# A Medium Vocabulary Urdu Isolated Words Balanced Corpus for Automatic Speech Recognition

Hazrat Ali
Department of Electrical Engineering
University of Engineering and Technology
Peshawar Pakistan
hazrat.ali@nwfpuet.edu.pk

Nasir Ahmad
Department of Computer Systems Engineering
University of Engineering and Technology
Peshawar Pakistan
n.ahmad@nwfpuet.edu.pk

Khawaja Muhammad Yahya
Department of Electrical Engineering
University of Engineering and Technology
Peshawar Pakistan
yahya.khawaja@nwfpuet.edu.pk

Omar Farooq
Department of Electronics Engineering
Aligarh Muslim University
Aligarh, India
omar.farooq@amu.ac.in

***Abstract*— The role of a standard database in conducting and evaluating the speech recognition research is two-fold. Firstly, it provides a standard platform for the research by providing a balance amongst various aspects of speech recognition such as gender, dialect, and age. Secondly, it provides a common platform for comparing the performance of various speech recognition approaches. This paper presents the development of a Medium Vocabulary Speech Corpus for Urdu Language. The Corpus comprises of 250 isolated words, including digits and the most frequently spoken words of the Urdu Language. The words have been selected from the 5000 most frequently words amongst the 19.3 million words of Urdu. The selected words have been uttered by 50 speakers in a noise-free acoustically balanced studio. The speakers comprises of both native and non-native, male and female, youngsters and aged persons. The corpus has been built for Automatic Speech Recognition of isolated words in Urdu Language.**

***Keywords-Corpus Development, Urdu Language Corpus, Urdu Automatic Speech Recognition***

## I. INTRODUCTION

Urdu is one of the major languages of the World, with more than 100 million [1] users around the World. Being the national language of Pakistan, it is understood by the majority of Pakistani population. It is worth mentioning that English is understood by only 5% of Pakistani population [2] while Urdu is understood by more than 75 % of the Pakistani population [3]. Urdu comprises of a large vocabulary, shared with many other Asian languages such as Arabic, Hindi, Farsi, Punjabi and Hindku. Thus, research on Automatic Speech Recognition (ASR) of Urdu Language will on one hand, lead to a natural human machine interaction (HMI) by the speakers of Urdu Language and on the other hand, will provide a background for the research in ASR of other akin languages. Despite its importance, the ASR research on Urdu Language is relatively new. This is partially due to the lack of research on Urdu linguistics studies and partially due to the unavailability of balanced speech corpora.

A balanced speech corpus is indispensable in developing an ASR system. Unlike TIMIT [4] for English, BREF for French [5] and ATR for Japanese [6], there is no standard corpus of Urdu Language available for ASR research. Sarfaraz et al. in [7], has presented the development of Urdu Corpus, but it covers speakers only from a single city and has been developed for continuous speech. Similarly, [8] has presented the development of a phonetically rich corpus for continuous speech recognition but these corpora are not yet available to public for research purpose. Moreover, both of these databases are designed for research on continuous speech recognition. However, to the best of author's knowledge there is no database available for research on isolated word or digit recognition. This paper presents the development of Medium Vocabulary Balanced Speech Corpus of isolated words in Urdu Language. The corpus developed will be available for research and academic purposes, free of cost.

## II. WORDS COLLECTION

The criteria for words selection can differ depending on the application for which the corpus is being developed. For example, a corpus for digit recognition application will contain the digits only. Similarly, other applications such as command and control will include words frequently used in that application. However, a database suitable for generic ASR research should ideally contain all the words of the language, used in common applications, but there is no such defined list of words. One approach is to select words containing all the 38 letters of the Urdu Language, at least once in the beginning, once in the middle and once at the end of the word. Yet another criterion is to select names of items used in daily life such as greetings, sports, news and weather etc. However, the approach that seems to be the most suitable is to select the most frequently used words, thus making the resulting isolated word corpora to be suitable for a general purpose research applicable to various applications.

978-1-4673-1850-1/12/$31.00 © 2012 IEEE

The later criterion was used in the development of the database reported in this paper.

