Linear Discriminant Analysis Based Approach for Automatic Speech Recognition of Urdu Isolated Words

Hazrat Ali[1], Nasir Ahmad[2], Xianwei Zhou[1], Muhammad Ali[3], and Ali Asghar[4]

[1] Department of Communication Engineering, School of Computer and Communication Engineering, University of Science and Technology Beijing, 10083, Beijing China.
engr.hazratali@yahoo.com

[2] Department of Computer Systems Engineering, University of Engineering and Technology Peshawar, 25120, Peshawar, Pakistan.
n.ahmad@nwfpuet.edu.pk

[3] Department of Electrical and Computer Engineering, North Dakota State University, 58108-6050, North Dakota, USA.
muhammadali.sahibzad@my.ndsu.edu

[4] Department of Computer Systems Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.
ali.manjotho@faculty.muet.edu.pk

Abstract. Urdu is amongst the five largest languages of the world and enjoys extreme importance by sharing its vocabulary with several other languages of the Indo-Pak. However, there has not been any significant research in the area of Automatic Speech Recognition of Urdu. This paper presents the statistical based classification technique to achieve the task of Automatic Speech Recognition of isolated words in Urdu. For each isolated word, 52 Mel Frequency Cepstral Coefficients have been extracted and based upon these coefficients; the classification has been achieved using Linear Discriminant Analysis. As a prototype, the system has been trained with audio samples of seven speakers including male/female, native/non-native and speakers with different ages while the testing has been done using audio samples of three speakers. It was determined that majority of words exhibit a percentage error of less than 33 %. Words with 100 % error were declared to be bad words. The work reported in this paper may serve as a strong baseline for future research work on Urdu ASR, especially for continuous speech recognition of Urdu.

Keywords: Urdu Automatic Speech Recognition, Mel Frequency Cepstral Coefficeints, Linear Discriminant Analysis, Isolated Words Recognition

1 Introduction

User friendly and natural interaction between man and machine has always been a complementary part of technological development. Speech is the most effective medium of communication between human and same is envisaged to be applicable for human-machine interaction. Therefore, Automatic Speech Recognition (ASR) has significantly grabbed the attention of researchers for the last five decades and has attained considerable success in noise-free environments. Successful ASR enables the computers to exhibit human-like behavior by understanding the voice input to them. Such hearing systems having been developed in various languages such as English, French, Japanese, Chinese and Arabic [1–5], and have wide-spread application ranging from data entry to security and surveillance. The research on ASR has enabled the communities with lower level of literacy to interact with machines, and similarly facilitated the interaction of blind and disabled people with the computers [6]. Despite the development of ASR systems in these languages, there has been no significant contribution to ASR of Urdu language, which is one of the largest languages of the world. Wiqas [7] has summarized the research work conducted on the ASR of the languages of the Indo-Pak, including the research work on Urdu ASR. A continuous speech ASR system for Urdu language has been presented in [8], however, no information on the use of a standard corpus of Urdu has been provided. The recognition rate is limited to 55 % accuracy for continuous speech. Furthermore, it lacks the information about the use of number of words/sentences and the training/test data. Azam [9] has proposed an Artificial Neural Networks (ANN) based Urdu speech recognition system however; this work is limited to digits recognition only. Moreover, the application of the system is limited to single speaker only. Ahad et al [10] has used a different class of ANN called multilayer perceptrons (MLP) however; they have achieved recognition of Urdu digits from 0 to 9 for mono-speaker database only. Hasnain et al [11] has made yet another effort to achieve the task of digits recognition for 0 to 9, based on the use of feed-forward neural network models developed in Matlab. A more recent contribution to isolated words recognition has been made by [12], developing a Hidden Markov Model (HMM) [13] based speaker-independent speech recognition system for Urdu. In this work the open source framework Sphinx-4 has been used for the classification. A wordlist grammar language model was adopted where each word was represented as a single phoneme instead of dividing into sub-units. An apparent limitation of this approach is that this may be applicable to shorter words but for longer words, the performance may degrade drastically. Huda [14] has used a relatively larger data set for the training purpose, however, the system developed is for continuous speech recognition task and the recognition results are yet modest. Research on ASR can be targeted at small, medium or large vocabulary applications; it may be for digits only, isolated words only or continuous speech applications. The applications of isolated words recognition are well known including the automated banking applications, automatic data and PIN codes entry applications, e-health monitoring and voice dialing phone applications etc. In this paper the ASR task for medium vocabulary isolated words has been undertaken containing 100 isolated words of Urdu. The three important components of an ASR system are the corpus i.e. the database of speech data, the features extraction and the classification. In Section II of this paper, the corpus used for this work has been discussed briefly. The features extraction approach and the major steps involved in the extraction of these features have been presented in Section III. The classification of the different words based upon the features obtained for each word, has been discussed in Section IV. Finally, the results have been summarized in Section V.

