Database Development and Automatic Speech Recognition of Isolated Pashto Spoken Digits Using MFCC and K-NN

Zakir Ali[1], Arbab Waseem Abbas[2], Thasleema T. M[3], Burhan uddin[1], Tanzeela[1]

[1] Institute of Business and Management Sciences

The University of Agricultural Peshawar

[2] Universities of Engineering and Technology Peshawar

Pakistan

[3] Dept. Computer Science, Central University of Kerala

{ali.musqan, aristocratarbab, burhan217, raaz_tann}@yahoo.com, thasnitm1@gmial.com

ABSTRACT: *Automatic recognition of isolated spoken digits is one of the most challenging tasks in the area of automatic speech recognition. In this paper, database development and automatic speech recognition of isolated Pashto spoken digits from sefer (0) to naha (9) has been presented. A number of 50 individual Pashto native speakers (25 male and 25 female) of different ages, ranging from 18 to 60 years, were involved to utter from sefer (0) to naha (9) digits separately. Sony PCM-M 10 linear recorder is used for recoding purpose in the office and home in noise free environment. Adobe audition version 1.0 is used to split the audio of digits into individual digits and result is saved in .wav format. Mel Frequency Cepstral Coefficients (MFCC) is used to extract speech features. K Nearest Neighbor (K-NN) classifier is used for the first time up to author knowledge in Pashto language to classify the features of speech. The experimental results are evaluated, and the overall average recognition accuracy of 76.8% is obtained.*

Keywords: KNN, MFCC, Pashto Digits

Received: 14 May 2014, Revised 26 June 2014, Accepted 2 July 2014

1. Introduction

Speech is a primary way of interaction for human beings. Human want to use this primary source of communication to make a convenient and user friendly interaction with computer. Machine oriented interfaces bound the computer usage because user must be computer and English literate. Speech recognition system which can recognize speech in native languages, enable layman to take benefits of information technology [1], and make interaction with systems simple and easier . In the field of computer science Automatic speech recognition is the ability of a machine to capture the sound waves of a speaker, convert sound waves into digital form, divided it into fundamental language phonemes, construct words from units and then analyze the words contextually, to check the spelling of the words according to the sound [2]. Many people have been done their research work on the Automatic Speech Recognition in Malayalam, Arabic, Tamil, Bangla languages[2], However it is impossible to built a generalize ASR system which is applied to all languages, because every language has its own stock of specific phonemes.

Therefore it is a need of time to built separate ASR system for local languages. To achieve the above goal, Abbas, Developed Pashto Spoken Digits database for automatic speech recognition using Mel Frequency Cepstral Coefficients (MFCC) for features extraction and Linear Discriminate Analysis (LDA) for features classification [3]. Sheena in [2] has presented an isolated digits recognition system in Malayalam language. In the first step, k nearest neighbor (KNN) was used to classify features, which was obtained through MFCC algorithm. In the next step Hidden Markov Model (HMM) was used to develop word modeling schema, which reduced 80% search time in the recognition process of first digit. In Arabic, Yousef Ajami has designed Arabic isolated digits recognition system for multi speaker and speaker independent. In this paper MFCC was used to remove noise and extract features from audio signals. An Artificial Neural Network was used to recognize the unknown digits from 0 to 9 [4]. In Bangla, Automatic speech recognition system was presented by Ghulam. In this research data from 0 to 9 was collected from bangle people. For features extraction MFCC algorithm and for recognition HMM classifier was used. The result was more than 95% for 0 to 5 digits, and for 6 to 9 digits the result was less than 90% [5]. Majid developed automatic speech recognition system for Malay digits. MFCC was used to extract feature vectors from speech data. After feature extraction, Hierarchical K mean algorithm was used for training and testing purpose and accuracy of system was 87.5% [6]. In Tamil, Karpagavalli designed Tamil isolated digits speaker independent recognizer. MFCC was used for features extraction and Linde Buzo Gray (LBG) victor quantization algorithm was used for code book generation for each digit. The recognition accuracy was 91% [7].

Pashto is the official language of Afghanistan and one of the local language of Pakistan [8], and it has about 50-60 million speakers throughout the world [9]. Pashto is the extended version of Arabic and Persian script and it has 43 consonants and 23 vowels sound out of 66 phonemes in Pashto script [10].

In this work, the rest of the paper is divided in the following sections. Section II describes the method for database development. In section III features extraction using MFCC algorithm is discussed. Sections IV explain recognition process using K-NN classifier and in section V experimental result are discussed. Finally in section VI conclusion and future work has been presented.

