Proceedings of the 18th International Conference on Automation & Computing, Loughborough University, Leicestershire, UK, 8 September 2012

Pashto Spoken Digits Database for the Automatic Speech Recognition Research

[1]Arbab Waseem Abbas, [1]Nasir Ahmad, [2]Hazrat Ali

1 Department of Computer Systems Engineering
2 Department of Electrical & Electronics Engineering
University of Engineering and Technology
Peshawar, Pakistan
aristocratarbab@yahoo.com, n.ahmad@nwfpuet.edu.pk

Abstract--**This paper presents the development of a Pashto Spoken Digits database for the automatic speech recognition research. This is, to the best of author's knowledge, the first Pashto isolated digits database. The database consists of Pashto digits from zero (sefer) to hundred (sul) uttered by sixty speakers, 30 male and 30 female. The speakers included are having ages, ranging from 18 to 60 years. The recordings are performed in a noise-free environment using Sony PCM-M 10 Linear Recorder. The audio is stored in .wav format and transferred to the laptop via an usb cable. After editing, the audio is split into individual digits using Adobe Audition ver. 1.0. The isolated digit recognition experiments are then performed on a subset of the database containing first 11 digits and 18 speakers. Mel Frequency Cepstral Coefficients (MFCC) were used as the feature vector while Linear Discriminent Analysis (LDA) based classifier was used for the classification.**

Keywords-Pashto speech recognition; Pashto acoustic database and Pashto islotated word database.

I. INTRODUCTION

The world is moving towards Computerization and there is an increased interaction of humans with the machines in their daily lives, so there is a strong need to make this human-computer interaction more natural and pervasive [1]. To achieve this goal, the development of listening [2], speaking [3] and viewing [4], [5] machine has became very hot areas of research. The area of research, conducting work on listening machine is popularly known as Automatic Speech Recognition (ASR). Obtaining such capabilities in local languages is a next stage of this technological advancement so that everyone can communicate with the computers in a natural way, and a number of research works on the ASR of local languages has been reported in the recent years. Hejazi, in [2] has presented an isolated digits recognition system in Persian languages. He has used Hidden Markov Model (HMM) to decompose a word into small parts for solving the problem arising from the pronunciation similarity of some Persian digits that are composed of very similar phonetic and spectral components. For recognition a Support Vector Machine SVM is used to segment the input word and to find an entry with the maximum number of similar segments. In Arabic, an automatic recognition of spoken digits was described by Yousef Ajami [6]. In this paper, the system was designed to recognize an isolated whole-word speech of the ten digits, zero through nine in Arabic. The Hidden Markov Model Toolkit (HTK) was used to implement the isolated word recognizer with phoneme based HMM models. In the training and testing phase of this system, isolated digits data sets are taken from the telephony Arabic speech database, SAAVB. This standard database was developed by KACST and it is classified as a noisy speech database. A hidden Markov model based speech recognition system was designed and tested with automatic Arabic digits recognition. In Urdu language, corpus development was discussed by Huda Sarfraz [7] while the development of isolated word database is presented in [8]. In [7], development of an Urdu speech database for speaker independent spontaneous speech recognition has been presented containing data from 82 speakers (42 male and 40 female), recorded over a microphone and a telephone line. The speech was collected from speakers ranging from 20 to 55 years of age. Recording sessions were conducted in office and home environments. In Indian languages, speech databases in Tamil, Telugu and Marathi were elaborated by Anumanchipalli [9]. In this work speech data was collected from about 560 speakers in these three languages. Preliminary speech recognition results using the acoustic models created on Sphinx 2 speech recognition toolkit, has also been presented. Speech translation system for American English and Pashto was developed by Andreas Kathol [10]. In this paper, issues encountered by Pashto in the areas of written representation, corpus creation, speech recognition, speech synthesis, and grammar development for translation have been discussed.

General rules for corpus development are discussed by Sue Atkins [11]. This paper, explains the principal aspects of corpus creation and the major decisions to be made when creating an electronic text corpus, basic standards and features for corpus design in the initial stages of corpus building.

