Research article

Pashto isolated digits recognition using deep convolutional neural network

Bakht Zada *, Rahim Ullah

Article info

Keywords:
Computer science
Pashto isolated digits recognition
CNN
MFCC
Speech recognition

Abstract

Speech recognition has become one of the most significant parts of human-computer interaction due to emergence of new technologies such as smartphone, smart watch and many modern technologies, therefore the need of an ASR for local languages is felt. The basic aim of this paper is to develop an isolated digits recognition for Pashto language, using deep CNN. The database of Pashto digits from 0 to 9 with 50 utterance for each digit is used. Twenty MFCC features extracted for each isolated digit and fed as input to CNN. The network has been used for the proposed system is deep up to 4 convolutional layers, followed by ReLU and max-pooling layers. The network has been trained on the 50% of data and the rest of the data was used for testing. The total average of 84.17% accuracy was achieved for testing which show 7.32% better performance as compared to existing similar works.

1. Introduction

Pashto is one of the most spoken languages of Pakistan and an official language of Afghanistan. It is spoken language of Khyber Pakhtunkhwa's people in Pakistan. The word “Pashto” is pronounced in three different ways; Pashto, Fukhto, Pushto [22]. About 50–60 million population of the world speak Pashto including countries such as Pakistan, Afghanistan, UAE, UK, USA, and India [22]. Adversely, the educational aspect of Pashtoon people (people speaking Pashto) are not advance enough [23], due to this issue most of the people face difficulties in the use of modern technology. To overcome this issue, an ASR is needed to be develop for Pashto language in such a way that it most robust and less error proven.

The basic aim of this research work is to develop an ASR for Pashto language by utilizing new machine learning technique such as deep learning. More particularly, the primary goal of this research study is to design Pashto isolated digit recognition system by using deep convolutional neural network (CNN). Originally, CNN is developed for image recognition and become more popular for handwritten digit recognition, however in the last few years it also used for speech recognition [3, 24]. To the best of our knowledge no CNN even no deep learning technique is used in the field of Pashto speech recognition – 60 million population of Pashto people in Pakistan is a formidable task due to the lack of resources such as corpus with enough vocabulary, dialectical variation, accent and so on. Many works have been done for local languages such as in Punjabi (spoken in Pakistan and India) [5, 6, 7], Gujrati (local language of India) [8], Urdu (national language of Pakistan and forth most widely spoken language in the world) [9, 10, 11, 12, 13], Hindi (official language of India) [1, 14, 15, 16], Marathi (spoken in India) [17], Arabic (official language of Arab and fifth widely used language in the world) [18, 19], Bengali (spoken language of Bangladesh) [20]. The current work also includes Pashto language but very little works has been done in the development of Pashto speech recognition system [4]. Fortunately, Pashto share some characteristics with other languages like Arabic, Urdu and Persian languages. Therefore, some of the existence work for these languages can be utilized for Pashto as well [21].

Pashto is one of the most spoken languages of Pakistan and an official language of Afghanistan. It is spoken language of Khyber Pakhtunkhwa's people in Pakistan. The word “Pashto” is pronounced in three different ways; Pashto, Fukhto, Pushto [22]. About 50–60 million population of the world speak Pashto including countries such as Pakistan, Afghanistan, UAE, UK, USA, and India [22]. Adversely, the educational aspect of Pashtoon people (people speaking Pashto) are not advance enough [23], due to this issue most of the people face difficulties in the use of modern technology. To overcome this issue, an ASR is needed to be develop for Pashto language in such a way that it most robust and less error proven.

