Vol. 8, No. 1 | January-June 2024

Design and Development of an Acoustic-Based Recognition System Using DNN

Saba Sultan[1], Humaira Ijaz[1], Bushra Jamil[1*]

Abstract:

Automatic speech recognition is a process of using computers to convert voice signals produced by human speech into reasonable format i.e. text or command that conveys the same meaning as the speaker intended to do. Many researchers are working on various languages including English and other European languages like Spanish, German, and French, etc. to develop an automated system for speech recognition (ASR). However, researchers have put little effort into developing ASR for the Urdu language. We have developed an Urdu speech recognition system using Deep Neural Network (DNN) on our developed corpus that contains some of the most frequently used words in Urdu like digits, season names, and month names. The accuracy rates of our ASR are very encouraging because 72% accuracy is achieved for 26 words and 92% accuracy is achieved separately for names of seasons.

Keywords: *Automatic Speech Recognition, Deep Neural Network, Multilayer Perceptron, Mel-frequency cepstral*

1. Introduction

The past decades have witnessed continuous development in computer technology along three dimensions i.e., increase in computing power, decrease in size and price of computers [1]. This advancement has changed the way we use computers. It has not only increased the use of computers multiple times in every facet of daily life but also at any time and everywhere. This transition from personal computing to ubiquitous and context-aware computing has also changed the human-computer interaction method. According to Mark Weiser ubiquitous environment enables us to perform computation and interaction more easily [2]. In the personal computing era, the mouse and keyboard were used as interaction tools between humans and computers. But now the modern ubiquitous computing environment uses speech as input to perform different tasks like searching for some information, controlling various IT appliances, etc [1].

Speaking is the most effective method of communication for humans to interact and share information. Computational linguistics with advanced computer technologies made it possible to develop an interaction tool between mankind and computers i.e. ASR (automatic speech recognition systems).

[1]Department of CS & IT, University of Sargodha, Sargodha, Pakistan

Corresponding Author: bushra.jamil@uos.edu.pk

Speech recognition is a process of converting voice signals produced by humans into the understandable format of text or command that conveys the same meaning as the speaker intends. ASR systems use computers to translate and identify words uttered by a human into reasonable text-like formats or commands. ASR is rapidly embedded in human life as it is used in home automation, automobiles, toy development and to help handicapped people in the educational field [3]. ASR also makes work processes and information retrieval more efficient because it produces words in documents as fast as they are spoken, which is generally much faster than a person can type.

There exist many ASR systems for English and other European languages based on different techniques like the Support Vector Machine (SVM), Hidden Markov Model (HMM), and Recurrent Neural Network (RNN) model,[4] however there are very few ASRs for the Urdu language. The Urdu language is one of the most extensively spoken languages because 588 million people around the world speak and understand the Urdu/Hindi language [5]. It is also the national language of Pakistan being recognized by 75% of the population. Despite these facts, little effort is made in the field of speech recognition for the Urdu language leading to the development of only a few ASR for the Urdu language. The existing studies focused on isolated word in controlled environments that are devoid of noise. These facts stresses the need for development of effective Urdu ASR that can address the challenges like background noise, non-continuous data as well as variability in pronunciation and speakers.

In the light of above challenges, our study aims to develop a medium scale Urdu Corpus with variability in speaker and pronunciation on the noisy and non-continuous data set. Moreover, we have developed a comprehensive ASRU system using SVM, DNN, as well as the generation of transcripts of spoken words. The main contributions of this article are summarized as follows:

1. Development of a medium-scale Urdu speech corpus with noisy and non-continuous data consisting of numbers, months, and seasonal names
2. An Urdu language ASR (ASRU) has been created using SVM and DNN
3. Generating transcripts of spoken Urdu words.

The remainder of this article is organized as follows. In Section II, we review the latest work on Urdu ASR. Section III presents the design models for ASRU using SVM, DNN, and implementation details about the training and test setup of both of the systems. An analysis of the results of training and testing is explained in Section IV. Section V presents the conclusion and prospects.

