Interspeech 2018
2-6 September 2018, Hyderabad

Rapid Collection of Spontaneous Speech Corpora using Telephonic Community Forums

Agha Ali Raza[1], *Awais Athar*[2], *Shan Randhawa*[1], *Zain Tariq*[1], *Muhammad Bilal Saleem*[1], *Haris Bin Zia*[1], *Umar Saif*[1], *Roni Rosenfeld*[3]

[1]Information Technology University, Lahore, Pakistan
[2]European Bioinformatics Institute (EMBL-EBI), Cambridge, UK
[3]Carnegie Mellon University, Pittsburgh, PA, USA

{agha.ali.raza, shan.randhawa, zain.tariq, bilal.saleem, haris.zia, umar}@itu.edu.pk,
awais@ebi.ac.uk, roni@cs.cmu.edu

Abstract

We present a novel technique for rapid collection of spontaneous speech data over mobile phone channel using telephonic community forums. Our public forum allows users to post audio messages, listen to messages posted by others, post votes and audio comments, and share content with friends through subsidized phone calls. The entertainment aspects and sharing features of the forum lead to its viral spread in Pakistan. Within 8 months, it reached 11,017 users and gathered 1,207 hours of speech data comprising 57,454 audio-posts and 130,685 audio-comments, spanning Urdu and 9 regional languages. We trained an ASR using just 9.5 hours of the corpus to obtain 24.19% WER. Community forums automatically overcome common spontaneous speech data collection challenges like speaker recruitment, natural speech elicitation, content diversity, informed consent, sampling real-world ambient noise, and reach (for geographically remote linguistic communities). This technique is especially useful for gathering speech corpora for under-resourced languages hence enabling the development of speech recognition, keyword spotting, speaker ID, and noise classification systems (among others) for such languages. It also allows rapid, automatic preservation of spoken languages and oral aspects of culture. This technique can be extended to collect speech data for endangered languages, oral cultures, and linguistic minorities.

Index Terms: Mobile phone channel, spontaneous speech corpus, automatic rapid corpus collection, speech recognition, under-resourced languages, oral cultures, telephonic community forums, preservation of language and oral culture.

1. Introduction

Recent years have seen a significant increase in the use of speech technologies such as speech recognition and synthesis to augment human machine interaction. Speech technologies find their way in useful applications like speaker identification and verification, audio forensics, threat preemption using keyword spotting, and spoken dialog systems. Unfortunately such benefits remain limited mostly to languages that are rich in linguistic resources as most of these techniques require large amounts of speech data for training. Further, such tools and techniques are also highly language dependent and as a result could not be efficiently and reliably used cross-linguistically.

Speech corpus collection is a major challenge for underserved languages. Speech systems typically require large amounts of training data matching target usage in terms of speech type, channel, environment, language, etc. This entails collection of naturally spoken speech data matching these parameters, from speakers who might be low-literate, tech-shy, and geographically remote. In addition, there are more specific hurdles including (a) speaker recruitment and providing incentives for their contributions, (b) speech elicitation ensuring natural pitch, tone, style, (c) content diversity, (d) informed consent, and (e) read speech collection from people who cannot read.

Due to these hurdles, collection of speech corpora is a slow and difficult process and such corpora are limited to a handful for such languages. On the demand side, speech technologies are very relevant in developing world context where speech-based human computer interfaces over mobile phones are proving an effective tool for providing information access and connectivity to people and overcoming literacy, tech naivety and visual impairment hurdles of technology usage. Successful examples include information dissemination regarding health [1], agriculture [2, 3], jobs [4], finance [5], public awareness [6] and many other domains [7, 8, 9]. However, with all their benefits, telephone-based speech services in developing countries are mostly limited to spoken output and push-button (DTMF) input from the user. While such situations could benefit from a two-way spoken interaction (dialog systems), the lack of linguistic resources to enable speech recognition, keyword spotting and text-to-speech for the involved languages pose a major hurdle. Languages facing extinction are an additional incentive to expedite efforts of preservation.

