DISTRICT NAMES SPEECH CORPUS FOR PAKISTANI LANGUAGES

Sahar Rauf, Asima Hameed, Tania Habib, Sarmad Hussain

Center for Language Engineering,
Al-Khawarizmi Institute of Compute Science,
University of Engineering and Technology,
Lahore
firstname.lastname@kics.edu.pk

ABSTRACT

This paper presents a speech corpus that is developed for Urdu automatic speech recognition (ASR) system. The corpus comprises of single word utterances fixed vocabulary consisting of district names of Pakistan. The data is recorded over a telephone channel from all over Pakistan to cover six major accents; Punjabi, Urdu, Saraiki, Pashto, Sindhi, and Balochi. The data was collected in challenging acoustic environments; the major issues were silence, background noise and alternate pronunciations, which can affect the performance of the system. In order to address these issues, comprehensive data verification and cleaning guidelines are presented. The proposed process serves as a data pre-processing step for the development of ASR, which is successfully integrated in an Urdu dialog system to provide weather information of Pakistan.

Index Terms— Automatic speech recognition system, annotated speech corpus, alternate pronunciation, six major accents

1. INTRODUCTION

Speech corpus is a collection of audio recordings that is a necessary element to build the ASR systems. ASR systems are developed to recognize the speech of a person speaking in a microphone or over a telephone channel and convert the speech into text [1]. The speech corpus can be a good source of capturing variability occurred due to age, gender, dialect, background noise and language of a speaker [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. The target user group is semi or non-literate people of Pakistan who are unable to access online information.

Pakistan is a country blessed with a variety of languages. Six major languages; Urdu, Punjabi, Pashto, Balochi, Sindhi and Saraiki along with 58 other languages are spoken in Pakistan [4]. Urdu is a national language with 11 million mother-tongue speakers and 105 million second language speakers. District names speech corpus is specifically designed to capture the accent variation of Urdu spoken in different areas of Pakistan. The data has been recorded through a telephonic channel. It is necessary to carefully handle the data on specific parameters for better recognition results. A Pakistan map is presented on the Center for Language Engineering (CLE) website that presents the color coded information of different districts from where the data was collected [5].

The organization of the paper is as follows: Section 2 overviews different speech corpora developed for different languages of the world, a detailed description of the proposed speech corpus and its cleaning process is presented in Section 3, Section 4 describes the procedure for the testing of data that is required to achieve the inter-annotator accuracy and finally, conclusion and dimensions for the future work are presented in Section 5.

2. LITERATURE REVIEW

Different kinds of speech corpora are being developed in many languages such as isolated words [6], continuous speech [7] in the field of ASR and natural language processing (NLP) [8]. Different acoustic models have been developed to minimize the cost rate, time and energy in developing the ASR systems. These models somehow have been successful for best recognition. Different Algorithms have been used for noise removal at surface level. However, no method is defined yet to deal with pronunciation errors, misplaced word boundaries, and different kinds of noise.

OGI telephonic speech corpus is a multi-language corpus used for automatic multi-language identification [9]. The data has been recorded through telephonic lines from the speakers of different languages including; Korean, Mandarin, Spanish and English etc. Development of this corpus includes; preliminary verifications (listening and deleting invalid calls), chopping (removing excess noise), evaluation (judgments on quality of speech) and assigning broad phonetic transcription. This data is consisted of total 2485calls.

978-1-4673-8279-3/15/$31.00 ©2015 IEEE

Telephonic speech corpus developed at Center for Spoken Language Understanding (CSLU) provides a source for researches on alphabet, word and large vocabulary identification and recognition of yes or no words [10]. The corpus is transcribed by 5-10 trained transcribers that specifically capture issues related to noise, silence and pauses. In spite of this corpus, different types of corpora have been developed at CSLU. These corpora include; spelled and spoken names corpus (3667 calls), stories corpus (50 sec of spontaneous speech), 21 language corpus from 200 native speakers, English census corpus from Census Bureau employees including their family members and family friends, cellular words, numbers and alphabet corpus (600 calls), OPERA corpus (10,000 different numbers) [11].

Thai speech corpus consisted of 5,000 vocabulary words has also been developed for ASR [12]. Good performance speech recognition systems even with the best algorithms cannot perform better with poor corpus. This corpus is recorded from 248 speakers and manually handled by the linguists. The vocabulary set has been used for Thai language construction. During the corpus development, automatic segmentation tool, automatic sentences distributor, and wave cutting tool have been used.

A Japanese corpus of spontaneous speech has been developed for linguistic/phonetic and NLP research and for ASR. The basic motivation of this corpus is to cover 800-1000 hour spontaneous speech with morphological transcription. The data is recorded at 48 kHz sampling frequency and 16 bit quantization [13]. A read speech corpus of Texas Instruments/Massachusetts Institute of Technology (TIMIT) is developed to cover 8 dialects of American English consisting of 630 speakers. The data is consisted of correct and word aligned transcriptions and phonemic dictionary [14].