2. Related Work

This section presents the research work done for automatic Urdu speech recognition systems up till now. National Language Authority (NLA) and Research Center for Processing Urdu Language CRULP are two famous research bodies that are performing dedicated research in natural language processing (NLP) for Urdu and have contributed to the development of Urdu speech corpus for continuous speech [6]. In a speech recognition system, the type of speech can be either continuous or non-continuous. In non-continuous speech recognition, words are recorded separately one by one while in continuous speech sentences are recorded without any pause between the words. Ahad et al. [7] worked on the non-continuous speech of the Urdu language and made initial efforts in this regard. They used the Multilayer Perception (MLP) model to develop ASR for isolated digits of Urdu from

0 to 9. The developed corpus contained the speech of a single speaker. Later on S.K. Hasnain and S.M. Azam [8] implemented the recognition system for isolated words of the Urdu language using the feed-forward ANN model. In addition effort in this regard was made by Javed Ashraf and co. in 2010 who developed a speaker-independent ASR system for small-size vocabulary using HMM [9]. Ali et al. developed a corpus of isolated Urdu words consisting of 250 words. In 2013, Ali [10] worked to use Linear Discriminant Analysis for the classification of words. Shaukat et al. in 2016 [11] used HMM to develop an automatic Urdu speech recognition system using a speech corpus developed by Ali et al. They claimed to achieve better accuracy using this approach. In 2016, Hazrat Ali et al. worked on an Urdu Speech Corpus containing 250 words and achieved 73% accuracy using the Support Vector Machine model [12].

Till 2008, all the Urdu ASRs were developed for isolated Urdu words only. Afterward, some researchers worked on the development of a continuous speech corpus. In 2009, Dr. Agha Ali Raza et al. developed an Urdu corpus for continuous speech. In 2010, Sarfraz et al. developed a speaker-independent Urdu speech corpus for spontaneous speech recognition [13]. They also made a continuous speech recognition system for large Urdu vocabulary by using an open-source toolkit i.e., CMU Sphinx for speech recognition. Data were collected from native speakers and the achieved accuracy was about 60% [14]. M.U. Akram and M. Arif developed another MLP-based speech recognition system. They analyzed the matching pattern for continuous Urdu speech with 55 to 60% accuracy [15]. Later on, in 2014 Hazrat Ali et al. used discrete wavelets transform (DWT) and Mel Frequency Cepstral Coefficient (MFCC) based features to perform a comparative analysis and found that MFCC was much better as compared to DWT for recognition [16].

In [17], the authors used a vocabulary size of 199K words recorded from different Urdu and Punjabi speakers. They applied Gaussian Mixture Models based on Hidden Markov Models (GMM-HMM), Time Delay Neural Networks, Long-Short Term Memory, and Bidirectional Long-Short Term Memory networks. They developed a speech recognition system with a minimum Word Error Rate of 13.50%.

Aisha and her fellows developed an automatic audio Urdu digit recognition system for digits ranging from 0 to 9 [18]. They took 25,518 samples from 740 persons and applied SVM, MLP, EfficientNet, and convolutional neural networks (CNN) for digit classification. They claimed that the proposed CNN is more efficient and accurate than SVM and MLP.

Arif and fellows presented a comprehensive evaluation of multiple Urdu ASR models that included Whisper, MMS, and Seamless-M4T [19]. They used Word Error Rate (WER) as metric for performance evaluation. In 2023, the authors proposed an end-to-end neural network for Urdu ASR [20]. They incorporated semi-supervised learning techniques like dropout, ensemble averaging, and Maxout units for enhancing model generalization. In [21], authors presented a robust ASR system for Urdu by leveraging multilingual training and XLSR architecture. The authors claimed to achieve competitive performance using a cross-lingual approach for enhancing generalizability and recognition accuracy for Urdu in a resource-limited environment.

In [22] Khan et al. applied Time-delay Neural Networks (TDNN) and Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) training for effective handling of the variability in audio quality and speaking styles across different genres and claimed to achieve accuracy in genre-specific broadcast content.

Another study investigated the efficacy of three hearing aid amplification strategies including peak clipping, compression limiting, and wide dynamic range compression (WDRC) for hearing-impaired Urdu-speaking children [23]. The results showed that WDRC provides the most significant benefit in speech understanding.

Table 1 presents a brief summary of Literature Review in the form of table.

