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

The Development of Isolated Words Pashto Automatic Speech Recognition System

Irfan Ahmed[1], Nasir Ahmad[2], Hazrat Ali[1], Gulzar Ahmad[1]
[1]Department of Electrical Engineering, University of Engineering and Technology
Peshawar, Pakistan
[2]Department of Computer System Engineering, University of Engineering and Technology,
Peshawar, Pakistan
Email: irfanahmed@nwfpuet.edu.pk, n.ahmad@nwfpuet.edu.pk, hazrat.ali@nwfpuet.edu.pk, gulzar@nwfpuet.edu.pk

Abstract— **The availability of standard speech database is of paramount importance in the automatic speech recognition (ASR) research in the context of providing a baseline for comparing the performance of automatic speech recognition approaches. This paper presents the development of a Medium-Vocabulary Speech Corpus for Pashto language and development of Pashto ASR system by using the corpus. The vocabulary encompasses 161 isolated words of Pashto language, consisting of most frequently used words of Pashto language, names of the days of the week and digits from 0 to 25. The words were uttered by 50 speakers of different ages and genders, including both native and non-native speakers of Pashto language. Recording of the corpus was performed in a noise free office environment. The Corpus developed is then used for the development of an automatic speech recognition system for Pashto language.**

Keywords- Automatic Speech Recognition; Pashto Speech Corpus; Human Computer Interaction.

I. INTRODUCTION

Pashto is one of the major languages of Asia with almost 50-60 million speakers around the world [1]. It is one of the two official languages of Afghanistan and spoken and understood in the western and north-western regions of Pakistan. Pashto alphabetic list is customized form that of Arabic alphabet [2]. The major population of Pakistan having Pashto as the first language is in the provinces of Khyber Pakhtunkhwa (KPK) and Baluchistan. The literacy rate of KPK is almost 38% and in rural areas where Pashto is the major spoken language, it is about 33.94% [3]. Thus the majority population of KPK is unable to interact with modern machines having an English interface. Therefore this research on ASR system aims at the development of a human machine interaction (HMI) resource for Pashto speakers. To the best of author's knowledge, no work has yet been done for the development of Pashto ASR and the primary reason behind this is because of the unavailability of standard corpus, which may be used to build, train and evaluate a generic ASR system for Pashto language. The primary aim of a balanced speech corpus development is to provide some customary speech database for recognizer training and testing with respect to some genuine phonetically balanced catalog. On the other hand, the availability of balanced corpus relieves the researchers of linguistic processing, in the sense of avoiding the strenuous task of recording the utterances of diverse speakers.

Speech corpus plays an important role in ASR system development. A standard corpus should consist of phonetically compact and randomly selected words to provide ease for later development. An initial work towards the development of a balanced speech database was done by MIT by providing TIMIT for English language [4]. In the same way, speech corpus in many other languages of the world has been developed such as BREF for French [5], ATR for Japanese [6], Spanish [7], Thai [8] and Bengali [9] etc. To develop an ASR system for Pashto language, it is indispensible to have an acoustically rich, phonetically balanced speech corpus. This research work aims to develop a medium-vocabulary isolated-word speech corpus for Pashto with a set of words used in most common applications and an ASR system development for Pashto. The corpus developed will be made available for future research on Pashto ASR system and other linguistic studies regarding Pashto language.

II. TEXT CORPUS FOR SPEECH DATABASE

In order to have phonetically rich words list, the selection of some optimal sources of words is of critical importance. This collection can be acquired on the basis of numerous conditions such as general purpose database or words assortment for some specific pragmatics such as security purpose, command and control purpose and so on. General purpose word-database might be developed on the basis of most frequently used words in the language, names of week days, names of seasons, months' names etc as has been done in Urdu [12]. The later approach which has also been adopted in this work is the most appropriate one for the development of a generic ASR system. The development of the general purpose Pashto database carried out in this work is explained in the preceding sections.

A. Words Selection Criteria

A balanced corpus should consist of phonetically compact and randomly selected words to provide ease for later development. Concerning the selection of phonetically balanced Pashto words, the Pashto Academy's primer of Pashto by Dr. Hidayt Ullah Naeem [10], has been consulted.

The words list consists of Pashto nouns and adjectives making part of the daily conversation. In addition, seasons' names, names of the days of the week and digits from 0 to 25 have also been included to widen the scope of the database. Besides these words, vigilantly filtered words from different sources like newspapers, magazines and daily conversation has been included in the database. Similarly, some of the words along with their antonyms have also been included. These are the most frequently used in daily life (such as “MARG” meaning DEATH and “JWAND” meaning LIFE).