2 Corpus Selection

One of the most important components of an ASR system is the use of a standard corpus covering a range of acoustic variations and different aspects of a language. In this work, the corpus developed in [15], has been used. The corpus contains 250 isolated words selected from the list of most frequently used words, developed by the Center for Language Engineering [16]. Audio files for one hundred isolated words have been selected from the corpus and used in the training and testing of the system. The one hundred words used contain the digits from 0 to 9, names of seasons, days of the week and the names of months. Besides this, for few of the words, their antonyms have also been included. The words are available in separate audio files with an average length of 500 milliseconds and stored in mono format with .wav extension. Based upon the attributes such as age, gender and origin, this corpus provides a balanced distribution. The files include the words uttered by both male and female speakers of different ages. Similarly, a variety of accents has been covered by including the audio recordings by both native and non-native speakers originating from different areas. For example, Pashto speakers from different regions of Pakistan differ in the pronunciation of Urdu words, thus, data from these speakers provide a variety of samples for training and testing purpose. A sample representation of the attributes of the speakers has been shown in Table 1.

Table 1. Sample of Representation of the Speech Data (as in [15])

S. No	Speaker Name	Age Group	Gender	Native Non-Native
1	AAMNG1	G1	Male	Non-Native
2	ABMNG1	G1	Male	Non-Native
3	ACMNG2	G2	Male	Non-Native
4	AEFYG1	G1	Female	Native
5	AFFYG1	G1	Female	Native
6	AGMNG1	G1	Male	Non-Native
7	AHMNG1	G1	Male	Non-Native

3 Feature Selection

Feature Extraction is one of the most important modules of an Automatic Speech Recognition System. For continuous speech recognition, the feature extraction is typically aimed to capture the distinguishing characteristics of the phonemes i.e. the smallest unit of sound. However, for isolated words recognition, each word is usually split into equal number of segments and features are extracted from each of the segments. In this work, each word is split into four segments and the Mel Frequency Cepstral Coefficients (MFCC) based features have been obtained for each segment.

3.1 Mel Frequency Cepstral Coefficients

The MFCC features are the most commonly used features for Automatic Speech Recognition as MFCCs most closely resembles the human hearing mechanism. The Mel scale is based on the fact that the frequency response of the humans ear to the audio signal is not a linear function of frequency. This response can be best modeled on a Mel scale where the spacing between frequencies above 1000 Hz is logarithmic [17]. The relation between the Mel scale frequencies and the Hertz frequencies can be represented by equation 1;

$$f_{mel} = 2595 \times \log{(1 + \frac{f}{700Hz})} \tag{1}$$

The Mel Frequency Cepstrum is the power spectrum of a speech signal for short term and is based upon a linear cosine transform of a log power spectrum on the Mel scale. The Mel Frequency Cepstrum comprises of the MFC coefficients. Several methods for MFCC extraction have been proposed by [17–19]. The major steps in the extraction of MFCC are shown in Algorithm 1.

In the pre-processing step, the segmentation of the words and noise removal have been achieved by using Adobe Audition Software. The sampling rate was set to 16000 Hz and the audio samples were saved as .wav files in mono format before being input to the algorithm. The Adobe Audition software has also been utilized for amplification or attenuation of the audio signal, as necessary, to obtain a uniform db level for all the samples. Besides, as the recording was performed in a controlled environment, this helps out in assuring minimum effect of noise. A snapshot of the Adobe Audition environment has been shown in Fig. 1.

