INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Improving Large Vocabulary Urdu Speech Recognition System using Deep Neural Networks

Muhammad Umar Farooq, Farah Adeeba, Sahar Rauf, Sarmad Hussain

Center for Language Engineering,
Al-Khawarizmi Institute of Computer Science,
University of Engineering and Technology, Lahore.
{umar.farooq, farah.adeeba, sahar.rauf, sarmad.hussain}@kics.edu.pk

Abstract

Development of Large Vocabulary Continuous Speech Recognition (LVCSR) system is a cumbersome task, especially for low resource languages. Urdu is the national language and lingua franca of Pakistan, with 100 million speakers worldwide. Due to resource scarcity, limited work has been done in the domain of Urdu speech recognition. In this paper, collection of Urdu speech corpus and development of Urdu speech recognition system is presented. Urdu LVCSR is developed using 300 hours of read speech data with a vocabulary size of 199K words. Microphone speech is recorded from 1671 Urdu and Punjabi speakers in both indoor and outdoor environments. Different acoustic modeling techniques such as Gaussian Mixture Models based Hidden Markov Models (GMM-HMM), Time Delay Neural Networks (TDNN), Long-Short Term Memory (LSTM) and Bidirectional Long-Short Term Memory (BLSTM) networks are investigated. Cross entropy and Lattice Free Maximum Mutual Information (LF-MMI) objective functions are employed during acoustic modeling. In addition, Recurrent Neural Network Language Model (RNNLM) is also being used for re-scoring. Developed speech recognition system has been evaluated on 9.5 hours of collected test data and a minimum Word Error Rate (%WER) of 13.50% is achieved.

Index Terms: Urdu, ASR, GMM-HMM, DNN-HMM, TDNN, BLSTM, RNNLM, LVCSR

1. Introduction

Automatic Speech Recognition (ASR) is one of the applications of speech and language technologies that converts speech into text. ASR has numerous applications in all fields of life such as agriculture [1], health care [2], banking sector and hotel management are a few to name. Urdu is the national language and lingua franca of Pakistan; bridging people speaking regional languages such as Balochi, Pashto, Punjabi and Sindhi with various dialects. It is spoken by more than a hundred million speakers in Pakistan, India, Bangladesh and the regions of Europe [3]. Urdu is a low resource language and very little transcribed speech data, text corpus and pronunciation lexicon is available publicly. A robust speech recognition system requires hundreds of hours of transcribed speech data, very large text corpus and lexicon.

Speech recognition for low resource languages has received scant attention in recent few years [6]. Hidden Markov Models (HMMs) [8] is the widely used technique to build acoustic models for speech recognition systems [9]. However, with the resurgence of deep learning, paradigm has been shifted towards Deep Neural Networks (DNN) based acoustic and language models. DNN based acoustic models are trained using alignments produced by HMM models [10]. Additionally, extensively used n-gram Language Models (LMs) are now being replaced by RNNLMs [24].

Over the years, limited efforts have been made to develop resources and speech technology related applications for Urdu. Developments in the domain of Urdu speech recognition started with development of an isolated speech recognition system [4]. Speech corpus was collected from 10 speakers and vocabulary size for this system was 52 words. A minimum Word Error Rate (WER) of 10.6% was achieved for unseen speakers.

Urdu continuous speech recognition system [32] was developed using spontaneous speech corpus collected from 82 speakers [25]. Speech corpus was recorded over telephone and microphone channels. Total duration of the corpus was 45 hours and vocabulary size was 14K words. Minimum WER of 68.8% was attained. Qasim et al. [34] developed a speaker independent Urdu speech recognition system for 139 district names of Pakistan. It covered the accent variation around the Pakistan and gained a minimum WER of 7.44% by building adapted ASR on field data.

First Urdu LVCSR was developed on 99 hours of Urdu broadcast data [33]. The vocabulary size of the system was 79K. To build a 5-gram Language Model (LM), a corpus of 266M words was collected from different newspapers. System was evaluated on an evaluation data set of 0.5 hours and a minimum WER of 32.6% was achieved by GMM-HMM based speaker adapted system.

A. Raza et al. [35] recorded about 1207 hours of speech data from 11017 speakers from all over the Pakistan. However, only 9.5 hours of data was annotated out of which 8.5 hours were used for ASR development of 5K words. System was evaluated on 1 hour of speech data and a minimum WER of 24.14% was attained.