The basic aim of this research work is to develop an ASR for Pashto language by utilizing new machine learning technique such as deep learning. More particularly, the primary goal of this research study is to design Pashto isolated digit recognition system by using deep convolutional neural network (CNN). Originally, CNN is developed for image recognition and become more popular for handwritten digit recognition, however in the last few years it also used for speech recognition [3, 24]. To the best of our knowledge no CNN even no deep learning technique is used in the field of Pashto speech recognition – 60 million population of Pashto people in Pakistan is a formidable task due to the lack of resources such as corpus with enough vocabulary, dialectical variation, accent and so on. Many works have been done for local languages such as in Punjabi (spoken in Pakistan and India) [5, 6, 7], Gujrati (local language of India) [8], Urdu (national language of Pakistan and forth most widely spoken language in the world) [9, 10, 11, 12, 13], Hindi (official language of India) [1, 14, 15, 16], Marathi (spoken in India) [17], Arabic (official language of Arab and fifth widely used language in the world) [18, 19], Bengali (spoken language of Bangladesh) [20]. The current work also includes

* Corresponding author.
E-mail address: bakhtzada8c@gmail.com (B. Zada).

https://doi.org/10.1016/j.heliyon.2020.e03372
Received 7 December 2018; Received in revised form 25 October 2019; Accepted 3 February 2020

© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
A Pashto isolated digits ASR along with database was developed by [4]. Features were extracted through MFCC and k-NN classifier was used for digits classification. Total average accuracy of 76.8% was obtained. Recently, a new corpus was developed by [23], containing Pashto digits from zero (Sefar) to nine (Naha) with 150 instances for each digit. Mel-frequency cepstral coefficient (MFCC) was used for feature extraction and SVM classifier for classification. The total average result was 91.5%. The database developed by Zakir Ali in [4], is used in this work which contain 50 instance for each digits (25 males and 25 females). MFCC is used to extract 20 features and CNN is used for recognition of digits. The total accuracy of the model is 84.17% which is quite satisfactory as compare to 76.8% obtained by [4].

All the works of the paper is divided in to the following sections. Section 2 discusses the related works done in the field of speech recognition. A brief explanation of Pashto digit database is presented in section 3. Section 4 describes an overview of ASR along with MFCC features extraction and CNN Model. Section 5 of the paper narrates the result analysis and finally the conclusion and future work are reported in section 6.

2. Related work

This section presents some of the literature work in the domain of speech recognition. Mohit Dua et al [5], have developed and ASR for Punjabi. MFCC were used for features extraction and HTK toolkit for recognition. Data was recorded from eight Punjabi speakers for 115 different words of Punjabi. Each word was uttered 3 time by each speaker. The performance of the system was evaluated for the speakers, who were involved in training and other were only involved in testing. For the second case, 35–50 words were recorded by 6 different speakers to test the system. Total average accuracy was obtained in the range of 94%-96%.

An automatic spontaneous live speech recognition system for Punjabi language corpus has been developed by [7]. Carnegie Mellon University (CMU) Sphinx4 was used for both MFCC feature extraction and speech recognition. The data was collected from different male and female. Java language is used to develop a GUI for recording speech, also train and test the system. The system was evaluated for 691 Punjabi sentences and 2115 Punjabi words. Total accuracy for sentence recognition was 91.17% and for word recognition was 85.38%.

In [20], an ASR for Bangla speech was designed. HTK was used for feature extraction as well as for recognition. Hundred native speakers (50 male and 50 female) were selected to record 10 digits with sampling rate of 1125 Hz. More than 95% accuracy was achieved for first six digits (0–5) but less than 90% for other 4 digits.

A comparative work of LPC and MFCC has been done by H. B. Chauhan et al [8]. Two methods were used for feature extraction MFCC and LPC. The comparative study of both of these methods were presented in which MFCC gives better result than LPC. Vector Quantization (VQ) was used for testing the model (comparison of trained data with new input data). The data was recorded from both male and female in closed room with noisy environment. Each speaker, utter 10 words, 5 time each. More than 85% accuracy was obtained with LPC and more than 95% with MFCC.

Hazarat Ali et al [12], have presented an ASR for Urdu language. Feature were extracted using MFCC and delta-delta features. For recognition purpose three classifier were used; SVM, RF, and LDA. Dataset was created by recording speech from ten Urdu speakers including native and non-native. SVM was train with MFCC features, the average accuracy of 73% was achieved. Classification was also done by LDA and RF classifier using the same data and 63% accuracy was achieved for both of them.