This paper focuses on the use of speech interfaces over simple mobile phones to rapidly gather spontaneous speech corpora. Speech over simple phones makes our platform inclusive to low-literate, poor, tech-shy and geographically remote people. Our telephonic community forum encourages people to contribute diverse content covering unconstrained genres and also allows them to explicitly share and comment on the posts. Due to its viral and organic nature, the platform automatically overcomes common speech data collection hurdles like reach, speaker recruitment, natural speech elicitation, speaker incentives, informed consent and topic diversity. This is achieved as the platform spreads from person-to-person via shared (forwarded) voice messages and offline discussions. Users are explicitly informed that the recorded data would be publically available and will also be used for research purposes. Their main incentive for using the service, and contributing speech data, is social connectivity and entertainment. Topic diversity is an outcome of the open-ended nature of the platform. As a result, within 8 months, and without any advertisement, we were able to reach 11,017 users and gather 1,207 hours of spontaneous speech data from 4,678 speakers, comprising Urdu and

10.21437/Interspeech.2018-1139

Table 1: *List of available Speech Corpora for Urdu, Pashto and Punjabi. None found for Sindhi and Saraiki. (* Estimated size is calculated assuming 10 utterances per minute where not stated, ** Speech type is I: Isolated, R: Read, C: conversational)*

Language	Region	Stated size	Est. size (hrs)*	Speech type**	Channel	# Speakers	Public	Vocabulary size	Reference
Pashto	Afgh/Pak	214 hrs	214	C	Telephone	-	Y	-	[10]
Pashto	-	150 hrs	150	C	Telephone	-	N	-	[11]
Pashto	Pak	34.3 hrs	34.3	C	Radio	-	Y	-	[12]
Pashto	-	34 hrs	34	C	Mic	-	N	10K	[13]
Pashto	Pak	101 utt	0.1	I	Mic	1	N	101	[14]
Pashto	-	12 hrs	12	R+C	Mic	80	N	-	[15]
Pashto	Pak	8,050 utt	13.4	R	Mic	50	Y	161	[16]
Pashto	Pak	500 utt	1	I	Mic	50	N	10	[17]
Punjabi	Ind	6,000 utt	10	I	Mic	6	N	200	[18]
Punjabi	Ind	2,760 utt	5	I	Mic	8	N	115	[19]
Punjabi	Ind	2 hrs	2	I	Mic	1	N	3085	[20]
Urdu	Pak	45 hrs	45	R+C	Mic+Tel	82	N	14K	[21]
Urdu	Pak	41.9 hrs	41.9	C	Radio	-	Y	-	[12]
Urdu	Pak	12,500 utt	21	I	Mic	50	N	250	[22]
Urdu	Pak	12 hrs	12	I	Telephone	300	Y	139	[23]
Urdu	Pak	5,200 utt	9	I	Mic	10	N	52	[24]
Urdu	Pak	3	3	R	Mic	1	Y	7K	[25]
Urdu	Pak	200 utt	0.3	I	Mic	10	N	20	[26]
Multiple	Pak	1,207 hrs	1,207	C	Telephone	4,678	N	20K+ (estimated)	This work

9 other languages (see [27] for a detailed analysis of the viral spread and usage of the service). To our knowledge, this is the first time *spontaneous* speech corpora for Punjabi, Sindhi and Saraiki are being made available. Further, our Urdu corpus is significantly larger (885 hours) than the ones currently available (maximum size: 45 hours, Table 1). We present an analysis of this corpus and show that it provides both phonetic cover and balance for Urdu. Finally, in order to investigate if the collected data is high quality enough to train speech systems, we used a very small subset of recordings (9.5 hours) to train an Urdu speech recognizer and achieved a word error rate of 24.19% (see Table 2. These are disparate Urdu corpora so the WERs are not comparable but give an overall idea of Urdu ASR accuracy). It is important to mention that collection of speech data is only part of the corpus collection challenge and its reliable annotation and transcription is required before it can be used for practical speech processing tasks. Our focus in this paper is only the rapid and automatic collection of hard-to-get spontaneous speech data.

2. Related Work

Table 1 shows a break down of the speech corpora available for Urdu, Punjabi and Pashto. We were unable to find any publicly available speech corpora for Sindhi and Saraiki. In the rest of this section we review available speech corpus collection techniques that have been used for other languages.