The pre-processing stage also includes the Pre-emphasis of the signal to increase the energy of the higher frequency contents. The pre-emphasis is achieved using filter of the form, as in Equation 2.

$$H(z) = 1 - 0.97z^{-1} \tag{2}$$

The pre-processing is followed by the windowing of the speech signal. A rectangular window as defined by equation for $w(n)$ in equation 3 has been used. For speech processing applications, hamming window is more commonly used to avoid information loss, however for isolated words processing, rectangular window is equally beneficial.

$$w(n) = \begin{cases} 1 & 0 \leq x \leq M-1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

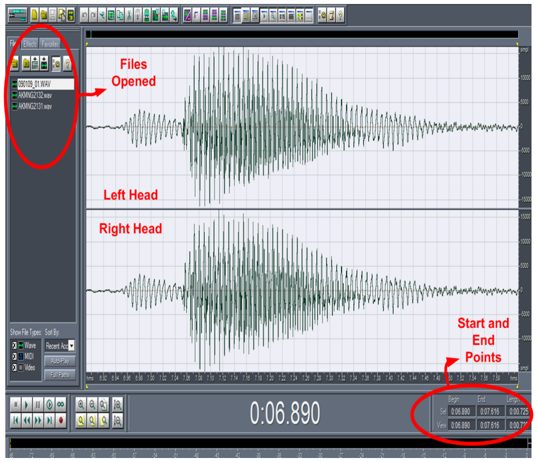
```
input  : An isolated word - audio file
output: 52 Mel Frequency Cepstral Coefficients
initialization;
Preprocessing
for $i \leftarrow 1$ to 4 do
    Segmentation;
    Noise Removal;
    Pre-Emphasis $(1 - 0.97z^{-1})$
end
Windowing
if isolated words then
    Rectangular Window else
        Hamming Window
    end
end
Discrete Fourier Transform
FFT Size = 552
Transformation to Mel Scale
linear scale → $logarithmicscale$
$f_{mel} = 2595 \times \log 1 + \frac{f}{700Hz}$
Discrete Cosine Transform for $j \leftarrow 1$ to 4 do
    processing;
end
A single word ;
$\Longrightarrow 52\ MFCC$
Dimensionality Reduction
```

Algorithm 1: Extraction Algorithm for Mel Frequency Cepstral Coefficients

where $M = 128$. Fast Fourier Transform [20, 21] is applied to the windowed frame of the signal. The size of FFT is $N = 512$. The spectrum, thus obtained, is transformed to the Mel scale, as defined by the equation for fmel. To imitate the logarithmic response of human ear, the output of the mel scale filters bank is subjected to base 10 Log. Finally, the application of Discrete Cosine Transform (DCT) [22] generates the MFCCs, i.e. 52 MFCCs for each isolated word.

4 Classification

The recognition on the basis of MFCCs requires a supervised classification technique for which Linear Discriminant Analysis is a strong candidate [23, 24]. The classification includes; *training of the system* and *testing of the system.* 70 % percent of the data has been used for training the ASR system and the remaining 30 % data has been used for testing of the system.



	Begin	End	Length
Sel	0:06.890	0:07.616	0:00.725
View	0:06.890	0:07.616	0:00.72

Fig. 1. Segmentation in Adobe Audition Enviroment

4.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a classification as well as dimensionality reduction technique. LDA can be class-dependent or class-independent, based upon maximization of the ratio of between class variance to within class variance or maximization of the ratio of overall variance to within class variance, respectively.

4.2 Training and Testing Data

To evaluate the performance of the ASR system, the MFCCs of a total of hundred words have been used for training and testing of the system. As a simple case, the training and testing has been done with the speech data of first ten speakers. The training set contains data from both native and non-native speakers of Urdu. Similarly, it also contains male as well as female speakers.