TABLE I. Literature Review Summary

Year	Contribution	Metrics	Datasets	Ref
2002	Isolated Urdu digit recognition using MLP.	Accuracy	Single speaker dataset	[7]
2009	Isolated word recognition with Feed-Forward ANN.	Accuracy	Custom dataset of isolated words	[8]
2016	Urdu ASR with HMM for a small vocabulary.	Word Error Rate (WER)	Small vocabulary Urdu dataset	[11]
2021	Urdu digit recognition using SVM, MLP, CNN.	Precision, Recall, F1 Score	Urdu digits dataset	[18]
2024	Benchmark for Urdu ASR models like Whisper, MMS.	WER, Accuracy	Conversational and read speech dataset	[19]
2023	Semi-supervised Urdu ASR using dropout, Maxout.	Word Error Rate (WER)	Urdu speech corpus with limited labeled data	[20]
2023	Multilingual ASR model for Urdu using XLSR.	WER, Accuracy	Multilingual Urdu dataset	[21]
2023	Multi-genre Urdu broadcast ASR using TDNN, GMM-HMM.	Word Error Rate (WER)	Multi-genre broadcast dataset	[22]

Most of the research work done to develop ASR for the Urdu language to date was for some isolated Urdu words, recorded in a noiseless environment. However, the objective of our research was to develop an Automatic Speech Recognition System of the Urdu language (ASRU) for using a medium vocabulary corpus consisting of non-continuous speech and noisy data.

3. Design and Implementation

This section explains the design model of an Automatic Speech Recognition System (ASRU) based on improved SVM and deep neural networks that use a medium-scale vocabulary corpus consisting of non-continuous speech and noisy data. ASRU is a novel, simple, yet efficient speech recognition system consisting of the following components:

1. Corpus Development
2. Data Preprocessing

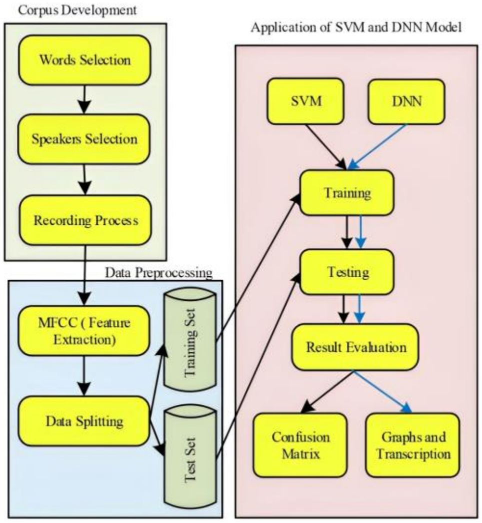


Fig I: Design model of ASRU

SVM and DNN model application for classification and recognition of words. The block diagram showing the entire process is shown in Figure 1.

3.1. Corpus Development

To develop an ASR, it is always necessary to have a speech corpus to which the speech recognition model is applied. Because there is no open-source speech corpus of the Urdu language, we need to construct our corpus. The main steps to develop a speech corpus are as follows:

1. Word Selection
2. Speaker Selection
3. Recording Process

The following subsections explain the details of each step mentioned above.

3.1.1 Word Selection

The choice of words is the first step toward speech corpus development. Selection is done with great care and two weeks have been spent selecting frequently used words from our daily lives. We chose 26 words, i.e., Urdu names of months are shown in Table 2, and Urdu numbers from 0 to 9, season names are shown in Table 3.

TABLE II. Urdu Vocabulary-name of Months

Labels	Urdu Words اردو کےالفاظ	Pronunciation or Urdu Roman
0	جنوری	Janvari
1	فروری	Farvari
2	مارچ	March
3	اپریل	Aprail
4	مئی	Mai
5	جون	Jun
6	جولائی	Julai
7	اگست	Agast
8	ستمبر	Sitambar
9	اکتوبر	Aktubar
10	نومبر	Navamvar
11	دسمبر	Disambar

TABLE III. Urdu Vocabulary-name of Digits, Seasons

Labels	Urdu Words اردو کےالفاظ	Pronunciation or Urdu Roman
12	صفر	Sifar
13	ایک	Aik
14	دو	Doe
15	تین	Teen
16	چار	Chaar
17	پانچ	Paanch
18	چھ	Chhay
19	سات	Saat
20	آٹھ	Aath
21	نو	Nau
22	سردی	Sardi
23	گرمی	Garmi
24	بہار	Bahaar
25	خزاں	Khizaan

3.1.2 Speaker Selection

The second step in corpus development was the selection of speakers. We performed casual and walk-through interviews to select appropriate speakers. Most respondents were undergraduate students, and graduate students, and some others were members of the faculty. Selected speakers had a clear voice, and accent and belonged to the 14-40 year-old age group. For recording purposes, we have chosen 22 male and 121 female native Urdu speakers.

