Assessing Performance of Bengali Speech Recognizers Under Real World Conditions Using GMM-HMM and DNN Based Methods

Soma Khan, Madhab Pal, Joyanta Basu, Milton S. Bepari, Rajib Roy

Centre for Development of Advanced Computing, Kolkata, India

{soma.khan,madhab.pal,joyanta.basu,milton.bepari,rajib.roy}@cdac.in

Abstract

Real world Automatic Speech Recognition (ASR) system development requires rigorous performance review under varying real world conditions. This paper reports our effort on ASR resource creation, transcription, system building and performance assessment for connected and continuous word applications in Bengali language (ranked seventh worldwide) using GMM-HMM and DNN framework on available open source toolkits. Baseline models are built from merging Bengali dataset hosted on government sites with application specific 100 hours indigenous audio collected under target deployment scenario. After feedback analysis of live systems by real users, novel Error Handling Techniques like Signal Analysis and Decision, Confidence based ASR output Polling and Runtime LM are implemented which results around 2%-11% overall gain in WER with encouraging task success rates in final field trials. Results suggest that recent approaches along with application, environment, target user and runtime resource specific appropriate strategies will yield better acceptability of live ASR systems in India.

Index Terms: Bengali speech recognizer, ASR Error Handling Techniques, Real world speech recognition

1. Introduction

India, with its multilingual and multicultural heritage serves a wide platform for field level implementation of Automatic Speech Recognition (ASR) related recent technological advancements. However, there remain some interesting India specific phenomena which create multi-faced challenges while developing real world live ASR systems in India, like:

- uneven distribution of languages across native language speaking communities. The year 2001 census indentified 122 major languages in India, out of which 29 languages have more than a million native speakers. While other 1,599 languages are spoken by smaller societies, local groups and tribes in the interior areas.
- non-availability of written script, non-standardized lexemes, phonemes, lexicons etc.
- concurrent use of multiple languages in spoken as well as written communications
- effect (accent, style, lexemes) of learned language (L2) over native language (L1)
- significant population within the target user group for many of the real time speech applications in India is non tech-savvy, semi literate or even illiterate

Thus, problems like “target language selection”, “resource constraints”, “zero or limited resources”, “code switching”, “multi-language learning” and “non-cooperative end users” are all here to add to the “technological challenges” for real world ASR development. This is why extremity of the same needs to be assessed carefully and thoroughly before starting the actual development. Recent advances in Deep Neural Network (DNN) based architectures [1,2,3,4], multilingual[5] and cross-lingual speech data processing techniques and machine-learning based models incorporating high level linguistic knowledge [6,7,8] are reported to be helpful in building systems for low and resource constraint languages. Most of this studies used benchmark datasets from worldwide known data sources (like LDC) which often fail to replicate real world field conditions. Scarcity of the required amount of Indian language audio and linguistic resources (transcript, lexicon et.) within this datasets also creates a major hindrance for applying recent data-driven model training approaches to ASR making.

2. Purpose of the study

Here in this paper, we present our work on development of field level ASR applications on Bengali language. Uniqueness of our work lies in the series of experiments that we conducted to assess Bengali language ASR performance:

i) for both connected and continuous word applications on TDIL-DC [9] hosted data resources along with indigenously collected around 100 hours Bengali field data under real world environments

ii) while varying the underlying core technology setup from baseline GMM-HMM [10], then LDA, MLLT with SAT training [11] to recent days DNN framework on popular open source toolkits, eg. HTK [12], Sphinx [13], Kaldi[14]

iii) under field conditions with live audio from real users and discussed novel Error Handling Techniques (EHT) introduced in the final system after feedback analysis

This paper reports our indigenous effort in Bengali language on conducting and reporting a series of ASR experiments on real world speech comprehensively within a single research paper. We acknowledge the privilege of using Bengali as our choice of Indian language in this study, as this is our native language and also have moderately available resources in terms of standardized script, lexemes, phonemes, Grapheme to Phoneme rules [15] that provided a good start point for ASR system development. Still it can be considered as under resourced in terms of digitally available transcribed speech data amount which is again essential for ASR model building.

