DISTRICT NAMES SPEECH CORPUS FOR URDU ASR

Presenter: Sahar Rauf

Center for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology Lahore, Pakistan

Contents

- Purpose of the work
- Introduction
- Literature review
- District names speech corpus
- Conclusion and future work

Purpose of the Work

- To develop a district names speech corpus for Urdu ASR.
- To collect speech data from different districts of Pakistan.
- To capture accent variations of Urdu.
- To address different environmental issues for the cleaning of the speech data.

Introduction (1/2)

- Speech corpus is a collection of audio recordings that is a necessary element to build the ASR systems.
- The speech corpus can be a good source of capturing variability occurred due to age, gender, dialect, background noise and language of a speaker[1].
- ASR systems are specifically developed to recognize the speech of a person speaking in a microphone or over a telephone channel and convert the speech into another medium[2].

[1] Y. K. Muthusamy et al., "Reviewing Automatic Language Identification," *Signal Processing Magazine, IEEE*, vol. 11, no. 4, pp. 33-41, Oct. 1994.

[2] P. Saini and P. Kaur, "Automatic Speech Recognition: A Review," International Journal of Engineering Trends and Technology, vol. 4, no. 2, pp. 1-5, 2013.

Introduction (2/2)

- ASR is a promising field of research and a part of services related to healthcare, agriculture, weather forecasting and mobile applications[3].

- The proposed corpus is specifically designed to build an Urdu ASR for a mobile based Urdu dialog system to provide weather information of Pakistan. This service is deployed at Pakistan Meteorological Department (PMD), Islamabad[4].

[3] P. Saini et al., "Hindi Automatic Speech Recognition Using HTK," International Journal of Engineering Trends and Technology, vol. 4, no. 6, pp. 1-7, June 2013.

[4] +92519250363, Service no. at PMD.

Literature Review (1/2)

- Different kinds of speech corpora are being developed in many languages such as isolated words[5] and continuous speech[6] in the field of ASR and natural language processing (NLP)[7].

- In recent years, many Urdu ASR systems have been proposed. These systems have been designed for limited vocabulary, large vocabulary, read Urdu speech, spontaneous Urdu speech, and for continuous speech.

[5] G. Raskinis, "Building Medium-Vocabulary Isolated Word Lithuanian HMM Speech Recognition System," Information Journal, vol. 14, pp. 75-84, 2003.

[6] H. Sarfraz et. al., "Large Vocabulary Continuous Speech Recognition for Urdu," in the Proc. International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, Dec. 2010, pp. 1-5.

[7] J. Ashraf et. al., "Speaker Independent Urdu Speech Recognition Using HMM," in International Conference on Informatics and Systems, Cairo, Egypt, Mar. 28-30, 2010, pp. 1-5.

Literature Review (2/2)

- For developing these Urdu speech corpora, phonemic transcriptions generated from Urdu orthography[6] and phonetic lexicon[8] have been used.

- Issues related to accent variation or alternate pronunciations are not discussed in the development of these corpora. Thus, the presented work aims to provide inclusive guidelines for the pre-processing of new speech corpus and developing new transcriptions according to the pronunciation variations in the data.

[6] H. Sarfraz et. al., "Large Vocabulary Continuous Speech Recognition for Urdu," in the Proc. International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, Dec. 2010, pp. 1-5.

[8] A. A. Raza et al., "An ASR System for Spontaneous Urdu Speech," in the Proc. Oriental COCOSDA, Nepal, 2010, pp. 1-6.

District Names Speech Corpus for Urdu ASR (1/2)

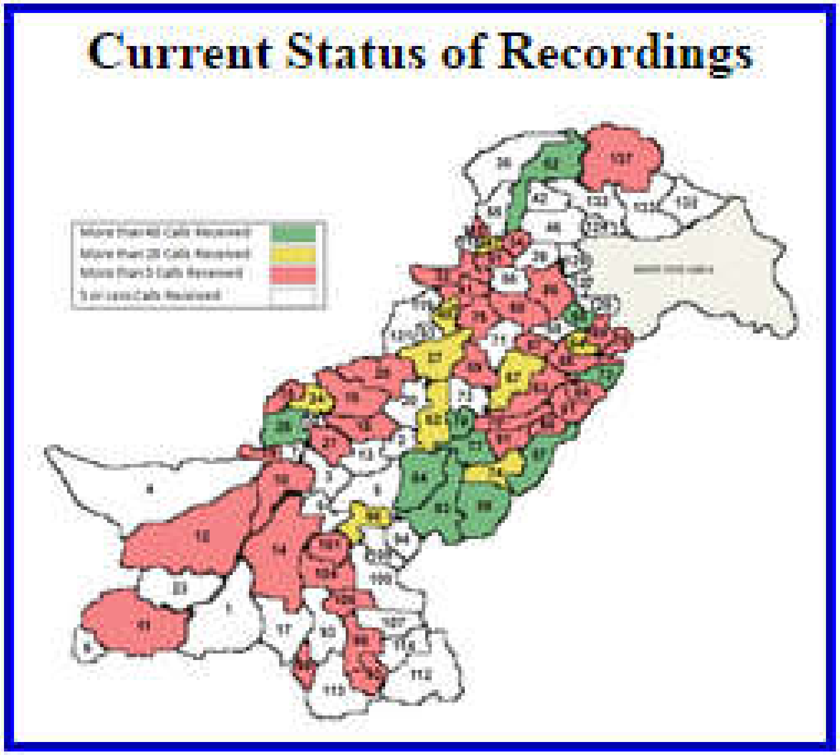
- It is a collection of single word utterances fixed vocabulary.
- Annotated at word level using CISAMPA which is directly mapped on Urdu IPA symbols[9].