3.1.3 Recording Process

The next phase after choosing the speaker was to record the chosen phrases spoken by the selected speakers. It was a hard task to dictate to each applicant the pronunciation of distinct Urdu phrases. Some speakers were shy and some were scared to use their voices. We made sure that their voices were used positively. Another challenge that could lead to distinct outcomes was major differences in accent and speaker style. Two months were spent in the selection and recording phase of speakers. 143 native speakers recorded the speaker's speech sounds. We saved recorded voices in.wav format for the digital representation of the spoken words.

3.2. Data Preprocessing

Following the development of the speech corpus, the next phase involves prepossessing the collected data set. Preprocessing consists of the following steps:

1. Feature Extraction
2. Speaker Selection

3.1.4 Feature Extraction

The first step in preprocessing is feature extraction from the speech dataset. We can extract the features from input speech to identify the relevant attributes, that can be useful in recognizing spoken words. The most commonly used feature extraction techniques for voice are MFCC and DWT. In ASRs, MFCCs are widely used for acoustic feature extraction to predict the vocal tract because their mechanism resembles human hearing. DWT features are used for non-stationary signals. We have used MFCC for feature extraction as it has proved much better than DWT [24]. Extracted features against each word were saved in .txt files as shown in Figure 2.

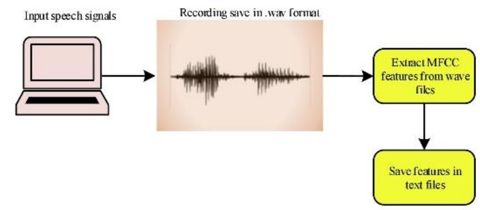

Input speech signals
Recording save in .wav format
Extract MFCC features from wave files
Save features in text files

Fig II: Feature Extraction Using MFCC Method

3.2.1 Splitting of Dataset

The next stage of preprocessing is to divide the dataset into a training dataset and a test dataset because we have applied our model in two stages namely the training and testing phase. In the first step, the model is initially trained on the training data set to learn some patterns by using SVM and DNN models. After training, the trained model is tested on the test dataset for an unbiased evaluation. Testing should always be performed on fresh datasets. Therefore, we have split our dataset into a training dataset and a test dataset and then labeled this training and testing dataset. 90% of speech sounds for each word are used for training purposes while the rest of the 10% is used for testing purposes.

3.3. SVM Model Application for Classification and Recognition of Words

SVMs (Support Vector Machines) are discriminative models well-adapted to classification issues. SVM can handle high-dimensional information and has excellent generalization capability, making it efficient for noisy speech data. We can use SVM classification for both linear and nonlinear problems. The main key concepts and hyperparameters in SVM are:

- Maximize the margin
- Selection of kernel(convert low dimension space(non-separable) into high dimension space(separable)

- Gamma(tries to fit the training data set correctly) c (cost function helps to achieve low training and testing error)

Classification is performed by finding a wide hyperplane used to differentiate classes, and the margin(the distance between samples of one class from the samples of another) plays an important role in drawing a hyperplane. We have used RBF as a kernel in our model. We have used Scikit-Learn, an open-source library for implementing SVM.