While telephone speech corpora have been used for research and development for more than two decades [32], methods for collecting them vary with target tasks such as language/speaker/speech recognition [33, 34]. Gathering media broadcasts [35] and read speech [36] have proven successful for collecting large scale spoken language data. Another approach is the use of smart phones. [37] collected over 3,000 hours of transcribed speech corpora in 17 languages using a smart phone app to record read speech. Similar apps were used by [38] to collect upto 100 hours of speech data for 3 less-resourced languages; Basaa, Myene and Embosi. Smartphone-based speech data collection is challenging for languages for which the majority of speakers is low-literate, poor and non-tech savvy. Such populations either do not have access to smartphones or have difficulty handling them. One of the early examples of using Interactive Voice Response (IVR) for speech data collection is by Lander et al. [39], who used prerecorded questions and prompts to gather a data set of responses ranging from single words to short topic-specific descriptions and up-to a minute of unconstrained spontaneous speech by 200 different speakers in 22 languages. Another relevant mechanism used to collect the Fisher Corpus [40] relied on system initiated calls to connect strangers. [41] created a community moderated voice forum called Sangeet Swara using an approach similar to ours. They focused on content belonging to three genres: jokes, songs and poems, and reported a total of 5,376 voice posts by 1,521 callers.

3. User Interface and Deployment

We have reported the user interface and HCI aspects of a sub-component of our social platform in [27]. User interaction with the forum begins when a user places a *missed call* to our phone number, i.e., they dial our number and hang up immediately as it rings (see [27] for the interaction flowchart). Since this is a common method in developing countries to signal call-back requests, we used it to subsidize call costs. When the system calls back, users are provided the options to record a new voice post, listen to posts recorded by others, or check the status (votes, comments) of their own previously recorded posts. A sub-component of the platform allows users to record questions instead of messages. Users are informed that they should avoid recording personal details like phone numbers, etc. as the recordings would be made available to public and would also be used for research purposes. Next, they are provided a randomly chosen set of suggestions regarding genre of content (e.g. discussions, grievances, news, jokes, poetry, songs, problems in their area etc.). Users are provided up to 60 seconds to record their message or question. They can terminate earlier by pressing #. Users are also asked to record their name (only once for each user). Users who choose to listen to content posted by others get to choose between hearing the recordings sorted by popularity (more up-votes), recency, or a mixture of the two (i.e. trending posts). After hearing each post, they are provided the options to vote it up or down, to report abuse or to record an audio comment. Users can also share their favorite posts with friends by entering their phone numbers. These numbers are called out by our system and the forwarded posts are played to the recipients after telling them the name of the sender.

Table 2: *Summary of Urdu ASR systems*

#	Speakers	Training size (hrs)	Vocab Size (# types)	Speech Type	Genre	Channel	Publicly Available	Technique	Best WER	Ref.
1	Multiple	-	52	Isolated	Frequently used words	Mic	No	GMM/HMM	10.60%	[24]
2	Single	3	6K	Spontaneous	Phonetically rich sentences	Mic	Partially	GMM/HMM	18.80%	[28]
3	Multiple	45	14K	Spontaneous	Interviews	Mobile+Mic	No	GMM/HMM	68.80%	[29]
4	Multiple	99	79K	Read/prompted	Broadcast news	Mic	No	GMM/HMM CMLLR+MPE	32.60%	[30]
5	Multiple	9.5	139	Isolated	District names	Mobile	Yes	GMM/HMM	7.13%	[31]
6	Multiple	9.5	5K	Spontaneous	Audio posts	Mobile	No	SGMM/MMI	24.19%	This work

The community platform was seeded via advertisement over a popular entertainment-based IVR service [4] for a month. Once seeded we have no subsequent control on the extent and nature of the spread. Over the next 8 months, we received 389,587 phone calls (involving 11,017 users) and accumulated 57,454 posts (messages and questions) and 130,685 audio comments contributed by 4,678 users. Over all, we accumulated 1,207 hours of audio data with 517 hours of posts and 690 hours of comments.

4. Corpus Analysis

This section presents an analysis of the gathered corpus and demographic details of our users based on telephonic surveys.