In domain of speech recognition, most of the work has been done on Hindi among south Asian languages [13] which is much similar to spoken Urdu. Isolated word recognition [14], connected digit recognition [15], statistical pattern classification [17], online speech to text engine [18] and large vocabulary speech recognition systems have been developed. Upadhyaya et al. [12] developed a Hindi speech recognition system using deep neural networks on AMUAV Hindi speech database. This database consists of 1000 phonetically balanced sentences recorded by 100 speakers and covers 54 Hindi phones. Minimum WER achieved was 11.63% using Karel's DNN [11]. Though Hindi speech recognition systems have achieved a low WER, these systems can not be used as an alternative of Urdu ASRs due to substantial lexicon differences [19].

This paper presents collection of large Urdu speech corpus and development of deep neural networks based Urdu LVCSR

http://dx.doi.org/10.21437/Interspeech.2019-2629

system. Urdu speech data of 292.5 hours is recorded from 1671 speakers and annotated at sentence level. Along with readily available Urdu data, 300 hours of speech data covering a vocabulary size of 199K words are used for development of the system. Different state-of-the-art techniques for acoustic modeling such as GMM-HMM, Time-Delay Neural Network (TDNN) [5], Long-Short Term Memory (LSTM) [20], Bidirectional Long-Short Term Memory (BLSTM) [21] and a combination of TDNN and BLSTM networks (TDNN-BLSTM) with cross entropy and LF-MMI [26] loss functions are investigated to get minimum WER. Lattice free MMI based TDNN-BLSTM network outperforms for AMI [22] and Switchboard corpora of English speech data [23]. Impact of this network for Urdu data is investigated in this work.

To make it accessible for development of further speech interfaces, Urdu speech recognition system is available for developers as a web service[1].

2. Urdu Speech Corpus Development

For Urdu speech collection, text corpus is designed covering some available Urdu corpora [28, 29, 30], proper nouns, dates, months, news from different categories, 11 digits long telephone numbers, national identity card numbers of 13 digits and addresses. A large corpus of news is extracted from different Urdu news channels' websites and tweets. English, being the official language of Pakistan, is frequently mixed with Urdu even in everyday conversations. To cover this mixing, code-switched sentences between Urdu and English are also included. Around one thousand Urdu news containing most frequently used English words are extracted. After verification and rephrasing, 779 code-switched sentences are included in corpus.

Collection and annotation of speech data is a cumbersome task. Initially, phonetically rich text corpus [30] is used to record around 50 hours of speech data from 182 speakers in supervised recording sessions. This data is recorded in a clean environment through an automated utility. User interface for recording utility is shown in Figure 1. Speakers' information (unique ID, gender and channel) is provided before start of recording session. A speaker is asked to record the sentences appearing on screen one by one. On completion of a sentence, a speaker may proceed to next sentence or re-record current sentence in case of mispronunciation or linguist's advice. All the recorded data is manually verified by linguists.

After collecting 50 hours of speech data, a baseline GMM-HMM based Urdu ASR system is developed. Another utility is used for further data recording. Interface of utility, used in this step, is shown in Figure 2. Speakers' metadata such as unique ID, gender and channel is stored before session start. During recording session, speaker is directed to record the sentences appearing on screen. On completion of a speech utterance, it is decoded by ASR. If a sentence is perfectly decoded, it gets separated out and rejected otherwise. However, sentences with 1 and 2 word errors are reconsidered by linguists and are accepted or rejected after manual verification.

During speech corpus collection, gender and channel balancing is taken into consideration. Data is recorded from more than 1650 male and female Punjabi and Urdu speakers from age group ranging from 18-50 years. All the audios are recorded in WAV format at sampling rate of 16 KHz using USB microphone, USB headsets, hands-free and laptop microphone.

[1]Available at: https://tech.cle.org.pk/services/speech/asr

Figure 1: *User interface of recording utility for first 50 Hours*

For evaluation, a test corpus of 9.5 hours of Urdu speech is collected from 62 speakers. It is ensured that both the text corpus and the speakers should be unseen. It is designed in the same way as training data to ensure a balanced text corpus. Speech corpus recorded for test data is also gender and channel balanced. Summary of corpus collection for training and evaluation is given in Table 1.