2. Pashto Digits Database Development

In Pashto digits database development section, 25 male and 25 female native Pashto speakers to record Pashto digits from Sefer (0) to naha (9) have been selected. Sony PCM-M recorder is used to record Pashto digits in the office and home in noise free environment. Speaker has been uttered digits from sefer (0) to naha (9) with a little gap. After recording, the recorded stuff was transferred to system through USB cable. Adobe audition version 1.0 is software used to edit audio signals by making sampling frequency of 16 kHz and quantized at 16 bits. The recorded Pashto digits files were split into isolated Pashto digits and also background information were also removed from speech signals and result was saved .wav format using this software. The splitting digits are saved in separate folders i-e the uttered sefer (0) from all speakers were placed in a separate folder. The same process repeated up to naha (9). And at the end 10 folders folders of splitting digits were obtained.

2.1 Pashto Isolated Digits ASR System

In the beginning of the 21st century, research work starts on Pashto recognition. Few experiments were carried out on small vocabulary of isolated word recognition, continuous speech recognition using different algorithms [5]. A small database is used for these experiments to perform recognition process. In this specified research a large size database of 50 speaker of 10 Pashto digits are used for testing and training purpose. Mel frequency Cepstral coefficients (MFCC) were used to extract features and K nearest neighbor (K-NN) based classifier was used for the recognition which is discussed below.

3. Features Extraction Using MFCC

Feature extraction means to point out the parts of the audio signal that are better for identifying the linguistic content and removing all other parts which carries background information like noise, emotion [11]. In this desired research in speech signals related studies the most important parametric representation technique Mel Frequency Cepstral Coefficients (MFCC) is used [12].

Here MFCC is used to extract features from speech signals. MFCC use a number of Mel filter bank according to the Mel scale which are linearly spaced up to 1kHz and logarithmically above 1kHz to smooth and capture the linguistic characteristics of the speech signal spectrum [2]. In signal analysis phase the isolated speech signals are provided as input to MFCC algorithm to perform the following operations.

1. Pre-emphasis the speech signal by using a below formula.

$$Y[n] = X[n] - a[n-1].$$

Where $Y[n]$ is the output signal, $X[n]$ is the input signal and the value of $a = 0.97$ [11].

2. Then the pre-emphasized signal was separated into small block called segmentation and the length of each frame was 256 samples i-e [(256/16000) * 1000] = 16 ms.

3. Windowing frame to minimize the signal discontinuities at the start and end of the each frame. For this propose hamming window was used to reduce the edge effect. Hamming window is describe through below equation.

$$Y(n) = X(n) * W(n) \text{ for } 0 \le n \le N-1$$

Where N = number of sample in each frame and $W(n) = 0.54 - 0.46\cos(2\pi / N - 1)$ [12].

$$y[n] = \sum_{n=-\infty}^{\infty} x(n)\, e^{-iwn}$$

4. Fast Fourier transform (FFT) was applied to each frame to convert each frame of N sample from time domain to frequency domain. equation of the fast Fourier transform is

$$mel(f) = 2595 \times log\, 10\,(1 + f1700)$$

5. Mel filter bank is applied to calculate the average energy in each block and take the logarithm of all filter bank energies by using a formula below [13]

6. Discrete cosine transform (DCT) convert log filter bank energies into time domain. The process of conversion is called Mel frequency cepstrum coefficients. Keep DCT coefficient 2-16, which are used for k-NN classifier. Block diagram of MFCC shown in below figure 1.

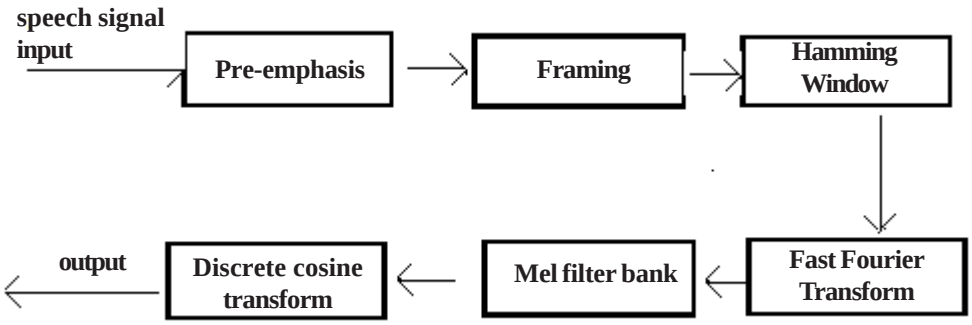

Figure 1. Block diagram of MFCC

Parameters	Values	Parameters	Values
Speech signal	16000Hz	Cepstral coefficient	16
Linear Filters	12	Window size	128
Log Spacing	100	Fast Fourier transform	512
Log filters	13	Sampling rate	16000Hz

Table 1. Parameters list of MFCC

The list of parameter which are used in MFCC to extract features from speech signals are listed in table 1. Eleven txt files were obtained, which contained the numerical values (features) of speech data after applying the MFCC algorithm on speech data. Table 2 also contains the parameter of MFCC algorithm but using these parameters to extract feature from speech signal have not give good result during classification [14].