4.1. Content Annotation

Two annotators listened to a random sample of 14,228 posts and 1,700 comments and annotated them for language, gender and genre. A subset of recordings are also annotated in Urdu Unicode to train an automatic speech recognition system. A random sample of 103 recordings was annotated by both to confirm inter-annotator agreement. The inter-annotator agreement around genre was found to be $\kappa = 0.97$ (Cohen's kappa, defined as $\kappa = \frac{P(A)-P(E)}{1-P(E)}$, where $P(A)$ is the proportion of agreement and $P(E)$ is the proportion of agreement by chance [42]). Perfect agreement was found for the gender attribute. Average annotation rate was 60 posts or comments in per hour.

4.2. Corpus Analysis and User Surveys

We calculated the distribution of languages in recordings based on a random sample of 15,928 audio files (14,228 posts and 1,700 comments). From this we have estimated the number of hours of speech of each language in our corpus. Table 3 shows this distribution proportionally among the posts and comments. Urdu is the most widely understood language in the country as well as the interface language of our platfrom. We believe that both these factors motivated our users to record content mostly in Urdu. We also conducted telephonic surveys of 415 randomly selected users of our platform and found that 66.49% of our users preferred to communicate in their local languages (other than Urdu) and only recorded content in Urdu so that wider community can understand it. Of the 415 survey participants, 33.5% stated Urdu as their preferred language, 26% stated Punjabi, 14.23% stated Pashto, 10.67% stated Saraiki, 7% stated Balochi, while the remaining mentioned other local languages. In our future deployments, we plan to experiment with modifying the interface language to elicit more local language content.

Table 3: *Distribution of languages in the corpus (estimated using a random sample of 15,928 recordings)*

Language	Weight	Length (hrs)	Language	Weight	Length (hrs)
Urdu	73.27%	885.07	Saraiki	0.42%	5.06
Unclear	15.07%	182.08	Sindhi	0.12%	1.39
Punjabi	7.33%	88.58	Balochi	0.01%	0.15
Pashto	2.04%	24.68	Hindko	0.01%	0.11
English	0.91%	10.99	Persian	0.01%	0.06
Arabic	0.81%	9.75	Other	0.00%	0.02

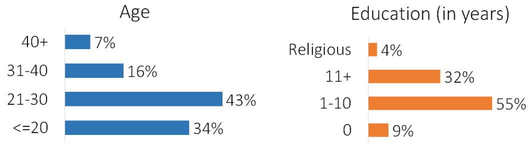

Figure 1: *Distribution of age and education in survey*

In the sample of 15,928 audio files, 88.19% were found to contain actual spoken content while 11.81% files were either empty or contained recorded noise (users hesitating, coughing, murmuring etc.) and silence. Of the files containing actual recorded content, 93.18% were male recordings while the remaining were female-contributed posts. Table 4 shows the diversity of genres of the posted content. 62.62% of the posted content is prose, comprising discussions, arguments and questions. These span a wide variety of topics ranging from casual greetings, discussions around current affairs and social problems, sayings and quotes, history, literature, sports, religion, general knowledge, health, jokes, and movies. Next most frequent category is content sung or recited by users e.g. recitation of poetry, hymns and songs. A much smaller (0.8%) category is content recorded by users from other devices, e.g. songs and music recorded from a TV, radio or another mobile phone in the background. 4.3% of the recorded files contained one or more offensive words. We did not encounter a lot of code-switching and only 1.8% posts contain words of more than one language.

Based on our telephonic surveys, we found that our users are mostly low educated (N=274, Figure 1), young (N=206, Figure 1) men (93%, N=276) with a large fraction (13%) having no formal education. However, there were also a significant number (32%) of users with more than 10 years of education. 75% (N=276) of survey participants owned simple or feature phones, 21% used smart phones, while 4% owned both. 56% (N=262) did not have access to internet.