4.3 Confusion Matrix

The number of correct matches from the testing data with the training data has been summarized in a Confusion Matrix. The confusion matrix is of size $N \times N$ for N number of words. It can be represented as shown by Mc.

$$M_c = \begin{array}{cccc} m_{11} & m_{12} & m_{13}... & m_{1N} \\ m_{21} & m_{22} & m_{23}... & m_{2N} \\ m_{31} & m_{32} & m_{33}... & m_{3N} \\ . & . & .... & . \\ . & . & .... & . \\ m_{N1} & m_{N2} & m_{N3}... & m_{NN} \end{array} \qquad (4)$$

The number of correct matches for a word i has been shown by the diagonal entries of the confusion matrix, i.e. m_{ij} for $i = j$. Number of confusions of word i with word j has been shown by non-diagonal entries, i.e. m_{ij} for $i \neq j$.

5 Results

The error in the recognition of any isolated word is calculated from the confusion matrix. For an isolated word i, the diagonal entry m_{ii}, divided by the sum of all the entries in row i, gives the fraction of test data correctly matched. The sum of all the entries in a row is always equal to the number of test signals. This ratio can be defined mathematically as;

$$Correct\ Match, C \equiv \frac{m_{ij}}{m_{i1} + m_{i2} + ...m_{iN}}, \text{for } i = j,\ j = 1, 2, 3...N. \tag{5}$$

Thus, the error is measured by using the following equation;

$$\%error = (1 - C) \times 100 \tag{6}$$

5.1 Results for first ten words

Fig. 2 shows the confusion matrix graph for the first ten words. The x-axis and y-axis represent the indexes for the words i.e. 001 to 010. The number of successful or incorrect matches is represented by the height of the bars. As already mentioned, the maximum possible height is 3 as the number of test signals used here is 3. The percentage error and number of fraction of test signals correctly recognized has been summarized for the first ten words in Table 2. As shown in this table, the first word gives 66 % correct match, also depicted by the confusion matrix graph, by the first bar having a height of 2. The test signals for word 004 has undergone a 0 % error and the bar for this word has a height of 3. Similarly, the results for other words are obvious from the confusion matrix graph in Fig. 2 and the corresponding Table 2.

5.2 Results for words 031 to 040

As a second sample of the result, confusion matrix graph for word 031 to 040 has been shown in Fig. 3. The corresponding fractional values for correct matches and percentage error have been summarized in Table 3. The results shown in Fig. 3 are very important and needs to be discussed. As shown in Table 3, it is obvious that there is a zero percent error for words 032 through word 034. On the other hand, a complete mismatch exists for word 031, resulting in a 100 % error.

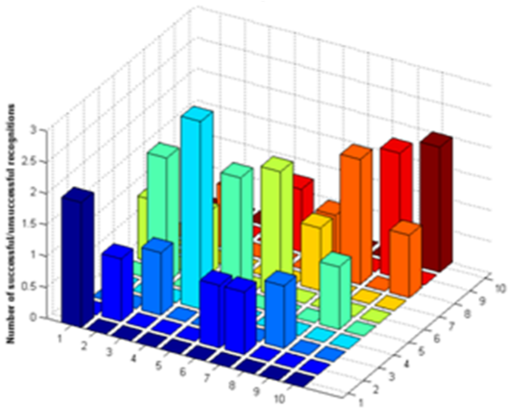


Fig. 2. Confusion Matrix Graph for First Ten Words

Table 2. Percentage Error for Words 001 to 010

S. No	Word Number	Value of C	% error
1	001	0.667	33.33%
2	002	0.333	66.67%
3	003	0.333	66.67%
4	004	1	0%
5	005	0.667	33.33%
6	006	0.667	33.33%
7	007	0.333	66.67%
8	008	0.667	33.33%
9	009	0.667	33.33%
10	010	0.667	33.33%

Table 3. Percentage Error for Words 031 to 040

S. No	Word Number	Value of C	% error
1	031	0	100%
2	032	1	0%
3	033	1	0%
4	034	1	0%
5	035	0.667	33.33%
6	036	0.333	66.67%
7	037	0.667	33.33%
8	038	0.667	33.33%
9	039	0.667	33.33%
10	040	1	0%

