2016 Conference of The Oriental Chapter of International Committee
for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)
26-28 October 2016, Bali, Indonesia

Urdu Speech Corpus for Travel Domain

Muhammad Qasim, Sahar Rauf, Sarmad Hussain
Center for Language Engineering, Al-Khwarizmi Institute of Computer Sciences
University of Engineering and Technology
Lahore, Pakistan
firstname.lastname@kics.edu.pk

Tania Habib
Computer Science and Engineering Department
University of Engineering and Technology
Lahore, Pakistan
firstname.lastname@uet.edu.pk

Abstract—**Speech corpus is a collection of recorded speech data. It is a fundamental component to develop any speech based applications. This paper presents the design and development of an Urdu speech corpus for travel domain. The corpus consists of a total of 250 vocabulary items including city names, days, time and numbers. The corpus has been used to develop a speech recognition system with a laboratory accuracy of 95.6% and a field accuracy of 87.21%. The corpus can be used for domestic flight queries, domestic flight reservation, train inquiry service and reservation and bus information and reservation.**

Keywords—speech corpus; travel domain; speech recognition; isolated word

I. INTRODUCTION

Information and communication technologies (ICTs) are being used to provide services to people for their socio-economic uplift. Despite the certain benefits offered by ICTs, illiteracy is a significant barrier to their effective use in developing countries. The use of speech as a medium can overcome this barrier as people can communicate in their local languages without any difficulty. Hence, speech interfaces can help in providing services to the people who are otherwise unable to benefit from them.

The idea of communicating with machines or computer systems using speech has always fascinated the mankind [1]. A significant development in this field has been done during the last few decades. Today, humans can communicate with machines in their local language to complete tasks such as information retrieval and navigational services. The development of speech technologies and applications for a language is dependent on the development of speech corpora for that language. Speech corpus of a language is a collection of recorded speech audio files of that language and their text transcriptions. There are two types of speech corpora: read and spontaneous speech [2]. Read speech includes the speech audio files of list of words, book excerpts and broadcast new while the spontaneous speech consists of dialog between two or more people and interviews.

Low literacy rate and lack of internet facilities restricts people of Pakistan from utilizing ICTs. Therefore, the use of speech interfaces designed in Urdu language can be helpful in disseminating information to the masses. The design and development of a read speech corpus in Urdu for travel domain is presented in this paper. The vocabulary items for corpus includes city names, date, time and numbers that are relevant to travel domain for reservation purposes. We describe the process of designing this corpus, recording setup used, cleaning process employed and quality assurance of the data. The developed dataset has been used for speech recognition in a bus reservation spoken dialog system.

The rest of the paper is organized as follows: related works in the field of speech corpora are discussed in Section 2. Section 3 describes the design and collection of speech corpus. Section 4 describes the cleaning and annotation process of speech corpus. Section 5 presents the use of corpus in a bus reservation system. Finally, In Section 6 we conclude our discussion.

II. LITERATURE REVIEW

Speech corpus is an essential component of any speech interface system. Speech corpora for many domains have been developed for different languages. The development of speech corpora for two travel domains L'ATIS (for air travel information) and MASK (for train travel information) is discussed in [3]. Acoustic and language training models are used to improve the performance of continuous speech recognizer. The corpus was recorded at regular basis including 1000 queries per month from 20 speakers. L'ATIS provides information about flights and fares within the cities of United States and Canada. MASK provides rail travel information for 500 cities of France which includes timetables, tickets and reservations for train. Different scenarios or sentences were selected for retrieving the vocabulary items; city names, dates and times of travel. Office environment was used for the recording with a noise cancelling Shure SM10 and a PCC160 microphone. For the better statistical understating of the data, the corpora was transcribed and classified into different sets. The performance of the system was also assessed through the filling of the questionnaires from the users.

Hindi language speech corpus in travel domain was developed for automatic speech recognition system [4]. The training data consisted of 26 hours of speech recordings was collected from 30 female speakers in a noise free environment. Total 74,807 words were recorded from the speakers ranging

978-1-5090-3516-8/16/$31.00 ©2016 IEEE

from 17 to 60 years of age. The recognition accuracy was 70.73% for the training data and for the testing data, it was 60.66%. Another speech to speech translation system for Indian languages was developed for the travel and emergency services [5]. The data was collected according to the possible usage scenario, so the above mentioned major domain were divided into sub domains including; tourism, emergency services, hotel transactions and local travel. The speech corpora have been developed for English, Telugu and Hindi and recorded from 15 speakers. A noise free environment was used for recordings through microphone and a laptop.