In recent years, many Urdu ASR systems have been proposed [15]. These Urdu ASR systems have been designed for limited vocabulary, large vocabulary, read Urdu speech, spontaneous Urdu speech, and for continuous speech. A large vocabulary (LVASR) is developed for spontaneous Urdu speech where read and spontaneous both speech data are used in training [15]. A Large Vocabulary Continuous Speech Recognition is another Urdu dialogue based system emphasizing on robust ASR [7]. Another work in the field of Urdu speech recognition is the method to collect minimally balanced speech corpus [16].

For developing these Urdu speech corpora, phonemic transcriptions generated from Urdu orthography [7] and phonetic lexicon [15] have been used. It has been observed that the recognition results can be improved by refining the transcriptions. Different issues specifically related to accent variation or alternate pronunciations are not discussed in the development of these corpora. Therefore, this paper 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.

3. SPEECH CORPUS

To develop an Urdu speech corpus, 139 district names of Pakistan [17] and 34 vocabulary items specific to the weather domain e.g., time, days, numbers have been used. The data is recorded through a telephonic channel at 8000 Hz and 16 bit digitization rate. The signal to noise ratio (SNR) varies for the recording purposes but it is preferably more than 20 dB. The data is recorded to cover accent variations of Urdu. The purpose is to collect data from those regions of Pakistan where Urdu and its 5 major accents; Punjabi, Pashto, Sindhi, Saraiki and Balochi have been spoken. The data is collected from around 300 speakers ranging from 18 to 50 years of age and their education level varies from semi-literate to literate.

PRAAT (Language Processing Software) has been used for processing the speech files. The data is annotated at word level using CISAMPA which is directly mapped on the Urdu IPA symbols [18]. A complete detail of duration of cleaned speech in hours is presented in Table 1. The methodology for processing the speech files is discussed in later sections.

Table 1: Duration of clean speech in hours

Accents	Duration of speech		
	Hours	Minutes	Seconds
Pashto	2	16	34
Saraiki	2	35	14
Balochi	1	57	56
Punjabi	1	19	30
Urdu	0	58	31
Sindhi	0	7	42
Others	2	45	9

3.1. Data verification

The first step includes the verification of the information used for labeling/naming of the speech files. During the recording, speaker has been asked different questions to get his/her personal information. Subsequently, the recorded speech file is labeled using information such as; number of the speaker, number of the district speaker belongs to, speaker's mother language, speaker's gender, number of the district spoken and number of data cleaning cycle represented by version. Therefore, the final convention of the file would be sp1100_z025_pun_M_dt001_ver01.

3.2. Labeling and cleaning process

A PRAAT utility is used to automatically label the speech files with CISAMPA transcription. Moreover, utility generates two folders; clean and incorrect and then files are moved into these folders after carefully analyzed by the expert linguists, as shown in Figure 1.

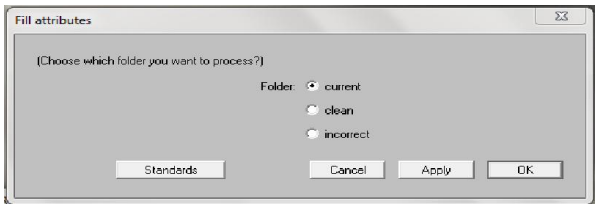

Figure 1 PRAAT utility for cleaning process

Three main aspects are considered in cleaning of the speech corpus. These are; silence, noise and pronunciation. Silence or closure period is only required in the case of consonant stops and affricates. At the word initial position, 100 ms of closure period is given before the burst of stops and affricates [19]. 77 ms of closure period is given to the stop when its burst is not visible at word final position [19]. In Figure 2, a closure period of 100 ms is given to stop /تھ/ /ʈʰ/ /t/ at word initial position.

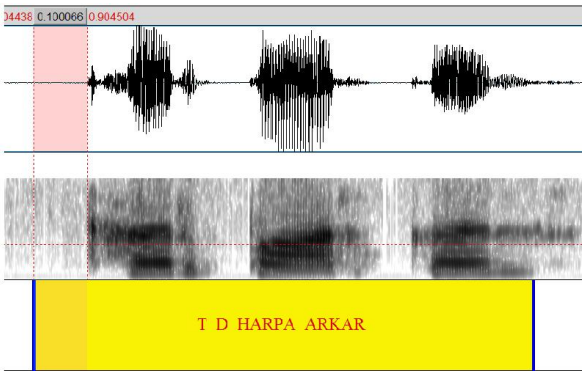

Figure 2 Silence marking in /تھرپارکر/ /ʈʰərpa:rkər/ /Tharparker/

Another important issue is noise; street noise, babble and traffic noise. Noisiness can affect the data accuracy. 10 dB threshold is set to check the level of noise when it is disturbing the speech signal. PRAAT utility also provides the information of SNR and when it is less than 10 dB than the file is discarded, as Figure 3 is showing SNR value which is 28 dB so the file will be processed.