3. Bengali: the language and its ASR

Bengali is a part of the Indic group of Indo-Aryan (IA) branch of the Indo-European family of languages. It is the official state language of East Indian state West Bengal (WB) and the national language of Bangaladesh. Over 300 million people across the world speak Bengali, making it the seventh most

10.21437/SLTU.2018-40

spoken native language in the world [16]. As per 2001 census, Bengali (8.86% population) is the second most spoken official language in India after Hindi (53.6% population). Colloquial Bengali language carries a wide range of pronunciation and accent variations that subsequently formed five major dialectal groups, known as Rarhi, Barendri, Jharkhandi, Kamrupi and Bangali. Mainly influenced from the South Western or West Central branch of Rarhi dialect, the Standard Colloquial form of Bengali (SCB) is spoken around WB state capital Kolkata and its surrounding areas [17] and is mostly used for official purposes. SCB phoneme inventory includes total 47 unique sounds with 33 consonants and 14 (7 orals + 7 nasals) vowels. IPA representation of all phonemes is given in Table 1.

Table 1: *Bengali phoneme inventory*

	Vowels	
Oral	/ɔ/, /a/, /æ/,/o/,/e/,/i/,/u/	7
Nasal	/ɔ̃/,/ã/,/æ̃/,/õ/,/ẽ/,/ĩ/,/ũ/	7
	Consonants	
Plosives	(Unaspirated Unvoiced) /k/,/ʈ/,/t/,/p/	4
	(Aspirated Unvoiced) /kʰ/,/ʈʰ/,/tʰ/,/pʰ/	4
	(Unaspirated Voiced) /g/,/ɖ/,/d/,/b/	4
	(Aspirated Voiced) /gʰ/,/ɖʰ/,/dʰ/,/bʰ/	4
Affricates	/ʧ/, /ʧʰ/, /ʤ/, /ʤʰ/	4
Nasals	/ŋ/, /ɳ/, /n/, /m/	4
Fricatives	/ʃ/,/s/,/h/	3
Lateral	/l/	1
Flaps	/ɽ/,/ɽʰ/	2
Trills	/r/	1
Glides	/J/ , /ʋ/	2
	Total Phonemes	47

Isolated efforts have been made in earlier studies to develop ASR systems for small task specific applications in Bengali language [18, 19]. A work on optimal text selection method to develop Bengali speech corpus for phone recognition task has been reported in [20]. A comprehensive review of ASR performance on different Indian language datasets using recent ASR training approaches can be found in [5], though there remains some uncertainty on the data sources and collection of the same under real world outdoor conditions beside usability of the same for task specific ASR applications. Thus, ASR works on application oriented real world speech data is rather scarce for Bengali language as per literature review.

4. Experiments on Bengali ASR systems

Two different applications for connected (two to three words) and continuous word ASR applications are developed in Agricultural and Travel domains respectively for real world usage. Specification of both the application is summarized in table 2. Following sections describe experiment details on finding the best suited ASR models for the two applications and unique strategies to make them ready for real world usage.

4.1. Connected word ASR application (Agri domain)

The real time application is actually a telephony spoken dialog system [21] designed to disseminate regional agricultural commodity market prices and weather information (with forecast) to the target users (mostly farmers) on recognition of spoken district, commodity and market names in Bengali language over telephone channel. Thus, real world transcribed audio, high accuracy yet low latency core ASR, technology illiterate real users were the obvious challenges for us.

