URDU SPEECH CORPORA FOR BANKING SECTOR IN PAKISTAN

Benazir Mumtaz, Sahar Rauf, Hafsa Qadir, Javairia Khalid, Tania Habib, Sarmad Hussain, Rukhsana Barkat, Ehsan-ul-Haq*

Centre for Language Engineering,
Al-Khawarizmi Institute of Compute Science,
University of Engineering and Technology Lahore, Pakistan
*Computer Science and Engineering Department
University of Engineering and Technology
Lahore, Pakistan

ABSTRACT

This research describes an effort to build Urdu speech corpora for the banking sector in Pakistan. We have designed speech corpora to develop debit card activation ASR and these corpora are comprised of eight types of corpora mainly debit card number corpus, expiry date corpus, last four digit corpus, months' name, date of birth corpus, account type and Urdu-counting corpus. These corpora contain telephone speech in read style obtained from more than 400 speakers specifically in Punjabi accent in both outdoor and indoor environments, including offices, homes, banks, and universities. The speech is automatically annotated and manually verified at sentence tier and reports 98% inter-annotator accuracy. In this paper, we report the design, recording and annotation process of speech corpora that serve as a data development step for ASR, and will be integrated in debit card activation service in banking sector of Pakistan.

Index Terms—Eight speech corpora, telephone speech, sentence tier annotation, outdoor and indoor environments, debit card activation ASR

1. INTRODUCTION

This paper presents speech corpora to develop a debit card activation ASR using Urdu spoken dialog system. Several digits speech corpora are already developed for multiple languages such as English [1], Swahili [2] and Hindi [3] which can be used in banking sector. However, very little effort is made to develop digits speech corpus for Urdu language. The previously developed Urdu digits corpus is very basic and limited to isolated words mostly ranging from 0-9 digits [4]. An ASR designed to activate debit card requires both isolated and connected digits speech covering all possible combination of digits in various sequences.

Spoken dialog system can provide Phonebanking and Internetbanking services to banking sector. These services are both time saving and cost effective. Most of these services are available in Pakistan in English language and require Internet connection. With a reported English literacy rate and Internet penetration of only 10% in Pakistan, we believe that a speech interface in a local language will be able to provide the users with important information by overcoming the barriers of literacy and connectivity. Moreover, the speech corpora reported in this research are recorded from Lahore district where people use Urdu, Punjabi and English languages in their daily life. Due to language versatility, the presented corpus also captures accent variability occurred due to language and dialectical differences in Lahore district. The proposed speech corpora also capture diversity in pronunciation, which will be helpful for the researchers in the field of linguistics.

In order to cover all possible information related to card activation process, multiple corpora are developed in this research covering two, four and sixteen digits' sequences. These speech corpora incorporate isolated as well as connected speech and can be used in numerous applications of number system including banking sector [2] travel domain [5], voice-dialing telephone [6], health services and agricultural domain [2].

The paper organization is as follows: Section 2 overviews different digits speech corpora developed for other languages. Section 3 presents the detailed description of corpora design, recording and annotation. Section 4 describes the quality assessment of the data. Section 5 discusses the issues faced during recording and annotation and finally, conclusion and dimensions for the future work are presented in Section 6.

2. LITERATURE REVIEW

Several efforts have been undertaken both at international and national levels to develop digits speech corpora. Leonard [1] has discussed English digits speech corpus developed for an independent digits speech recognizer. The speech corpus is comprised of 11 digits and recorded in 25

thousand digits sequences by 326 speakers (male/female) ranging from 6 to 70 years of age. In order to obtain the digits data, each speaker has to record 253 digits in a noise free environment over a microphone. Aaron et al. [2] have also worked on numeric speech recognition for Swahili Language. They collect the digits data from 60 Swahili speakers (30 native and 30 non-natives) between the age group of 20 to 50 years. Their speech data is consisted of isolated digits (0-9) and is recorded in a noisy environment.