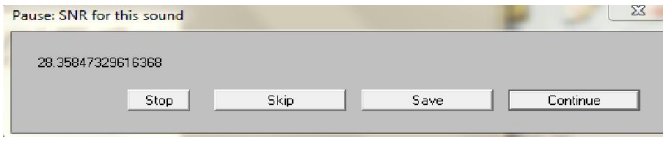

Figure 3 Utility showing SNR value

Third important aspect is pronunciation that includes; mispronunciation and alternate pronunciation. The file is discarded on mispronunciation of a word. Alternate pronunciations are those pronunciations which are acceptable because of general trend in accent variation e.g. Figure 4 shows that people from different accents replace /جعفرآباد/ /ʤa:fəra:ba:ɖ/ /Jafarabad/ with /جعفرآباد/ / ʤa:fra:ba:ɖ/ /Jafrabad/.

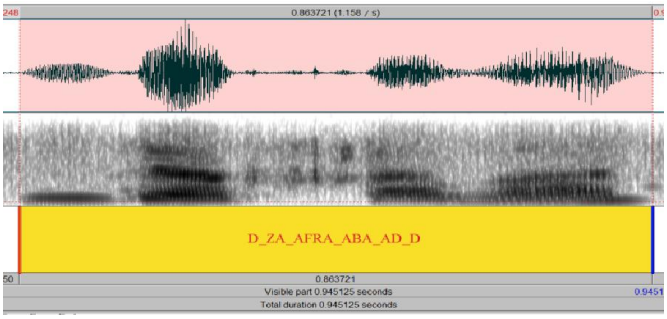

Figure 4 Jafrabad: an alternate pronunciation of Jafarabad

Variations in data can be occurred due to variations of consonants and vowels. Consonantal and vowel substitution, addition and deletion are the important causes of variations or errors. As the Figure 5 is describing the variation in vowel, people from different accents replace /استوُر/ /əsʈu:r/ /Astur/ into /استور/ /əsʈo:r/ /Astor/.

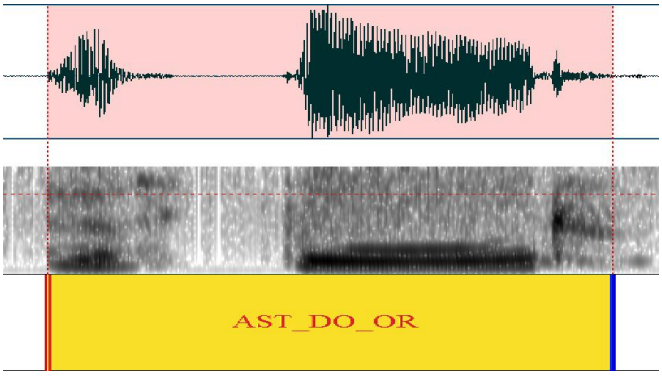

Figure 5 Astor: an alternate pronunciation of Astur

3.3. Description of error flags

Information regarding variations in the data has been logged in an excel sheet. Main flags and sub flags are used to mark

the variations occurred in the data. Six main flags; IF (incorrect file name), SIL (issues related to silence), NSE (issues related to the noise), AP (alternate pronunciation) and RM (removed files) are used for marking. A complete description of error flags is given in Table 2.

Table 2: Description of error flags

No	Main flags		Description of sub flags	
1	IF		Incorrect file name	
2	SIL	Silence	1. SLB	Silence less than the defined threshold at beginning
			2. SLE	Silence less than the defined threshold at end
3	NSE	Noise	3. NSB	Noise found at the background
4	AP	Alternate pronunciation	4. CSP/ M	Consonant substitution on the basis of place or manner of articulation
			5. VS	Vowel substitution
			6. VDM	Vowel deletion at middle position
			7. VI	Vowel insertion
5	DLY	Delayed due to pronunciation		
6	RM	Removed	8. CSD	Consonant substitution discarded
			9. CDD	Consonant deletion discarded
			10. CID	Consonant insertion discarded
			11. VSD	Vowel substitution discarded
			12. VDD	Vowel deletion discarded
			13. VID	Vowel insertion discarded
			14. EM	(Empty) Files contain nothing
			15. OOV	Out of vocabulary word
			16. SNR	Signal to noise ratio discarded