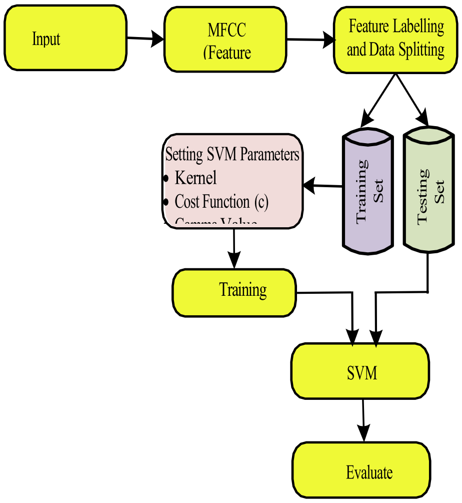


Fig III: SVM Architecture

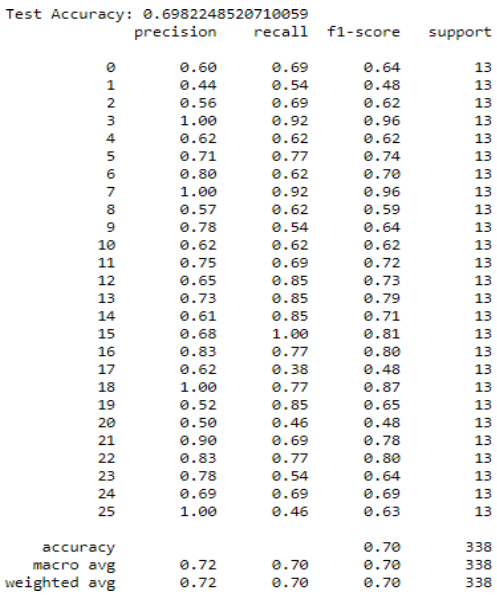
3.3.1 Training and Testing Process of SVM

After developing the SVM model the next phase includes a training process in which a speech recognition system is trained on the training dataset. For the training process, we have performed two sets of experiments. In the first set of experiments, we trained our SVM model on recorded words of our speech corpus consisting of 26 words i.e. digits, months, and seasons names, with 3380 voice samples. In this data set, some words have similar sounds like November and December which can affect the accuracy of our ASRU. Therefore we performed another set of experiments in which we selected another dataset consisting of totally dissimilar sounds like names of the four seasons with 520 speech sound samples separately. After this, we trained our SVM model using this dataset. To check the accuracy and performance of our ASRU, we performed testing on our test dataset consisting of 838 fresh speech sound samples. For the second experiment, testing is performed on the season names which consist of 130 unseen speech sound samples. We have tuned Hyperparameters like kernel, gamma, and value of c to obtain good training and testing results. When the model gets trained, we give test data to the SVM classifier to predict words accurately. The results of the training and testing process of our ASRU with SVM are shown in section 4.

3.4 DNN Model Application for Classification and Recognition of Words

We applied the DNN model for word classification and recognition because it has outperformed speech recognition [25]. A deep neural network is a subset of machine learning that imitates the working of human brains in processing and analyzing data in a structured way. DNN model contains three types of layers: input layer, hidden layer, and output layer. We have used one input and output layer with three dense and dropout layers as hidden layers. Each layer consists of neurons or nodes connected like a human brain. In ASRU the input layer takes .txt files of recorded words as input, processes it, and passes it on to the next layer as input. It is a feed-forward artificial neural network because each hidden layer receives input from the prior layer, applies a logistic function that utilizes weights, and biases, and maps its output as input to the next layer.

Weights give strength to each connection to the next node and bias values are used in the activation function to change the position of the line curve to leftward or rightward accordingly. The activation function is defined at each layer to help output nodes find optimal results from the given resources. ReLU is used as an activation function in input and hidden layers while Softmax is used in the output layer for multi-class classification of words. The architecture of the used DNN is shown in Figure 4.



```
Test Accuracy: 0.6982248520710059
              precision    recall  f1-score   support

           0       0.60      0.69      0.64        13
           1       0.44      0.54      0.48        13
           2       0.56      0.69      0.62        13
           3       1.00      0.92      0.96        13
           4       0.62      0.62      0.62        13
           5       0.71      0.77      0.74        13
           6       0.80      0.62      0.70        13
           7       1.00      0.92      0.96        13
           8       0.57      0.62      0.59        13
           9       0.78      0.54      0.64        13
          10       0.62      0.62      0.62        13
          11       0.75      0.69      0.72        13
          12       0.65      0.85      0.73        13
          13       0.73      0.85      0.79        13
          14       0.61      0.85      0.71        13
          15       0.68      1.00      0.81        13
          16       0.83      0.77      0.80        13
          17       0.62      0.38      0.48        13
          18       1.00      0.77      0.87        13
          19       0.52      0.85      0.65        13
          20       0.50      0.46      0.48        13
          21       0.90      0.69      0.78        13
          22       0.83      0.77      0.80        13
          23       0.78      0.54      0.64        13
          24       0.69      0.69      0.69        13
          25       1.00      0.46      0.63        13

    accuracy                           0.70       338
   macro avg       0.72      0.70      0.70       338
weighted avg       0.72      0.70      0.70       338
```

