Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System

Huda Sarfraz*, Sarmad Hussain*, Riffat Bokhari**, Agha Ali Raza**, Inam Ullah*, Zahid Sarfraz**, Sophia Pervez**, Asad Mustafa*, Iqra Javed*, Rahila Parveen*

Abstract

This paper reports the design and development of an 82 speaker Urdu speech corpus for speaker independent spontaneous speech recognition using the CMU Sphinx Open Source Toolkit for Speech Recognition. The corpus consists of 45 hours of spontaneous and read speech data from 82 speakers (42 male and 40 female), recorded over a microphone and a telephone line. The speech was collected from speakers ranging from 20 to 55 years of age. Recording sessions were conducted in office and home environments.

1. Introduction

Urdu, the national language of Pakistan, has over 100 million speakers in Pakistan and other regions [1]. This paper presents the development of a spoken language corpus for Urdu, specifically for a Lahore suburban accent. A spoken language corpus is defined as "any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data in-house, or by scientists in other organizations" [2]. As noted in the literature review section next, this work represents one of the few speech corpora available for Urdu. The speech corpus has been released freely under an open content license and is envisioned to play a significant role in Urdu speech processing research in the future. It contains speech from 82 adult native Urdu speakers, with Lahore suburban dialect, ranging in age from 20 to 55 years.

The corpus was specifically designed to be used for speaker independent spontaneous speech recognition using the CMU Sphinx Open Source Toolkit for Speech Recognition [3].

The next section gives an overview of the current state of speech corpora development for Urdu, and also looks at some standards for spoken language resources. After a description of the methodology adopted for this work, key corpus statistics are reported, and critical issues encountered and resolved during the development process are discussed.

2. Literature review

This section will give a brief overview of currently available Urdu speech corpora, and will then present some of the spoken language resource standards in use for speech corpora.

2.1. Urdu speech corpora

Previous work done for Urdu speech corpora development includes the work described in [4], which is similar in content to the corpus presented here, but includes data for a single speaker only. It was focused towards designing a phonetically rich speech corpus for speech recognition purposes. The work presented here is an extension of the corpus described in [4].

The U.S. Army Research Laboratory (ARL) Urdu Speech Database, a collection of recorded speech from 200 adult native Urdu speakers from Pakistan and Northern India, is available through the Linguistic Data Consortium (LDC) [5]. The corpus was released in February 2007. Its data is divided into training and testing sets, and has speech of multiple Urdu dialects including South Sindh, North Sindh, South Punjab, North Punjab, North West regions and Baluchistan.

The Enabling Minority Language Engineering (EMILLE) corpus [6] available through the Evaluations and Language Resources Distribution Agency (ELDA) [7] contains over 2 million words of transcribed spoken data for five languages including Urdu.

 * Affiliation: Center for Language Engineering, Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan; Email: *firstname.lastname@kics.edu.pk*.
** Affiliation: Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Lahore, Pakistan. Email: *firstname.lastname@nu.edu.pk*.

Speech corpora development in Urdu and other Indian languages for speaker recognition is also reported in [8].

2.2. Standards for spoken language resources

With respect to speech standards used for corpora, two approaches are followed as reported in [2]. In the first approach, information pertaining to the speech is stored within the speech file itself, and in the second approach, the information is stored externally. This information may include utterance identification, channel details, sampling rates, speaker information, recording conditions, etc.

As described in [2], the widely used NIST SPHERE format is an example of the first approach, where information is stored within the speech file. It consists of an ASCII structure that is pre-pended to speech data. The format has been developed by the National Institute of Standards and Technology [9], and is followed widely in the US. Most of the speech corpora available through the LDC are in the NIST SPHERE format, with the accompanying text files in SGML format [5].

An example of the second approach given in [2], of storing information externally to the speech file, is the SAM format which is widely used in Europe for multilingual and national databases. The format was defined by the ESPRIT Speech Assessment Methods (SAM) project, which ran from 1987-1994 [10]. It consists of a speech waveform file and an associated ASCII description file [2]. The ARL Urdu Speech Database mentioned earlier uses the SAM format.