A speech translation application in travel domain has been developed for Bengali [6]. The system is used for speech translation from English to Bengali language. Phonetic transcriptions have been developed and statistical analysis is given for Bengali BTEC text. The phone coverage for the system is 70% including the mono-phones, di-phones and tri-phones. Total 20952 words have been analyzed for the corpus. A rule based grapheme to phoneme converter has been used to phonetically transcribe the Bengali text followed by a manual verification.

For Urdu language, a single word and fixed vocabulary speech corpus consisting of district names of Pakistan has been developed [7]. The data was recorded from 300 speakers ranging from 18 to 50 years of age. The data was recorded from different districts of Pakistan covering six major accents; Punjabi, Urdu, Sindhi, Balochi Pashto and Saraiki. A telephonic channel was used for the recording with a sampling rate of 8 kHz. The data was annotated at word level using PRAAT [8] and also verified manually for the better accuracy of the data. A 95% accuracy level has been achieved on the speech annotation. Inclusive guidelines have been developed for the verification and cleaning of the data from different environmental issues. The cleaned corpus is used in an Urdu dialogue system that provides the weather information of Pakistan. The speech recognition accuracy at this data is 88%. In this paper, we focus on the design and development of a travel domain speech corpus for Urdu language.

III. Speech Corpus Design and Collection

Different types of fields like city names, days and time are required for any travel domain reservation system. Daewoo[1], a high quality transport service of Pakistan, has been taken as the case study. Through the study, the basic fields required for the corpus design were finalized. This section provides details about the design of speech corpus, setup used to collect data and information about collected data.

A. Corpus Design

To develop the speech corpus for travelling domain, the vocabulary items have been selected from different dimensions. It has been specifically noticed during the development of the corpus that such vocabulary items should be selected which can cover the possible fields of traveling. As the concern is to benefit the people so it kept in mind that corpus should contain such words which are frequently used by the user. So the first selected field for travelling is destination. The major cities names of Pakistan have been considered from which the people of Pakistan usually travel. Thus the proposed corpus includes; 44 major cities names of Pakistan e.g., /ڈسکہ/ /جام پور/ /Daska/ /Jam Pur/ /dəska:/ /ʤa:mpu:r/. Second important field is day and the corpus includes names of the days in both Urdu and English language as people can use days' names in both Urdu and English e.g., /منڈے/ /پیر/ /Monday/ /Pir/ / mənde:/ /pi:r/. The information of number of seats is another important field which can be important for a user to make reservations e.g., /ایک/ /دو/ / /One/ /Two/ /e:k/ /ɖo:/. The information of time is one of the crucial fields. The presented corpus divides time into further 2 sections; time that includes part of the day; /صبح/ / شام//Morning/ /Evening//suba:/ /ʃa:m/ and time in hours; /سوا/ /ساڑھے//Quarter past/ /Half past/ /səva:/ /sa:ʈʰe:/. Other categories have also been selected depending on the usage by Pakistani people like; affirmation /ہاں/ /نہیں/ /Yes/ /No/ /hɑ̃:/ /nəhĩ:/ and special key words; /فرسٹ/ /سیکنڈ/ /First/ /Second/ /fərst/ /sækɪnd/, /غلط/ /Wrong/ /ɣələʈ/. The complete information of the data and the number of the vocabulary items selected are given in Table 1. It was noticed during the data collection that time is a complex field. Apart from the above mentioned vocabulary items for time field, it also includes the combination of these items i.e. time (part of the day) and time (hours). For this reason, it has been noted that there is a requirement of small phrases for time e.g., /رات آٹھ/ /Eight at night/ /ra:ʈ a:tʰ /, /رات ساڑھےتین / /Half past three/ /ra:ʈ sa:ʈʰe: ʈi:n/ etc.

Table 1: Traveling fields and number of vocabulary items selected

Fields	No. of Vocabulry Items
Destination City names	44
Days	22
Time (Part of Day)	4
Time (Hour)	17
Time (phrases)	148
Number of Seats	10
Affirmation	2
Special key words	3

B. Data Collection Method

We used an Interactive Voice Response (IVR) system for recording of speech data from the users over a telephone line. A telephone landline was connected to a computer system using Cisco SPA 3102 VOIP gateway device [9]. The computer system ran on CentOS and used Asterisk PBX [10] which handles calls using a dialplan. Dialplan controls the flow of dialog with the user. A dialplan was designed and deployed for recording of speech data from users. The sampling rate of recorded data was 8 KHz and it was saved in "wav" format with a quantization rate of 16 bits.