Table 1: *Statistics of Urdu speech corpus*

	Training Data	**Testing Data**
Total Duration (in Hours)	292.5	9.5
Number of speakers	1586	62
Channels	USB microphone, USB headset, hands-free, laptop microphone	USB microphone, USB headset, hands-free, laptop microphone
Age Group	18-50	18-50

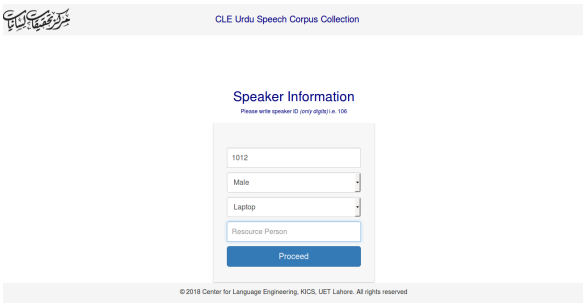
3. Experimental Setup

3.1. Lexicon

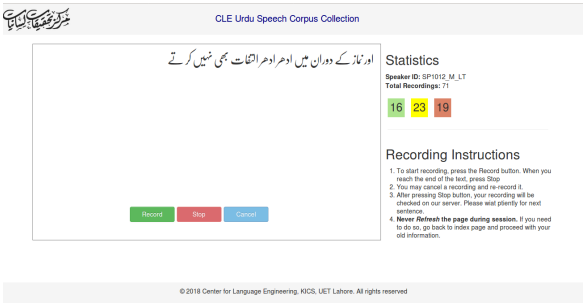
Urdu ASR is developed with a vocabulary size of 199K words. It includes 106K words from readily available Urdu lexicon [30] and 93K words added during corpus development.

3.2. Acoustic Modeling

For training of acoustic model, 300 hours of speech data including readily available Urdu speech corpora [25, 27, 30, 31] (8.5 Hours) and 292.5 hours of newly recorded data are used. Using this speech data, a baseline GMM-HMM acoustic model is built using Mel-Frequency Cepstral Coefficients (MFCC) features with 40 coefficients (high resolution MFCCs). This model is used to get alignments for DNN training. TDNN, LF-MMI based TDNN, LSTM, BLSTM and hybrid TDNN BLSTM deep networks are investigated. The best model among all is selected and fine tuned by varying different parameters such as number of hidden layers and cell dimensions. On each step, the best



(a) *Speaker information form for recording*



(b) *Recording interface*

Figure 2: *User interface of recording utility*

configuration is opted to proceed further. At the end, output lattice from the finest tuned configuration is selected for re-scoring using recurrent neural network language model. Kaldi toolkit [36] is used for Urdu ASR development.

3.3. Language Modeling

SRI Language Modeling (SRILM) toolkit [7] is used for building trigram language model. A very large corpus is collected by crawling a huge number of Urdu websites covering a number of categories such as news, magazines, books and blogs. Corpus is sentence tokenized and cleaned for only Urdu and code-switched sentences. This collected corpus along with readily available Urdu corpora [28, 29] is used for language modeling. It contains around 154 million Urdu words forming 35 million trigrams. This corpus is also used for training RNNLM to combine with best acoustic model.

4. Experimental Results

Various experiments with different configurations of deep neural networks for acoustic modeling are performed. Results are reported in terms of Word Error Rate (WER).

4.1. Acoustic Modeling

4.1.1. GMM-HMM

Using 300 hours of training data, a baseline system on GMM-HMM using high resolution MFCCs is built. Furthermore, speaker independent Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) transforms are investigated. On top of LDA+MLLT, Speaker Adapted Training (SAT) is done and its alignments are used for DNN training. Summary of WER of baseline systems is shown in Table 2.