In addition to these, there are many more formats, specifically created by projects to suit their own needs. One example is the Verbmobil Project's [2, 11] format designed to handle dialog. Another example is the CHAT transcription and coding format developed through the Child Language Data Exchange System (CHILDES) [12], and also used in the C-ORAL-ROM Corpus [13] available through ELDA [7].

The Urdu speech corpus described in this paper was developed for speech recognition using the CMU Sphinx speech recognition toolkit [3], so the design was made to conform to the format required by the toolkit for processing, but was also kept flexible enough for convenient usage in other areas.

3. Corpus content selection

The content of the suburban Urdu corpus has been designed based on the premise of using a combination of read and spontaneous speech for spontaneous speech

recognition presented in [14]. The spontaneous content serves to ensure phonemic balance and the read content serves to ensure phonemic coverage. The collection of spontaneous speech data is considerably more difficult than the collection of read speech data, so this served as a good strategy for rapid corpus development for speech recognition purposes, especially in the case of a low resourced language like Urdu.

3.1. Read content

The following read content was designed to be included in speech recording sessions.

3.1.1. Phonemically rich sentence list. To ensure that the entire phonetic inventory of Urdu was covered in the speech corpus, a sentence set consisting of 775 phonemically rich sentences was constructed derived from an 18 million word Urdu corpus collected from recent electronic and print media [15] as described in [4]. Each speaker was then requested to read out 45 sentences from the set.

3.1.2. Open content Urdu text. To increase the amount of continuous speech, some normal everyday text was also included in the content to be recorded. Six passages of Urdu text from Wikipedia [16] were selected such that they included numerous proper names, numbers and dates (topics included cricket, women writers etc.). The average length of each passage was about 1,200 words. Each volunteer speaker had to read 2 of the 6 selected passages.

3.2. Spontaneous content

The spontaneous content was designed to be obtained through a series of questions that a volunteer speaker would answer during a recording session. The questions were designed to minimally obtain the following types of speech data from volunteer speakers: proper names (including names of people and places), numbers (including ordinals and cardinals), dates, and times as used in normal day-to-day speech. Five question sets were designed to obtain the spontaneous speech.

3.2.1. Bio-data. This was a set of questions designed to obtain the biographical data of the volunteer speaker. It was not meant to be released as part of the speech corpus, as the volunteers were disclosing personal information. This set was made up of a total of 10 questions.

3.2.2. Daily routine and past experience. This question set consisted of questions relating to the daily routine of the volunteer speaker, and aimed to elicit content including place names, and times through questions about the volunteer speaker's daily routine. This set was made up of a total of 22 questions.

3.2.3. Hobbies and interests. This question set consisted of questions designed around the perceived common hobbies and interests of the speaker set, and was aimed at eliciting common conversational topics and also at increasing the vocabulary coverage of the recorded data. Questions were included about the volunteer speaker's favorite books, films, TV channels etc. This set was made up of a total of 32 questions.

3.2.4. Short sentences. This question set was designed to elicit answers consisting of complete sentences. This set was designed after noticing that most speakers tended to halt and trail off frequently while speaking when answers were long (often in the case of questions described in 3.1.2 and 3.1.3). This question set would ensure that at least a minimum amount of relatively clear speech was obtained from each speaker. Speakers were asked to name their favorite fruits, flowers and vegetables for example, using complete sentences, as opposed to only naming an item. For example, in response to the question, "What is your favorite color?", the volunteer was instructed to respond with a sentence of the type "My favorite color is blue", instead of just "Blue". This question set also helped to increase the vocabulary coverage of the corpus. This set was made up of a total of 37 questions.

3.2.5. Multiple topics. The question sets described above fell short of eliciting the desired amounts of speech per speaker in some cases during test recording sessions. The problem was that not all speakers had an interest in all the common topics targeted (e.g. television, shopping, films, cricket etc.), and would end up speaking very little for some questions. So, two additional question sets were designed to include numerous questions on a variety of situations, such that any speaker would be sufficiently interested in at least some of them, and would be able to maintain a steady flow of spontaneous speech for a sufficient amount of time. For example the following questions were included, "Describe your favorite breakfast", "Describe the last time you had to ask someone for help" and "If you could go back into the past to change your profession, what would you change it to?". In addition to this, many more questions on a broad scope of topics

and situations were also included. These two sets included a total of 119 questions (59 in the first set and 60 in the second).