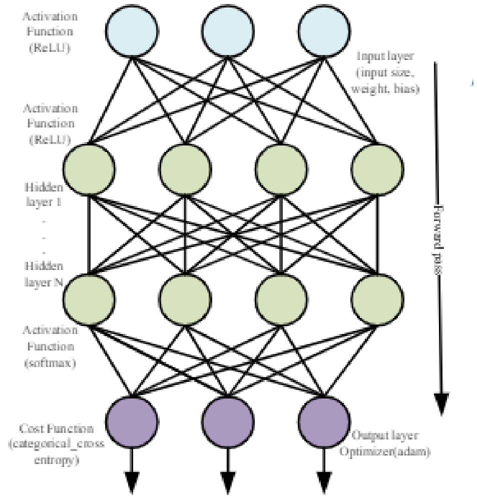

Activation Function (ReLU)
Input layer (input size, weight, bias)
Activation Function (ReLU)
Hidden layer 1
Hidden layer N
Forward pass
Activation Function (softmax)
Cost Function (categorical_cross entropy)
Output layer Optimizer(adam)

Fig IV: Layered Architecture

We have used TensorFlow and Keras to implement our model using the Anaconda environment. TensorFlow is a machine learning library used as a platform, and on top of TensorFlow, a range of libraries perform multiple tasks. Keras is a Python library used for fast and easy development of the deep neural network. Anaconda is an open source free distribution having thousands of packages to support scientific programming.

After developing the DNN model the next phase includes a training process in which a speech recognition system is trained on the training dataset. For the training process, we conducted two sets of experiments on the same dataset used in SVM for the training phase. After this, we performed testing on the same test dataset used for SVM to verify the accuracy and performance of our ASRU.

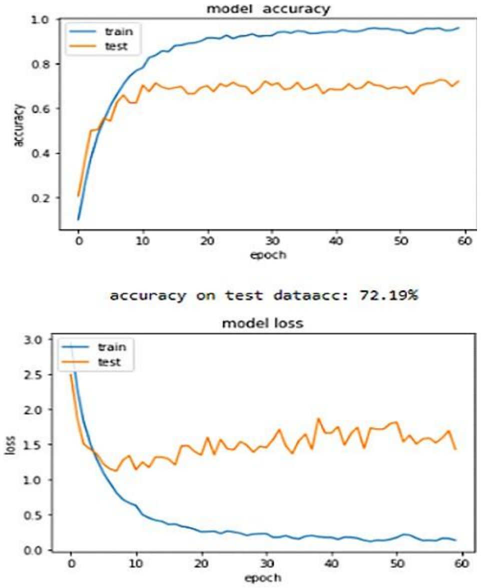
4. Results and Discussion

This section presents the results of the training and testing process of our developed ASRU. The results of the ASRU training and testing method with SVM are shown by the confusion matrix of the SVM model for speech recognition in Figure 5.

Fig V: SVM Confusion Matrix

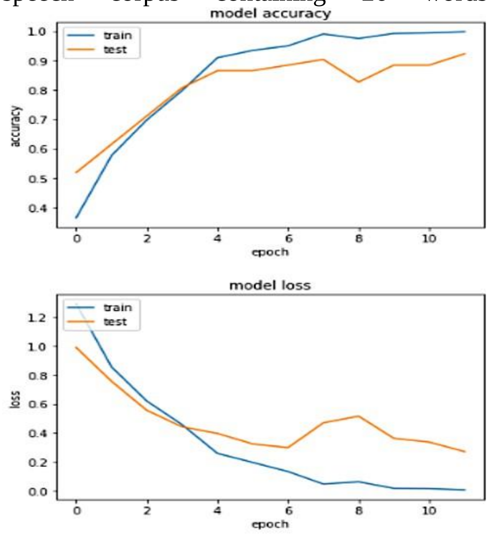
70% of accuracy is achieved on Urdu Speech Corpus. In the confusion matrix, the following rates are defined to show the performance of the SVM model.

