Improving Recognition Accuracy of Urdu Weather Service by Identifying Out-of-Vocabulary Words

Saad Irtza[1], Aneek Anwar[2], Sarmad Hussain[3]

1. Electrical Engineering Department, University of Engineering & Technology, Lahore, Pakistan, saad.irtaza@kics.edu.pk
2. Center for Language Engineering, KICS, University of Engineering & Technology, Lahore, Pakistan, aneek.anwar@kics.edu.pk
3. Center for Language Engineering, KICS, University of Engineering & Technology, Lahore, Pakistan, sarmad.hussain@kics.edu.pk

Abstract

Mobile based dialogue systems in local languages provide a very suitable information delivery channel. Many of the tasks can be addressed by designing and developing small vocabulary systems. However, as small vocabulary systems generally try to match each input word onto one of the words in the vocabulary, if inadvertently out of vocabulary (OOV) words are spoken, they are also mapped onto the closed set of words in vocabulary and reduce the accuracy. The current work addresses this issue. We present the development of mobile based dialogue system in local language (Urdu) to provide weather information to urban and rural populations. Performance of this speaker independent automatic speech recognition system (ASR) is evaluated by offline and online testing. In offline testing, based on unseen dataset limited to the speakers used for training the system, 100% accuracy is achieved. In online testing, 74.79% accuracy is achieved. Analysis shows that a significant reduction in accuracy is caused by out-of-vocabulary words (OOV) spoken by users. Phone-based model is then added to detect and reject OOV words and system accuracy improves to 88.24%.

Key Words: Out of vocabulary words, Urdu dialogue system, Weather information system

1. Introduction

Access to information is vital for socio-economic development in today's age. But there are multiple barriers to access this information for Pakistanis. First, high illiteracy rate of 45% prevents a major part of the population to access information online. Even those who are literate have very limited access to computers (which have only 11% penetration). The solution is to use mobile based speech interfaces as they bypass the literacy barrier and use mobile services, which have 70% penetration. Mobile based dialogue systems are especially useful for rural populations, where illiteracy and lack of internet access is more acute.

A dialogue system takes speech as input from the user, processes it to extract the query been made and generates the appropriate response in the form of speech [1]. The output speech can be pre-recorded or generated at runtime using a text to speech system. A spoken dialogue system can be used for multiple tasks such as remote banking, weather information, travelling reservation, information enquiry, taxi booking, stock transactions and route planning etc. In this paper, we will discuss a limited vocabulary Urdu spoken dialogue system for weather information. The system prompts the user to speak the name of the district for which weather information is desired, recognizes the district name spoken, looks up in the database for weather information of the recognized district name and responds to the user with the relevant information. Users often speak out-of-vocabulary words (OOV) which are recognized incorrectly and consequently incorrect information is provided. The current paper also discusses special measures put in place to address this aspect in this system.

The rest of the paper is organized as follows. First a literature survey is presented for the existing dialogue systems in English and other major languages. Then the architecture of the dialogue system is discussed. Finally, the results of the field testing of the original system and the system enhanced to handle OOVs are discussed.

2. Literature Review

Spoken dialogue systems have been developed on different domains including weather services in different languages, e.g., in English, Dutch, German, Japanese, Chinese, etc. [2, 3, 4, 5, 6, 7, 8]. JUPITER is a telephone based conversational interface for weather information [4]. Its corpus consists of approximately 3500 read utterances collected from a variety of local telephone handsets and recording environments. Its acceptance rate has been found to be 97.2%. Another multilingual spoken dialogue application has been developed to provide pervasive access to weather, wind and water conditions for domestic and international tourists [5]. The accuracy of the system has been found to be varying from 95% to 98%. A Croatian weather domain spoken dialogue system [6] provides information about weather in different regions of Croatia, consisting of 2300 different words. The word error rate (WER) for the weather forecast task has been 20% for the telephone speech.