4. Speech data collection

This section describes the speech data collection process and includes the speaker recruitment and recording process.

4.1. Speaker recruitment

The speaker recruitment process was designed to select speakers who were native Urdu speakers with a Lahore suburban accent, no speech impediments and who were comfortable with the idea of having their speech recorded. This was deduced during a short recruitment interview in which volunteers were asked questions to determine their linguistic background (language spoken at home, area of residence, schooling etc.). Volunteers were asked to read a sample Urdu sentence, which contained the entire phonemic inventory of Urdu, in order to detect any speech impediments. In addition, volunteers were also asked to read some sample sentences from the phonemically balanced set of sentences, in order to check if they were able to read comfortably at the required level. This had to be done because some of the words included in the sentences were low frequency words, and not used in regular, spoken Urdu.

Volunteers who passed the recruitment took part in a recording session. Volunteers were required to sign a contract before starting the recording session, according to which they agreed to have their speech publicly released for research purposes. At the end of the recording session, volunteers received an honorarium for their participation.

4.2. Recording setup

During recording sessions, speech data was recorded simultaneously through a microphone and a telephone line.

During recording sessions, volunteer speakers were seated at a table with a telephone set and a microphone (connected to a laptop) and were required to speak into the microphone and telephone simultaneously. The microphone rested on the table near the speaker's mouth and the telephone receiver had to be held up to the speakers ear by hand. These were depicting the situation in which the ASR system being developed will eventually be used.

4.2.1. Recording hardware. A Dell Latitude E5400 laptop was used to record speech through a Logitech USB Desktop microphone. A Linksys SPA400 telephony gateway was used to capture recordings over a telephone line. The telephone calls were made through an extension, and a combination of land lines and extensions were used at the receiving end.

4.2.2. Recording software. Praat [17] was used on the laptop to capture and manage the speech received over the microphone. Microphone speech was recorded at 16 kHz and stored in .wav format. Telephone speech was recorded at 8 kHz and managed through Trixbox, an Asterisk-based PBX phone system [18].

4.2.3. Recording locations. Office rooms and a student lab were used to conduct the recordings. External noise in the office environment was contributed by the opening and shutting of doors and drawers, people talking, printers, telephones etc. Recording sessions were conducted in the student lab almost always when it was completely empty. In addition to these environments, some recording sessions were also conducted in home environments.

4.3. Recording session

Selected volunteer speakers spent up to three hours for the recording session. It was ensured that the volunteer was seated comfortably within reach of the microphone and telephone set to be used for the recording. Objects that the volunteer could use to produce noise unintentionally were removed from the reach of the volunteer e.g., chairs that creaked when rocked and pens that could produce clicking sounds. The session conductor and volunteer decided on hand signals to communicate with during the recording, e.g., if the session conductor wanted the volunteer to repeat a sentence. The recording session was divided into 12 sub-sessions, covering the content described in Section 3. The content was arranged such that the speaker was required to read and speak spontaneously in alternate sessions. This helped distribute the stress and monotony of the recording procedure throughout the session. The speaker was given breaks between the sub-sessions and plenty of drinking water.

5. Speech data processing

This section describes how the speech data was processed after it had been acquired through the recording session.

5.1. Speech segmentation

Recorded speech from volunteer speakers was manually split into smaller portions, about 10 seconds long, using Praat [17], such that they were suitable for use as training data for CMU Sphinx speech recognition toolkit [3].

The basic rule followed during this process was to only mark a boundary during silence (though desired, it was not always aligned with a phrase or a sentence boundary). Thus smaller .wav files (not more than 10 seconds long) were produced.

Any portions that included disruptive noises, such as a telephone ring, a drawer opening or closing, or someone else speaking, in close proximity to the speaker, were marked as unusable for the training process.

5.2. Speech transcription

The segmented speech files were transcribed orthographically in Urdu script manually by a team of linguists. Each speech segment file name therefore had a corresponding transcription string. The orthographic transcription was later converted into phonemic transcription using a transcription lexicon for use by the CMU Sphinx speech recognition toolkit. This phonemic transcription process will be described in a subsequent paper, but the general transcription rules have been based on [21].