Channel	Mobile phones/ All telecom operators of Pakistan
Environment	Natural
Sampling Rate	8 kHz
Vocabulary Items	139 district names, 34 vocabulary items; days, date, no & affirmation (yes/no)
Speaker's Education Level	Semi literate to literate
Speaker's Age	18-50

[9] B. Mumtaz et al., "Multitier Annotation of Urdu Speech Corpus," in the Proc. Conference on Language and Technology (CLT14), Karachi, Pakistan, Nov. 13-15, 2014, pp. 1-8.

District Names Speech Corpus for Urdu ASR (2/2)

- The data is recorded from different districts of Pakistan.
- The data is collected from 6 major accents; Urdu, Punjabi, Pashto, Sindhi, Saraiki and Balochi.
- A Pakistan map is presented on the CLE website that presents the color coded information of different districts from where the data was collected[10].


Current Status of Recordings
More than 40 Calls Received 🟩
More than 20 Calls Received 🟨
More than 5 Calls Received 🟥
5 or Less Calls Received □

[10] Center for Language Engineering. [Online]. http://cle.org.pk/dialog1/images/pakistan-district.gif

Data Verification

The first step includes the verification of the information used for labeling/naming of the speech files. The information includes;

- Number of the speaker
- Number of the district speaker belongs to
- Speaker’s mother language
- Speaker’s gender
- Number of the district spoken
- Information of the version.

Thus the naming convention of a file would be **Sp1100_z025_pun_M_dt001_ver01**.

Corpus Cleaning (First Pass)

Accents	No. of speakers			File count
	Males	Females	Total	
Saraiki	415	316	731	12260
Pashto	441	99	540	8576
Punjabi	408	219	627	4989
Urdu	164	202	366	3853
Balochi	24	21	45	693
Sindhi	25	17	42	424

Corpus Cleaning (Second Pass)

Accents	No. of speakers			File count
	Males	Females	Total	
Saraiki	415	316	731	12204
Balochi	645	78	723	8849
Pashto	440	99	539	7936
Sindhi	314	88	402	5718
Punjabi	407	218	625	4499
Urdu	163	202	365	3428

District Coverage Saraiki Accent

Bahawalpur	Rahimyar Khan	Multan	Muzzafargarh	Rajanpur
Lodhran	Dera Ghazi Khan	Vehari	Chakwal	Lahore
Tobataik Singh	Mianwali	Muzzafarabad	Quetta	Loralai
Layya	Bhakkar	Jaccobabad	Dera Ismail Khan	Barkhan

District Coverage
Balochi Accent

Qalat	Noshki	Khuzdar	Kharan	Quetta
Panjgur	Dera Ghazi Khan	Chaghi	Mastung	Washuq
Nasirabad	Kech	Lasbela	Awaran	Sibbi
Jaffrabad	Junubi-waziristan	Jacobabad	Loralai	

District Coverage
Pashto Accent

Swat	Quetta	Pishawar	Loirdeer	Pishin
Mardan	Karak	Swabi	Bannu	Mansehra
Malakand	Kohat	Charsadda	Qilla Abdullah	Nowshehra
Loralai	Bunair	Dera Ismail Khan	Bajor	Zhob
Lakki Marwat	Tank	Ziarat	Qilla Saifullah	Khaibar
Harnai	Musakhail	Qurram	Nowshehro feroz	Junubi Waziristan

District Coverage
Sindhi Accent

Hedrabad	Sanghar	Badin	Tharparkar	Shahid Benazirabad
Khairpur	Jamshoro	Dadu	Sakhar	Larkana
Umarkot	Shikarpur	Ghotki	Qambar Shahdadkot	Karachi
Okara				

District Coverage
Punjabi Accent

Bahawalnagar	Quetta	Lahore	Jhang	Gujranwala
Gujrat	Shaikhupura	Sargodha	Pakpatan	Okara
Sialkot	Sahiwal	Narowal	Faisalabad	Rahimyar Khan
Khaniwal	Chakwal	Bahawalpur	Wihari	Tobataik Singh
Qasur	Hafizabad	Rawalpindi	Nankana Sahib	Layya
Khoshab	Chinjot	Muzaffarabad	Multan	Mianwali

District Coverage Urdu Accent

Lahore	Quetta	Bahawal nagar	Pishawar	Gujranwala
Gujrat	Faislabad	Karachi	Rawalpindi	Qasuur
Okara	Rahimyar khan	Sialkot	Jhang	Jehlam
Sargodha	Sahikhupura	Haidrabad	Zhob	Kohat
Bahawalpur	Chiniot	Dera Ghazi Khan	Miawali	Jaffrabad
Khuzdar	Musakhail	Nasirabad	Noshki	Pishin

Challenges

The data is collected in challenging acoustic environments; the major issues that can affect the accuracy of the system are;

- Silence
- Background noise
- Alternate pronunciations

Conclusion and Future Work

- The current work describes the development and the use of Urdu district names speech corpus. .
- 95% inter-annotator accuracy has been achieved at this data.
- The further perspective is to develop CSR in different domain; weather, flood etc.
- Moreover, the presented work would be helpful in developing speech corpus for ASR's of other Pakistani languages.

Thank You

Any Questions?