A Large Multilingual Corpus of Pashto, Urdu, English for Automatic Spoken Language Identification

Aamer Zahoor
University of Engineering and Technology
Peshawar, Pakistan

Nasir Ahmad
University of Engineering and Technology
Peshawar, Pakistan

ABSTRACT

The availability of a standard and phonetically rich speech corpus provides a common platform for comparing the performance of different speech recognition approaches and therefore is the first step for the research in a language. This work presents the development of a large multilingual speech corpus of Pashto, Urdu and English. Recordings have been made from a total of 194 speakers in the three languages, covering diverse dialects, age groups, genders and professions. Pashto and Urdu both native and non-native speakers have been considered while for English, all the speakers were non-native. The corpus comprises of three categories of phonetically rich spoken data in each language, that is, short questions regarding speaker's personal information, read speech and spontaneous speech from the domain of tourism. Although the corpus is developed primarily for research on Automatic Spoken Language Identification purpose, nevertheless, it can also be used for research on other topics such as Automatic Speech Recognition, Accent Recognition, Automatic Speaker Identification and the study of effects of non-nativeness on Language and Speaker Identification.

Keywords

Corpus Development, Pashto Language Corpus, Urdu Language Corpus, Pashto Automatic Spoken Language Identification, Automatic Speaker Identification, Automatic Speech Recognition.

1. INTRODUCTION

Speech is the most natural and convenient [1] way of communication among humans and that is why since long the researches are conducted to develop computer/machines that can discern and speak like them [2]. In this modern era, where computers have become the core component of almost every run of life, it is strongly needed to make this human-computer interaction further natural and ubiquitous [3]. The development of such automatic speech recognition systems need some previously stored data for training and evaluation purposes commonly referred as a speech corpus. Similarly, for developing a spoken language identification system, a phonetically rich and balanced speech corpus is imperative. Various speech corpora have been produced for the developed languages, such as BREF for French [4], TIMIT for English [5], and CSJ for Japanese [6], but no such standard speech corpus exist for the spoken language identification research of Pashto and Urdu languages. Ali in [7], Sarfaraz in [8], and Raza in [9] have made great efforts in developing Urdu corpora but the former is application specific corpus as it lacks spontaneous and read speech, while in latter two, the speakers were selected from a one city and non-native speakers of Urdu were not taken into account. Similarly Abbas in [10] has developed Pashto spoken database but it contains only spoken digits and hence it is not suitable for the problem of spoken language identification and speaker identification research areas. To the best of author's knowledge there is no standard Pashto spoken corpus that can cater the problem of automatic spoken language identification research and so as for Urdu language. This paper presents the development of a standard and phonetically rich large multilingual spoken corpus for Pashto, Urdu and English to serve as baseline corpus for automatic spoken language identification research.

Pashto is Afghanistan's national language and one of the main languages of Pakistan being spoken in Khyber Pakhtunkhwa and Baluchistan provinces with 50 – 60 million speakers around the world [10] [11]. Pashto is also spoken in UAE, Saudi Arabia, U.S, U.K, Qatar, Sweden, Thailand, Russia, Japan, Canada, New Zealand, Ireland and Australia by significant Pashtun population [12]. Similarly Urdu is the national language of Pakistan and widely spoken throughout the world with over 100 million speakers [7]. It is worth mentioning over here that 75% of Pakistani population understands Urdu [13] while only 5% of Pakistanis understands English [7].

2. CORPUS DEVELOPMENT

The first step in natural language processing of any language is the development of a standard resource [10], to be used for training as well as to serve as a basic platform for performance evaluation of various approaches. The prime aim of this corpus development is to provide a standard database for the development, evaluation and testing of algorithms for Pashto and Urdu automatic language identification systems. This development process comprise of many phases discussed in the subsequent sections below.

2.1 Content and Domain Selection

The content should be phonetically rich and should contain all the sounds [9] occurring in that language. To do so, one approach given in [7] [11] is to select words which are most frequently used by speakers in daily life and the same is followed in writing our scripts. It is very difficult to cover all the words of a given language but as reported by [7], there is no such defined list of words available for a given language. So in order to limit the very broad pool of words for a given target language, the domain of tourism is selected for selection of content in each language.