### A. Words Selection Criteria

The research conducted in the center of Language Engineering (CLE) has identified the top 5000 words having the highest frequency of occurrence among 19.3 million words, collected from a variety of domains [9]. The domains covered include sports/games, news, finance, culture/entertainment, consumer information and personal communication. In the work reported in this paper, words have been selected amongst these 5000 most frequent words after careful filtering. Besides these words, digits from 0 to 9, names of months, names of days and seasons of the year have also been included. Similarly, for some of the words, their antonyms have also been included e.g زندگی (zindagi - meaning life) and موت (moath - meaning death). Similarly, for some of the words, other words related in terms of meaning or significance have been added, such as, for چھوٹا (chota - meaning small), بڑا (barra - meaning large) and درمیانہ (darmyana - meaning medium) have been added. Words making less sense when used in isolation such as prepositions were discarded. Examples include نے، سے، کا، کے (nay, say, ka, ky) etc.

### B. Recording Setup

The recording was done in a sound proof studio. The studio as designed with intent to be acoustically balanced and sound proof provided a very noise-free environment. The recording was done using a Sony Linear PCM Recorder, PCM-M10. It was ensured that minimum manual handling be done during the utterances of the words and thus a remote was used for start and pause purpose, instead of touching the recorder buttons during the process. In case of any mistake during the recording, the mistake was removed by re-recording the mistaken or badly pronounced words. The process of recording was completed over a period of 5 weeks.

### C. Recording Specifications

Initially, the utterances were recorded in stereo with a sampling rate of 44100 Hz and saved in .wav files. Then, the files were converted to mono having a sampling rate of 16000 Hz as used by [10] and saved in Windows PCM .wav format. This was done using Adobe Audition software. Each file had an average length of half a second and an average size of 16 kB. A total of 250 words have been uttered by each one of the 50 speakers.

## III. DATABASE DESCRIPTION

The corpus for generic ASR research should ideally contain all possible acoustic variabilities, however development of such an ideal corpus is practically impossible. Yet, there are a number of attributes that needs to be taken into consideration in corpus development. The key attributes leading towards a standard corpus development are mentioned in [11] and [12]. Amongst other attributes, origin of the speaker, age, first language and mood are the important factors which determine the suitability of the corpus for a particular application. The design criterion also alters with the target application of the ASR system or the target group of users of the corpus.

In this database, most of the speakers have been selected amongst the students, faculty and supporting staff of the University of Engineering and Technology, Peshawar. These speakers belong to different regions of Pakistan, thus, covering all the major accents of Urdu Language including both native and non-native speakers. Hence, the database is not limited to speakers belonging to a single city, unlike that in [7]. In this way, a uniform distribution of native and non-native, male and female speakers has been achieved. Similarly the speakers' ages have been selected randomly, ranging from 20 to 50 years.

### A. Attributes Distributios

Gender, age, native/non-native are the key attributes which have been considered in the development of this corpus. The graph in Figure 1 provides the distribution of Male and Female Speakers. Similarly, the proportion of speakers of different ages has also been provided in Figure 2. To achieve a uniform distribution of native and non-native speakers, utterances of equal number of native and non-native speakers have been recorded, as shown in Figure 3. Most of the non-native speakers have Pashto as their first language. Being from different regions, they differ in the accent of Pashto and thus in accent of Urdu as well. This covers various accents of Urdu Language. It is a common observation that some of the Pashto speakers cannot pronounce the phoneme ف ( /f/ ) and instead pronounce پ (/p/ ). Such speakers have been excluded as this may lead to a completely meaningless set of words.

### B. Letters Distributions

For a corpus developed for isolated word recognition, the corpus should be balanced and should ideally contain all the letters of the language. The equal distribution of the letters is not required as the letters normally have different frequencies of occurrence. For example, in English Language, the most frequent letter is 'E' [13]. Such statistics for frequency of occurrence of letters are not available for Urdu Language. However, as we have partially selected the most frequent words of Urdu Language, the letter count can provide such a distribution of letters. This analysis shows that all the letters of Urdu Language are included in the corpus, with letter ا "alif" being the most frequent.

### C. Phonemes Inclusion

Urdu includes the phonemes of English lanuage, besides some additonal phonemes. However, a standard set of phonemes in Urdu Language is yet to be finalized [14]. Some of the phonemes of Urdu Language does not exist in English and this is one of the reasons for the need of research on ASR of Urdu Language. For example, غ ('ghain' - /□/) as in the word باغ "bagh" (meaning garden) does not exist in

English, ق (‘qā’f - /q/) in the word حقوق “huqooq” (meaning rights) which sounds similar to /k/ but is actually different from /k/ [15]. The words included in this corpus contain a suitable collection of these distinct phonemes.