Table 2: *%WER of GMM-HMM Urdu ASR*

Model	%WER
GMM	46.87
GMM+LDA+MLLT	37.26
GMM+LDA+MLLT+SAT	32.24

4.1.2. Deep Neural Networks (DNNs)

On alignments of SAT training, various deep neural networks are trained. DNN training is done with high resolution MFCCs and i-vector of dimensionality 100 for each sample. Network consists of 7 hidden layers. First layer is fixed affine layer whereas rest of the hidden layers are TDNN with cell dimensionality of 1024. *Relu-renorm* [16] is used as activation function. Experiment with LF-MMI based TDNN is also done and results show that it performs better than cross-entropy based TDNN. So, LF-MMI based networks are used for rest of the DNN experiments. Lattice free MMI based TDNN, LSTM and BLSTM networks are trained. Furthermore, reduced frame rate is used in decoding to speed up the process. These networks are termed as *chain models* in Kaldi. Comparison of all these experiments is shown in Table 3.

Table 3: *Comparison of %WER of differnet deep neural networks for Urdu ASR. Number of epochs fixed to 2*

Model	No. of hidden layers	Cell dim.	%WER
TDNN	7	1024	21.33
Chain TDNN	8	625	19.92
Chain TDNN-LSTM	7	512	19.18
Chain BLSTM	3	1024	19.38
Chain TDNN-BLSTM	5	1024	18.64

It is evident from Table 3 that chain TDNN-BLSTM outperforms for Urdu speech data also. In further experiments, it is optimized by varying different parameters to achieve best configuration.

4.1.3. Number of layers

Several experiments are done to optimize number of hidden layers for chain TDNN-BLSTM network. By default, network consists of 5 hidden layers among which first two are time delay neural layers while rest of 3 are bidirectional long short-term memory layers. Number of hidden layers are varied on this stage and results are shown in Table 4.

4.1.4. Hidden layers' size

For experiments shown in Table 4, number of neurons in hidden layers are fixed to 1024 per layer which means 1024 nodes per memory cells are used for each forward and backward direction. Table 5 compares the different layer sizes for chain

Table 4: *Comparison of hidden number of layers. Cell dimensions fixed to 1024, recurrent and non-recurrent projection dimension as 256, delay=-3, decay-time=20, No. of epochs=2*

No. of hidden layers	No. of param (M)	%WER
4 (2 TDNN + 2 BLSTM)	51.2	19.80
5 (2 TDNN + 3 BLSTM)	51.2	18.64
6 (2 TDNN + 4 BLSTM)	62.7	18.88
7 (2 TDNN + 5 BLSTM)	74.3	18.92

TDNN-BLSTM with best number of hidden layers.

Table 5: *Comparison of cell dimensionality (layer size). Number of hidden layer fixed to 5, recurrent and non-recurrent projection dimension as 256, delay=-3, decay-time=20, Number of epochs=2*

Layer size	No. of param (M)	%WER
512	26.8	28.6
1024	51.2	18.64

4.2. Language Modeling

All word error rates reported in last section are decoded using 3-gram language model. After choosing the best configuration of acoustic model, trigram LM is replaced with recurrent neural network based language model. For RNNLM training, TDNN-LSTM network is used with 3 TDNN and 2 LSTM hidden layers. Layer size is fixed to 1024 cells. Best lattice is re-scored using this model and WER is improved from 18.64% to 16.94% which is shown in Table 6.

5. Discussion

A post analysis is done on decoded output by aligning hypothesis and reference texts. It is found that digits are being decoded into words and vice versa. Sometimes, ASR decodes digit 6 as /چھ/ /ʧʰeː/ (six in Urdu) and /چھ/ /ʧʰeː/ as 6 or /٦/ (digit six in arabic script). In case of a single digit, error computation penalizes it as one substitution. However, for larger numbers, penalty goes higher. For instance, decoding year *2021*, ASR decodes it as /دو ہزار اکیس/ /d̪oː həzaːr ɪkkiːs/ (two thousand twenty-one in Urdu) that raises a penalty of three words (one substitution and two insertions). Conversely ASR is penalized as one substitution and two deletions.

Similarly, some Urdu words can be written in two alternate ways. And lexicon contains alternate orthographic representations of same pronunciation. For instance, there is a proper noun in Urdu /ابراہیم/ /ɪbraːhiːm/ which can be written as /ابراہیم/ or /ابراھیم/. If the decoded one is different than the one in reference text, ASR is penalized as one substitution.

Additionally, ASR intermittently inserts space in some words that are correct with or without space. Such words are correct acoustically and semantically in both conditions but may contradict to reference text. For example, the word /حقدار/ /həqd̪aːr/ is sometimes decoded as /حق دار/ /həq d̪aːr/ that is correct both acoustically and semantically but not correct for WER calculation.