- Accuracy(show overall accuracy of model)
- Precision(when the prediction is 4, how many times is it correct?)
- Recall(when the actual result is 4 how many times it is predicted as 4)
- f1-score(weighted average of the recall and precision)



```
accuracy on test dataacc: 72.19%
```

Figure 6 shows the results of the first experiment of the DNN model on the same speech corpus containing 26 words.



```
acc: 92.31%
```

Fig VI: Model Accuracy and Loss Graph

The first graph shows the relationship between epochs and accuracy. The horizontal axis in the graph represents several epochs and the vertical axis shows accuracy. Two colors are used in the graph, the blue line represents the results of the training process, and the orange line represents testing results. The accuracy graph shows that accuracy increases after 10 epochs and goes above 72% for the testing process as shown in Figure 6.

The second graph in Figure 6 shows the relationship between epochs and loss. The loss rate decreases after 10 epochs. Here, the blue line is again used to represent the results of the training process while the orange line represents testing results.

Fig VII: Accuracy and Loss Graph for Seasons

Figure 7 shows the results of the second set of experiments. We achieved 92% accuracy for the testing process because the sounds of words were dissimilar. Again blue and orange lines are used to present the accuracy of training data, and test data respectively. The second graph represents the model loss that starts decreasing after 10 epochs.

The above results indicate that our ASRU can perform more accurately for dissimilar voices even in noisy environments.

5. Conclusion and Future Work

We have developed ASRU for medium scale corpus of the Urdu language which contains the 26 most frequently used words in our daily life. For corpus development, only native male and female candidates were involved in the recording process. ASRU is implemented using two classifiers; SVM and DNN. Experiments delineate that DNN provides better results than SVM. Although the resources were limited, no proper recording environment was available and testing was also performed on a CPU with low computational power, still we achieved 70% accuracy using the SVM model and 92% accuracy y using the DNN model. Our work is the baseline for developing more accurate automatic speech recognition systems. There are a lot of prospects of such an experiment ASR for a large vocabulary corpus consisting of continuous data of Urdu sentences can be developed also.

AUTHOR CONTRIBUTION

Saba Sultan executed the research, whereas Bushra Jamil and Humaira Ijaz conceived the idea and supervised the work.

DATA AVAILABILITY STATEMENT

Not applicable.

CONFLICT OF INTEREST

The Authors declare that there is no conflict of interest.

FUNDING

(No funding available)

REFERENCES

[1] Furui, S. (2005). Speech recognition technology in multimodal/ubiquitous computing environments. Spoken Multimodal Human-Computer Dialogue in Mobile Environments, 13-36.

[2] Ballagas, R., Borchers, J., Rohs, M., & Sheridan, J. G. (2006). The smart phone: a ubiquitous input device. IEEE pervasive computing, 5(1), 70-77.

[3] Masui, T., Tsukada, K., & Siio, I. (2004). Mousefield: A simple and versatile input device for ubiquitous computing. In UbiComp 2004: Ubiquitous Computing: 6th International Conference, Nottingham, UK, September 7-10, 2004. Proceedings 6 (pp. 319-328). Springer Berlin Heidelberg.

[4] Yu, Y. (2012, March). Research on speech recognition technology and its application. In 2012 international conference on computer science and electronics engineering (Vol. 1, pp. 306-309). IEEE.

[5] Rabiner, L. R., Wilpon, J. G., & Soong, F. K. (1989). High performance connected digit recognition using hidden Markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(8), 1214-1225.

[6] Noack, R., & Gamio, L. (2018). The world's languages, in 7 maps and charts.[online] Washington Post.

[7] Ahad, A., Fayyaz, A., & Mehmood, T. (2002, August). Speech recognition using multilayer perceptron. In IEEE Students Conference, ISCON'02. Proceedings. (Vol. 1, pp. 103-109). IEEE.

[8] Beg, A., & Hasnain, S. K. (2009). A Speech Recognition System for Urdu Language. In Wireless Networks, Information Processing and Systems: International Multi Topic Conference, IMTIC 2008 Jamshoro, Pakistan, April 11-12, 2008 Revised Selected Papers (pp. 118-126). Springer Berlin Heidelberg.