There is no robust speaker independent automatic speech recognition system in Urdu language that can be integrated with spoken dialogue system. Different domain specific Urdu ASR systems have been developed on limited vocabulary. Artificial neural networks [9] and HMM [10] have been used to develop the systems. An effort has been made to develop speaker independent Urdu speech recognition system [10] on continuous and read speech using HMM technique. In another system, a corpus has been recorded from 82 speakers including 41 males and 41 females. CMU sphinx open source toolkit has been used to develop the three acoustic models on incremental basis. The performance of acoustic models have been evaluated and overall word error rate, on combined speakers, has been found to be 60.2%. Another effort has been made to develop small vocabulary ASR system on ten speakers using CMU sphinx toolkit. Training data consists of 5200 utterances of speech data from ten speakers. The WER comes out to be 5.33% [11]. An experiment has been developed to evaluate the performance of acoustic model by mixing of read and spontaneous speech utterance combined in various ratios of speech corpus [12]. Training data of 800 utterances has been used to develop single speaker medium vocabulary ASR system. Two ASR systems have been developed on phonetically rich corpus to investigate the accuracy issues, one on single speaker [13] and second on ten speakers [14]. Through the recognition accuracy and analysis, it is reported that the amount of training data of each phoneme plays an important role in performance of acoustic model.

The detection of OOV words and misrecognitions in ASR system has been addressed by using many different approaches e.g. by measuring different confidence score of recognized word [15 16 17]. Phone and filler models have been used to detect and measure confidence scores respectively [16]. The likelihood score of a word from ASR decoder and phone model have been used to define threshold for confidence measure of OOV. An approach based on phone posterior probability has been developed to reject OOV words from large vocabulary English ASR systems [17]. Bigram model has been developed to decode in parallel with word level decoding. Viterbi alignment using phone models is used to compute the confidence measure [15]. Word accuracy of such system has been found to be 96.61%. Language model (LM) is also used to detect the OOV words [18].

3. Methodology

The spoken dialogue system developed for Urdu weather service consists of four major components: (1) speech recognizer, (2) dialogue manager, (3) speech synthesizer, and (4) information database. The flow of the system is shown in Figure 2. The user dials the designated number from the phone. The incoming call is received by a VoIP gateway which transfers the call to the server. The server processes the input speech and generates the response which is played back on the same call.

Development of speaker independent automatic speech recognition system (ASR) for spoken dialogue system is a challenging task. There are several variables involved in automatic speech recognition system that affect the performance, e.g. accent of speakers, OOV, age, channel and background noise level. The effect of these variables becomes prominent when ASR is used in spoken dialogue system.

The development of Urdu spoken dialogue system on weather domain over mobile channel has been divided in two phases. In first phase (Experiment-1), baseline system has been developed on 19 districts of Punjab, Pakistan. Performance of the system has been evaluated by using field testing.

In second phase (Experiment-2), ASR system has been improved by resolving the issues recognized in field testing.

In Experiment-1, the ASR system has been developed on the data presented in Table-1. The training data has been recorded in an office environment through mobile phone to include small variation of background noise. Testing of the system has been performed in two stages. In the first stage, the system has been tested in a normal office environment with little background noise. In the second stage, field testing has been done in the open environment with varying levels of background noise. In both cases performance of ASR system has degraded due to out of vocabulary words and low signal to noise ratio (SNR) of spoken words.

Table 1 Data for ASR System

	Data of Experiment-1	Data of Experiment-2
No of speakers	80	
Vocabulary size	19 District Names	
Per word training data	15	
Total training data	21216	
Offline testing data of ASR	1476	
Online testing data	246	613

In Experiment-2, all-phone model has been developed to filter the OOV words [15]. Two ASRs have been developed, including one which decodes at word level and the other which decodes at the phone level. The output string of the two decoders is compared and based on their percentage match, it is decided whether the decoded word is OOV. The performance of improved system has been evaluated on data presented in Table-1. Districts names and their corresponding pronunciation is given in Appendix-A.

Training data has been collected on mobile phones. Praat has been used to clean and transcribe the data. ASR has been developed using CMU sphinx open source toolkit. Centos OS and asterisk platform have been used to develop dialogue manager. Fig. 1 shows the architecture of complete dialogue system.

4. Speech Corpus Collection

A dialogue system has been developed to record the speech data from different mobile phones. Figure 1 shows the recording dialogue.