When a user calls the system, the system greets the user and gives a brief introduction of the corpus development. Then, the system tells some guidelines and instructions for

[1] http://www.daewoo.com.pk

recording of corpus to the user. These instructions included what, when and how a user should record the speech data.

The vocabulary items were divided into several lists where each list consisted of 20-30 words. In order to record a list, user has to enter that list number using the keypad when asked by the system. Once a correct list number has been entered, the system narrates numbers from one to the number of total vocabulary items in that list with a gap of 4 seconds between each number. The user has to speak the vocabulary item written against the number spoken by the system. Once all the vocabulary items for that list are recorded, the system thanks the user and terminates the call. The flow of a call during a recording session is shown in Figure 1.

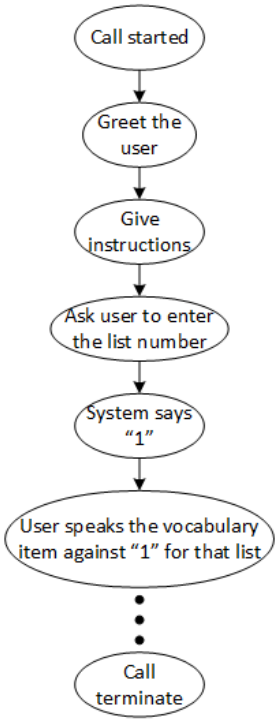

Figure 1: Flow of call during a recording session

C. Recording Process

The recordings took place in office environment at Center for Language Engineering. Speakers were provided with a mobile phone for calling the system. The speakers were also given lists of all vocabulary items. The speaker would call the system and enter the list number for each list and get its vocabulary items recorded. The speakers were asked to take 3-4 minute breaks after recording of each list so that user does not get tired and the speech does not get affected.

D. Corpus Details

The speech corpus was collected from students, teaching staff and non-teaching staff of a University of Engineering and Technology, Lahore. The age for speakers ranged from 18 to 40 years. Speech corpus was collected from a total of 60 speakers of which 32 were female and 28 were male. Table 2 shows the total amount of data recorded from the users.

Table 2: Statistics of recorded data

Dimension	Vocabulary size	Total utterances
Destinations	44	2532
Day of Reservation	23	1144
Time of Reservation	150	7491
Number of Seats	11	548
Bus Preference	2	100
Confirmation	2	100

IV. Speech Corpus Cleaning And Annotation

The collected corpus was cleaned and verified by a team of expert linguists. The following subsections provide details about the cleaning and quality assurance of the data.

A. Data Cleaning

Conclusive guidelines have been followed for the cleaning of the recorded data. PRAAT (speech processing software) utilities have been used for automatically aligned the transcriptions to the spectrograms of the wave/speech files. Then the each and every file is listened very carefully and manually checking has been done on the crucial issues occurred in the data. These significant environment issues were; noise, silence and alternate pronunciations.

The issue of noise was handled as the file was rejected if the level of noise was overlapping the signal of the speech. A threshold of signal to noise ratio (SNR) is also used for handling this issue. The issue of silence was the difficulty of marking the whole property of phones (voiceless stops and affricates) at the start and at the end of the speech file. It was noted that the issue might be arising from the recording duration in which the user gets less time to completely utter the words. For this reason, the duration of the recording window has been increased to 4 sec and it has found that the problem of silence is lessened due to this amendment.

The third issue was of alternate pronunciation that was the variation of pronunciation among the people. These pronunciations have been transcribed manually by an expert team of linguists. Each and every issue in the data has been marked manually in an excel sheet to keep a log of each variation. Different error codes have also been generated according to the issues occurred in the data.

B. Quality Assessment

To check the quality of the data, gold corpus or reference corpus has been generated on the data marked by the linguist. The gold corpus has also been generated by another linguist. The linguist data was then compared with the gold data to check the mismatches between the both. PRAAT utility has been used to compare the data and then the mismatches have been observed manually by an expert linguist. The reasons of the mismatches have been found out and if the mistake was due to linguist data then the linguist had to correct that

problem. Thus a threshold of 95% was being placed on the cleaned data as to get the inter-annotator accuracy at this level.