For the research on natural language processing of any language, the first step is the development of a standard language resource, both for training and a basic platform for performance evaluation. In the ASR research, isolated word and digit recognition tasks are generally the first stages posing a relatively simple and well defined tasks and a baseline for onward research [2], [6], [8]. This paper presents the development of an isolated digits database in Pashto language. Initial ASR experiments have been carried out on the database using Mel-Frequency Cepstral Coefficients (MFCC) based features and Linear Discriminant Analysis (LDA) classifier. The contents of the database, recording setup and the results of initial experiments on the database are discussed in details in the following sub-sections.

II. PASHTO DIGITS DATABASE

Pashto, the national language of Afghanistan and one of the major languages of Pakistan, has an estimated 50-60 million speakers all over the world [12]. Pashto is divided into three dialects, Northern Pashto, Central Pashto and Southern Pashto [13]. Northern Pashto is spoken in most of Khyber Pakhtoonkhwa and its adjoining areas in Afghanistan, the Provinces of Kunar and Nangarhar. It is also most common dialect among the Pashtun Diaspora mostly in Canada, India, UAE, UK and the USA. Its alternate names are Pakhto, Pukhto and Yusufzai Pukhto. Pashto central is spoken in Wazirstan, Bannu and Karak region of the Khyber Pakhtoonkhwa province of Pakistan. Its alternate names are Waciri (Waziri) and Bannuchi (Bannochi, Bannu). Southern Pashto is spoken in Quetta area of Balochistan province and also in Afghanistan, Iran, Tajikistan, United Arab Emirates, and United Kingdom. In this paper Northern Pashto (Yusufzai dialect) has been used.

A. *Database content*

The first step in database development is the selection of content. A database for the general purpose research should contain all the sounds in the language called phonemes.

A phoneme is the smallest meaningful unit of sound utterance [14]. Although no standard set of Pashto phonemes has yet been identified, in [15], 41 phonemes in Pashto have been reported including 35 consonants and 6 vowels phonemes. For a database for ASR, all phonemes of the language need to be included along with maintaining a statistical balance between their occurrences. However, for the digit recognition application the set of words is fixed i.e. digit which may or may not contain all the phonemes of the language.

TABLE I. PASHTO DIGITS PATTERN

Digits	Pashto (in figure)	Pashto (words)	Pronunciation
0	٠	صفر	Sefer
1	١	يو	Yaw
2	٢	دؤه	Dwa
3	٣	درے	Dray
4	٤	څلور	Celour
5	٥	پنځه	Penza
6	٦	شپږ	Shpeg
7	٧	أوؤه	Owa
8	٨	أته	Ata
9	٩	نهه	Naha
10	١٠	لس	Las
11	١١	يو لس	Yawo- Las
12	١٢	دؤه لس	do- Las
13	١٣	دیأر لس	Dyar- Las
14	١٤	څوأر لس	Swaar- Las
15	١٥	پنځه لس	Penza- Las
16	١٦	شپاړس	Shparh-as
17	١٧	أوؤه لس	Owa- Las
18	١٨	أته لس	Ata- Las
19	١٩	نو لس	Noo- Las
20	٢٠	شل	Shul
21	٢١	يویشت	Yaw-Vesht
22	٢٢	دؤه ویشت	Dwa—Vesht
23	٢٣	درے ویشت	Dray-Vesht
24	٢٤	څلیر ویشت	Celour-Vesht
25	٢٥	پنځه ویشت	Penza-Vesht
26	٢٦	شپږ ویشت	Shpag-Vesht
27	٢٧	أوؤه ویشت	Owa-Vesht
28	٢٨	أته ویشت	Ata-Vesht
29	٢٩	نهه ویشت	Naha-Vesht
30	٣٠	دیرش	Dairsh
40	٤٠	څلویښت	Celwaikht
50	٥٠	پنځوس	Panzoos
60	٦٠	شپیته	Shpeta
70	٧٠	اویه	Away
80	٨٠	اتیه	Atya
90	٩٠	نوی	Nawi
100	١٠٠	سل	Sul

The pattern of Pashto digits from zero to hundred is shown in Table 1.

In the number system of Pashto, like most other languages such as Persian [2], Arabic [6] and other languages [16], each digit from zero (sefer) to ten (las) has different name. The numbers following las, yawo-las(11) to noo-las(19) contains common post-fix las with each digit from yaw(one) to naha(nine) with slight modification, while from yaw-vesht(21) to naha-vesht(29) the common post-fix is vesht with digits from yaw(1) to naha(9) as a pre-fix. Same pattern is followed for all digits till sul(100). There are some variations of this pattern among different regions/dialects. The numbers included here are those of the Yousafzai dialect.