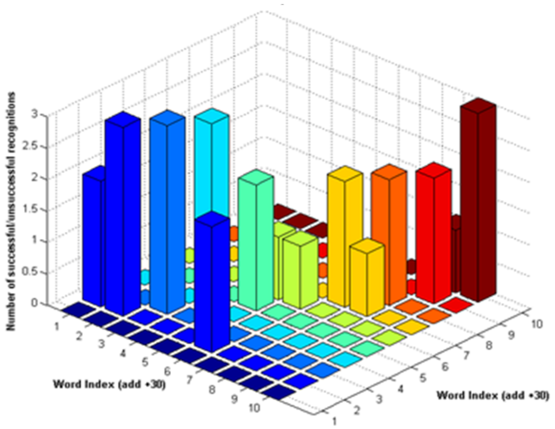


Fig. 3. Confusion Matrix Graph for Words 031 to 040

5.3 Overall Percentage Error

Fig. 4 shows the proportion of the words with 100 %, 66.67 %, 33.33 % and 0 % error, respectively. The analysis shows that the percentage error is either zero or 33.33 % for majority of the words. However, for few of the words, the value is larger approaching the maximum possible value i.e. 100 %. The overall error, E, can be measured as;

$$E = \frac{100\% \; of(10 \times 3) + 66.67\% \; of(13 \times 3) + 33.33\% \; of(32 \times 3) + 0\% \; of(45 \times 3)}{(100 \times 3)} \tag{7}$$

From this calculation, $E = 29.33\%$. This is comparable with so many existing ASR systems as developed for other languages with audio-only based features. This value, however, can be reduced further by increasing the amount of training data.

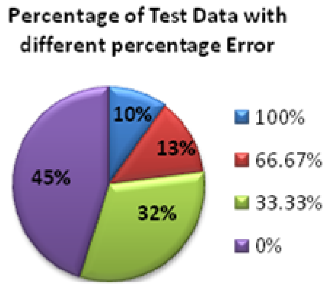


Fig. 4. Percentage of Test Data having different percentage error

5.4 Bad Words

The words having a 100 % error rate are referred to be the Bad Words. The primary reason for such a poor performance of the ASR system for these words, is the poor quality of recording which was determined through manual analysis of the audio files. Besides this, as each word has been divided into four segment, there is a possibility that more than one segment are matching exactly with segments of other words and the ASR framework is confused.

6 Future Work

This ASR system has been developed for speech recognition of isolated words only. This is a medium vocabulary application limited to a hundred words and can be extended to several thousand words. However, in that case, an even larger data for the training of the system will be required. Thus, there is need to increase the corpus size. This paper has provided a baseline for future research on ASR of Urdu language and can be extended to Continuous Speech Recognition of Urdu. This is an audio-only based feature extraction for ASR. The system can be evaluated by using audio-visual features which should result in the enhancement of the performance.

Acknowledgment

The authors are thankful to the supporting staff of the Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan. It is due to their efforts which they put into keeping the Computer Lab open for extra hours and providing the authors with opportunity to use it. The authors also extend their gratitude to Engr. Salman Ilahi, Department of Electrical Engineering and Engr. Irfan Ahmad, Department of Industrial Engineering, UET Peshawar, for their valuable input and suggestions throughout this work.

References

1. H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 26, no. 1, pp. 43-49, Feb 1978.
2. L. Gagnon, S. Foucher, F. Laliberte, and G. Boulianne, A simplified audiovisual fusion model with application to large-vocabulary recognition of French Canadian speech, Canadian Journal of Electrical and Computer Engineering, vol. 33, no. 2, pp. 109-119, Spring 2008.
3. S. Morii, K. Niyada, S. Fujii, and M. Hoshimi, Large vocabulary speaker-independent Japanese speech recognition system , in IEEE International Conference on Acoustics, Speech and Signal Processing, 1985, pp. 866-869.
4. Tohru Shimizu, Yutaka Ashikari, Eiichiro Sumita, and Jinsong Zhang, NICT/ATR Chinese-Japanese-English speech-to-speech translation system, Tshingua Science and Technology, vol. 13, no. 4, pp. 540-544, August 2008.