Table 2: *Specification of the two ASR applications*

Spec.	Connected word	Continuous Word
App. type	Telephony IVR system in Agriculture domain	Standalone GUI based system in Travel domain
Data type	District, Market and Commodity names, Right, Wrong etc.	Sentence level general query on transportation
Data size	198 hrs	99.4 hrs
Sampling encoding	8000 Hz, 16 bit mono, ms wav PCM	22050 Hz, 16 bit mono, ms wav PCM
Environment	Outdoor data with all kind of field noises	Real world, mostly in house and less noisy

To include as many as pronunciation variations in the baseline acoustic models, real time speech of 3000 speakers are indigenously collected from all dialectal regions of West Bengal state and then transcribed by native transcribers. A threefold cross validation testing is done using Sphinx [13] toolkit on the three orthogonal partitions of development dataset, that yields around 80% percent correct, but 38% average Word Error Rate (WER) mostly due to insertions for untreated real world noises. Table 3 presents the list of real world noises (with individual occurrence percentage) that we marked in the transcribed train data. AM parameter tuning experiment on entire dataset finally reports the lowest WERs at 3000 senones and 32 mixtures as 13.92% and 11.25% using Finite state Grammar and statistical tri-gram Language Models (LM) respectively. With similar parameters, LDA, MLLT, SAT and SGMM [22] training is further performed on the baseline AM of entire development set following *timit* recipes under KALDI [14, 23] toolkit. The initially designed ASR with multiple models is then field trialed on the telephony IVR system by real users in field conditions. Figure 1 shows a photograph of the live ASR system field trial.

Table 3: *Noise tags: glimpse of real world field speech*

Noise tag	% Occurrence	Noise tag	% Occurrence
<babble>	15.1	<beep>	5.59
<bn>	13.91	<bird>	5.22
<laugh>	9.08	<vehicle>	4.8
<air>	7.92	<ct>	4.03
<cough>	6.9	<animal>	3.72
<ring>	6.08	<breath>	3.4
<line>	6.06	<bang>	1.09
<horn>	6.02	<hesitate>	1.08

*<bn> → general back noise, <ct> → clearing of throat

Figure 1: *A snapshot from live ASR field trial*

The field trial process reveals some interesting feedbacks on the system and overall performance, as stated below:

- First time users were quite hesitant, speakers often remained silent, spoke too fast, before the beep or spoke even after the specified time. These lead to truncation; recognition errors or too many insertions while silence.
- Co-operative speakers often get dissatisfied with confirmation at each recognition stage in the call flow as it needs extra effort from user and delays the process
- Too much information often confused speakers like for the query commodity, when price information is provided for all the four-five listed (in Agmarknet) markets within a district reporting on latest date.
- In actual field trials, recorded speech data contain loud speaking, continuous background speech, heavy noises of vehicles' horn, air flow etc.

Aligned results obtained after transcription of live test speech (collected during field trial) is presented in table 4 within "before EHT" column.

Table 4: *ASR results of Connected word application*

Model description		**Best %WER before EHT**	**Best %WER after EHT**
Model name	**Method**		
Sphinx AM + trigram LM	Baseline GMM-HMM with MFCC+D+DD	21.05	10.4
Kaldi tri1	Baseline GMM with MFCC + D+DD	17.92	13
Kaldi tri2	tri1with LDA + MLLT	16.35	12.73
Kaldi tri3	tri2+ SAT	14.52	9.2
Kaldi SGMM	tri3 + SGMM	9.25	7.9
Amount of data: Train.: 238775uttr (198 hrs), Test: 10327 uttr		Train Lexicon: 4859 words, Decode Lexicon: 2800 words	

4.2. Error Handling Techniques for Real world live ASR

Following Error Handling Techniques are designed and incorporated into the final system as per feedback analysis:

4.2.1. Signal Analysis and Decision (SAD):

This filter like special module before ASR decoding, detect and remove undesirable speech inputs at runtime, so that miss recognitions due to silence, truncation, clipping, channel drop, extreme noise can be avoided to a certain extent. Incoming speech signal is treated based on some extracted temporal (zero crossing rate, short time energy etc.) and spectral (formants) features and then the signal accept-reject decision is taken by comparing those features with a previously stored knowledge base [24] created from transcribed development data. For rejection decision, audio is discarded and user is prompted to speak again. Otherwise the recoded speech is directly sent to core ASR engine. Table 5 and table 6 shows offline testing results of SAD with field audio.