Moreover, Fachrie and Harjoko explained Indonesian digits speech corpus in [7]. The isolated speech of 10 digits (0-9) is recorded from 20 native speakers including 10 males and 10 females. During recording, each digit is spoken 5 times by each speaker. The recordings are conducted using a microphone in a quiet room environment. The digits data is used to develop a robust digits recognizer. Similarly, Gedam et al. [8] describe the numeral speech corpus for a digits recognizer in Marathi language. The database comprises of ten isolated digits ranging from 0-9. They have obtained recording from 100 native speakers (50 males and 50 females) ranging between the ages of 18 to 30 years. The speakers record each digit thrice a time using headsets and PRAAT in a noisy environment. Thus, the total number of 3000 utterances have been collected and stored in ".wav" format. Saxena and Wahi [3] also worked on digits speech recognition for Hindi language. The reported corpus is comprised of 10 isolated Hindi digits that have been recorded from 10 speakers ranging from 20 to 40 year of age. The data is also recorded using headset.

Along with that, an important work is also found on Pashto language. Ahmed et al. [9] presented the isolated digits speech corpus for Pashto language using the most frequent vocabulary of 161 words and digits (0 to 25). The text data for digits corpus is filtered from several sources such as magazines, conversations, newspapers etc. The data is collected from 50 students and faculty members (both males and females) and is recorded in a noise free office environment. This corpus is used for the development of Pashto automatic speech recognition system. Some works are also found on digits speech corpus for Urdu ASR. Ali et al. [4] have described the speech corpus containing 250 high frequent words including digits 0-9, seasons, months and days' names. Qasim et al. [5] have also discussed Urdu speech corpus that is comprised of 250 vocabulary items including numbers (1-10), time, days and city names.

The review of the literature suggests that lots of work is done to develop digits speech corpora in different languages but only a limited set of digit vocabulary is recorded for Urdu. Hence, our main objective of this research is to design, collect and annotate a speech database, which covers all possible digits in all possible combination of Urdu language and can be used to develop a robust debit card activation spoken dialog system. Details of proposed Urdu speech corpora are given in section 3.

3. URDU SPEECH CORPORA

At the time of debit card activation, multiple fields of information are required from the debit cardholders such as his debit card number (DCN), debit card last four digit (DCLFD), debit card expiry date (DCED) and date of birth (DOB). Moreover, in order to develop spoken dialog system other various fields are also required such as affirmations (yes/no) and account type (AT). We have developed corpus for each type of information. The reason for developing different corpora for each type of field is that every field requires different length and different input formats of digits e.g. DCN is comprised of 16 digits with a blank space after every 4 digits, DCLFD is composed of four digits, DCED is consisted of 4 digits separating the information of month and year using a "/". DOB is comprised of 8 digits with the information of date/month/and year. Along with main corpora, sub-corpora are also developed including 0-100 Urdu English counting (UEC) corpus, months' name corpus (MNC) and special digits corpus (SDC) i.e. 1st, 2nd, 3rd etc. Details of vocabulary count and input format for each corpus are presented in Table1.

Table 1 Digit corpora vocabulary counts and formats

Corpus names	Vocabulary count	Input format
DCN	10 Urdu digits	16 digits card number (e.g. 4289 1234 5678 9010)
DCLFD	107 Urdu digits	4 digits (e.g. 4996)
DCED	101Urdu digits	4 digits (12/18)
DOB	120	8 digits (04/06/1990 or 04.06.1990)
YN&AT	17	Urdu words
UEC	105	0-100 digits, 1000, double, triple & tetra
MNC	12 Urdu months' names & 13 English months' names	January to December & Feb
SDC	32	1st to 31st & jəkəm

Moreover, there is a possibility that different speakers give same information in different ways. For example, date 20/12/16 can be read as twenty/twelve/sixteen, two zero/one two/one six or twenty/December/sixteen. In order to capture this diversity, a small experiment was conducted.