4.3. Phonetic Cover and Balance

We phonetically transcribed 922 posts (N=77,190; V=4,473) and compared their phonemic distribution with a large newspaper corpus (N=3,470,130; V=81,255). The posts provide phonetic cover and balance for all common Urdu phonemes (Figure 2). The missing phonemes were r^h, l^h, ɔ̃, and ʒ that occurred 5,

Table 4: *Distribution of genres*

Genre	*f*	%
Prose (discussions and questions)	8,909	62.62
Poetry and hymns (recited by participants)	3,461	24.33
Silence and noise	1,017	7.15
Other	727	5.10
Prerecorded songs and music	114	0.8
Total	**14,228**	**100**

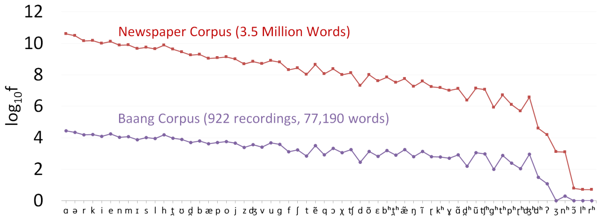

Figure 2: *Phonetic cover and balance*

Table 5: *WER with various models*

Model	Word Error Rate
GMM	30.52%
GMM + LDA + MLLT	29.74%
GMM + fMLLR	28.19%
SGMM + fMLLR	24.59%
SGMM + MMI	24.19%

5, 6 and 1,267 times in the news corpus respectively. This shows that they are indeed very rarely used in Urdu.

5. Urdu Speech Recognition

To assess the usefulness of the gathered corpus, we trained a speech recognition system for Urdu using a small subset of the corpus. Our annotators transcribed 7,230 Urdu utterances (132 speakers, about 5 min/hr), constituting 9.5 hours of speech data (Avg. SNR 7.73 dB). We split the data randomly into a training set (6,500 segments, 8.5 hrs) and test set (730 segments, 1 hr).

5.1. Acoustic Model

We used Gaussian Mixture Model (GMM) and Subspace Gaussian Mixture Model (SGMM) and applied Feature-space Maximum Likelihood Linear Regression (fMLLR) to make our models speaker adaptive [43]. We also used a sequence discriminative training technique, Maximum Mutual Information (MMI) to improve word error rate. We trained all of our acoustic models using KALDI ASR toolkit [44]. All of our models are standard triphone models [45]. Our GMM system has 10K Gaussians for 2K HMM states while SGMM system has 9K Gaussians for 7K HMM states.

5.2. Language Model and Pronunciation Lexicon

We use a trigram language model with Kneser-Ney discounting, based on training transcripts, built using SRILM toolkit [46]. Our LM has 74K tokens (5K types), an OOV rate of 3.64% and perplexity of 37.04 on test data. We used PronouncUR [47] to generate a pronunciation lexicon containing most of the popular pronunciation variants of each word. It has 60 phonemes: 59 for speech and 1 for modeling silence.

5.3. Results

The word error rates (WER) of various acoustic models are shown in Table 5. For comparison, Table 2 shows the state-of-the-art in terms of Urdu speech recognition systems.

6. Conclusion and Future Work

In this paper we show that telephonic community forums can be effectively used to rapidly collect spontaneous speech data. This technique overcomes common spontaneous speech data collection challenges for under-resourced languages and low-literate and non tech-savvy populations. We also presented an analysis of the gathered speech corpus and its use for speech recognition tasks. Telephonic community forums can also be used for spoken language preservation as it involves collection of large amounts of speech data. About 46% of the languages spoken globally have no written form [48] and an estimated 43% of all world languages, spoken by 136 million people, have been declared endangered, including 26 languages in Pakistan and 197 languages in India [49]. As a next step, we plan to provide regional language support to the platform to encourage linguistic sub-communities via *channels*. We plan to use this to try to collect speech data for endangered languages in South Asia. We also plan to transcribe the remaining speech data that we have gathered and get it annotated for prosody, accent, sentiment etc. This would enable localization of speech and natural language resources and tools for Urdu and other languages of South Asia. We also plan to release our trained models to facilitate further research.

7. Acknowledgments

We thank ITU for supporting this work and Hira Ejaz, Abdullah Sabri, Aleena Haider, Mehar Usama, Mamoona Riaz, Gulnar Khan, Arooj Saleem and Mukhtar Ahmad for annotating data, moderating posts and performing surveys.

8. References

[1] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld, “Healthline: Speech-based access to health information by low-literate users,” in *ICTD*. IEEE, 2007.