B. Recording Setup and Specifications

Recording of the selected phonetically compact Pashto words was done in an office environment in an acoustically balanced room with the aim of avoiding any noise which can creep in during the recording process. The recordings were carried out using Sony linear PCM recorder, PCM-M10.

The Recording phase is a very sensitive step and the recording process can potentially be disturbed by starting or pausing the recording in case any error take place during the recording. In order to avoid such disturbances, an intention of minimal manual interaction with the recorder was achieved by using a remote control. Similarly, mistakes could be made during the recording process, when any word is pronounced improperly. In such a case, the recorder is paused and words were uttered again to be recorded in a proper accent.

Recordings of the words were carried out by using a sampling rate of 44100 Hz. The audio files were saved in .wav format. After the recording phase, recorded files were converted to mono and any erroneous files were separated from the other files with the help of Sony Vegas Pro 9. These mono files were also saved in .wav format.

III. Speech data collection

The main task in the development of a generic purpose speech database is to obtain a corpus which ideally contains all the acoustic variabilities. However, it is practically impossible to develop such a corpus of ideal attributes. To develop a phonetically balanced speech database for Pashto, it should include all of the phonemes of Pashto, at least once at the beginning once in the middle and once in the end. The main feature of the words used in this research work is that it contains phonetically balanced set of words, acquired from [10]. The other main attributes that needs to be addressed in the database development for ASR are gender, age and accent of the speaker. These factors in the context of this work are discussed one by one in the following sub-sections.

A. Speaker Selection

Speaker selection is a vital task in the development of speech database since the age, accent, gender and origin of the speaker describe the viability of corpora for a certain application. For the development of Pashto speech database, most of the speakers were students and faculty members of University of Engineering and Technology Peshawar. These included both native and non-native speakers of Pashto. For equal distribution of attributes, utterances of both males and females have been recorded. In this way, the aim of equal distribution of native/ non-native and male/female speakers has been achieved. Age is also an important attribute in the development of speech database because of the fact that aging causes change in the voice due to the changes occurs in larynx with age [11]. By keeping this phenomenon in mind, the speakers were selected from different age groups ranging from 16 to 40 years.

B. Data Statistics

The collected data was uttered by 50 speakers, which included native/non-native speakers of Pashto, males/females having ages in the range from 16 to 40 years. Each speaker uttered all the vigilantly filtered 161 Pashto words, including the season’s names, week days and digits from 0 to 25, nouns along with their antonyms and some masculine along with their feminine.

C. Ponunciation Correction

To carry out the recording, speakers which are either native or non-native, has been selected which could be able to utter the words with correct pronunciation. In case of any ambiguity or mistake in the pronunciation, the words were re-recorded from the same speaker. The recording process was accomplished in a noise free environment to assure that no error is appended to the audio files. However, if there was still some error found in any file, it was eliminated manually by discarding those particular words and replaced by its re-recorded audio file.

IV. Distributions

While developing a speech corpus for Pashto, it was intended to make it balanced regarding the significant attributes causing acoustic variability. Speaker’s age, mother tongue, accent and gender were some specific attributes which were considered in the development of this corpus. Care was taken to distribute these attributes in a nearly uniform manner. The table shown in Fig.1 shows the proportion of native and non-native speakers of Pashto language. It is apparent that equal number of native and non-native speakers of Pashto has been selected for the recording. The non-native speakers selected, consist of speakers having different first languages such as Hindko, Urdu and Punjabi, but were able to speak Pashto as a second language. The non-native speakers were able to speak Pashto easily since they have studied in Pashto as language of instruction but in a slightly different accent. These were included with the intention to develop corpus covering different accents of Pashto language, to make it available for a generic ASR system.

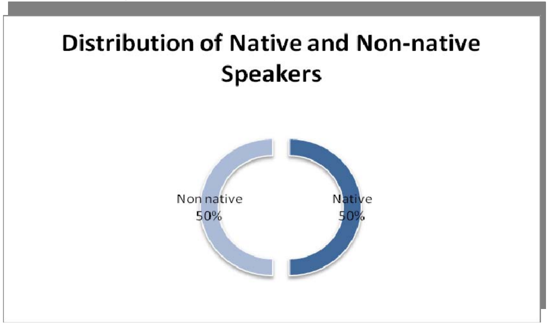

Figure 1. Distribution of native and non-native Speakers

Similarly, to have uniform distribution among the gender, similar number of male and female speakers has been deployed for the recording of selected Pashto words as shown in Table 1.

TABLE I. NUMBER OF MALE AND FEMALE SPEAKERS

Total Speakers	Male	Female
50	28	22

As, the age of the speakers is also another important attribute causing acoustic variability, the distribution of different age groups has also been considered in the development of this database. The age wise distribution of the speakers is depicted in Fig.2.