In this experiment, mock debits cards were designed and recorded by 15 speakers. The results illustrated that all the speakers used single digits ranging from 0-9 in case of DCN. In case of DCLFD, DOB and DCED, speakers used both single and double-digit utterances ranging from 0-100. Moreover, speakers used months' names frequently in DOB. Similarly, people showed affirmation using various vocabulary items such as ʤi: hɑ̃: /جی ہاں/'yes', hɑ̃:/ہاں/'yes' and hɑ̃: ʤi:/ہاں جی/'yes' etc. It is also observed that some

speakers use English and Urdu digits simultaneously in an utterance. Therefore, sub corpora such as Urdu English counting, Urdu English months' name and special digits are recorded to strengthen the main digits corpora.

Based on the results of the experiment, the digits corpus containing all the possible formats that are likely to be uttered by the users are developed accordingly. Table 2 presents various utterances possibilities in DOB corpus.

Table 2: Various utterances possibilities in DOB corpus

1/6/1 9 9 0
0 1/06/1 9 9 0
1/06/1 9 9 0
0 1/June/1 9 9 0
1/June/1 9 9 0
1/6/19 100 90
0 1/06/19 100 90
0 1/6/19 100 90
1/06/19 100 90
0 1/June/19 100 90
1/June/19 100 90

.

The details of corpus design, recording and annotation procedures are discussed in following sections.

3.1. Corpus design

The important step in designing the speech corpus is the development of text corpus to be recorded. The text corpus should be designed in such a way that it can provide the coverage of speech units that are likely to be uttered by the users. The most commonly used speech units for building acoustic models are complete words, di-phones, and tri-phones based models. For digits text corpora design, we have used tri-phones as a basic speech units because ASR systems use tri-phones for acoustic modeling. A Greedy search algorithm is used for selecting sentences in such a way that it contains coverage of all valid tri-phones. Hence, a list of 42 unique sentences is generated for DCN, 374 sentences for DCLFD, 218 sentences for DCED and 281 sentences for DOB using Greedy search algorithm. The valid tri-phones extraction for each main digits corpus is discussed below:

3.1.1 DCN text corpus

The text corpus for DCN is generated by randomly concatenating all possible values of digits with each other to develop a corpus of size 10K sentences. The corpus contains 18 unique phones. The total number of tri-phones generated from the unique phones set is 5832. Of all the generated tri-phones, 223 are valid tri-phones.

3.1.2 DCLFD text corpus

The total size of generated DCLFD corpus is 29400 sentences. The corpus contains 35 unique phones. The total number of tri-phones generated from the unique phones set is 42875. Of all the generated tri-phones, 1057 are valid tri-phones.

3.1.3 DCED text corpus

The text corpus is generated by concatenating all possible values of months with every possible value of year separated by a "/". In this way, the total size of generated corpus is 4104 sentences. The corpus contains 35 unique phones. The total number of tri-phones generated from the unique phones set is 42875. Of all the generated tri-phones, 552 are valid tri-phones.

3.1.4 DOB text corpus

The text corpus is generated by randomly concatenating all possible values of days, months and years with each other to develop a corpus of size 30K sentences. The corpus contains 40 unique phones. The total number of tri-phones generated from the unique phones set is 64000. Of all the generated tri-phones, 853 are valid tri-phones.

3.1.5 YN&AT text corpus

The vocabulary of this corpus is a closed list of 17 Urdu words, which is given in Table 2 below.

Table 3: Vocabulary for YN&AT corpus

Serial No.	IPA	Words
1	hɑ̃:	ہاں
2	nəhi:	نہیں
3	ʤi:	جی
4	ʤi: hɑ̃:	جی ہاں
5	hɑ̃: ʤi:	ہاں جی
6	nəhi: ʤi:	نہیں جی
7	ʤi: nəhi:	جی نہیں
8	jæs	یس
9	no:	نو
10	kərənt əkɑ:u:nt	کرنٹ اکاونٹ
11	se:vɪŋg əkɑ:u:nt	سیونگ اکاونٹ
12	sɪŋgəl əkɑ:u:nt	سنگل اکاونٹ
13	ʤʊɑ:ɪnt əkɑ:u:nt	جوائنٹ اکاونٹ
14	kərənt	کرنٹ
15	se:vɪŋg	سیونگ
16	sɪŋgəl	سنگل
17	ʤʊɑ:ɪnt	جوائنٹ