2.2 Scripts Development

The first step in developing a corpus is to write a script for each language which contains the contents to be recorded [14]. The scripts developed in this work contain three types of data:

- General questions regarding the speaker (name, gender, age, city belonging to, etc.)
- Read speech (phonetically rich sentences, time, date, spoken and spelled words and fixed prompts such as "what is the date today?")
- Spontaneous speech ("say something about tourism and economy")

2.3 Recording Setup and Specification

Recording of the corpus was carried out in a noise free acoustically balanced studio environment. Some of the recordings were also conducted in home environment due to the unavailability of sufficient native female speakers however case was made to keep the conditions similar to that of studio environment. Recordings were conducted with Sony Linear PCM Recorder (PCM-M10) at the sampling rate of 44.1 KHz, with a bit depth of 24 bits and channel stereo. The recorded files were saved in Wave PCM (.wav) format. Sensitivity level of the recorder was adjusted to manage the Signal-to-Noise Ratio for a clear recording.

2.4 Splitting and Trimming of Recorded Files

Splitting and trimming of long recorded files into appropriate size i.e. word, sentence or spontaneous speech, was carried out using "Adobe Audition CC 2014.2". Some files having extraneous background noise was fixed by applying noise removing adaptive filters or manual techniques. The split files were saved into new files in an uncompressed form.

3. DATABASE DESCRIPTION AND COLLECTION

A main task in developing a standard spoken corpus is to achieve a database ideally covering all the acoustic variabilities [11], however the development of such corpus with ideal attributes is practically impossible. The other attributes that need to be taken into account while developing a spoken corpus for automatic language identification are keeping balance in gender, age, dialects, nativeness/non-nativeness of the speakers. These attributes are discussed one by one in the context of this work in the succeeding sections.

3.1 Phonemes Inclusion

Pashto contains the phonemes of Urdu along with some additional phonemes of its own, and similarly Urdu contains phonemes of English with some additional phonemes of its own. Words consisting of phonemes peculiar to each language are included, like in Pashto ژبه (žābā - meaning life), نیټه (nyṭā – meaning date) and ډنډونه (ḍānḍūnā - meaning lakes) contain phonemes /ژ/, /ټ/ and /ډ/ respectively which are unique to Pashto language only. Similarly in Urdu, words included like گھڑی (ghāṛī - meaning wristwatch) and قدرتی (qūdrātī - meaning natural) contain phonemes /ڑ/ and /ق/ respectively which are lacking in English language.

3.2 Speakers Selection

The selection of speakers to be recorded is one of the most vital tasks in the development of a speech corpus because the age, gender, dialect, native or non-native, origin, and education level of the speaker describe the viability of a corpus for a particular application. For developing this corpus, the speakers were selected from diversified regional, professional and educational backgrounds. Most of the students as well as faculty members of Journalism Department University of Peshawar, University of Engineering and Technology Peshawar, Agricultural University Peshawar, University of Malakand, Khyber Girls Medical College Peshawar were recorded. The graph in Figure 1 provides the region wise distribution of speakers. To have gender balance, recording of both male and female speakers were considered, as depicted by Figure 2. As age is an important attribute and voice varies with age due to occurrence of changes in larynx with age [15], so speakers with age ranging from 18 – 50 years were considered, illustrated by the graph in Figure 3. Both native and non-native speakers of Pashto and Urdu were considered while for English all the speakers recorded were non-native due to the non-availability of native ones, as revealed in Figure 4.

3.3 Statistics

A total of 194 speakers were recorded, out of which 56 for Pashto, 70 for Urdu and 68 for English. Male to female ratio for Pashto, Urdu and English was 29:27, 31:39, and 31:37 respectively. The recordings were completed in a lapse of four months. The splitting, editing and refinement of the recorded data took one and a half month. Approximately a total of 49.62 hours of recording was captured, out of which 4.85 hours of useful data was extracted for corpus. On average, each speaker recording took 4 minutes out of which 1.5 minutes of useful data was extracted.

3.4 Correction of Pronunciation

In order to pronounce the script of each language accurately, each speaker native as well as non-native was first guided for about 5mins about the contents of the script. Despite guidance if a speaker still pronounced a word incorrectly, were asked to re-record correctly. If in a continuous speech utterance some error was found, it was fixed manually by discarding those particular utterances of words or by replacing them by their re-recorded utterances.