In addition to the orthographic transcription of speech in segments, the *Silence, Vocalization* and *Breath* tags were defined to represent non-speech areas in the segments. All silences or pauses during speech as audible or viewable in the waveform displayed on Praat [17] were marked with a silence tag, in particular at the start and end of segmented portions. Sounds produced by the speaker that could not be classified as speech were marked by a vocalization tag within the orthographic transcription. Breath sounds identified within segments were marked with a breath tag.

6. Results

Most of the speakers were recruited from a university campus (including students, faculty and staff) and nearby residences. Recruited speakers were between 20 and 55 years of age.

At the end of the speech data collection and microphone data processing procedure, 44.5 hours of processed microphone speech data was collected from a total of 82 speakers. This included 20.7 hours of speech data from 40 female speakers, and 23.8 hours of

speech data from 42 male speakers. An average of 0.5 hours of speech was elicited from each speaker.

The minimum amount of processed speech obtained from a speaker was around seven minutes. The maximum amount was above an hour.

Table 1 shows the amount of speech obtained during the recording process that was discarded. This included speech segments that were interrupted by disruptive environment noise, e.g., telephone ringing in the background, but it also significantly included speech segments where words were un-recognizable. This could happen, for example, in cases where volunteer speakers were not speaking clearly.

Table 1. Percentages of discarded speech

	Processed speech (hrs)	Discarded speech (hrs)	Discarded %
Spontaneous speech	30.6	2.7	8.1
Read sentences	5.6	1.2	18.1
Read passages	9.7	0.3	3.5

Table 1 also shows that the discarded percentage is the highest for reading phonemically rich sentences, primarily because they are normally awkward and contain low frequency words. When speakers encountered a non-familiar word, they would either 1) pronounce it correctly, 2) pronounce it incorrectly, where the incorrect pronunciation corresponded to another Urdu word, 3) pronounce it incorrectly where the mispronunciation did not correspond to any known Urdu word. Segments where the third case occurred were discarded. In the second case, the segment was processed as spoken by the speaker, even though it did not map directly to the sentence.

The total processed data from 82 speakers resulted in a vocabulary set of over 14,000 words. This included phonemically rich words used in the phonemically rich sentence set and also included numerous words introduced by volunteer speakers during spontaneous speech.

The complete corpus has been released freely under an open content license.

7. Discussion

Problems encountered during spontaneous speech elicitation included speakers not speaking freely because they were too conscious of the process, and not sure about how the speech was going to be used. This was somewhat alleviated by explaining the use of the speech in complete detail, and answering any questions that speakers may have about the process. Some speakers were uncomfortable speaking in the office environment with other people nearby. These sessions resulted in speakers speaking in a very low volume, sometimes rendering the acquired speech data almost un-usable. This was solved through seating speakers within the office environment, but separated from the rest of the room with a partition. This gave the impression that the recording process was not disrupting the office, and allowed the speakers to speak more freely, while at the same time the ambient noise was also captured during the recordings.

In contrast to this, some volunteer speakers were quite enthusiastic about the process, and produced unusable speech, for example, by giggling while relating an incident. This was addressed by explaining the type of speech needed before the start of the session. Reminders through hand-signals were required during the recording sessions as well, because most speakers would inadvertently forget.

With read speech, the problem, as reported in the results section, arose due to the inclusion of words that were not part of the vocabulary. This was somewhat alleviated through having the speakers listen to correctly pronounced pre-recorded sentences, and also reading through the set once through recording. It did not completely solve the problem however, as reflected in Table 1, and further refinement of the phonemically rich sentence list with respect to unfamiliarity of words and naturalness of sentences is suggested.

One of the issues encountered during transcription verification of speech segments was the usage of some meaningless words during spontaneous speech. These are referred to as generalization words in [19], and can be defined as meaningless words spoken along with a meaningful word to convey a generalization effect. For example, a speaker responded "kʰɑnɑ vɑnɑ" as part of a response to a question about their daily routine, to convey that their weekend activity includes going out to dinners and other similar engagements. Here, the word "kʰɑnɑ" means food, but "vɑnɑ" is not actually a word, and only has significance when spoken after "kʰɑnɑ". Additionally, the generalization pair for a word cannot be predetermined. As reported in [19], for example, some speakers may reduplicate it as "vɑnɑ" while others as "ʃɑnɑ". A decision was made to include these types of words when they came up in speech as legitimate words in the vocabulary and transcribe them as per the normal process for regular Urdu words, because they were occurring quite frequently during spontaneous speech elicitation.