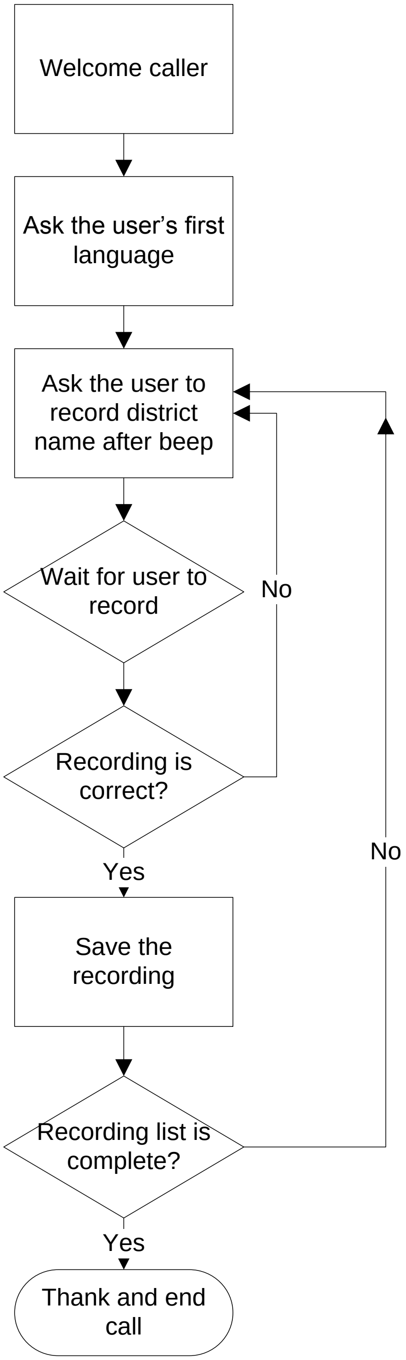

Fig.1 Flow Chart of Recording Dialogue System

Speech data has been collected largely from the students of University of Engineering & Technology, Lahore. Table 2 shows the recordings details.

Table 2 Number of Recordings of Speech Data

City Name of User	Number of Recordings
Lahore	49
Nankana Sahib	1
Jhang	1
Toba Tek Singh	2
Sahiwal	1
Gujranwala	4
Vehari	1
Sargodha	2
Chakwal	1
Faisalabad	4
Multan	7
Rajanpur	1
Attock	1
Sialkot	1
Rahim Yar Khan	1
Bahawalpur	1
Rawalpindi	1
Khanewal	1
Total Number of Speaker	80

5. Spoken Dialogue System Architecture

The architectural diagram of Urdu weather spoken dialogue system is shown in Figure 2. This is a centralized framework and flow of dialogue is controlled by dialogue manager unit. Call management is designed on asterisk framework. LinkSys SPA4000 Voip gateway is used as a communication box between end user and asterisk. Query from user is sent to ASR module which decodes the district name. Weather information of specific district is retrieved from weather database. Text to speech conversion is done by text to speech module which combines the pre synthesized units to formulate complete response. Figure 3 shows the dialogue of weather information retrieval system.

6. Experiment-1 Results and Discussion

Performance of ASR system has been evaluated in two ways: first by selecting non-overlapping testing data set (offline testing) and second after integration with spoken dialogue system (online

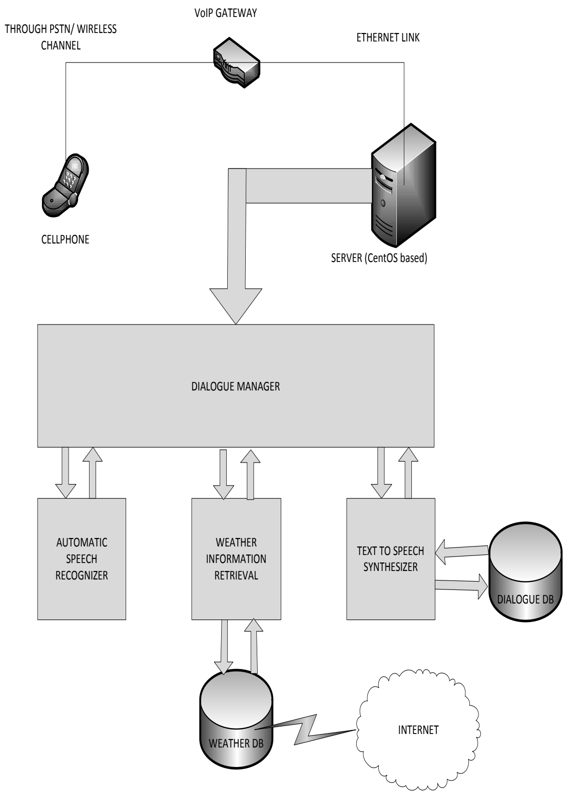
THROUGH PSTN/ WIRELESS CHANNEL
VoIP GATEWAY
ETHERNET LINK
CELLPHONE
SERVER (CentOS based)
DIALOGUE MANAGER
AUTOMATIC SPEECH RECOGNIZER
WEATHER INFORMATION RETRIEVAL
TEXT TO SPEECH SYNTHESIZER
DIALOGUE DB
WEATHER DB
INTERNET