Parameters	Values	Parameters	Values
Speech signal	16000 Hz	Cepstral coefficient	12
Linear Filters	11	Window Size	128
Log Spacing	100	Fast Fourier Transsform	512
Log Filters	12	Sampling rate	8000 Hz

Table 2. Parameters list MFCC

4. Pashto Speech Recognition Using K-NN Classifier

Classification is third step of Pashto automatic speech recognition. K-Nearest Neighbor algorithm (k-NN) is used for supervised learning that has been used in many areas such as in the field of statistical pattern recognition, data mining, image processing and many others areas [15].

K-NN classifier categorize the unknown sample s to a predefine class based on previously classified samples (training data). K-NN is useful for such a data that changes or updates rapidly [15]. K-NN classifier was used for the first time up to the author knowledge in Pashto language to classify the feature of speech. K-NN classifier works in the following steps.

4.1 KNN Algorithms

Step 1: provide features of speech signals to K-NN classifier for classification to train the system

Step 2: Measure the distance by using in [15] the Euclidean Distance formula shown in figure 2, between the new observation s and training data.

$$Euclidean\ Distance,\ D\ (x_s . y_s) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

Step 3: Sort the Euclidean distance values as $d_i \leq d_i + 1$, select k smallest samples.

Step 4: Apply voting or means according to the application.

The above K-NN algorithm can be explain below through a simple general example

x	y	Labels
2	3	2
7	8	2
5	7	1
5	5	2
4	4	1
1	8	1

Table 3. Training dataset for classification

For example we have training data with class label $C = \{1, 2\}$ listed above in table 3. The new observation $S = \{3, 4\}$ need to be classify, which is show in fig. the k-NN algorithm indicate the nearest neighbor with $k = 3$ and $k = 5$ according to the new observation s. Distance can be calculated through Euclidean distance formula and the calculation process is shown in table 3.

K-NN used voting method where k value equal to 3, the number of 2 labels > 1 labels that's why the new observation S will be classified as 2. If k value changes from 3 to 5 then the number of 1 label > 2 labels so the new observation will be classified as 1.The voting process is highlighted in figure 2.

x	y	Euclidean distance formula	Distance	Labels
2	3	$\sqrt{(3-2)^2+(4-3)^2}$	1.41	2
7	8	$\sqrt{(3-7)^2+(4-8)^2}$	5.66	2
5	7	$\sqrt{(3-5)^2+(4-7)^2}$	3.61	1
5	5	$\sqrt{(3-5)^2+(4-5)^2}$	2.24	2
4	4	$\sqrt{(3-4)^2+(4-4)^2}$	1.00	1
1	8	$\sqrt{(3-1)^2+(4-8)^2}$	4.47	1

Table 4. Distance calculation among training data and new observation using Euclidean distance formula

Serial no	Distance	Label
5	1.00	1
1	1.41	2
4	2.24	4
3	3.61	1
6	4.47	1
2	5.66	2

Table 5. Sorted list of distance values

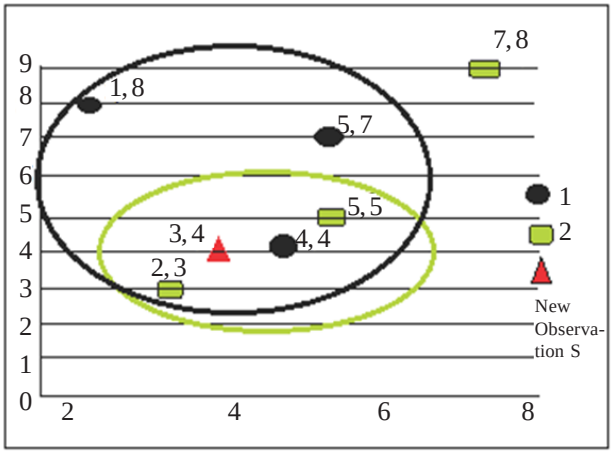

Figure 2. Voting based method in K-NN algorithm

V. Experimental results

In this experiment, a system that recognizes Pashto isolated digits from sefer (0) to naha (9) is implemented. First database was developed by collecting speech data from 50 native Pashto speakers then MFCC algorithm was used to extract features from speech data and as a result 10 text files, which contain features of speech data, were obtained. In third step, K-NN classifier was used to classify the features vectors of Pashto isolated speech signals. For this purpose features vectors are divided into two parts for training and testing, 25 speaker's (Male and Female) data were provided to K-NN classifier to train the system

and next 25 speaker's (Male and Female) data were used to test the Pashto ASR system and check the average accuracy of the system. Both MFCC and K-NN algorithms were run in MATLAB environment. To extract features using MFCC with table 1 parameters then the K-NN classifier take less processing time to classify the data but the recognition accuracy are not too good while MFCC using table 2 parameter for features extraction, so the average accuracyof Pashto ASR system is good but take much time during processing. The confusion Matrix of test data of Pashto ASR system shown in fig. 3 is evaluated and the overall average recognition accuracy of 76.8% is obtained.