Table 5: *Filtration statistics of offline SAD module*

#Test utterances	**Channel Reject**	**Silence Reject**	**Truncate Reject**	**General reject**
10327	659	795	248	458

Table 6: *Accuracy of decision of SAD module*

Types of Rejection	**Channel Reject**	**Silence Reject**	**Truncate Reject**	**General reject**
% correct decision	65.85	86.98	87.26	32.65

4.2.2. Confidence Measure (CM) and Polling:

System prompted confirmation in each of the successive ASR nodes (district, commodity, market) in connected word ASR application annoyed the target user in initial field trials. To avoid this confirmation, system needs certain confidence in its performance, so that ultimate ASR output is reliable to proceed further. Previous studies on this can be found in [25]. However, we followed a novel method, where three decoders are built around the entire development dataset using Sphinx (with trigram LM), Kaldi tri3, and Kaldi SGMM methods respectively. From parallely generated decoded output from the three decoders, relative occurrence frequencies of hypothesis words are calculated.

Comparing the highest relative frequency with empirically set thresholds of confidence level, a decision of confidence and next plan of action is decided whether to skip explicit confirmation, ask for output confirmation or to prompt for re-record by the user. The polling decision logic is given below:

If,

Confidence of j th word is, $C_{w(j)}$ = (count w_j)/(count w_{all}), then

Case 1: if (*count* (*max* (C_w)) > 1 || *max* (C_w) < 0.2), then loop back to current node and prompt for re-record

Case 2: else if (*max* (C_w) >= 0.2 && *max* (C_w) < 0.5), then take confirmation from user

Case 3: else proceed to next ASR node without confirmation

Table 7 shows the results of offline polling experiment with transcribed test audio. Using the confidence measure and polling logic a slight decrease in %WER (14.2) is noticed and hence the same was applied to the final system.

Table 7: O*ffline polling experiment results on Live audio*

Results	**Decoder 1 (Sphinx) % WER**	**Decoder 2 (Kaldi tri3) % WER**	**Decoder 3 (Kaldi SGMM) % WER**	**Polled outpu t**
%WER	33.5	25.2	18.7	14.2
ASR output confidence %	**All matched**		**Any two matched**	**No match**
	62.0		27.0	11

4.2.3. Runtime LM generation:

Idea here is to organize the searched information in an intelligent way so that, user will never get confused with too much disbursed information at a time. In our final system, commodity price information disbursement for up to two markets (within the recognized district) at a time is allowed. For more than two markets, system will play out respective market names and prompt user to say any one. At this stage, system will go for market name recognition with runtime generated LM out of the available markets only. Instead of having a large static LM of all market names, this little change made the decode process faster, easier and more accurate.

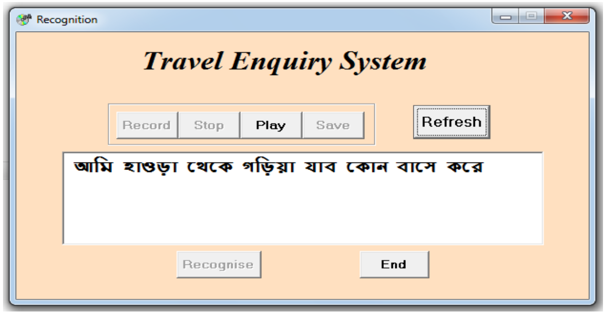
Table 8: *Task success statistics in field trials*

Specification	Before EHT	After EHT
Total calls to live ASR system	907	823
Valid calls (clear speech	753	733
Successful calls (correct info retrieved)	559	601
Success % of valid calls	74.24	81.99
Success % of total calls	61.63	73.03

From table 4 last "after EHT" labeled column, and table 8 "After EHT" column, it is noticeable that, applied EHT within the target application framework not only decreases the %WER, but also helps in achieving good task success rate. EHT application on GMM-HMM gives comparable ASR results to that of recent methods on field audio under real world conditions.