Fig.2 Architecture of the Mobile Based Urdu Speech Dialogue System for Weather Service

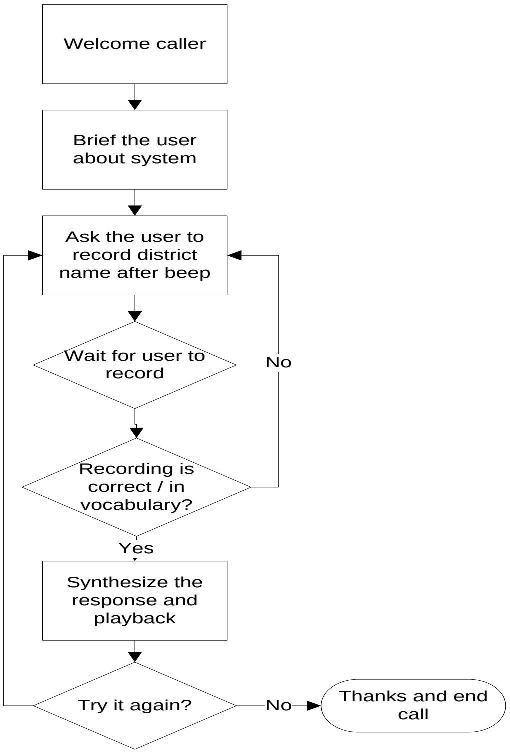
Welcome caller
Brief the user about system
Ask the user to record district name after beep
Wait for user to record
No
Recording is correct / in vocabulary?
Yes
Synthesize the response and playback
Try it again?
No
Thanks and end call

Fig.3 Flow Chart of Weather Information Retrieval System

testing). There are no recognition issues in offline decoding of ASR system and 100% accuracy has been achieved. Accuracy of ASR system is significantly reduced to 74.79% after integration with spoken dialogue system. The testing data is analyzed to investigate the recognition issues. The field testing data shows that performance of dialogue system largely depends on user input. Out of 50 words spoken, 28 are OOV and 11 are with very high background noise. Recognition accuracy is significantly affected by these two reasons. Example of an OOV is that the user has spoken “Paris” as a possible city/district name, which is not in the vocabulary of the system. Closest match is found by the system, causing an error. Classroom recordings have high background noise that causes mis-recognition.

7. Experiment-2 Results and Discussion

In order to improve the performance of dialogue system, there is a need to find the confidence measure of district name to reject mis-recognized and out-of-vocabulary words. In this regard, an all-phone model [15] has been developed on the same data presented in Table-3.

Table 3 Example of Detection of OOV Words

City name spoken by user	ASR Word model output	ASR Phone model output	Phone-Word model comparison	OOV Decision
Paris	Attak	k ae p i s	1/5=20%	OOV
Islamabad	Xuushaab	s aan b a d_d_h	2/8=25%	OOV
Attak	Attak	a a tt a p	4/5=80%	Not OOV
Mianwali	Miiaanvaali	m n m ii aan v aa l t i	7/8=87%	Not OOV
Bhakar	Bhakar	b b_h a k a l	4/5=80%	Not OOV
Pakistan	Multan	t a k ii s t_t a m	1/6=16.7%	OOV
Lahore	Lahore	l aa h o r r	5/6=83.3%	Not OOV