	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	
Zero	20	1	0	1	2	0	1	0	0	0	
One	0	21	1	1	0	0	0	2	0	0	
Two	0	5	15	0	0	0	0	0	2	4	
Three	0	1	0	21	22	0	0	0	1	0	
Four	1	4	0	2	14	2	0	0	0	3	
Five	2	0	3	0	0	17	0	0	3	0	
Six	2	0	0	0	1	2	10	0	0	1	
Seven	1	6	2	1	1	0	0	14	0	0	
Eight	1	1	0	1	1	6	1	0	13	1	
Nine	0	0	3	2	0	0	1	1	0	18	
Average Accuracy											76.8%

Figure 3. Confusion matrix of Test data

6. Conclusion and Future work

The experiment has been successfully performed to design speaker independent, large vocabulary, Pashto isolated digit recognition system using MFCC and K–NN algorithm and simulated in MATLAB environment. The average recognition accuracy was good for isolated Pashto digits.

In future, the experiment can be performed using other different algorithm like Support Vector Machine, Victor Quantization and K-means algorithms to achieve better result. The work can be extended for isolated numbers and continuous speech recognition instead of isolated digits.

References

[1] Shah, F. (2010). Isolated Malayalam digit recogntion using Support Vector Machines. *In*: 2010 International Conference on Communication Control and Computing Technologies. p. 692-695.

[2] Sheena, C. V., Thasleema, T. M., Narayanan, N. K. Search Time Reduction Using Hidden Markov Models for Isolated Digit Recognition .

[3] Abbas, A. W., Ahmad, N., Ali, H. (2012, September). Pashto Spoken Digits database for the automatic speech recognition research. *In*: Automation and Computing (ICAC), 2012 18th International Conference on (p. 1-5). IEEE.

[4] Alotaibi, Y. A. (2003). High performance Arabic digits recognizer using neural networks. *In:* Neural Networks, 2003. Proceedings of the International Joint Conference on (V. 1, p. 670-674). IEEE.

[5] Muhammad, G., Alotaibi, Y. A., Huda, M. N. (2009). Automatic speech recognition for Bangla digits. In *Computers and Information Technology,* 2009. *ICCIT*'09. 12th *International Conference on* (p. 379-383). IEEE.

[6] Majeed, S. A., Husain, H., Samad, S. A., Hussain, A. (2012). Hierarchical K-Means Algorithm Applied On Isolated Malay Digit Speech Recognition. *International Proceedings of Computer Science & Information Technology*, 34.

[7] Karpagavalli, S., Rani, K. U., Deepika, R., Kokila, P. (2012). Isolated Tamil Digits Speech Recognition using Vector Quantization. *International Journal of Engineering*, 1 (4).

[8] Prasad, R., Tsakalidis, S., Bulyko, I., Kao, C. L., Natarajan, P. (2010, March). Pashto speech recognition with limited pronunciation lexicon. In: *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* (p. 5086-5089). IEEE.

[9] Abdur, S., Abid, R., Ahmad, N., Khan, M. A. A., Zuhra, F. T. (2013). Concatenative based Pashto Digits and Numbers Synthesizer. *International Journal of Computer Applications*, 72.

[10] Halpern, J. (2007). The challenges and pitfalls of Arabic romanization and arabization. *In:* Proc.Workshop on Comp. Approaches to Arabic Scriptbased Lang.

[11] Meseguer, N.Alcaraz (2009). Speech Analysis for Automatic Speech Recognition.

[12] Muda, L., Begam, M., Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*.

[13] Poonkuzhali, C., Karthiprakash, R., Valarmathy, S., Kalamani, M. An Approach to feature selection algorithm based on ant colony optimization for automatic speech recognition.

[14] Pei, J. I. A. (2010). Automatic Speech Recognition.

[15] Jan, Z., Abrar, M., Bashir, S., Mirza, A. M. (2009). Seasonal to inter-annual climate prediction using data mining KNN technique. In *Wireless Networks, Information Processing and Systems* (p. 40-51). Springer Berlin Heidelberg.