[2] N. Patel, S. Agarwal, N. Rajput, A. Nanavati, P. Dave, and T. S. Parikh, “A comparative study of speech and dialed input voice interfaces in rural india,” in *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009.

[3] A. B. M. Vashishtha24, “Innovative ivr system for farmers: enhancing ict adoption,” 2014.

[4] A. A. Raza, F. Ul Haq, Z. Tariq, M. Pervaiz, S. Razaq, U. Saif, and R. Rosenfeld, “Job opportunities through entertainment: Virally spread speech-based services for low-literate users,” in *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013.

[5] B. Rocheleau and L. Wu, “E-government and financial transactions: Potential versus reality,” *The Electronic Journal of e-Government*, vol. 3, no. 4, 2005.

[6] N. Wolfe, J. Hong, A. A. Raza, B. Raj, and R. Rosenfeld, “Rapid development of public health education systems in low-literacy multilingual environments: Combating ebola through voice messaging,” in *ISCA Special Interest Group on Speech and Language Technology in Education (SLaTE)*. INTERSPEECH, 2015.

[7] A. Raza, C. Milo, G. Alster, J. Sherwani, M. Pervaiz, S. Razaq, U. Saif, and R. Rosenfeld, “Viral entertainment as a vehicle for disseminating speech-based services to low-literate users,” in *ICTD*, vol. 2, 2012.

[8] R. Roche, E. Hladilek, and S. Reid, “Disaster recovery virtual roll call and recovery management system,” 2006, uS Patent 7,026,925.

[9] A. Zainudeen, R. Samarajiva, and N. Sivapragasam, “Cellbazaar, a mobile-based e-marketplace: Success factors and potential for expansion,” 2010.

[10] N. Adams, A. Bills, J. G. Fiscus, B. Gillies, M. Harper, T. J. Hazen, A. Jarrett, K. K. Khugyani, W. Lin, J. Ray, A. Rytting, W. Shen, T. E. Strahan, and E. Tzoukermann, “Iarpa babel pashto language pack iarpa-babel104b-v0.4b,” 2016. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2016S09

[11] X. Cui, J. Xue, P. L. Dognin, U. V. Chaudhari, and B. Zhou, “Acoustic modeling with bootstrap and restructuring for low-resourced languages,” in *11th Annual Conference of ISCA*, 2010.

[12] K. Walker, X. Ma, D. Graff, S. Strassel, S. Sessa, and K. Jones, "Rats speech activity detection ldc2015s02." 2015. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2015S02

[13] R. Prasad, S. Tsakalidis, I. Bulyko, C.-l. Kao, and P. Natarajan, "Pashto speech recognition with limited pronunciation lexicon," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010.

[14] S. A. R. Abid, N. Ahmad, M. A. A. Khan, and F. T. Zuhra, "Concatenative based pashto digits and numbers synthesizer," *International Journal of Computer Applications*, vol. 72, no. 6, 2013.

[15] A. Kathol, K. Precoda, D. Vergyri, W. Wang, and S. Riehemann, "Speech translation for low-resource languages: The case of pashto," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[16] I. Ahmed, N. Ahmad, H. Ali, and G. Ahmad, "The development of isolated words pashto automatic speech recognition system," in *Automation and Computing (ICAC), 18th International Conference on*. IEEE, 2012.

[17] Z. Ali, A. W. Abbas, T. Thasleema, B. Uddin, T. Raaz, and S. A. R. Abid, "Database development and automatic speech recognition of isolated pashto spoken digits using mfcc and k-nn," *International Journal of Speech Technology*, vol. 18, no. 2, 2015.

[18] W. Ghai and N. Singh, "Phone based acoustic modeling for automatic speech recognition for punjabi language," *Journal of speech sciences*, vol. 1, no. 3, 2013.

[19] M. Dua, R. Aggarwal, V. Kadyan, and S. Dua, "Punjabi automatic speech recognition using htk," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 4, 2012.

[20] P. Singh and G. S. Lehal, "Text-to-speech synthesis system for punjabi language," in *International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain*, 2006.

[21] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, and R. Parveen, "Speech corpus development for a speaker independent spontaneous urdu speech recognition system," *O-COCOSDA, Kathmandu, Nepal*, 2010.