To compensate such errors, retraining is done after text normalization of training and test sets' transcriptions. In text normalization process, all the numbers are standardized to same format (in words). All the words with same pronunciation but different orthographic representations are replaced with one of the representations. Words that are correct with and without spaces are replaced with the one having spaces. Furthermore, for all cases, redundant entries from lexicon are removed. After retraining the acoustic model, WER is further reduced to 13.50% from 16.94% (shown in Table 6).

Table 6: *Comparison of WER after RNNLM and text normalization*

	%WER
3-gram LM	18.64
RNNLM	16.94
Text normalized acoustic model+RNNLM	13.50

6. Conclusion

This paper presents collection of Urdu speech corpus of 292.5 hours from 1586 speakers. A large vocabulary Urdu speech recognition system is developed using 300 hours of microphone speech data from 1671 speakers. A text corpus of 154 million words is developed for 3-gram and neural network based language modeling. For evaluation of speech recognition system, a test data set of 9.5 hours is collected from 62 unseen speakers. Different state-of-the-art modeling techniques are investigated to develop Urdu LVCSR system. After evaluation of various techniques for acoustic modeling, TDNN-BLSTM network is chosen to develop the system. Decoded output lattice is re-scored using RNNLM. To compensate error due to insertion or deletion of spaces and alternate orthographic representations of same pronunciation, text normalization is done and acoustic model is retrained. A minimum WER of 13.50% is achieved on test data set. Speech corpora, collected in this work, can also be used for development of various other speech technologies such as Urdu speakers' recognition, age estimation and gender identification systems.

7. References

[1] N. Patel, S. Agarwal, N. Rajput, A. Nanavati, P. Dave, and T. S.Parikh, "A comparative study of speech and dialed input voice in-terfaces in rural india," in SIGCHI Conference on Human Factorsin Computing Systems. ACM, 2009.

[2] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim,R. Tongia, and R. Rosenfeld, "Healthline: Speech-based access tohealth information by low-literate users," in ICTD. IEEE, 2007.

[3] British Broadcasting Coorporation (BBC)., [Online] Available: http://www.bbc.co.uk/languages/other/urdu/guide/facts.shtml (Last Accessed on March 12, 2019).

[4] J. Ashraf, N. Iqbal, N. S. Khattak, A. M. Zaidi, "Speaker Independent Urdu Speech Recognition," in International Conference on Informatics and Systems (INFOS), Cairo, Egypt, 2010.

[5] V. Peddinti, D. Povey, S. Khudanpur, "A time delay neural networkarchitecture for efficient modeling of long temporal contexts," In Sixteenth Annual Conference of the International Speech Communication Association 2015.

[6] L. Besacier, E. Barnard, A. Karpov, T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," Speech Commun. , vol. 56, pp. 85–100, Jan. 2014

[7] Andreas Stolcke, "SRILM – an extensible language modeling toolkit," In Proceedings of the International Conference on Spoken Language Processing, Vol. 2, pages 901–904.

[8] L. E. Baum, J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," Bulletin of American Mathematical Society, vol. 73, pp. 360–363, 1967.

[9] M. Gales and S. Young. "The application of hidden markov models in speech recognition," Founds. Trends Signal Process., 1(3):195–304, January 2007.

[10] V. Manohar, D. Povey, and S. Khudanpur, "Semi-supervised maximum mutual information training of deep neural network acoustic models," in Proc. Interspeech, Dresden, Germany, Sep. 2015, pp. 2630–2634

[11] K. Vesely, A. Ghoshal, L. Burget, D. Povey, "Sequence-discriminative training of deep neural networks," In Proceedings of Interspeech. 2013

[12] P. Upadhyaya, S. K. Mittal, O. Farooq, Y. V. Varshney, M. R. Abidi, "Continuous Hindi Speech Recognition Using Kaldi ASR Based on Deep Neural Network," In: Tanveer M., Pachori R. (eds) Machine Intelligence and Signal Analysis. Advances in Intelligent Systems and Computing, vol 748. Springer, Singapore (2019)

[13] D. Dash, M. Kim, K. Teplansky, J. Wang, "Automatic Speech Recognition with Articulatory Information and a Unified Dictionary for Hindi, Marathi, Bengali and Oriya," in Interspeech 2018.