Most of the remaining orthography issues were resolved as per the specification in [20], including the orthography of English words that were used by speakers during spontaneous speech, and were also included in the phonemically rich sentences.

8. Conclusion

This paper presented the design and development of an Urdu speech corpus containing 44.5 hours of transcribed microphone speech (and also telephone speech) data from a total of 82 speakers. Future work would include increasing the corpus by adding speech from new speakers and also improving the process in order to capture more speech per recording session. One suggestion is to use a set of objects or pictures during the recording session and to ask speakers to describe them. Volunteers may speak more freely with this method as opposed to the question-answer style adopted in this work.

9. Acknowledgements

9. References

[1] Lewis, M. Paul (ed.). Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International, 2009. Online version: http://www.ethnologue.com/.

[2] Handbook of standards and resources for spoken language systems. Dafydd Gibbon, Roger Moore, Richard Winski. Walter de Gruyter, 1997. Berlin, Germany.

[3] CMU Sphinx Open Source Toolkit for Speech Recognition Project by Carnegie Mellon University, http://cmusphinx.sourceforge.net/, accessed June 2010.

[4] A. Raza, S. Hussain, H. Sarfraz, I. Ullah, Z. Sarfraz, "Design and development of phonetically rich Urdu Speech Corpus", proceedings of O-COCOSDA 2009, School of Information Science and Engineering of Xinjiang University, Urunqi, China.

[5] The Linguistic Data Consortium, www.ldc.upenn.edu, accessed June 2010.

[6] The Enabling Minority Language Engineering (EMILLE) Corpus, www.lancs.ac.uk/fass/projects/ corpus/emille, retrieved June 2010.

[7] Evaluations and Language resources Distribution Agency, www.elda.org, accessed June 2010.

[8] H.A. Patel and T.K. Basu, "Development of speech corpora for speaker recognition research and evaluation in Indian languages", *International Journal of Speech Technology*, Springer Netherlands, May 2009, pp. 17-42.

[9] National Institute of Standards and Technology, www.nist.gov, accessed June 2010.

[10] Spoken language system and corpus design. Dafydd Gibbon, Roger Moore, Richard Winski. Walter de Gruyter, Berlin, Germany, 1997.

[11] Verbmobil, http://verbmobil.dfki.de/.

[12] MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 4rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates

[13] E. Cresti, F.B. do Nascimento, A.M. Sandoval, J. Veronis, P. Martin, K. Choukri, The C-ORAL-ROM Corpus: A Multilingual Resource of Spontaneous Speech for Romance Languages. Proceeedings of the 4th International Conference on Language Resources and Evaluation. 26-28 May, 2004, Lisbon, Portugal.

[14] A. Raza, S. Hussain, H. Sarfraz, I. Ullah, Z. Sarfraz, "An ASR System for Spontaneous Urdu Speech".

[15] M. Ijaz and S. Hussain, Corpus Based Urdu Lexicon Development, Conference on Language Technology, 2007.

[16] Wikipedia, Urdu version, http://ur.wikipedia.org, accessed June 2010.

[17] Praat: doing phonetics by computer, www.fon.hum.uva.nl/praat, accessed June 2010.

[18] Fonality trixbox CE, an Asterisk-based PBX Phone System, www.trixbox.org, accessed June 2010.

[19] H. Sarfraz, Formation of Generalization Words ("Mohmil") in Urdu. Akhbar-e-Urdu, April-May 2002.

[20] Urdu LSP for LC-STAR II, June 2006, www.lc-star.org/pccdocs/Pakistan_Urdu_LSP_V1.0.pdf, accessed June 2010. LC-STAR II: Lexica and corpora for speech-to-speech translation components.

[21] S. Hussain, , "Letter to Sound Rules for Urdu Text to Speech Sytem", proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages, COLING 2004, Geneva, Switzerland, 2004.