III. DATABASE DEVELOPMENT

For the digits database the content is naturally fixed. As Pashto speakers also use other number, such as Urdu and English in their daily lives, to make it easy for their recording, numbers from sefer (zero) to sul (100) were written on a sheet using Liwal Pashto/Dari support software [17].

A. *Speaker selection*

For the purpose of recording, the speaker who could fluently & accurately pronounce the written digits in Pashto was selected. A page with Pashto digits written on it was given to each speaker for reading and was asked to read. The pronunciation of the speaker was checked whether his/her the reading was correct and that the pronunciation was in the Yousafzai dialect.

B. *Recording environment*

The recording environment is an important aspect of the database such as office environment and in car etc. In this database the recording was performed in an office environment either in faculty room or in the library of the University. Further, the signal to noise ratio was controlled by adjusting the recorder sensitivity.

C. *Recording hardware*

The recording of speakers has been done using Sony PCM-M 10 Linear Recorder and the recorded data was then transferred to Sony vaio laptop via usb port for editing and splitting. The digits which were either wrongly pronounced or with noise were re-recorded. This was done either by pausing during the recording process and re-recorded or by playing back the recorded file after complete recording and re-recording in case of error in recording. The Adobe audition software ver. 1.0 was used for editing and splitting.

D. *Recording setup*

Before starting recording the sensitivity level of recorder was adjusted by uttering a few test digits and fixing the signal to noise ratio. The actual recording was performed after adjusting sensitivity level of recorder and the speaker would say his/her name and then utter the digit from 0 to 100 in Pashto with little gaps in between each digit to make clear distinction between digits.

To develop a balanced digits database, recording had done from equal number of male & female speakers.

E. *Editing software*

Adobe Audition ver. 1.0 software is used for editing & splitting of recorded digits. Recorded audio file were opened in Adobe Audition ver. 1.0 and edited/splitted, in zoom-in mode for easy splitting/editing. The utterance of the specified length was played and checked for the spoken word and created new word files.

F. *Splitting/editing of recorded digits*

The original recording data is provided in the folder named ‘recording’. The recorded data in the .wav format is transferred from recorder to laptop via usb cable for splitting/editing of connected digits of each speaker into isolated digits. Adobe Audition ver. 1.0 software was used for the purpose of editing and splitting.

The wave files from the recording folder were zoomed in and each recorded digit was extracted. To make the splitting easier, the speakers were asked to utter each digit with little delay. New files with sample rate of 16 kHz, channels mono and 16 bit resolution were created by copying and pasting the contents for each uttered word. The first word for all speakers is the speaker name and is saved in a unique format specifying the speaker number such as 1sp meaning 1st speaker. The remaining words are the digits and are saved in same the format but with ending characters specifying the digit uttered, for example 1sp0 the 1 in the beginning means speaker one while the 0 at the end means the utterance of zero by the first speaker. In the same way all the digits from the same speaker are stored in the same folder from zero till 100. Same procedure was adopted for all the speakers in the database. Thus each folder name uniquely specifies the speaker and each file in the folder specifies the uttered word. This makes it easy to use these files in different grouping for the purpose of training and testing. Also it was also tried to keep the length of each digit the same as is needed by many feature extraction/classification techniques.

Although each folder also contains a file with the speaker’s names these have been removed in the final database which is created for the distribution purpose.

Besides the arrangement based on the speaker, the database has been rearranged on the base of digit number carrying the same digit utterances by all the speakers.

IV. INITIAL ASR EXPERIMENTS

Initial Automatic Speech Recognition (ASR) experiments were performed on a subset of the database containing the first eleven digits from sefer (0) to las (10). Mel frequency Cepstral coefficients (MFCC) were used as the feature vector and Linear Discriminant Analysis (LDA) based classifier was used for the recognition.