[14] U. G. Patil, S. D. Shirbahadurkar, A. N. Paithane, "Automatic Speech Recognition of isolated words in Hindi language using MFCC," in 2016 International Conference on Computing, Analytics and Security Trends (CAST), Dec 2016, pp. 433–438.

[15] A. Mishra, M. Chandra, A. Biswas, S. N Sharan, "Robust features for connected Hindi digits recognition," International Journal of Signal Processing, Image Processing and Pattern Recogni-tion, vol. 4, no. 2, Jun 2011.

[16] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," In ICML, 2010.

[17] R. K. Aggarwal, M. Dave, "Using gaussian mixtures for hindi speech recognition system," in International Journal of Signal Processing, Image Processing and Pattern Recognition, 2011.

[18] B. Venkataramani, "SOPC-based speech-to-text conversion," in Nios II Embedded Processor Design Contest Outstanding Designs, 2006.

[19] K. V. S. Parsad and S. M. Virk, "Computational evidence that Hindi and Urdu share a grammar but not the lexicon," In the 3rd Workshop on South and Southeast Asian NLP, COLING (2012).

[20] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in Proc. INTERSPEECH, 2014, pp. 338–342.

[21] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in Proc. Int. Conf. Artif. Neural Netw.: Formal Models Their Appl., 2005, pp. 799–804.

[22] I. McCowan et al., "The AMI meeting corpus," in Proc. 5th Int. Conf. Methods Tech. Behav. Res., 2005, vol. 88.

[23] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," IEEE Signal Processing Letters, 2017.

[24] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018.

[25] H. Sarfraz, S. Hussain, R. Bokhari, A.A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, R. Parveen." Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System," in O-COCOSDA 2010.

[26] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na,Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networksfor ASR based on lattice-free MMI," in Proc. Interspeech, pp. 2751–2755, 2016.

[27] Raza A., Hussain S., Sarfraz H., Ullah I., and Sarfraz Z.,"Design and Development of Phonetically Rich Urdu Speech Corpus," in Proceedings of IEEE Oriental COCOSDA International Conference on Speech Database and Assessments, Urumqi, pp. 38-43, 2009.

[28] S, Urooj, S, Hussain, F, Adeeba, F. Jabeen, R. Perveen, "CLE Urdu Digest Corpus", in the Proc. of Conference on Language and Technology 2012 (CLT12), Lahore, Pakistan, 2012.

[29] F. Adeeba, Q. Akram, H. Khalid, and S. Hussain, "CLE Urdu BooksN-gram," in Proc. Conf. Lang. Technol., Karachi, Pakistan, 2014, pp. 87–92.

[30] F. Adeeba, S. Hussain, T. Habib, E. Ul-Haq, K. S. Shahid, "Comparison of Urdu text to speech synthesis using unit selection and HMM based techniques", Presented at the Oriental COCOSDA Bali (Indonesia, 2016)

[31] B. Mumtaz, A.Hussain, S. Hussain, A.Mehmood., R. Bhatti, M. Farooq, S. Rauf.,"Multitier Annotation of Urdu Speech Corpus", Conference on Language and Technology (CLT14), Karachi, Pakistan, 2014

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

[33] M. A. B. Shaik, Z. Tukse, M. A.Tahir, M. Nubaum-Thom, R. Schluter, H.Ney,"Improvements in RWTH LVCSR evaluation systems for Polish, Portgeese, English, Urdu and Arabic", in Sixteenth Annual Conference of the ISCA, 2015.

[34] M. Qasim, S. Nawaz, S. Hussain, T. Habib, "Urdu speech recognition system for district names of Pkaistan: Development, challenges and solutions", in O-COCOSDA, 2016.

[35] A. A. Raza, A. Athar, S. Randhawa, Z. Tariq, M. B. Saleem, H. B. Zia, U. Saif, R. Rosenfeld, "Rapid Collection of Spontaneous Speech Corpora Using Telephonic Community Forums," in Proc. Interspeech 2018 (2018), 1021–1025.

[36] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, "The Kaldi Speech Recognition Toolkit," In IEEE 2011 Workshop on Automatic Speech Recognitionand Understanding. IEEE Signal Processing Society, 2011.