[22] H. Ali, N. Ahmad, K. M. Yahya, and O. Farooq, "A medium vocabulary urdu isolated words balanced corpus for automatic speech recognition," in *2012 international conference on electronics computer technology (ICECT 2012)*, 2012.

[23] S. Rauf, A. Hameed, T. Habib, and S. Hussain, "District names speech corpus for pakistani languages," in *Oriental COCOSDA (O-COCOSDA/CASLRE)*. IEEE, 2015.

[24] J. Ashraf, N. Iqbal, N. S. Khattak, and A. M. Zaidi, "Speaker independent urdu speech recognition using hmm," in *Informatics and Systems (INFOS)*. IEEE, 2010.

[25] A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah, and Z. Sarfraz, "Design and development of phonetically rich urdu speech corpus," in *Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on*. IEEE, 2009.

[26] B. Rehmam, Z. Halim, G. Abbas, and T. Muhammad, "Artificial neural network-based speech recognition using dwt analysis applied on isolated words from oriental languages," *Malaysian Journal of Computer Science*, vol. 28, no. 3, 2015.

[27] A. A. Raza, B. Saleem, S. Randhawa, Z. Tariq, A. Athar, U. Saif, and R. Rosenfeld, "Baang: a viral speech-based social platform for under-connected populations," in *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2018.

[28] A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah, and Z. Sarfraz, "An asr system for spontaneous urdu speech," *the Oriental COCOSDA*, 2010.

[29] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, and R. Parveen, "Large vocabulary continuous speech recognition for urdu," in *8th International Conference on Frontiers of Information Technology*. ACM, 2010.

[30] M. A. B. Shaik, Z. Tüske, M. A. Tahir, M. Nußbaum-Thom, R. Schlüter, and H. Ney, "Improvements in rwth lvcsr evaluation systems for polish, portuguese, english, urdu, and arabic," in *Sixteenth Annual Conference of the ISCA*, 2015.

[31] M. Qasim, S. Nawaz, S. Hussain, and T. Habib, "Urdu speech recognition system for district names of pakistan: Development, challenges and solutions," in *Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Conference of The Oriental Chapter of International Committee for*. IEEE, 2016.

[32] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992.

[33] A. Canavan and G. Zipperlen. (1996) Callfriend american english-non-southern dialect. [Online]. Available: https://catalog.ldc.upenn.edu/LDC96S46

[34] A. Martin and M. Przybocki, "The nist 1999 speaker recognition evaluationan overview," *Digital signal processing*, vol. 10, no. 1-3, 2000.

[35] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, 2014.

[36] Appen Pty Ltd, Sydney, and Australia., "Arl urdu speech database, training data ldc2007s03." 2007. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2007S03

[37] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. J. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *11the Annual Conference of ISCA*, 2010.

[38] G. Adda, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. E. Bonneau-Maynard, E. Gauthier, P. Godard, F. Hamlaoui, D. Idiatov *et al.*, "Innovative technologies for under-resourced language documentation: The bulb project," in *Workshop CCURL 2016-Collaboration and Computing for Under-Resourced Languages-LREC*, 2016.

[39] T. Lander, R. A. Cole, B. T. Oshika, and M. Noel, "The ogi 22 language telephone speech corpus." in *Eurospeech*, 1995.

[40] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004.

[41] A. Vashistha, E. Cutrell, G. Borriello, and W. Thies, "Sangeet swara: A community-moderated voice forum in rural india," in *33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[42] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, 1960.

[43] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance gaussians," in *Ninth International Conference on Spoken Language Processing*, 2006.

[44] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[45] L. Deng, M. Lennig, F. Seitz, and P. Mermelstein, "Large vocabulary word recognition using context-dependent allophonic hidden markov models," *Computer Speech & Language*, vol. 4, 1990.

[46] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[47] H. B. Zia, A. A. Raza, and A. Athar, "PronouncUR: An urdu pronunciation lexicon generator," *arXiv:1801.00409*, 2018.

[48] "Ethnologue: Languages of the world," 2017, https://www.ethnologue.com/country/US.

[49] C. Moseley, *Atlas of the world's languages in danger*. Unesco, 2010.