2022 IEEE 19th India Council International Conference (INDICON) | 978-1-6654-7350-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/INDICON56171.2022.10039815

On Improving Performance of NeMo ASR system for Indian Accent English

Sourav Chattopadhyay, Junaid Hamid Bhat, Sowmya Rasipuram, Aditi Debsharma, Anutosh Maitra
Accenture Research Labs
Bengaluru, India
sourav.chattopadhyay, junayd.jmi@gmail.com, sowmya.rasipuram, aditi.debsharma, anutosh.maitra@accenture.com

***Abstract*—While many speech recognition applications are quite commonplace now a days, applicability of the recognition techniques to accented languages is still a major challenge. Here, in this paper we have fine-tuned a state-of-the-art convolutional ASR system on regional accented Indian English to improve it's performance for accented English. We used NVIDIA NeMo[1] as our base framework and used QuartzNet 1D Convolutional model to explore their ability to understand language specific phonology. We also fine-tuned the convolutional network on regional accented Indian English. The tests were carried out on a dataset containing male and female English speech samples collected over 13 different regional accents. The experiments were bench-marked against the output generated by well known speech services and the effect of the convolution was found to be positive.**

***Index Terms*—speech recognition, ASR systems, Indian English**

I. Introduction

Automatic Speech Recognition (ASR) is a field of computer science that aims to design the computer system to recognize human speech. It has been an active area of research since many decades. The first works on ASR were based on statistical modeling techniques like Hidden Markov Models (HMM) [1] where phonemes were used to represent distinct sounds that make up the word. With the increase in the availability of data in the recent few years, deep learning techniques have become very popular [2], [3]. With the help of various complex deep learning architectures, many end-to-end systems have been introduced. End-to-end ASR systems convert the speech signal to characters/words without pronunciation dictionary, acoustic or language model. Developing such end-to-end ASR systems will be difficult for low-resource languages where the availability of aligned annotated data is limited. Linguistic diversities also makes it difficult to adapt models across different languages.

The use of end-to-end ASR systems for adapting to low-resource language accents is possible with the help of transfer learning. We also found that, in the recent past only a very few research work has shown some analysis of ASR performance for regional accented Indian English. In this paper, we investigated the suitability of well known ASR system for few regional accented Indian English. The main contributions of this work are as follows, 1) Use of NVIDIA Nemo QuartzNet ASR system to assess performance on a subset of English language accents from Indic database. 2) Fine-tune QuartzNet model and analyze the differences in performance for various language accents. 3) Perform in-the-wild experiments with speech samples from one male and female speaker in a noisy environment. These experiments help us understand the model that fits well for the specific language accent.

This paper is organized as follows. In Section II, we present literature survey on different ASR systems available. Section III describes our methodology and details about the QuartzNet model. In Section IV, we present detailed summary of performance of ASR systems. Finally, we conclude in Section V.

II. Related Work

Prior works on ASR that were based on HMM-GMM (Gaussian Mixture Model) frameworks [4] and were built on three models namely acoustic model, pronunciation Model and language Model. SPHINX system developed by Kai Fu Lee used HMM to model the speech state over time and GMM to model the HMM states [5]. HMM-GMM based approaches dominated the speech recognition systems for many years. But, these statistical approaches do not perform well when tested on conversational speech [6]. The training process is also complex and followes independent optimization for different modules.

One of the first deep learning methods for ASR was proposed by Dahl et al. [7]. The proposed method was based on a combined HMM-DNN (Deep Neural Network) that achieved significant improvement in the performance over conventional HMM-GMM based approaches. End-to-end ASR systems that map audio signal to a sequence of words without the designing of intermediate states, have gained attraction. These end-to-end ASR systems does not require hard alignment of states with the audio signal. The end-to-end ASR systems can be categorized into CTC-based, RNN-transducer and Attention-based [6].

There are many modern end-to-end ASR systems that are available to use without the need of training from scratch. Many cloud-based solutions such as Google Cloud Speech API[2], IBM Watson Text to Speech[3], Microsoft Azure Bing Speech API[4], Amazon Transcribe[5] etc. are available. These cloud-based solutions are not only built on heavy training data

[1] https://github.com/NVIDIA/NeMo
[2] https://cloud.google.com/speech-to-text
[3] https://cloud.ibm.com/apidocs/speech-to-text

978-1-6654-7350-7/22/$31.00 ©2022 IEEE

but also support to customize for specific domains. Along with cloud-based solutions, many opensource resources that provide robust solutions such as Deepspeech [8], Deepspeech2 [9], Wav2Letter [10] are available. Deepspeech uses a well optimized RNN based training and is found to work robustly in noisy environments. Wav2Letter is an end-to-end system that is based on convolution network for acoustic modeling and a graph-based decoding [10]. Luo et al. proposed cross-language transfer learning and domain adaptation for an end to end ASR [11].