3.1.6 Sub-corpora text

We have three sub corpora i.e. Urdu English counting corpus, Urdu English months' name corpus and special digits corpus. The vocabulary for sub-corpora is closed list

containing 105 digits in Urdu English counting case, 25 months' names in months' name corpus case, and 32 digits in special digit corpus case. (See Table 1 for details).

3.2. Corpus recording

After designing the corpus, recording of the data is carried out. The data is recorded over the telephone channel at the sampling rate of 8 kHz with 16 bit resolution. The data is saved in ".wav" format. Corpus is recorded through all the possible telecom operators in Pakistan. Different dialogs are designed to record the different corpora. The dialogs include the following information:

- A welcome note
- Corpus ID
- List ID of corpus
- Speaker's information (language, district and name)
- A list of Urdu digits
- Thank you note

3.2.1. Environment

The digit corpora are recorded in both indoor and outdoor environments. All the digit corpora are recorded from Lahore city in Pakistan. Lahore is the second major city of Pakistan [10]. People come from different cities for seeking the employment in Lahore [11]. Therefore, Lahore is a best place for capturing different accents of Urdu and Punjabi depending on the speakers' mother language.

3.2.2. Speaker information

The data is recorded from more than 400 speakers (both males and females) ranging from 18-50 years of age. The recording is obtained from various university students, bank employees and the family members of data collection team. Data statistics are given in Table 4. The speakers' personal information is also captured during recording such as speaker's gender, age, mother tongue, district and name.

3.2.3. Corpus details

The inclusive information regarding the corpus names and their data lengths in minutes is given in Table 4.

Table 4 Digit corpora statistics

Corpus names	Duration (minutes)
DCN	258
DCLFD	47
DCED	55
DOB	98
YN&AT	21
Counting	15
Month's names	44
Special words	23
Total	561 (9 hours)

3.3. Corpus annotation

The process of data annotation is carried out after the recording process. The data annotation is a semi-automatic process and composed of two steps. At first step, a PRAAT utility is used to automatically label the speech files with CISAMPA transcription at sentence level. The utility places two boundary markers with silence (SIL) labels at the start and end of the speech file. Moreover, utility generates three folders: correct, incorrect and alternate pronunciation (AP). At the second step, pre-labeled files are carefully listened and analyzed by the expert linguists. The annotators ensure that there is no mismatch between the speech and the annotated data. Moreover, any type of insertion or deletion of phonemes, hesitation and disfluency in speech are discarded by the annotators. Furthermore, sentence boundary markers are manually placed on correct positions. Afterwards, the discarded and accurate files are moved into incorrect and correct folders respectively.

In case of DOB, DCN and DCED, the pauses "PAU" are marked automatically using a PRAAT utility. The pauses are aligned with blank space, "/" and "." present in corpora text. However, during the annotation, it is observed that the speaker can add extra pauses or delete the automatically marked pauses. This phenomenon shows that continuous speech is a natural process and speakers can take pauses according to their ease. Because of this issue, the pauses are listened and analyzed very carefully by the annotators and are marked/ removed accordingly. Figure 1 shows marking of pauses in DCN data.

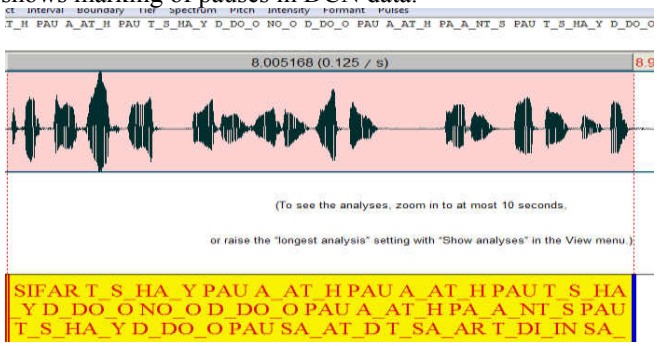

Figure 1 Pause marking in DCN

Multiple types of noises were also handled during the data annotation process. Files containing babbling and traffic noises, which disturbed the speech signals are moved into the incorrect folder. Moreover, minor clicks and breathing sounds in an utterance are labeled as non-speech sounds (NSS). Figure 2 shows marking of NSS in the data.