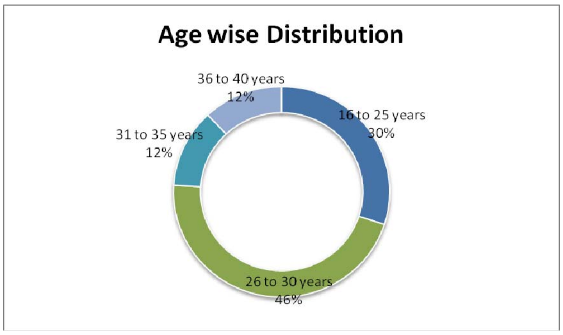

Figure 2. Age wise distribution of the Speakers

Phoneme distribution was also an important aspect considered while developing the speech corpus. Pashto consists of a number of phonemes which are different those of many other languages. Numerous phonemes available in Pashto are unique to Pashto and are not included in many other languages such as 'ړ' and 'ڼ'. Such sort of unique collection of phonemes and their distribution is one of the reasons which pledges for the development of a generic database for Pashto ASR system. The database developed consists of a careful assortment of all these phonemes.

V. REPRESENTATION

In the recording phase, 50 master audio files were collected, one file each recorded from one of the 50 speakers. Each of these files consists of 161 words included in the database. These speech files are saved in .wav format and stored in a folder named MAIN. The individual words from each of the master file were then extracted, were given a unique name and saved in the mono format. The name given to each mono file uniquely species the information about the speaker's name, age, gender and shows whether a speaker is a native speaker of Pashto or not. One of the modest approaches adopted for the representation of the mono files is discussed in [12], where each file is given nine characters names such as AAFYG2001, specifying all of the four attributes of the speaker and a unique identification of the word. Here the first two characters AA describes the speaker's name; third character F gives information about the gender of the speaker, where F stands for female and M for the male speaker. The fourth character Y depicts that the speaker is a native Pashto speaker while N is used instead if the speaker is a non-native speaker. To provide the age information, speakers' ages were divided into four age groups. Fifth and sixth letters indicates the age groups of a speaker while the final 3 digits represent the unique number of a word.

In this paper a seven character name convention is used instead to facilitate the name assignment to each individual word file by eliminating the separate age transcript and writing speakers' names in such a way that it also provides the age information. The ages of the speakers are divided in four groups shown in Table 2. In the seven letters plan adopted in this work, the first two letters specifies both the name and age group of the speaker. The speakers of age group 1 have both the letter as small, that of age group 2 first letter small, age group 3 second letter small and age group 4 as both letters capital. In this way, 'aa' represents the first speaker of age group 1. This scheme is depicted in Table 3, and can be extended up to any combination, depends upon the number of speakers in any age group. So now, a first word of database uttered by a male, native Pashto speaker of age group G2 can be represented by AaMY001, which is shorter than the name specified by using first approach.

The method for the speakers' names representation devised in this paper is very convenient and flexible. Each speaker was represented by the combination of two alphabets also giving the age information. As each of this combination can stretch from first letter 'a' or 'A' to 'z' or 'Z' thus covering a range of 26*26=676 speakers for each age group. In this research work, as recording was carried out for 50 speakers but can be extended to more number of speakers using the same name structure.

The third and fourth characters represent the gender and speaker's origin information respectively while the last three characters represent the word number. The three letter representation of the word number gives flexibility for extending the number of words up to 1000 and can also be extended further by adding more characters to the right.

TABLE II. AGE WISE GROUPING OF SPEAKERS

Ages of Speakers	Age Group
16 to 25 Years	G1
26 to 30 Years	G2
31 to 35 Years	G3
36 to 40 Years	G4

TABLE III. TABLE 3: AGE GROUP WISE NAMING OF SPEAKERS

Age Group	Transcription	Meaning
G1	aa	1st speaker of G1
	ab	2nd speaker of G1
	ac	3rd speaker of G1
	ad	4th speaker of G1
	.	.
	.	.
	.	.
	ao	15th speaker of G1

G2	Aa	1st speaker of G2
	Ab	2nd speaker of G2
	Ac	3rd speaker of G2
	.	.
	.	.
	.	.
	Aw	23rd speaker of G2
G3	aA	1st speaker of G3
	aB	2nd speaker of G3
	aC	3rd speaker of G3
	aD	4th speaker of G3
	aE	5th speaker of G3
	aF	6th speaker of G3
G4	AA	1st speaker of G4
	AB	2nd speaker of G4
	AC	3rd speaker of G4
	AD	4th speaker of G4
	AE	5th speaker of G4
	AF	6th speaker of G4