4. CORPUS ORGANIZATION

The master folder contains three main folders with one folder for each language i.e. Pashto, Urdu and English. Each folder further contains sub folders corresponding to each speaker of the respective language. The folder names of these sub folders provide information of the speaker. For instance, consider **L1S1MNTG2** where L1 shows that speaker has uttered the script of language1 of the corpus, S1 shows that it is first speaker of this language, M reveals that the speaker is male, NT describes that the speaker is native and finally G2 gives age information that speaker belongs to the age group 2.

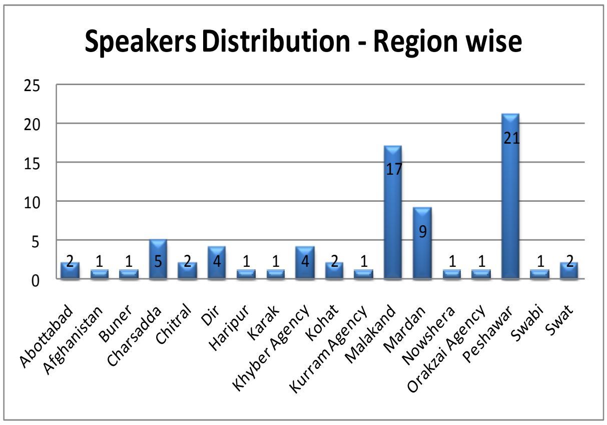

Figure 1. Speakers Distribution – Region wise

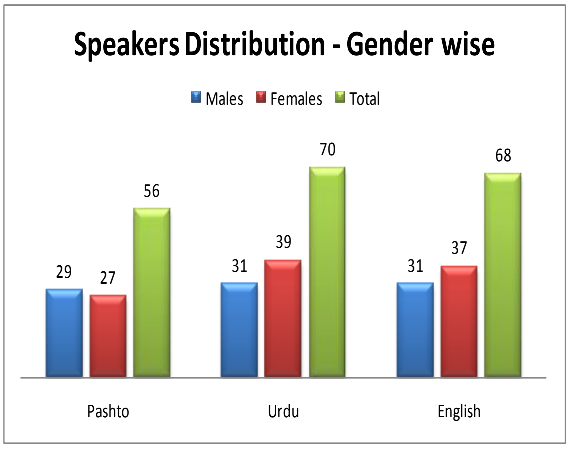

Figure 2. Speakers Distribution – Gender wise

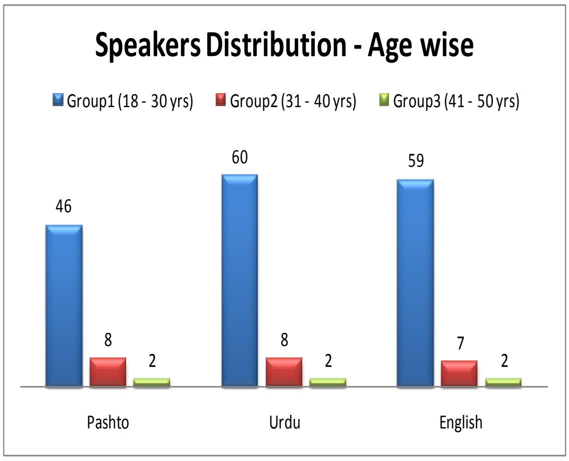

Figure 3. Speakers Distribution – Age wise

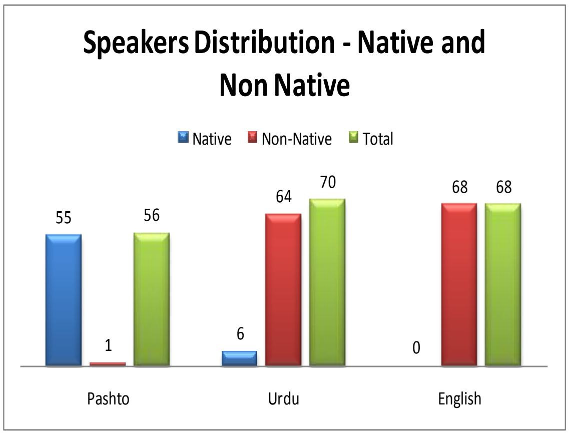

Figure 4. Speakers Distribution – Native/Non-Native wise

Each of these subfolders contain 21 files corresponding to different contents of the script being uttered by the speaker. The file name gives speaker information as well as the unique number of the uttered content. For instance in **L1S1MNTG2_001**, the first part has been inherited from the parent sub folder's name which it is included in, while the last combination of three digits represents the uttered content. A detailed guide to the representation scheme is provided in Table 1. This representation scheme is quite flexible and can be extended easily for increase in the size of corpus with the passage of time.