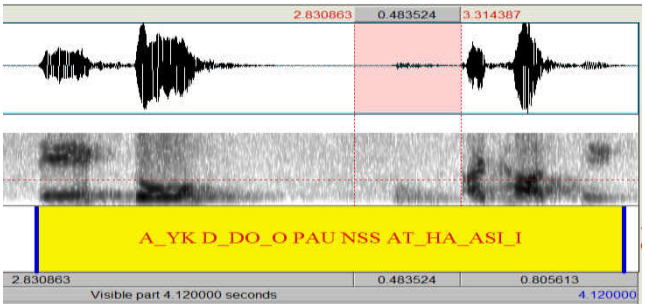

Figure 2 NSS marking in DCED

The mispronunciation and alternate pronunciation (AP) are another problems faced by the annotators while annotation. The data is discarded if the word is mispronounced. For example if ʈæɳʈɑ:lɪs/تینتالیس/ 'forty three' is spoken as ʈərʈɑ:lɪs/ترتالیس/'forty three', it is considered incorrect. However, APs are accepted if they are frequently used among the speakers. Thus, the files having mispronunciation are moved into incorrect folder and the files having alternate pronunciation are moved into AP folder.

4. QUALITY ASSESSMENT

The finest quality of the data is necessary alongside with the corpus collection and annotation. Therefore, meticulous measures are taken to ensure the quality of the data. PRAAT utilities generated by technical experts are used along with the manual effort for the testing processes. For testing purpose, we have two teams: annotation team and reference team. Annotation team annotates the data whereas reference team checks the labeling of annotation team. For testing purposes, reference team selects 20% data randomly from the source data and annotates it independently. The source data is than compared automatically with the reference data. During comparison, four levels assessment is conducted: mismatched files assessment, matched files assessment, sentence boundary assessment and lexicon generation.

Comparison of mismatched files includes the manual review of mismatches between the reference and annotation teams' data. After this review, only those packages are selected whom AP and correct files are 98% accurate. Comparison of matched files includes further 2 steps: verification of phones using CISAMPA list and verification of phones using reference files. Only those files are considered correct whose phoneme labels are 100% accurate.

At sentence boundary assessment level, the expert linguist checks the mismatches in initial and final boundary markings of an utterance. The sentence boundary marking must be 95% accurate to pass the test. The fourth level is to generate a lexicon to compile all the pronunciation in one Notepad++ file to check that data does not contain any incorrect pronunciations. APs are removed from the lexicon if they occur less than 3 times in the data.

5. DISSCUSSION

Several issues are faced during the corpus recording and annotation process. This section briefly discusses these issues and their respected solutions.

During the process of recording, speakers were very reluctant to record the data especially in case of bank employees. To solve this issue, the data collection team is instructed to explain the usage and purpose of this exercise to every speaker. It was also instructed that the team should consider the speaker's opinion or time and do not force the participants to record the data.

Similarly, during annotation process, annotators have faced many issues due to alternate pronunciation (AP). In final pronunciation lexicon, we have 57 APs in Urdu counting and 49 APs in English counting and 21 APs in months' name corpus. The numbers are very notable and denote our findings that some APs are used more frequently than the standard pronunciation. For example, the digit 9 is labeled as /nɔ:/ in dictionary [12] but most speakers pronounce it as /no:/. Similarly, digit zero is used more frequently instead of Urdu sɪfər/ صفر/ 'zero'. Therefore, due to the higher frequency of /no:/ and zero they are made part of Urdu digit pronunciation lexicon. Nasalization of /ɑ:/ vowel is another issue faced by annotators. It is observed that final /ɑ:/ vowel change in to /ɑ̃:/ in digits like bɑ:rɑ:/بارہ/'twelve' and ʧɔ:dɑ:/چودہ/'fourteen' due to the influence of Punjabi language.

