN model due to the complexity of its architecture.

Table 2 Training and testing accuracy and loss

Model	Average Training Accuracy	Average Training Loss	Average Testing Accuracy	Average Testing Loss
MLP	99.73	0.028	97.02	0.129
CNN	99.93	0.018	98.037	0.105
LSTM	100	0	97.712	0.09

For this study, we evaluate the best performance of the five different models by decreasing and increasing the number of dense layers, pooling layers, dropouts, testing, and training iterations, batch size, and kernel size. Finally, the best options were selected for this research as described in previous sections. Even though the LSTM represented better average accuracy during training (100%) but this value was lower during testing as depictedin table 2.

This research outperformed the previous related works on Dari isolated words speech recognition. The researchin (Dawodi et al., 2020) also used MFCC and CNN to create Dari ASR for isolated terms. Our model gives more than 10% test accuracy as compared to the existing work. Moreover, other methods showed lower evaluation loss and higher accuracy as well.

8. CONCLUSION AND FUTURE WORK

This paper demonstrates the impact of three different deep neural network models: the MLP, CNN, and LTSMalong with sensible training techniques for recognizing one-word Dari speech. It extracts audio signal features using MFCC. According to this study, the CNN model performed better in the field of isolated ASR for Dari words. It is a commencement study on Dari natural language processing and supplementary research needs to be done. The future work of this study is to implement continuous and more accurate Dari ASR models using hybrid models and novel techniques. However, a larger and richer set must be created for this proposal.

Table 3 Confusion-matrix MLP model

MLP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	169	0.5	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
2	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	5	0	220	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	135	0	0	2	0	0	0	0	0	0	0	0	0	0	3	0	0
5	0	0	0	0	131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	135	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	227	0	0	0	0	0	0	0	0	0	0	0	1	0
8	0	0	2	0	0	0	0	108	0	0	0	0	2	4	0	0	2	0	0	0
9	0	0	0	0	0	0	0	0	132	0	0	0	0	0	0	0	0	0	0	0
10	0	2	0	0	0	2	0	1	0	120	1	0	3	0	0	5	0	0	0	0
ML P	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	169	0.5	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
2	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	5	0	220	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	135	0	0	2	0	0	0	0	0	0	0	0	0	0	3	0	0
5	0	0	0	0	131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	135	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	227	0	0	0	0	0	0	0	0	0	0	0	1	0

8	0	0	2	0	0	0	0	108	0	0	0	0	2	4	0	0	2	0	0	0
9	0	0	0	0	0	0	0	0	132	0	0	0	0	0	0	0	0	0	0	0
10	0	2	0	0	0	2	0	1	0	120	1	0	3	0	0	5	0	0	0	0

Table 4 Confusion-matrix LSTM model

LSTM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	174	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	90	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	218	2	0	0	0	0	0	0	0	0	0	4	2	0	0	0	0	0
4	0	0	0	136	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0
5	0	0	0	0	131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	135	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	228	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	114	0	0	0	0	2	2	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	132	0	0	0	0	0	0	0	0	0	0	0
10	2	0	0	0	0	0	0	2	0	122	0	0	2	2	0	0	2	0	0	2
11	0	0	0	0	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	90	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	2	2	0	0	168	0	0	0	0	0	0	0
14	0	2	0	0	0	0	0	2	0	0	0	0	2	108	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	94	0	0	0	
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	0	
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136	0	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128	
20	0	0	4	0	0	0	2	0	0	4	2	0	2	0	2	0	0	0	0	

Table 5 Confusion-matrix CNN model

CNN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	171	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	1	0	0	0
2	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	213	2	0	0	2	0	0	2	0	3	2	0	0	0	0	0	1	1
4	0	0	0	140	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	135	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	2	0	226	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	114	0	2	0	0	0	0	0	0	0	0	0	2
9	0	0	0	0	0	0	0	0	132	0	0	0	0	0	0	0	0	0	0	0
10	0	1	2	0	0	0	3	0	0	128	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	2	0	88	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	172	0	0	0	0	0	0	0
14	0	0	0	0	0	0	4	0	0	0	0	0	4	106	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	94	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128	0
20	0	0	0	0	0	0	6	0	0	2	0	0	0	2	0	0	0	0	0	108

REFERENCES

[1] Abdel-Hamid, O., Mohamed, A., 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. https://doi.org/10.1109/TASLP.2014.2339736.

[2] Ali, H., Jianwei, A., & Iqbal, K. (2015). Automatic Speech Recognition of Urdu Digits with Optimal Classification Approach. International Journal of Computer Applications, 118(9), 1–5. https://doi.org/10.5120/20770-3275.

[2] Dahl, G. E., Dong Yu, Li Deng, & Acero, A. (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 30– 42. https://doi.org/10.1109/TASL.2011.2134090.