5. Mao Jiaju, Chen Qiulin, Gao Feng, Guo Rong, and Lu Ruzhan, SHTQS: A telephone-based Chinese spoken dialogue system, Journal of Systems Engineering and Electronics, vol. 16, no. 4, pp. 881-885, December 2005.
6. S. Khadivi and S. Ney, Integration of Speech Recognition and Machine Translation in Computer-Assisted Translation, IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 8, pp. 1551-1564, November 2008.
7. Wiqas Ghai and Navdeep Singh, Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages: Punjabi A Case Study, International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no. 1, pp. 379-385, March 2012.
8. M. U. Akram and M. Arif, Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling, in 8th International Multitopic Conference, 2004, pp. 91-96.
9. S. M. Azam, Z. A. Mansoor, M. Shahzad Mughal, and S. Mohsin, Urdu Spoken Digits Recognition Using Classified MFCC and Backpropgation Neural Network, in Computer Graphics, Imaging and Visualization, CGIV'07, 2007, pp. 414-418.
10. Abdul Ahad, Ahsan Fayyaz, and Tariq Mehmood, Speech recognition using multilayer perceptron, in Proceedings of IEEE Students Conference, ISCON'02, 2002, pp. 103-109.
11. S. K. Hasnain and M. S. Awan, Recognizing spoken Urdu numbers using fourier descriptor and neural networks with Matlab, in Second International Conference on Electrical Engineering, (ICEE 2008), 2008, pp. 1-6.
12. Javed Ashraf, Naveed Iqbal, Naveed Sarfraz Khattak, and Ather Mohsin Zaidi, Speaker Independent Urdu speech recognition using HMM, in The 7th International Conference on Informatics and Systems (INFOS 2010), March 2010, pp. 1-5.
13. Lawrence R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feburary 1989.
14. Huda Sarfraz et al., Large Vocabulary Continuous Speech Recognition for Urdu, in 8th International Conference on Frontiers of Information Technology, (FIT'10), 2010.
15. Hazrat Ali, Nasir Ahmad, Khawaja M. Yahya, and Omar Farooq, A Medium Vocabulary Urdu Isolated Words Balanced Corpus for Automatic Speech Recognition, in 2012 International Conference on Electronics Computer Technology (ICECT 2012), 2012, pp. 473-476.
16. Center for Language Engineering. (2012, May) [Online]. http://www.cle.org.pk/
17. Sirko Molau, Michael Ptiz, Ralf Schluter, and Herman Ney, Computing Mel-frequency cepstral coefficients on the power spectrum, in IEEE International Conference on Acoustics, Speech, and Signal Processing, ((ICASSP '01), 2001, pp. 73-76.
18. Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, and Kong-Pang Pun, An efficient MFCC extraction method in speech recognition, in IEEE International Symposium on Circuits and Systems, (ISCAS 2006), 2006.
19. Bojan Kotnik, Damjan Vlaj, and Bogomir Horvat, Efficient Noise Robust Feature Extraction Algorithms for Distributed Speech Recognition (DSR) Systems, International Journal of Speech Technology, vol. 6, no. 3, pp. 205-219, 2003.
20. John G. Proakis, and Dimitris G. Manolakis, Digital Signal Processing; Principles, Algorithms & Applications, 4th ed. Pearson Education, Inc, 2007.
21. Vinay K. Ingle and John G. Proakis, Digital Signal Processing Using Matlab, 3rd ed. Standford, USA: Cengage Learning, 2010.

22. David Salomon, Data Compression; The Complete Reference, 4th ed. London, United Kingdom: Springer, 2007.
23. S. Balakrishnama, A. Ganapathiraju, and J. Picone, Linear discriminant analysis for signal processing problems, in Proceedings of the IEEE Southeastcon, March 1999, pp. 36-39.
24. S. Balakrishnama and A. Ganapathiraju. (Accessed: 2012, March) Linear Discriminant Analysis; A Brief Tutorial. [Online]. http://www.music.mcgill.ca/ ich