TABLE IV. GENDER AND FIRST LANGUAGE WISE TRANSCRIPTION

Transcription	Meaning
M	Male Speaker
F	Female Speaker
N	Non-native Speaker
Y	Native Speaker

VI. ASR SYSTEM DEVELOPMENT

After the development of Pashto corpus, the primary task is to develop a Pashto speech recognition system. The pivotal steps in developing an ASR system are the extraction of salient audio features and the selection of suitable classifier. To develop Pashto isolated word ASR system, mel-frequency cepstral coefficient (MFCC) and their first and second derivatives were used as the audio feature vector while linear discriminant analysis (LDA) based classifier was used for recognition. In the experiments reported here, a subset of the database consisting of 17 speakers with 7 female and 10 male speakers has been used. Out of these, 70% of the data was used for the training purpose and the remaining 30% was used for testing. The recognition results for the first ten most commonly used Pashto words are shown in Table V. The results shown are obtained from confusion matrix for the ten words.

TABLE V. PASHTO ASR RESULTAS

Word number	Percentage Error
1	0
2	33.33
3	0
4	60
5	33.33
6	33.33
7	50
8	33.33
9	33.33
10	33.33

The higher error rate for some of the words is due to the limited training data while including speakers of different accents and speakers with Urdu as mother tongue. The ASR system is trained using only 11 speakers; however the recognition performance can further be improved by increasing the training data.

ACKNOWLEDGMENT

The authors show a great gratitude towards the Department of Journalism and Mass Communication, University of Peshawar, for facilitating the recording process. We are also thankful to Engr. Muhammad Fida for his technical support in the recording task. A great thankfulness to Engr, Imran Ahmad, Lab Engineer department of Industrial Engineering and Engr. Salman Ilahi for providing great help during the recording and editing phase and for providing a suitable place to perform the strenuous task of recording. Finally special thanks to all those who volunteered to carry out the recording.

REFERENCES

[1] Available online on [http://en.wikipedia.org/wiki/Pashto_language] accessed 7 /5 /2012.

[2] R. Prasad, S. Tsakalidis, I. Bulyko, C.- L. Kao and P. Natarajan (2010), "Pashto speech recognition with limited pronounciation lexicon", *Acoustic Speech and signal processing (ICASSP)*, pp.5086-5089.

[3] http://www.srsp.org.pk/srsp_main/regional-data/kohat.html

[4] V. Zue, S. Seneff, and J. Glass (1990), "Speech database development at mit: Timit and beyond" *Speech communication*, vol. 9, no.4, pp.351-356.

[5] J. Gauvain, L. F. Lamel, and M. Eske´nazi (1990), "Design Considerations and Text Selection for BREF, a large French read-speech corpus", *proceedings of ICSLP,* Orsay cedex, FRANCE, pp. 1097-1100.

[6] Y. Yamazaki, and T. Morimoto (1994), "ATR research activities on speech translation", *proceedings of Second IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Kyoto, Japan, pp. 61-66.

[7] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J. M. Pardo and A. Rubio (1991), " Development of Spanish Corpora for Speech Research (ALBAYZIN)", *proceedings of the workshop on international cooperation and standarization of speech database and speech I/O assesment methods*, 26- 28 sept. 1991, Chiavari, Italy.

[8] C. Wutiwiwatchai, P. Cotsomrong, Sinaporn Suebvisai, and Supphanat Kanokphara (2002), "Phonetically Distributed Continuous Speech Corpus for Thai Language**"** *Proceedings of LREC*,vol. 3, pp. 869-872 .

[9] B. Das, S. Mandal, P. Mitra (2011), "Bengali Speech Corpus for Continuous Automatic Speech Recognition System", *The Oirental COCOSDA 2011 International conference on Microelectronics and Information Systems*, Oct 26-28 2011, Research Center, National Chiao Tung University, Hsinchu, Taiwan, pp. 51-55.

[10] H. Naeem, "New Puxto Primer; the 21st century updated and augmented Puxto phonetic alphabet", available online on [www.upesh.edu.pk/academics/Departments/pushto/books/puxtophoeticbook.pdf] accessed 7 /5 /2012.

[11] http://www.entnet.org/HealthInformation/Voice-and-Aging.cfm

[12] H. Ali, N.Ahmad, K.M. Yahya, and O. Farooq (2012), "A Balanced urdu isolated words corpus for Automatic speech recognition", 4th *international conference on electronics and computer technology ICECT 2012*, April 6th to 8th, 2012, Kenyakumari, India, pp. 473-476.