[4] Dawodi, M., Baktash, J. A., Wada, T., Alam, N., & Joya, M. Z. (2020). Dari Speech Classification Using Deep Convolutional Neural Network. 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 1–4. https://doi.org/10.1109/IEMTRONICS51293.2020.9216370

[5] Dua, M., Aggarwal, R. K., Kadyan, V., & Dua, S. (2012). Punjabi Automatic Speech Recognition Using HTK. 9(4), 6..

[6] Good fellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning (Illustrated edition). The MIT Press.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947.

[7] Grozdic, D. T., & Jovicic, S. T. (2017). Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12), 2313–2322. https://doi.org/10.1109/TASLP.2017.2738559.

[8] Hasanabadi, H., Rowhanimanesh, A., Yazdi, H. T., & Sharif, N. (2008). A Simple and Robust Persian Speech Recognition System and Its Application to Robotics. 2008 International Conference on Advanced Computer Theory and Engineering, 239–245. https://doi.org/10.1109/ICACTE.2008.125.

[9] Ide, H., & Kurita, T. (2017). Improvement of learning for CNN with ReLU activation by sparse regularization. 2017 International Joint Conference on Neural Networks(IJCNN),2684–2691. https://doi.org/10.1109/IJCNN.2017.7966185.

[10] Kumar, K., Aggarwal, R. K., & Jain, A. (2012). A Hindi speech recognition system for connected words using HTK. International Journal of Computational Systems Engineering, 1(1),25.https://doi.org/10.1504/IJCSYSE.2012.044740.

[11] Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., Saltzman, E., & Tiede, M. (2017). Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. Speech Communication, 89, 103–112. https://doi.org/10.1016/j.specom.2017.03.003.

[12] Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., Saltzman, E., & Tiede, M. (2017). Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. Speech Communication, 89, 103–112. https://doi.org/10.1016/j.specom.2017.03.003.

[13] Muhammad, G., Alotaibi, Y. A., & Huda, M. N. (2009). Automatic speech recognition for Bangla digits. 2009 12th International Conference on Computers and Information Technology, 379–383. https://doi.org/10.1109/ICCIT.2009.5407267.

[14] Namnabat, M., & Homayounpour, M. (2006). A Letter to Sound System for Farsi Language Using Neural Networks. 2006 8th International Conference on Signal Processing, 4128933. https://doi.org/10.1109/ICOSP.2006.345518.

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

[16] Price, R., Iso, K., & Shinoda, K. (2016). Wise teachers train better DNN acoustic models. EURASIP Journal on Audio, Speech, and Music Processing, 2016(1), 10. https://doi.org/10.1186/s13636-016-0088-7.

[17] Qasim, M., Nawaz, S., Hussain, S., & Habib, T. (2016). Urdu speech recognition system for district names of Pakistan: Development, challenges and solutions. 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 28–32. https://doi.org/10.1109/ICSDA.2016.7918979.

[18] Ranjan, S. (2010). Exploring the Discrete Wavelet Transform as a Tool for Hindi Speech Recognition. International Journal of Computer Theory and Engineering, 642–646. https://doi.org/10.7763/IJCTE 2010.V2.216.

[19] Sameti, H., Veisi, H., Bahrani, M., Babaali, B., & Hosseinzadeh, K. (2009). Nevisa, a Persian Continuous Speech Recognition System. In H. Sarbazi-Azad, B. Parhami, S.-G. Miremadi, & S. Hessabi (Eds.), Advances in Computer Science and Engineering (pp. 485–492). Springer. https://doi.org/10.1007/978-3-540-89985-3_60.

[20] Sharma, A., Shrotriya, M. C., Farooq, O., & Abbasi, Z. A. (2008). Hybrid wavelet based LPC features for Hindi speech recognition. International Journal of Information and Communication Technology, 1(3/4), 373.https://doi.org/10.1504/IJICT.2008.024008.

[21] Sinha, S., Agrawal, S. S., & Jain, A. (n.d.). Continuous Density Hidden Markov Model for Hindi Speech Recognition. 7.

[22] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(56), 1929–1958..

[23] Unnibhavi, A. H., & Jangamshetti, D. S. (2017). LPC based speech recognition for Kannada vowels. 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 1–4. https://doi.org/10.1109/ICEECCOT.2017.8284582.

[24] Veisi, H., & Haji Mani, A. (2020). Persian speech recognition using deep learning. International Journal of Speech Technology. https://doi.org/10.1007/s10772-020-09768-x.

[25] Wahyuni, E. S. (2017). Arabic speech recognition using MFCC feature extraction and ANN classification. 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 22–25. https://doi.org/10.1109/ICITISEE.2017.8285499.

[26] Zada, B., & Ullah, R. (2020). Pashto isolated digits recognition using deep convolutional neural network. Heliyon, 6(2), e03372. https://doi.org/10.1016/j.heliyon.2020.e03372.