V. Application: Bus Reservation System

The speech corpus was used to develop speech recognition system to be used in an Urdu language bus reservation spoken dialog system [11]. It is a mobile based spoken dialog system where users call for reservation of seats in their desired bus. The users can also choose the location of the bus for the 44 cities of Pakistan from Lahore. The system asks the caller questions about his desired destination, date of departure, time of departure and number of seats. Once all the required information is retrieved, the system performs the reservation and informs the user of a reservation number to be used for future reference. For speech recognition, Hidden Markov Model (HMM) based CMU's Sphinx [12] toolkit was used. Training data consisted of 70% of corpus while 30% of corpus was used for testing. Table 3 is showing laboratory testing results of speech recognition system.

Table 3: Laboratory results of speech recognition system

ASR	Training Utterances	Testing Utterances	Accuracy (%age)
Destinations	1738	794	96.98
Day of Reservation	804	340	95.59
Time of Reservation	5246	2245	95.14
Number of Seats	385	163	94.48
Bus Preference	70	30	100
Confirmation	70	30	96.67
Total	8243	3572	95.6

The system was deployed at a landline number and field testing was performed from 22 speakers who were students, shopkeepers and drivers. Table 4 is showing field testing results of the system.

Table 4: Field results of speech recognition system

Total utterances	Correct Decoded	Incorrect Decode	Accuracy (%age)
172	150	22	87.21

VI. Conclusion

Design and development of an Urdu speech corpus for travel domain has been discussed in this paper. All the possible vocabulary items that may be used in a travel domain system have been covered in this corpus. The data collected from speakers was cleaned using comprehensive guidelines. The corpus has been used to develop speech recognition system that resulted in an accuracy of 95.6% in laboratory and 87.21% in the field. The developed corpus can be used to develop bus, train and domestic flight information and reservation systems. The subsets of corpus like days and time can be used in other systems that may involve the information related to date and time. Similarly, the speech data of counting may be used to decode any sequence of numbers like a telephone number, a bank account number or a credit card number. The presented corpus is restricted to the commute or travel domain but for the future prospect, other fields related to tourism; art, culinary and names of the famous places etc. can also be studied.

Acknowledgment

This research was supported by National ICT RnD Fund, Pakistan through the project "Enabling Information Access through Mobile based Dialogue Systems and Screen Readers for Urdu".

References

[1] S. Levinson, D. Davis, S. Slimon and J. Huang, "Articulatory speech synthesis from the fluid dynamics of the vocal apparatus," *Synthesis Lectures on Speech and Audio Processing,* vol. 8, no. 1, pp. 1-116, 2012.

[2] D. Gibbon and R. Winski, Spoken language system and corpus design, Hawthorne, NJ, USA: Walter de Gruyter, 1998.

[3] L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers and J. L. Gauvain, "Development of spoken language corpora for travel information," in *Eurospeech*, Madrid, Spain, 1995.

[4] S. Arora, B. Saxena, K. Arora and S. S. Agarwal, "Hindi ASR for travel domain," in *OCOCOSDA*, Kathmandu, Nepal, 2010.

[5] P. P. Shrishrimal, R. R. Deshmukh and V. B. Waghmare, "Indian language speech database: a review," *International Journal of Computer Applications,* vol. 47, no. 5, pp. 17-21, June 2012.

[6] S. Khan, J. Basu, T. Basu, M. S. Bepari, M. Pal and R. Roy, "Bengali basic travel expression corpus: A statistical analysis," in *OCOCOSDA*, Phuket, Thailand, 2014.

[7] S. Rauf, A. Hameed, T. Habib and S. Hussain, "District names speech corpus for Pakistani languages," in *OCOCOSDA*, Shanghai, China, 2015.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," [Online]. Available: praat.org. [Accessed 27 June 2016].

[9] "Cisco SPA3102 voice gateway with router," Cisco, [Online]. Available: http://www.cisco.com/c/en/us/products/unified-communications/spa3102-voice-gateway-router/index.html. [Accessed 27 June 2016].

[10] "Asterisk.org," Digium, [Online]. Available: asterisk.org. [Accessed 27 June 2016].

[11] M. Qasim, A. Anwar, T. Habib and S. Hussain, "Development of multiple automatic speech recognition systems in the galaxy framework," in *OCOCOSDA*, Shanghai, China, 2015

[12] K. F. Lee, H. L. Hon and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 38, no. 1, pp. 35-45, 1990.