In this experiment, two ASR systems (word level and all phone level) are used concurrently (in parallel). Speech input from user is decoded from the two ASR systems. Since phone-level decoding is highly sensitive to any noise in the input speech, it may skip some phonemes or include some erroneous ones, e.g. if speaker does not speak all the phonemes in the word, the phone level decoding will output only the spoken phonemes whereas word level decoding will give a word consisting of all the phonemes. So there is not always one-to-one matching for the two decoded strings; simply comparing them and rejecting words which do not match the phone-level decoding degrades the system accuracy. Thus, the maximum overlapping of two decoded strings is determined. If the two strings overlap within a certain threshold they are considered accurate, and the corresponding word level output is considered correct. Else, the word is considered to be OOV and rejected.

The performance of system is evaluated by using fixed and different threshold levels depending on length of words. Figure-4 shows the graph between accuracy of the system (which includes the recognition accuracy of ASR and rejection accuracy of OOV detector) and threshold value. It can be seen that the peak value of system accuracy is 77%, which is quite low. Analysis of individual words shows that the large value of threshold is good for matching words of larger length (more than 6 letters) but it is too high for words of small length (6 letters or less)

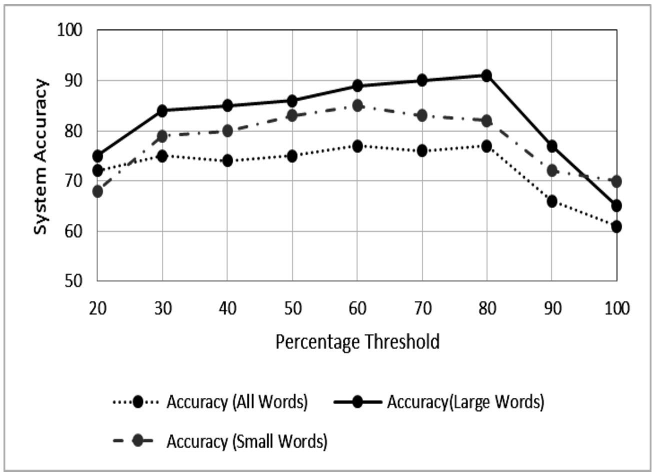

Fig.4 System accuracy for various thresholds

and often they are considered as OOV and rejected; hence overall accuracy drops. There is no single value of threshold which satisfies both conditions. Table-2 shows some real output examples.

Thus, separate thresholds are developed for the two categories of words. Optimum values come out to be 80% overlap for longer words and 60% overlap for smaller words. The system accuracy is found to be 91% for large words and 85% for small words at these threshold values. The cumulative accuracy of the system for both categories turns out to be 88.24%. These results are summarized in Table-4.

Table 4 Summary of Accuracy Results

Experiment-1	Experiment-2	
	Single threshold	Multiple thresholds
74.79%	78%	88%

8. Conclusions

Using phone model concurrently with the word model developed on the same data set significantly improves the detection of OOV words and reduces misrecognition. Filtering the OOV words improves the accuracy of ASR system and the overall performance of dialogue system. It is shown that using two thresholds for different length of words give better results instead of using single threshold. For larger vocabulary systems, optimum number of thresholds can be determined by analyzing the length of words in dictionary.

References

[1] Huang, X., Acero, A. and Hon, H.W.: 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall.

[2] Chen, J., Wu, J. and Wang, Z.: 2003. A Chinese spoken dialogue system for train information. In: Proc of IEEE SMC'2003 [C], Washington D.C., USA, (EI: 2003487750883, ISTP: BX83D).

[3] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T.J. and Hetherington, L.: JUPITER: 2000. A telephone-based conversational interface for weather information. In: *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 1.

[4] Bick, E. and Hansen, J.A.: 2007. The Fyntour Multilingual Weather and Sea Dialogue System. In: Ron Artstein and Laure Vieu (eds.), Proceedings of DECALOG - The 2007 Workshop on the Semantics and Pragmatics of Dialogue, May 30 – 1, pp. 157-158.

[5] Mestrovic, A., Bernic, L., Pobar, M., Ipsic, S.M. and Ipsic, I.: 2010. A Croatian Weather Domain Spoken Dialog System Prototype. In: Journal of Computing and Information Technology - CIT 18, 4, 309–316 doi:10.2498/cit.1001916.