4.3. Continuous word ASR application (Travel domain)

This simple real time application is our recent course of activity where, user will speak complete sentence like general queries on road travel, and after continuous word recognition system will show the output onscreen as shown in Figure 2.



আমি হাওড়া থেকে গড়িয়া যাব কোন বাসে করে

Figure 2: *A screenshot of Continuous word Bengali ASR*

This primary application will be included within a live "Travel Enquiry Chat App", where after recognition of spoken query, and valid keywords like source, destination place names, desired arrival and departure time, mode of transport etc. in successive ASR nodes, system will play out proper response and finally provide available travel options using Bengali TTS. Understanding the underlying valid query structures and designing high accuracy continuous word ASR models are the initial challenges for which we built a variety of ASR models with rule based as well as statistical LM using all the popular toolkits and also applied DNN based training. Baseline development dataset is built by merging data from TDIL hosted Bengali sentence corpus [9] of 9.4 hrs and indigenously collected travel domain data of around 90 hrs. Table 5 shows results of our experiments on 3hrs unseen test data of real world. We trained DNNs on top of the tri3 models and used recipes of *nnet2,* popularly known as Dan' DNN [2] with 4 hidden layers each of 256 dimensions. The initial learning rate is 0.015 and the final learning rate is 0.002.

Table 5: *ASR results of Continuous word application*

Model description		Best % WER
Model name	Method	
HTK results	Baseline GMM-HMM with MFCC + D + DD	21.63
Sphinx results	Baseline GMM-HMM with MFCC + D + DD	13.28
Kaldi tri1	Baseline GMM with MFCC + Delta + Double Delta	10.84
Kaldi tri2	Delta + Double Delta + LDA + MLLT	13.40
Kaldi tri3	Delta + Double Delta + LDA + MLLT + SAT	13.73
DNN with tri3	tri3+ Dan's DNN	9.74
Data amount: Train: 99.4 hr (160475uttr), Test: 17300 uttr	Lexicon size: 46033 words	

In our recent efforts, the continuous word standalone application is getting integrated within the mobile app and then it will be used for measuring ASR performance on more real world data under outdoor environment for real time usage. For, both the applications it is seen that ASR models built on recent methods (using Kaldi toolkit) have outperformed GMM-HMM based baseline models as a whole.

5. Conclusion

This paper summarizes our indigenous efforts on design and development of real world deployable ASR systems in Bengali language for task specific connected and continuous word live applications. Experiments are conducted to fix up the core ASR technology and also design novel ASR error handling methods for applicability of the same under target real world scenario. Resource scarcity is managed by adding adequate amount of indigenous field collected audio along with TDIL site hosted publicly available Bengali language transcribed speech data. Developed applications in our work, faced representative live speech from both non tech-savvy semi-literate and literate end users under real world condition and yield promising results. We are currently collecting resources from other Indian languages for applying cross-lingual model training methods to improve recognition performance specially for continuous word ASR application. Once adequate resources are available, we also like to conduct comparative performance analysis under multi-language experiment settings to see whether the proposed approach generalizes for other Indian languages so that it comply more with India specific multi-language speaking scenario.

6. Acknowledgements

This work is a part of ASR consortium project funded by MeitY, Ministry of Electronics and Information Technology, Government of India. The authors are thankful to CDAC, Kolkata, India for necessary financial and infrastructural support in further carrying out present research activities.

7. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.R. Mohamed, N. Jaitly, A.Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, "Deep Neural Networks for Acoustic

Modeling in Speech Recognition, The Shared Views of Four Research Groups", In IEEE Signal Processing Magazine, Vol. 29, No. 6. , November 2012, pp. 82-97

[2] D. Povey, X. Zhang and S. Khudanpur, "Parallel training of Deep Neural Networks with Natural Gradient and Parameter Averaging", in proc. of ICLR Workshop 2015

[3] V. Peddinti, D. Povey and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts", in *INTERSPEECH 2015*

[4] K. Vesel´y, A. Ghoshal, L. Burget, D. Povey, "Sequence-discriminative training of deep neural networks", in *INTERSPEECH 2013*

[5] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein and K. Rao, "Multilingual Speech Recognition With A Single End-To-End Model," ICASSP 2018

[6] Xu, H., Su, H., Ni, C., Xiao, X., Huang, H., Chng, E.S. and Li, H., Semi-Supervised and Cross-Lingual Knowledge Transfer Learnings for DNN Hybrid Acoustic Models Under Low-Resource Conditions. Proc. *INTERSPEECH* 2016, 1315-1319.