Another issue is to mark the /s/ sound specifically at the start of the utterance. /s/ is light fricative and shows very disperse energy. Therefore, sometimes annotators can hear the /s/ sound but its properties are invisible in the spectrogram. In such cases, the annotators assign the silence area of 70 ms to the sound.

6. CONCLUSION AND FUTURE DIRECTIONS

This paper reports the development of eight speech corpora specifically developed in the fields of DCN, DCLFD, DCED and DOB to capture the diverse digit combinations occurring in different length and formats. These corpora are recorded to develop ASR to be used in card activation service. Our future perspective is to use these corpora to generate a robust ASR that can decode any sequence of numbers such as credit card number, National identity card (NIC) numbers and date of birth and to deploy this system in various local banks of Pakistan. Moreover, spontaneous speech will be included as the part of speech corpus in order to capture real situation in future.

10. REFERENCES

[1] R. Gary Leonard, "A database for speaker-independent digit recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84*, Dallas, Texas,

USA, 1984, pp. 328-331.

[2] A. M. Oirere, R. R. Deshmukh, and P. P. Shrishrimal, "Development of isolated numeric speech corpus for Swahili language for development of automatic speech," *International Journal of Computer Applications (0975 – 8887)*, vol. 74, no. 11, pp. 20-22, July 2013.

[3] B. Saxena and C. Wahi, "Hindi digits recognition system on speech data collected in different natural noise," in *International Conference on Computer Science, Engineering and Information Technology (CSITY)*, Bangalore, India, 2015, pp. 14-15.

[4] H. Ali, N. Ahmad, K. M. Yahya, and O. Farooq, "A medium vocabulary Urdu isolated words balanced corpus for automatic speech," in *4th International Conference on Electronic Computer Technology, ICECT*, Kanyakumari, India, 2012, pp. 473-476.

[5] M. Qasim, S. Rauf, S. Hussain, and T. Habib, "Urdu speech corpus for travel domain," in *Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, Bali, Indonasia, 2016, pp. 237-240.

[6] M. Kalith, D. Ashirvatham, and S. Thelijjagoda, "Isolated to connected Tamil digit speech recognition system based on hidden markov model," *International Journal of New Technologies in Science and Engineering*, vol. 3, no. 4, pp. 1-11, April 2016.

[7] M. Fachrie and A. Harjoko, "Robust Indonesian digit speech recognition using Elman recurrent neural network," in *Konferensi Nasional Informatika (KNIF)*, Bandung, Indonesia, 2015, pp. 49-54.

[8] Y. K. Gedam, S. S. Magare, A.i C. Dabhade, and R. R. Deshmukh, "Development of automatic speech recognition of Marathi numerals - a review," *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 3, no. 9, pp. 198-203, March 2014.

[9] I. Ahmed, N. Ahmad, H. Ali, and G. Ahmad, "The development of isolated words Pashto automatic speech recognition system," in *18th International Conference on Automation & Computing, Loughborough University,* Loughborough, UK, 2012, pp. 1-4.

[10] K. Mahmood et al., "Groundwater levels susceptibility to degradation in Lahore metropolitan," *Sci.Int(Lahore)*, vol. 25, no. 1, pp. 123-126, Oct 2013.

[11] F. Mazhar and T. Jamal, "Temporal population growth of lahore," *Journal of Scientific Research*, vol. 39, no. 1, pp. 1-6, June 2009.

[12] S M Salim-ud-din and S.l Anjum, *Oxford Urdu-English dictionary*, 1st ed., Parekh Rauf, Ed. Karachi, Pakistan: Oxford University Press, 2013.