[6] Eckert, W., Kuhn, T., Niemann, H., Rieck, S., Scheuer, A. and Schukat-Talamazzini, E.G.: 1993. A spoken dialogue system for German intercity train timetable inquiries. In: EUROSPEECH, Berlin, 1993, pp. 129-132.

[7] Baggia,P., Kellner,A., Prennou,G., Popovici, C., Sturm, J. and Wessel, F. 1999. Language Modelling and Spoken Dialogue Systems - the ARISE experience. In: EUROSPEECH.

[8] Narayanan, S., Ananthakrishnan, S., Belvin, R., Ettaile, E., Gandhe, S., Ganjavi, S., Georgiou, P. G., Hein, C. M., Kadambe, S., Knight, K., Marcu, D., Neely, H. E., Srinivasamurthy, N., Traum, D. and Wang, D.: 2004. The transonics spoken dialogue translator: An aid for English-Persian doctor-patient interviews. In: AAAI Fall Symposium.

[9] Akram, M.U. and Arif, M.: 2004. Design of an Urdu Speech Recognizer based upon acoustic phonetic modelling approach. In: IEEE INMIC 2004, pp. 91-96, 24-26.

[10] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, R. Parveen, 2010. "Large Vocabulary Continuous Speech Recognition for Urdu", in the *Proceedings of International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, 21-23.

[11] Ashraf, J., Iqbal, N., Khattak, N.S. and Zaidi, A.M.: 2010. Speaker Independent Urdu Speech

Recognition Using HMM. In: INFOS, IEEE, Cairo, 28-30.

[12] A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah and Z. Sarfraz, 2010. “An ASR System for Spontaneous Urdu Speech”, *In the Proc. of Oriental COCOSDA*, Kathmandu, Nepal. 24-25.

[13] Irtza, S. and Hussain, S.: 2012. Error Analysis of Single Speaker Urdu Speech Recognition System. In: CLT-12, University of Engineering and Technology, Lahore, Pakistan.

[14] Irtza, S. and Hussain, S. 2013. "Minimally Balanced Corpus for Speech Recognition", in the Proceedings of 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13), IEEE, Sharjah.

[15] Young, S. R.: 1994. Recognition Confidence Measures: Detection of Misrecognitions, and Out-Of-Vocabulary Words. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-94, Adelaide, Australia.

[16] S. Maria da Feira, 2009. "Out-Of-Vocabulary and Confidence Measures for Speech Recognition Using Phone Models", Proc *Conf. on Telecommunications* - *ConfTele* , Portugal, Vol. 1 , pp. 457 - 460.

[17] Kombrink, S. Burget, L., Matejka, P., Karafiat, M., Heřmansky, 2009. "Posterior-based Out of Vocabulary Word Detection in Telephone Speech", In: Proc. Of INTERSPEECH 2009, Brighton, GB, ISCA, p. 80-83, ISSN 1990-9772.

[18] M. Thomae, T. Fábián, R. Lieb, and G. Ruske, 2005. "Lexical out-of-vocabulary models for one-stage speech interpretation", In: Proc. of INTERSPEECH, pp.441-444.

Appendix-A

Vocabulary and Phone Specification of the Dialogue System

District Name	Lexical Entry (Pronunciation)
ATTAK	A TT A K
BAHAWALPUR	B A H AA V A L P U R
BHAKAR	B_H A K A R
CHAKWAL	T_SH A K V AA L
FAYSLABAD	F AY S L A B AA D_D
GUJRANWALAN	G U D_ZZ R AAN V AA L AAN
GUJRAT	G U D_ZZ R AA TT
JHANG	D_ZZ_H A NG
JEHLAM	D_ZZ AE H L A M
KASUR	K A S UU R
KHUSHAB	X UU SH AA B
LAHORE	L AA H O R
MIANWALI	M II AAN V AA L I
MULTAN	M U L T_D AA N
RAWALPINDI	R AA V A L P I N DD II
SAHIWAL	S AA H II V AA L
SARGODHA	S A R G OO D_D_H AA
SHEIKHUPURA	SH AE X UU P U R AA
SIALKOT	S I A L K OO TT