In [25], speaker independent isolated speech recognition model is presented for three oriental languages i.e. Pashto, Urdu and Persian. Features were extracted by using discrete wavelet transform (DWT) and classification of digits are carried out through feed-forward artificial neural network (FFANN) with back-propagation. Three types of techniques were applied; haar, db-8 and sym-8. System was trained and tested for two, five, ten, fifteen, and twenty Urdu words. For two and five Urdu words, the model show 100% accuracy for all DWT filters. For 10, 15, and 20 Urdu words, db-8 level-5 DWT filters show the accuracy of 98.40%, 95.73%, and 95.20%, respectively. The accuracy of haar level-5 DWT filters is 97.20%, 94.40%, 91% and sym-8 level-5 filters is 95.20%, 94.67% and 89.40% for 10, 15, and 20 Urdu words respectively.

Zakir Ali et al [4], developed an ASR for Pashto language. The same dataset was used, as I used in the presented work. Features were extracted through MFCC algorithm. The overall average accuracy of 76.8% was achieved with k-NN classifier.

Another similar work on Pashto digits recognition was presented by Shibli Nisar et al [23]. MFCC and Prosodic Method were used for feature extraction. SVM and k-NN were used for digits classification. Data was collected from 150 Pashto native speakers includes both male and female. The accuracy with k-NN was 87.75% and 91.5% with SVM classifier.

This paper shows the design of an ASR for Pashto isolated digits. Very similar works have been performed by Zakir Ali et al [4] and Shibli Nisar et al [23]. However, the works presented in this paper is based on deep learning.

3. Pashto digits database

Database is developed by [4], and has 50 Pashto native speakers (25 male and 25 female) data. Each speaker utters digits from zero (Sefar) to nine (Naha); total of 10 digits uttered by 50 speakers (50 × 10 = 500 utterances). Each digit is saved in separate file with its name. The complete development process of database is presented in [4]. Database is split into two pieces one for training and other for testing. Fifty percent of the data is stored in a separate file named, “training” and the rest of 50% is stored in a file named, “testing”. Data in the “training” file is used to train the system and “testing” file data is used to test the system.
All of the speakers’ utterances are randomly distributed between training and testing, however, the number of male and female are the same in both training and testing.

4. ASR system - an overview

The block diagram of Pashto speech recognition using CCN is given in the Figure 1. The major parts of the system are speech signal as input, feature extraction from speech signal and CNN model for actual recognition. The model trained with five hundred iterations. After completing training phase, the trained model is obtained. The trained model then tested with 50% of data. There are only twelve iteration for testing phase and in the last the final test accuracy is obtained. The output of the model is recognized digits.

4.1. Feature extraction

Theoretically, it is also feasible to recognize speech directly from digitized waveform, but due to the large variance of speech signal, it is a better way to extract some features in order to minimize variability [26]. Particularly, to eliminate some unnecessary information like change of environmental conditions, change in transmission channel, background noise, and change in properties of microphone and so on [27]. The most dominant and widely used method to extract features is MFCC [28, 29]. To take out main and key attributes present in the speech, it is completely rest on frequency domain of Mel scale. MFCC degrade the total magnitude of an acoustic signal without interrupting its variability [30]. The work presented in the paper is based on MFCC features extraction algorithm. Block diagram of MFCC is given in Figure 2 [4], [31].

Figure 2 shows all the major phases of MFCC algorithm. Each of these phase are briefly discussed below.

**Step 1: Pre-emphasis**

In this step, the signal is passes through a filter to emphasis or compensate the high frequency part that was suppressed during the production of speech. Eq. (1) show the relationship between the input and output signal.

$$Y[n] = X[n] - \alpha X[n-1]$$  \hspace{1cm} (1)

where $Y[n]$ is the output pre-emphasis signal and $X[n]$ is the input signal. The value of $\alpha$ is in the range of 0.95–0.97, however the default value is 0.97 so, we consider its default value for our research work.

**Step 2: Framing**

In this step, the signal is divided into small segments called frames. Usually the size of each frame is between 20msec to 40msec. The size of each frame is taken as 20msec i.e. $\frac{320}{16000 \times 1000} = 20 \text{msec}$ ; where 320 is the sampling rate per frame, 16000 is the sampling rate of input speech and there are 1000 microseconds in one second.