## IV. DATABASE ORGANIZATION

The master folder contains 50 files, each one having all the 250 words spoken by a particular speaker. Then, each word has been extracted and saved as an isolated file. In this way, we get 50 secondary folders, each one containing the 250 isolated files. The folders’ names provide information about the speaker. For example, consider **AAMNG1** where the speaker number has been represented by the first two letters i.e. **AA**. The speaker gender is male, represented by the third letter **M**. The speaker is a non-native speaker of Urdu, represented by the fourth letter **N**. Finally, the age information is provided by the fifth and sixth letters i.e. the speaker belongs to age group **G1**.

Similarly, the file name contains speaker information as well as the unique number of the word. For example, in **AAMNG1001**, the first six letters, including one digit, have been borrowed from the folder name and the last three digits represent the word. A complete guide to representation has been provided in Table 1. This is a very flexible form of representation and can easily be extended for increase in the corpus size with the passage of time. The AA representation format facilitates corpus developers to represent a total of 26×26 = 676 unique speakers. This is indeed a very large value for the number of speakers. Even, if a need of extension beyond this value arises, a third letter can be added to the extreme left position of the name, thus increasing the value to 26×676 = 17576. Similarly, if extension is required to represent additional attributes, letters can be added to the right of the folder names and just before the word number position in a file name.