A. *Mel-Frequency Cepstral Coefficients (MFCC)*

Mel-Frequency Cepstral Coefficients (MFCCs) are most commonly used features in the speech recognition [8], and the speaker verification application [18]. The common procedure for MFCCs derivation as used in [19] and [20] is as follows. The Fourier transform of (a windowed excerpt of) the speech signal is taken. The powers spectrum obtained above is map onto the mel scale, using triangular overlapping windows. The MFCCs are then obtained by taking the discrete cosine transform (DCT) of the logs of the powers at each of the mel frequencies.

B. *Linear Discriminant Analysis (LDA)*

Linear Discriminant Analysis (LDA) is a method to discriminate between two or more groups of samples. LDA transforms the data such as to maximize the inter-classes variance while minimizing the intra-class variance [21]. LDA based classifier utilizes the supervised classification approach to classify unknown samples using the known training data and their class labels [22], so it is the best option for isolated digits recognition.

C. *Training and test data*

The subset of the database used in these experiments consists of digits from sefer(1) to las(10) uttered by 18 speakers. Two-third of the speakers (12 speakers) were used for the training while the remaining one-third (6 speakers) were used for testing. MFCC algorithm with the parameters shown in Table 2 was used to extract 52 dimensional feature vectors both from the training and test samples.

TABLE II. MFCC PARAMETERS

Parameters	Value	Parameters	Value
Lowest Frequency	0	Fft Size	512
Linear Filters	11	Cepstral Coefficients	13
Linear Spacing	100	Window Size	128
Log Filters	13	Sampling Rate	8khz
Log Spacing	1.148	Frame Rate	62.5

V. RESULTS & DISCUSSION

For the initial ASR experiments reported here, each Pashto digit in test data is classified using the LDA classifier. The results are obtained both on the basis of percentage of correct word classification as well as the confusion matrix. Moreover the results are obtained for the classification of both the training data and a different set of test data. The percentage of word correct obtained for the test data was up to 70% while that for the training data is up to 93%. The performance of the ASR for the training and testing is shown in Table 3 and Table 4, respectively.

TABLE III. CONFUSION MATRIX FOR TRAINING DATA (12 SPEAKERS)

	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten	
Zero	11	1	0	0	0	0	0	0	0	0	0	
One	0	12	0	0	0	0	0	0	0	0	0	
Two	0	1	11	0	0	0	0	0	0	0	0	
Three	0	0	0	11	0	0	1	0	0	0	0	
Four	0	0	0	0	12	0	0	0	0	0	0	
Five	0	0	0	0	0	12	0	0	0	0	0	
Six	0	0	0	0	0	0	12	0	0	0	0	
Seven	0	2	0	0	0	1	0	9	0	0	0	
Eight	0	0	0	1	0	0	0	0	11	0	0	
Nine	0	0	0	0	1	0	0	0	0	11	0	
Ten	0	0	0	1	0	0	0	0	0	0	11	
Per_corr	0.91	1.0	0.91	0.91	1.0	1.0	1.0	0.75	0.91	0.91	0.91	
Average												93%

TABLE IV. CONFUSION MATRIX FOR TEST DATA (06 SPEAKERS)

	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten	
Zero	5	0	0	0	1	0	0	0	0	0	0	
One	0	3	0	1	0	0	0	1	0	1	0	
Two	0	0	3	0	0	0	1	1	0	1	0	
Three	0	0	0	6	0	0	0	0	0	0	0	
Four	2	1	0	0	2	0	1	0	0	0	0	
Five	0	0	0	0	0	5	1	0	0	0	0	
Six	0	0	0	0	0	0	6	0	0	0	0	
Seven	0	2	0	0	0	0	0	4	0	0	0	
Eight	0	1	0	0	0	2	0	1	2	0	0	
Nine	0	0	0	1	0	0	0	2	0	3	0	
Ten	0	0	0	0	0	0	1	0	0	0	5	
Per_corr	0.83	0.5	0.5	1.0	0.33	0.83	1.0	0.66	0.33	0.5	0.83	
Average												67%

The recognition results for training and test set are quite different, most probably due to the small amount of the training data.

VI. CONCLUSION

In this research the development of an isolated Pashto spoken digits database has been presented. The database is suitable for the research on automatic speech recognition of digits Pashto. The database will be made available to the researcher working in automatic speech recognition in Pashto language.