[7] P. Bell, J. Driesen, and S. Renals, "Cross-lingual adaptation with multi-task adaptive networks", In proc. *INTERSPEECH* 2014

[8] S. Thomas, M. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in IEEE ICASSP 2013.

[9] Language Resource Development, (www.tdil-dc.gov.in), link: http://www.tdil-dc.in/index.php?option=com_vertical&parentid=36&lang=en

[10] M. Gales and S. Young, *The Application of Hidden Markov Models in Speech Recognition*, Foundations and Trends in Signal Processing, Vol. 1, No. 3 (2007) 195-304, 2008, DOI: 10.1561/2000000004

[11] S. P. Rath, D. Povey, K. Vesel´y and Jan H. Cernock´y, "Improved feature processing for Deep Neural Networks," *INTERSPEECH 2013*

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book Version 3.4, Cambridge University Press, 2006.

[13] CMU SPHINX available at: http://www.speech.cs.cmu.edu/

[14] KALDI [Online], Available at: http://kaldi-asr.org/

[15] J. Basu, T. Basu, M. Mitra, S. Kr Das Mandal, "Grapheme to Phoneme (G2P) Conversion for Bangla", Year of Pub. (2009), O.COCOSDA, 2009 IEEE Explorer of Oriental COCOSDA-2009, pp. 66 – 71.

[16] https://en.wikipedia.org/wiki/Bengali_language

[17] S. K. Chatterji, *The Origin and Development of the Bengali Language*, published by Rupa & Co., New Delhi, 1926.

[18] M. R. A. Kotwal, T. Halim, M. M. H. Almaji, I. Hossain and M. N. Huda, "Extraction of Local Features for Tri-Phone Based Bangla ASR," 2012 Ninth International Conference on Information Technology - New Generations, Las Vegas, NV, 2012, pp. 668-673.

[19] F. Hassan, M. R. A. Kotwal, G. Muhammad, MLN-based Bangla ASR using context sensitive triphone HMM, International Journal of Speech Technology, 2011, Volume 14, Number 3, Page 183

[20] S. Mandal, B. Das, P. Mitra, and A. Basu, "Developing Bengali speech corpus for phone recognizer using optimum text selection technique", In 2011 International Conference on Asian Language Processing (IALP), pp. 268–271, November 2011

[21] J. Basu, S. Khan, R. Roy and M. S. Bepari, "Commodity Price Retrieval System in Bangla: An IVR Based Application", India HCI 2013, 24-27 September, 2013, Bangaluru, India.

[22] D. Povey et al., "The subspace Gaussian mixture model: A structured model for speech recognition," Comput. Speech Lang., vol. 25, no. 2, pp. 404–439, Apr. 2011.

[23] D. Povey et. al., "The Kaldi Speech Recognition Toolkit", in IEEE 2011 workshop on Automatic Speech Recognition and Understanding", ASRU, Dec. 2011.

[24] J. Basu, M. S. Bepari, R. Roy and S. Khan, "Real Time Challenges to Handle the Telephonic Speech Recognition System", in Proc. of ICSIP 2012, (pub. at Lecture Notes in Electrical Engineering, Vol. 222, DOI: 10.1007/978-81-322-1000-9_38, @ Springer India 2013), , Coimbatore, India. Dec. 2012, pp. 395-408.

[25] H. Jiang, "Confidence measures for speech recognition: A survey", Speech communication, Volume 45, pages 455-470, 2003