Table 1. Corpus Representation Guide

Parameter	Representation	Meaning
Language	L1 L2 L3	1st Language 2nd Language 3rd Language
Speaker	S1 S2 S60	1st Speaker 2nd Speaker 60th Speaker
Gender	M F	Male Female
Native / Non-Native	NT NN	Native Speaker Non-Native Speaker
Age Group	G1 G2 G3	18 – 30 Years 31 – 40 Years 41 – 50 Years
Content	001 002 021	1st content 2nd content 1st content

5. FUTURE WORK

In this work parallel corpus has been developed for two low resource languages i.e. Pashto and Urdu. In future more local languages being spoken in the region such as Hindko, Khowar, Farsi, Siraiki etc, can be added to this corpus. The contents of this corpus are selected from the tourism domain. In future, the contents can be extended to include other domains such as sports, transportation, entertainment, finance, health, agriculture etc. Moreover, the approach for developing a standard multilingual corpus, presented in this paper can be used as a guideline for developing corpora of other local languages being spoken in this region.

6. ACKNOWLEDGEMENTS

The author is exceptionally grateful to the Department of Journalism and Mass Communication, University of Peshawar, Pakistan for allowing using the studio. Very special thanks to Engr. Fida Muhammad for his kind guidance and help in the process of recording and for arranging the speakers to be recorded. Last but not the least, the author is indebted to all the individuals who volunteered for conducting the recordings.

7. REFERENCES

[1] Vyas, G. and Dutta, M.K., 2014, August. An integrated spoken language recognition system using support vector machines. In 2014 Seventh International Conference on Contemporary Computing (IC3) (pp. 105-108). IEEE.

[2] Shrishrimal, P.P., Deshmukh, R.R. and Waghmare, V.B., 2012. Indian language speech database: A review. International journal of Computer applications, 47(5), pp.17-21.

[3] Shriberg, E., 2005. Spontaneous speech: How people really talk and why engineers should care. In Ninth European Conference on Speech Communication and Technology.

[4] Larnel, L.F., Gauvain, J.L. and Eskenazi, M., 1991. BREF, a large vocabulary spoken corpus for French. In Second european conference on speech communication and technology.

[5] "LDC - Linguistic Data Consortium" [Online]. Available at: *https://catalog.ldc.upenn.edu/LDC93S1*

[6] Maekawa, K., 2003. Corpus of Spontaneous Japanese: Its design and evaluation. In ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition.

[7] Ali, H., Ahmad, N., Yahya, K.M. and Farooq, O., 2012, April. A medium vocabulary Urdu isolated words balanced corpus for automatic speech recognition. In 2012 international conference on electronics computer technology (ICECT 2012) (pp. 473-476).

[8] Sarfraz, H., Hussain, S., Bokhari, R., Raza, A.A., Ullah, I., Sarfraz, Z., Pervez, S., Mustafa, A., Javed, I. and Parveen, R., 2010. Speech corpus development for a speaker independent spontaneous Urdu speech recognition system. Proceedings of the O-COCOSDA, Kathmandu, Nepal.

[9] Raza, A.A., Hussain, S., Sarfraz, H., Ullah, I. and 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.

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

[11] Ahmed, I., Ahmad, N., Ali, H. and Ahmad, G., 2012, September. The development of isolated words pashto automatic speech recognition system. In 18th ICAC (pp. 1-4). IEEE.

[12] Abbas, A.W., Ali, Z. and Uddin, B., 2014, December. Analyzing the Impact of MFCC and LDA for the Development of Isolated Pashto Spoken Numbers ASR. In 2014 12th International Conference on Frontiers of Information Technology (pp. 350-354). IEEE.

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

[14] Lamel, L., Adda, G., Adda-Decker, M., Corredor-Ardoy, C., Gangolf, J.J. and Gauvain, J.L., 1998, May. A multilingual corpus for language identification. In 1st International Conference on Language Resources and Evaluation (Vol. 1, pp. 1115-1122).