**Step 3: Windowing**

By keeping continuity of the first and last points in the frame, each frame has to be multiply with a humming window. The result of windowing signal is defined by Eq. (2).

$$Y(n) = X(n) \times W(n)$$  \hspace{1cm} (2)

where $Y(n)$ is the output signal and $X(n)$ in input signal multiplied with hamming window; $W(n)$ is defined by Eq. (3)

$$W(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right), \hspace{0.5cm} 0 \leq n \leq N - 1$$  \hspace{1cm} (3)

where N is the number of sample per frame and W(n) is hamming window.

**Step 4: Fast Fourier transform**

Fast Fourier transform is use to convert each frame from time domain into frequency domain as defined in the Eq. (4).
The input of CNN is 2-dimensional feature map representing MFCC features. Hidden layers. The architecture of our CNN model is encapsulated in it. Chain of adjacent convolution and pooling layers are referred to connected hidden layers, a layer is only connected to a small region before called convolution and pooling layers, in which, instead of using fully CNN layer for speech recognition. According to Abdel-Hamid et al., with the use of ANNs, has been used for more than three decades and attracts scientists due to performance improvement in the sphere of decades the hybrid model of GMM-HMM and ANN-HMM have a successful contribution. A few decades ago, ANN (more particularly DNN) attracts scientists due to performance improvement in the sphere of speech recognition. ANNs, has been used for more than three decades for speech recognition. According to Abdel-Hamid et al., with the use of CCN in the field of speech recognition, the error rate can be reduced by 6% – 10% as compared to other deep neural network (DNN).

Convolutional neural network (CNN) is special type of standard feed-forward artificial neural network. CNN introduces a special structure called convolution and pooling layers, in which, instead of using fully connected hidden layers, a layer is only connected to a small region before it. Chain of adjacent convolution and pooling layers are referred to “one CNN layer”. Deep CNN is the one, which has two or more than two hidden layers [3]. The architecture of our CNN model is encapsulated in Figure 3, having four hidden layers, therefore referred to deep CNN. The input of CNN is 2-dimensional feature map representing MFCC features.

**4.2. CNN and their use for ASR**

Many works had been done in the domain of speech recognition for many languages including local languages such as Urdu, Hindi, Arabic, Mulayam, Bangla and Pashto etc. Many techniques and methods were tried to make ASR most robust and less error prone. HMM is one of the most dominated method used for many decades. However, in the last few decades the hybrid model of GMM-HMM and ANN-HMM have a successful contribution. A few decades ago, ANN (more particularly DNN) attracts scientists due to performance improvement in the sphere of speech recognition. ANNs, has been used for more than three decades for speech recognition. According to Abdel-Hamid et al., with the use of CCN in the field of speech recognition, the error rate can be reduced by 6% – 10% as compared to other deep neural network (DNN).

**4.2.1. Convolution**

A sequence of acoustic feature values $X$ such that $X \in \mathbb{R}^{c \times b \times f}$ where $c$ is number of channel, $b$ is number of frequency band and $f$ is time length. The convolutional layer convolves $X$ with $k$ filter $(W_i)$ where $W_i \in \mathbb{R}^{c \times m \times n}$ is a 3D vector or tensor such that $m$ is equal to width along the frequency axis and $n$ is equal to length along time axis. $H_i \in \mathbb{R}^{k \times b_o \times f_u}$ is a 3D tensor, in which each feature map $H_i$ calculated in Eq. (6) [23].

$$H_i = W_i \ast X + b_i, \quad i = 1, \ldots, k \tag{6}$$

As shown in the Figure 3, CNN has four convolutional layers followed by max-pooling layers (except from layer 4) and one fully connected layer in the last. A 2D tensor of $44 \times 20$, representing MFCC features, is fed as an input to first convolution layer and convolve with 25 convolutional kernel. Each convolutional kernel produced a feature map based on input. Second layer convolve with 22 convolutional kernel to generate 22 features map, similar layer three and layer 4 produced 44 features map. The output of the preceding layer is given as input to next convolutional layer and so on. First and second convolution layers convolve with $10 \times 10$ filter. Third layer convolve with $3 \times 3$ and fourth with $2 \times 2$. A larger filter size gives more accuracy and a smaller filter size gives more speed that is why the first two convolutional layers use larger filter size and the last two layers use smaller filter size. All the required parameters for CNN are listed in the Table 1.