IF flag is used for those files in which the vocabulary codes of the files do not match with the spoken utterances. The codes are then corrected after carefully analyzing the vocabulary code lists. Sub flags of SIL are used when the silence period according to the defined threshold is not found in the files. When the variations in the speech corpus are accepted these are assigned with the AP and its sub flags. CSP/M is used when the consonant substitution due to place or manner is accepted e.g. /شیخُوپُورہ/ /ʃe:xu:pu:ra:/ /Sheikhupura/ into /شیکھُوپُورہ/ /ʃe:kʰu:pu:ra:/ /Sheikhupura/, the variation of /خ/ /x/ /kh/ into /کھ/ /kʰ/ /kh/ is accepted in this example. VS, VDM and VI are used for the accepted variations of substitution of vowel as /قِلات/ /kɪla:ʈ/ /Kilat/ into /قلات/ /kəla:ʈ/ /Kalat/, deletion of vowel at middle position as /جعفرآباد/ /ʤa:fəra:ba:ɖ/ /Jafarabad/ into /جعفرآباد/ ʤa:fra:ba:ɖ/ /Jafrabad/ and insertion of vowel as /لودھراں/ /lo:ɖʰrɑ̃/ /Lodhran/ into /لودھراں/ /lo:ɖʰərɑ̃/ /Lodharan/ respectively. DLY flag is assigned to the file where a delay is found within a compound word e.g. /قلعہ عبداللہ/ /qɪlɑ: əbɖulla:h/ /Qilla Abullah/**.** All the consonant and vowel substitutions, deletions and insertions which are not accepted are logged with RM flag and their respected sub flags.

4. QUALITY ASSESSMENT

Gold corpus has been generated by reference sources to assess the quality of the ASR data. For the testing purpose, gold corpus has been made up of 100% of the source data. The testing of the data consists of two steps;

- At first step, log files generated by the source and the reference are compared.
- At second step, utility checks the mismatches of boundary marking which are than manually analyzed.

A utility is used to compare the log files that check the mismatches of error flags among the log files. The mismatched files are then moved to another folder by the utility and then manually checked by the linguist to find out the reasons of the mismatches. A 95% similarity score thus been placed on the comparison of the log files to achieve inter-annotator accuracy.

It is also difficult to judge a pronunciation as a mispronunciation or alternate pronunciation. To cater this issue, it is decided to mark 5 similar pronunciation variations as alternate pronunciation and discard those variations which are less than 5. In spite of that, standard transcriptions written by the expert Urdu linguist are used for marking the data. Standard transcriptions help to differentiate between the variations in pronunciations.

5. CONCLUSION AND FUTURE WORK

The current work describes the development and the use of Urdu district names speech corpus. A complete process of cleaning is described to get the maximum accuracy of the data and the system. Inter annotator accuracy is also

considered important in this regard. 88% speech recognition accuracy has been achieved at this data. The further perspective is to cover more data from those districts where minimum recordings have received. In addition another data for a location based service is under development which consists of 238 names of the cities, day/date, time and number of seats. Moreover, the presented work would be helpful in developing speech corpus for ASR's of other Pakistani languages.

6. AKNOWLEDGEMENTS

This work has been conducted through Enabling Information Access through Mobile Based Dialog Systems and Screen Readers for Urdu project supported through a research grant from ICT RnD Fund, Pakistan.

7. REFERENCES

[1] 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.

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

[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] T. Rahman, "Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift," in *SCALLA Conference on Computational Linguistics*, vol. 99, 2004, pp. 1-19.

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

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

[7] 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] 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.

[9] Muthusamy et. al., "The OGI Multi-Language Telephonic Speech Corpus," in *International Conference on Spoken Language Processing, ICSLP*, Alberta, Canada, Oct. 13-16, 1992, pp. 1-9.

[10] R. A. Cole et. al., "Telephonic Speech Corpus Development at CSLU1," in *The 3rd International Conference on Spoken Language Processing, ICSLP*, Japan, 1994, pp. 1-4.

[11] R. A. Cole et. al., "New Telephone Speech Corpora at CSLU," in *EUROSPEECH*, 1995, pp. 1-4.

[12] S. Kasuriya et. al., "Thai Speech Corpus for Thai Speech Recognition," in *the Proc. Oriental COCOSDA*, Japan, June 2003, pp. 1-9.

[13] K. Maekawa, "Spontaneous Speech Corpus of Japanese," in *the Proc. LREC2000 (Second International Conference on Language Resources and Evaluation)* , vol. 2, Athens, Greece, 2000, pp. 1-7.

[14] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon Technical Report N 93, 1993.

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

[16] S. Irtza and S. Hussain, "Minimally Balanced Corpus for Speech Recognition.," in *the Proc. 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13)*, Sharjah, UAE, Feb. 12-14, 2013, pp. 1-6.

[17] Pakistan Bureau of statistics. [Online]. http://www.pbs.gov.pk/population-tables

[18] 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.

[19] S. Hussain, "Phonetic Correlates of Lexical Stress in Urdu," Ph.D. dissertation, Northwestern Univ., USA, Dec. 1997.