REFERENCES

[1] E. Shriberg (2005), "Spontaneous speech: how people really talk and why engineers should care", *procedings of the conference Interspeech*, Lisbon, Portugal, pp. 1781-1784.

[2] Hejazi, S.A., Kazemi, R., Ghaemmaghami, and S.,Sharif (2009), "Isolated Persian digit recognition using a hybrid HMM-SVM", *proceeding of the International Symposium on Intelligent Signal Processing and Communications Systems, (ISPACS 2008)*, Bangkok,Thailand pp.1-4.

[3] B. P. Douglas and M. B. Janet (1992), "The Design for the Wall Street Journal-based CSR Database", *Proceedings of the DARPA SLS Workshop,* Association for Computational Linguistics Stroudsburg, PA, USA

[4] O. Matan, J. C. B. Christopher, Y. L. Cu and S. D. John, (1992) "Multi Digit Recognition using A Space Displacement Neural Network", *procedingss of conference of NIPS*, vol. 4, Denver, Colorado, USA, pp. 488-495.

[5] M. Miciak (2008), "Character Recognition Using Radon Transformation and Principal Component Analysis in Postal Applications", *procedings of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008*, Wisla, Poland, pp. 495 – 500.

[6] Y. A Alotaibi, M. Alghamdi and F. Alotaiby (2010), "Speech Recognition System of Arabic Digits based on A Telephony Arabic Corpus", *procedings of the ICISP 2010*, Canada.

[7] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed and R. Parveen (2010), "Speech Corpus Development for a Speaker Independent Spontaneous Urdu Recognition System", *Proceedings of O-COCOSDA*, Kathmandu, Nepal.

[8] H. Ali, N. Ahmad, K. M. Yahya and O. Farooq (2012), "A Balanced Urdu Isolated Words Corpus for Automatic Speech Recognition", *proceedings of 4th International Conference on Electronics Computer Technology - ICECT 2012*, Kanyakumari, India.

[9] G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S. P. Singh and R. N. V. Sitaram (2005), "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems" *Proceedings of International Conference on Speech and Computer (SPECOM),* Patras, Greece.

[10] A. Kathol, K. Precoda, D. Vergyri, W. Wang and S. Riehemann (2006), "Speech Translation for Low-Resource Languages: The Case of Pashto", *procedings of conference. INTERSPEECH-2005*, Lisbon, Portugal, l2273-2276.

[11] S. Atkins, J. Clear and N. Ostler (1991), "Corpus Design Criteria", Oxford University Press UK.

[12] Herbert and I. Sloan (2009), "A Grammar of Pashto a Descriptive Study of the Dialect of Kandahar, Afghanistan", Ishi Press Intl. pp. 210.

[13] L. M. Paul (ed.), (2009), "Ethnologue: Languages of the World", (16th eddition), Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

[14] 21 Century Eloquence: Voice Recognition Glossary.

[15] H. U Naeem (2006), "New puxto primer (The 21st century updated & augmented puxto phonetic alphabet".

[16] http://en.wikipedia.org/wiki/List_of_numbers_in_various_languages.

[17] Liwal, available online on http://www.liwal.com/translate/pashto.htm

[18] T. Ganchev, N. Fakotakis, and G. Kokkinakis (2005), "Comparative evaluation of various MFCC implementations on the speaker verification task", *proceedings of SPECOM 2005*, vol.1, pp.191–194.

[19] M. Xu, L. Duan, J. Cai, L. Chia, C. Xu, and Q. Tian (2004), "HMM-based audio keyword generation", *Proc. of 5th Pacific Rim Conf. on Multimedia,* vol. 3333, pp. 566–574

[20] Sahidullah Md., G. Saha(2012), "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition", *published in Speech Communication journal*, vol. 54, no. 4, Elsevier Science Publishers, The Netherlands, pp. 543-565.

[21] Balakrishnama, S. and Ganapathiraju, A. (1998), "Linear Discriminant Analysis - A Brief Tutorial", Institute for Signal and Information Processing, http://www.isip.msstate.edu/publications/reports/isip_internal/1998/linear_discrim_analysis/.

[22] H. Lohninger (1999), "Teach Me Data Analysis", Springer-Verlag, Berlin-NewYork-Tokyo,ISBN3-540-14743-8. http://www.vias.org/tmdatanaleng/cc_lda_intro.htm.