**4.2.2. Activation function**

Number of nonlinear activation function are present, such as Tanh or Sigmoid for neural network and often give a well optimization during training of DNN. However, sometime it badly suffered with a problem, called gradient vanishing problem [32]. Fortunately, some activation function offer solution to this problem, few of them are rectified linear unit (ReLU), leaky rectified linear unit (leaky ReLU), parametric rectified linear unit (PReLU) and randomized rectified linear unit (RReLU) [32]. In the presented work, we prefer to use ReLU as activation function. The graph of ReLU is shown in the Figure 4 and define by Eq. (7) where $x$ inputs to $f(x)$, if the value of $x$ is negative then $f(x)$ will be zero otherwise the value of $f(x)$ will be equal to $x$.

$$f(x) = \max(x, 0) \tag{7}$$

**4.2.3. Max-pooling**

After applying activation function, max-pooling is applied on the feature. Pooling layer down-sampled the feature maps. To do this, the feature map divided into small rectangular region, which are not overlapping each layer [33]. These small rectangles usually referred to window. Maximum unit $g_{\max}(x) = \max(x_i)$ is selected (in the case of max-pooling) and average $g_{\text{avg}}(x) = \frac{1}{M} \sum_{i=1}^{M} x_i$ unit is selected (in the case of average-pooling) from all units in a specified window. In presented research study, the max-pooling and window of $2 \times 2$ dimension.

**Table 1. The parameters used in CNN.**

| Convolutional layer | Number of Filters | Filter size | Generated feature map | Max-pooling |
|---------------------|-------------------|-------------|-----------------------|-------------|
| First layer         | 25                | (10,10)     | 25                    | Yes         |
| Second layer        | 22                | (10,10)     | 22                    | Yes         |
| Third layer         | 44                | (3,3)       | 44                    | Yes         |
| Fourth layer        | 44                | (2,2)       | 44                    | No          |

**Table 2. Total training and testing accuracy.**

| Data             | Number of Speakers | Accuracy  |
|------------------|--------------------|-----------|
| Training         | 25 (both male and female) | 90.14%    |
| Testing          | 25 (both male and female) | 84.17%    |
are used because the max-pooling has a better result with compare to other pooling methods [33].

4.2.4. Dropout

When the input data is limited then unnecessary noise sample during the training that does not exist in real test data, therefore the performance of the network is degraded. Dropout offer a solution to such type of problem. Dropout is a way to prevent a neural network from overfitting [34]. Different probability values (in the range of 0 and 1.5) are tried for dropout in order to enhance accuracy and minimize the loss but the most effective probability value is 1.05 because our dataset contain limited data i.e. only 50 instances for each digit.

5. Results and discussion

We implemented an isolated Pushto speech recognition system that contain digits from zero (Sefar) to nine (Naha). Firstly, dataset was obtained from [4] that contain 50 utterance for each digit. Dataset divided into two parts, one for training and the other for testing. Twenty-five speakers (both male and female) data stored in a file named “training” and rest of the 25 speakers’ (both male and female) data in stored in file named “testing”. Features were extracted from data using MFCC algorithm. CNN was used for actual speech recognition. Data in the “training” file was used to train the network, similarly the model tested with data of “testing” file. All the isolated digits were recognized with the accuracy of 84.17%. Zakir Ali et al in [4], developed an isolated Pushto digit recognition system using MFCC for feature extraction and k-NN classifier for actual classification with accuracy of 76.85%. Our model gives 7.32% more accuracy as compare to the existence work.

MFCC feature extraction algorithm and CNN were implemented in python 3.6 using computer with Intel core i5, 2.50 GHz processor and in window 10 environment. Librosa version 0.6.0 library was used for feature extraction and Tensorflow version 1.6.0 for CNN implementation.

Total training and testing accuracy of the system is summarized in Table 2.