[9] Ashraf, J., Iqbal, N., Khattak, N. S., & Zaidi, A. M. (2010, March). Speaker independent Urdu speech recognition using HMM. In 2010 The 7th International Conference on Informatics and Systems (INFOS) (pp. 1-5). IEEE.

[10] Ali, H., Ahmad, N., Yahya, K. M., & Farooq, O. (2012, April). A medium vocabulary Urdu isolated words balanced corpus for automatic speech recognition. In Proceedings of 4th International Conference on Electronic Computer Technology, ICECT, Kanyakumari, India (pp. 473-476).

[11] Shaukat, A., Ali, H., & Akram, U. (2016, August). Automatic Urdu speech recognition using hidden Markov model. In 2016 International Conference on Image, Vision and Computing (ICIVC) (pp. 135-139). IEEE.

[12] Ali, H., Ahmad, N., & Hafeez, A. (2016). Urdu speech corpus and preliminary results on speech recognition. In Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17 (pp. 317-325). Springer International Publishing.

[13] Raza, A. A., Hussain, S., Sarfraz, H., Ullah, I., & Sarfraz, Z. (2009, August). Design and development of phonetically rich Urdu speech corpus. In 2009 oriental COCOSDA international conference on speech database and assessments (pp. 38-43). IEEE.

[14] Sarfraz, H., Hussain, S., Bokhari, R., Raza, A. A., Ullah, I., Sarfraz, Z., ... & Parveen, R. (2010). Speech corpus development for a speaker independent spontaneous Urdu speech recognition system. Proceedings of the O-COCOSDA, Kathmandu, Nepal, 24.

[15] Sarfraz, H., Hussain, S., Bokhari, R., Raza, A. A., Ullah, I., Sarfraz, Z., ... & Parveen, R. (2010, December). Large vocabulary continuous speech recognition for Urdu. In Proceedings of the 8th International Conference on Frontiers of Information Technology (pp. 1-5).

[16] Akram, M. U., & Arif, M. (2004, December). Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling approach. In 8th International Multitopic Conference, 2004. Proceedings of INMIC 2004. (pp. 91-96). IEEE.

[17] Farooq, M. U., Adeeba, F., Rauf, S., & Hussain, S. (2019). Improving Large Vocabulary Urdu Speech Recognition System Using Deep Neural Networks. In INTERSPEECH (pp. 2978-2982).

[18] Chandio, A., Shen, Y., Bendechache, M., Inayat, I., & Kumar, T. (2021). AUDD: audio Urdu digits dataset for automatic audio Urdu digit recognition. Applied Sciences, 11(19), 8842.

[19] Arif, S., Khan, A. J., Abbas, M., Raza, A. A., & Athar, A. (2024). WER We Stand: Benchmarking Urdu ASR Models. arXiv preprint arXiv:2409.11252.

[20] Ali Humayun, M., Hameed, I. A., Muslim Shah, S., Hassan Khan, S., Zafar, I., Bin Ahmed, S., & Shuja, J. (2019). Regularized urdu speech recognition with semi-supervised deep learning. Applied Sciences, 9(9), 1956.

[21] Mohiuddin, H., Ahmed, Z., Kasi, M., & Kasi, B. (2023, November). UrduSpeakXLSR: Multilingual Model for Urdu Speech Recognition. In 2023 18th International Conference on Emerging Technologies (ICET) (pp. 217-221). IEEE.

[22] Khan, E., Rauf, S., Adeeba, F., & Hussain, S. (2021, November). A multi-genre Urdu broadcast speech recognition system. In 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA) (pp. 25-30). IEEE.

[23] Ali, W., Saqulain, G., Rehman, A. U., Patafi, J., & Rehman, A. (2024). Impact of different hearing aid amplification strategies on speech recognition in hearing impaired Urdu speaking children: A comparative study. Journal Riphah College of Rehabilitation Sciences, 12(3).

[24] Ali, H., Ahmad, N., Zhou, X., Iqbal, K., & Ali, S. M. (2014). DWT features performance analysis for automatic speech recognition of Urdu. SpringerPlus, 3(1), 1-10.

[25] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 29(6), 82-97.