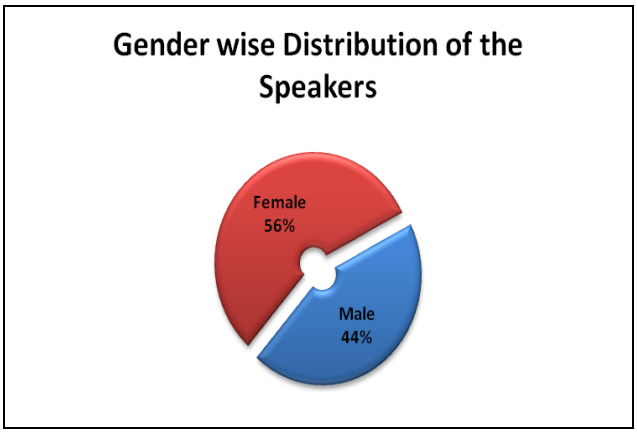


*Figure 1. Gender wise Distribution of the Speakers*

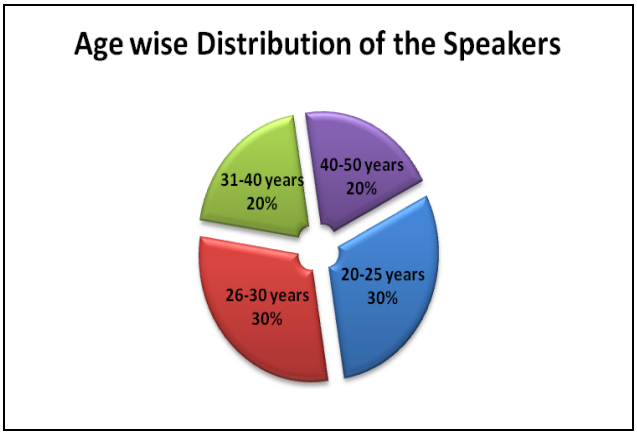


*Figure 2. Age wise distribution of the Speakers*

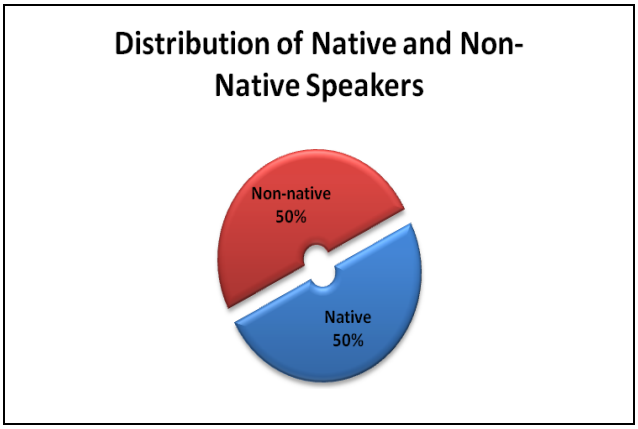


*Figure 3. Distribution of Native and Non-Native Speakers*

TABLE 1. DATABASE REPRESENTATION GUIDE

| Specification | Representation | Meaning |
|---|---|---|
| **Speaker** | AA<br>AB<br>AC<br>.<br>.<br>AZ<br>BA<br>.<br>.<br>BX | 1st Speaker<br>2nd Speaker<br>3rd Speaker<br>.<br>.<br>26th Speaker<br>27th Speaker<br>.<br>.<br>50th Speaker |
| **Gender** | M<br>F | Male<br>Female |
| **Native or Non-Native** | Y<br>N | Native Speaker<br>Non-native Speaker |
| **Words** | 001<br>002<br>.<br>.<br>250 | 1st word<br>2nd word<br>.<br>.<br>250th word |

## V. FUTURE WORK

This database is limited to 250 words only, covering a medium size vocabulary and can be extended to a large vocabulary database. As the research on Urdu ASR and Urdu linguistics is relatively new, the inclusion of new words will incorporate the findings of research by the authors and the study on Urdu language resource development, underway in CLE. This will lead to a larger and standard corpus of Urdu language for recognition applications. Furthermore, the corpus development approach presented in this paper can serve as a guideline for development of corpora of other local languages of Pakistan.

## ACKNOWLEDGMENTS

The authors are greatly indebted to the Department of Journalism and Mass Communication, University of Peshawar, Pakistan, for providing the opportunity to use the studio. We are also thankful to Mr. Muhammad Fida, alumni of the UET Peshawar, for providing his full support during the process of recording and guidance in the editing process. Lastly, special thanks to all those faculty members, students and staff of UET Peshawar who spared their time and volunteered for speech recording.

## REFERENCES

[1] "Ethnologue." [Online]. Available: http://www.ethnologue.com/show_country.asp?name=PK.

[2] J. Ashraf, N. Iqbal, N. S. Khattak, and A. M. Zaidi, "Speaker Independent Urdu Speech Recognition Using HMM," in *Proceedings of The 7th International Conference on Informatics and Systems (INFOS) , Cairo*, 2010, pp. 1-5.

[3] "Languages of Pakistan." [Online]. Available: http://countriesquest.com/asia/pakistan/the_people_of_pakistan/languages.htm.

[4] "Linguistic Data Consortium." [Online]. Available: http://www.ldc.upenn.edu/.

[5] J. L. Gauvain, L. F. Lamel, and M. Eskenazi, "Design Considerations and Text Selection for BREF, a large French Read-Speech Corpus," in *1st International Conference on Spoken Language Processing, ICSLP*, 1990, pp. 1097-1100.

[6] Y. Yamazaki and T. Morimoto, "ATR research activities on speech translation," in *Second IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 61-66.

[7] H. Sarfraz et al., "Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System," in *Proceedings of the O-COCOSDA, Kathmandu, Nepal*, 2010.

[8] A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah, and Z. Sarfraz, "Design and development of phonetically rich Urdu speech corpus," in *Proceeding of International Conference on Speech Database and Assessments, COCOSDA*, 2009, pp. 38-43.

[9] "Center for Language Engineering." [Online]. Available: www.cle.org.pk.

[10] J. S. Garofolo et al., *Documentation for the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. 1993.

[11] S. Atkins, J. Clear, and N. Ostler, "Corpus design criteria," *Oxford Journal of Literary and linguistic computing*, vol. 7, no. 1, pp. 1-16, 1992.

[12] D. Biber, "Representativeness in Corpus Design," *Oxford Journal of Literary and linguistic computing*, vol. 8, no. 4, pp. 243-257, 1993.

[13] W. Stallings, *Cryptography and Network Security: Principles and Practice*, 4th ed. Prentice Hall, 2006, p. 40.

[14] S. Hussain, "Computational Linguistics ( CL ) in Pakistan□: Issues and Proposals," in *Proceedings of EACL 2003 (Workshop in Computational Linguistics for Languages of South Asia)*, 2003, no. Cl.

[15] M. U. Akram and M. Arif, "Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling approach," in *Proceedings of INMIC 2004, 8th International Multitopic Conference*, 2004, pp. 91-96.