Graph for total accuracy and loss during the training are shown in the Figure 5 (a) and Figure 5 (b). Similarly graph for accuracy and loss of testing are shown in the Figure 5 (c) and Figure 5 (d). In all of the figures i.e. Figure 5 (a), (b), (c) and (d), x-axis show the number of iterations and y-axis show accuracy.

This study poses some limitations which are because of the immature platform for Pashto language. The dataset used in this study is very short; however, we aim to extend this. Secondly; the existing study is performed with small dataset on PC though the speed of PC is good but we aim to execute the same experiment with a huge dataset on cloud computer. Thirdly, the dataset consists of only isolated digits; however, in future; we aim to make a dataset consists of large vocabulary, connected words and continue words.

6. Conclusion and future works

In this paper, we have worked on Pashto isolated digit recognition using deep learning. The basic ambition of the work is to enrich the performance of Pashto ASR with the use of CNN. CNN is used for speech recognition and MFCC for feature extraction. The current study shows that both of these are powerful approaches in the area of speech recognition. The CNN is trained and tested with 20 coefficients extracted by MFCC feature extraction algorithm. The accuracy is enhanced by 7.32% as compared to previous work. The current works is only bounded into isolated digits recognition. The additional work can be achieve for large vocabulary, connected word recognition and continue word recognition.

Declarations

Author contribution statement

B. Zada, R. Ullah: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] S. Sinha, S.S. Agrawal, A. Jain, Continuous density hidden markov model for Hindi speech recognition, GSTF J. 3 (2) (2018).
[2] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, G. Penn, Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, in: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE, 2012, March, pp. 4277–4280.
[3] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu, Convolutional neural networks for speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (10) (2014) 1533–1545.
[4] Z. Ali, A.W. Abbas, T.M. Thaseema, B. Uddin, T. Raaz, S.A.R. Abid, Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and k-NN, Int. J. Speech Technol. 18 (2) (2015) 271–275.
[5] M. Dua, R.K. Aggarwal, V. Kadyan, S. Dua, Punjabi automatic speech recognition using HTK, IJCSI Int. J. Comput. Sci. Issues 9 (4) (2012), 1694-0814.
[6] K. Sharma, P. Singh, Speech recognition of Punjabi numerals using synergetic HMM and DTW approach, Indian J. Sci. Technol. 8 (27) (2015).
[7] V. Kumar, N. Singh, An automatic spontaneous live speech recognition system for Punjabi language corpus, Int. J. CTA 9 (20) (2016) 9575–9595.
[8] H.B. Chauhan, B.A. Tanawala, Comparative study of MFCC and LPC algorithms for Gujarati isolated word recognition, Int. J. Innovat. Res. Comput. Commun. Eng. 3 (2) (2015) 822–826.
[9] J. Asif, N. Iqbal, N.S. Khattak, A.M. Zaidi, March, Speaker independent Urdu speech recognition using HMM, in: Informatics and Systems (INFOS), 2010 the 7th International Conference on, IEEE, 2010, pp. 1–5.
[10] A.A. Raza, S. Hussain, H. Sarfraz, I. Ullah, Z. Sarfraz, An ASR system for spontaneous Urdu speech, Proc. Orient. COCOSDA (2010) 24–25.
[11] A.A. Raza, S. Hussain, H. Sarfraz, I. Ullah, Z. Sarfraz, Design and development of phonetically rich Urdu speech corpus, in: Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on, IEEE, 2009, August, pp. 38–43.
[12] H. Ali, A. Jianwei, K. Iqbal, Automatic speech recognition of Urdu digits with optimal classification approach, Int. J. Comput. Appl. 118 (9) (2015).
[13] M. Qasim, S. Nawaz, S. Hussain, T. Habib, Urdu speech recognition system for district names of Pakistan: development, challenges and solutions, in: Coordination and Standardization of Speech Databases and Assessment Techniques (O- COCOSDA), 2016 Conference of the Oriental Chapter of International Committee for, IEEE, 2016, October, pp. 28–32.
[14] A. Sharma, M.C. Shrotriya, O. Farooq, Z.A. Abbas, Hybrid wavelet based LPC features for Hindi speech recognition, Int. J. Inf. Commun. Technol. 1 (3-4) (2008) 373–381.
[15] S. Ranjan, Exploring the discrete wavelet transform as a tool for Hindi speech recognition, Int. J. Comput. Theor. Eng. 2 (4) (2010) 642.
[16] K. Kumar, R.K. Aggarwal, A. Jain, A Hindi speech recognition system for connected words using HTK, Int. J. Comput. Syst. Eng. 1 (1) (2012) 25–32.
[17] S. Kayte, M. Mundada, D.C. Kayte, Implementation of Marathi language speech databases for large dictionary, IOSR J. VLSI Signal Process. 5 (2015) 40–45.
[18] N. Hammami, M. Bedda, N. Farah, Probabilistic classification based on Gaussian copula for speech recognition: application to Spoken Arabic digits, in: Signal Processing: Algorithms, Architectures, Arrangements and Applications (SPA), 2013, IEEE, 2013, September, pp. 312–317.
[19] M. Hassine, L. Bousaid, H. Massoud, Hybrid techniques for Arabic Letter recognition, Int. J. Intell. Inf. Syst. 4 (1) (2015) 27–34.
[20] G. Muhammad, V.A. Alothai, M.N. Huda, Automatic speech recognition for Bangla digits, in: Computers and Information Technology, 2009. ICCIT’09. 12th International Conference on, IEEE, 2009, December, pp. 379–383.
[21] B. Rehman, Z. Halim, M. Ahmad, ASCII based GUI system for Arabic scripted languages: a case of Urdu, Int. Arab J. Inf. Technol. 11 (4) (2014).
[22] S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, A. Hafeez, KNN and ANN-based recognition of handwritten Pashto letters using zoning features, Mach. Learn. 9 (10) (2018).
[23] N. Nisar, I. Shahzad, M.A. Khan, M. Tariq, Pashto spoken digits recognition using spectral and prosodic based feature extraction, in: Advanced Computational Intelligence (ICACI), 2017 Ninth International Conference on, IEEE, 2017, February, pp. 74–78.
[24] Y. Zhang, M. Pezeshk, P. Brakel, S. Zhang, C.L.Y. Bengio, A. Courville, Towards End-To-End Speech Recognition with Deep Convolutional Neural Networks, 2017 arXiv preprint arXiv:1701.02720.
[25] B. Rehman, Z. Halim, G. Abbas, T. Muhammad, Artificial neural network-based speech recognition using dwt analysis applied on isolated words from oriental languages, Malays. J. Comput. Sci. 28 (3) (2015) 242–262.
[26] U. Shrawankar, V.M. Thakare, Techniques for Feature Extraction in Speech Recognition System: A Comparative Study, 2013 arXiv preprint arXiv:1305.1145.
[27] S.A. Majeed, H. Hussain, S.A. Samad, T.F. Ibeda, Mel frequency cepstral coefficients (MFCC) feature extraction enhancement IN the application OF speech recognition: a comparison study, J. Theor. Appl. Inf. Technol. 79 (1) (2015).
[28] A. Milton, S.S. Roy, S.T. Selvi, SVM scheme for speech emotion recognition using MFCC feature, Int. J. Comput. Appl. 69 (9) (2013).
[29] N. Dave, Feature extraction methods LPC, PLP and MFCC in speech recognition, Int. J. Adv. Res. Eng. Technol. 1 (6) (2013) 1–4.
[30] N. Zerari, S. Abdelhamid, H. Bourgos, C. Raymond, Bi-directional recurrent end-to-end neural network classifier for spoken Arab digit recognition, in: 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), IEEE, 2018, April, pp. 1–6.
[31] L. Muda, M. Begam, I. Elamvazuthi, Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, 2010 arXiv preprint arXiv:1003.4083.
[32] X. Wu, R. He, Z. Sun, A lightened cnn for deep face representation, in: 2015 IEEE Conference on IEEE Computer Vision and Pattern Recognition (CVPR), Vol. 4, 2015, June.
[33] H. Ide, T. Kurita, Improvement of learning for CNN with ReLU activation by sparse regularization, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE, 2017, May, pp. 2684–2691.
[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.