Hannun et al. proposed a fully convolutional sequence-to-sequence architecture that produced a notable performance improvement on LibriSpeech with an Word Error Rate (WER) of 3.28 on clean validation set [8]. The time-depth separable convolutions significantly reduced the number of parameters by keeping a large receptive field. QuartzNet uses a similar convolution based approach that uses depth-wise separable convolutions [12]. The input data is considered in a time-channel format that completely decouples the time and channel convolutions. This approach reduced the number of parameters significantly while obtaining a huge improvement in the WER of 2.68 on clean test data. Both of these approaches use convolutional methods and obtain state-of-the-art performance on well established datasets. But, their adaptability to regional accented Indian English has not been studied. Phonological differences in various regional accents in India make it difficult to adapt existing systems with less available data. Also, not many prior works have focused on developing models for Indian accented English. Das et al. presented an approach of domain adversarial training (DAT) along with transfer learning to improve the performance of an ASR system. They proposed a semi-supervised training strategy by implementing DAT and transfer learning together to improve the accent robustness of a trained ASR model [14]. Dubey et al. showed an approach to train an end-to-end ASR system for many different Indian English accents. They presented this approach to improve Deep Speech ASR system by fine-tuning [15].

In this work, we made an attempt to perform transfer learning on some of the regional accented Indian English dataset. We used QuartzNet model to perform our experiments on Indic dataset as the model shows promising results on different accented benchmarked datasets with a relatively low number of parameters. Our experiments prove the suitability of the models for regional accented Indian English languages and in-the-wild experiments.

III. Methodology

In this section, we describe details about our experiments on Nemo ASR systems on various Indian accented English files. We measure WER on test data and show our experiments on fine-tuned QuartzNet model.

TABLE I
Distribution of number of hours of data for various English Language Accents

Accent	Type	Gender	No. of hours
Assamese	English	Male	11.30
Assamese	English	Female	12.05
Bengali	English	Male	10.03
Bengali	English	Female	5.20
Gujarati	English	Male	10.13
Gujarati	English	Female	10.00
Hindi	English	Male	7.10
Hindi	English	Female	7.22
Kannada	English	Male	7.48
Kannada	English	Female	7.50
Tamil	English	Male	10.90
Tamil	English	Female	12.70
Telugu	English	Male	10.15
Telugu	English	Female	10.28

A. Indic Dataset

The Indic corpus from IITM[6] contains more than 50GB of speech samples with different speakers from 13 states in India. It has speech samples from both male and female native speakers with more than 10000+ utterances. Utterances are spoken in mono (corresponding language) and English. The audio files are available with the transcript in the dataset. We considered only English spoken utterances in this paper. We performed experiments on 7 language accents - Assamese, Bengali, Gujarati, Hindi, Kannada, Tamil and Telugu. The distribution of data w.r.t number of hours for male and female speakers for each language accent is given in Table I.

B. NVIDIA NeMo

NeMo provides two end-to-end ASR systems called Jasper [13] and QuartzNet [12], that are based on series of 1D Convolutionals. Jasper (Just Another Speech Recognizer) model consists of repeated blocks of 1D convolutions. A KxR Jasper model is made up of R sub-blocks (each consisting of a 1D convolution, batch normalization, Rectified Linear Unit (ReLU), and dropout) repeated K number of times. Jasper achieves significant improvement in the performance over state-of-the-art models on LibreSpeech data.

QuartzNet is a better variant of Jasper that uses time-channel separable 1D convolutions. It is based on Jasper architecture and uses Connectionist Temporal Classification (CTC) loss [2]. QuartzNet model has a 1D convolutional layer in the beginning followed by a sequence of blocks. Each block B_i is repeated S_i times and has residual connections between the blocks. Each block consists of 4 types of layers that perform depth-wise convolution, point wise convolution, normalization and ReLU activation. And the last part of the model again consists of three convolutional layers. Figure 1 shows the block diagram of the QuartzNet architecture. QuartzNet model shows similar performance as that of Jasper with significant reduction in the number of parameters. And hence, we utilized QuartzNet model for our experiments. It is trained on 6

[4] https://azure.microsoft.com/en-in/services/cognitive-services/speech-services/

[5] https://aws.amazon.com/transcribe/

[6] https://www.iitm.ac.in/donlab/tts/database.php

TABLE II
PERFORMANCE (WER) OF NEMO PRE-TRAINED, FINE-TUNED MODELS ON INDIC DATASET

	Male				Female			
Accent	NeMo	Fine-tuned	Third-party	Accent type	NeMo	Fine-tuned	Third-party	Accent type
Assamese	16.16	**14.27**	25.21	Heavy	13.12	**9.15**	22.45	Heavy
Kannada	9.41	**7.60**	22.38	Moderate	9.30	**8.50**	20.90	Moderate
Bengali	10.04	**8.79**	20.34	Moderate	**8.80**	**8.84**	20.56	No influence
Tamil	**8.12**	8.90	22.35	No influence	**6.43**	6.64	20.32	Faint
Telugu	**10.15**	11.30	30.21	No influence	9.84	**8.90**	26.31	Heavy
Gujarati	**4.31**	6.90	14.25	Moderate	**6.68**	7.56	15.37	Faint
Hindi	**5.60**	6.86	16.26	Faint	**4.45**	6.42	15.87	Moderate

datasets namely LibriSpeech, Mozilla Common Voice , WSJ, Fisher, Switchboard, and NSC Singapore English that included 6000+ hours of data. It shows a very low WER of 2.69 on clean LibriSpeech test data.

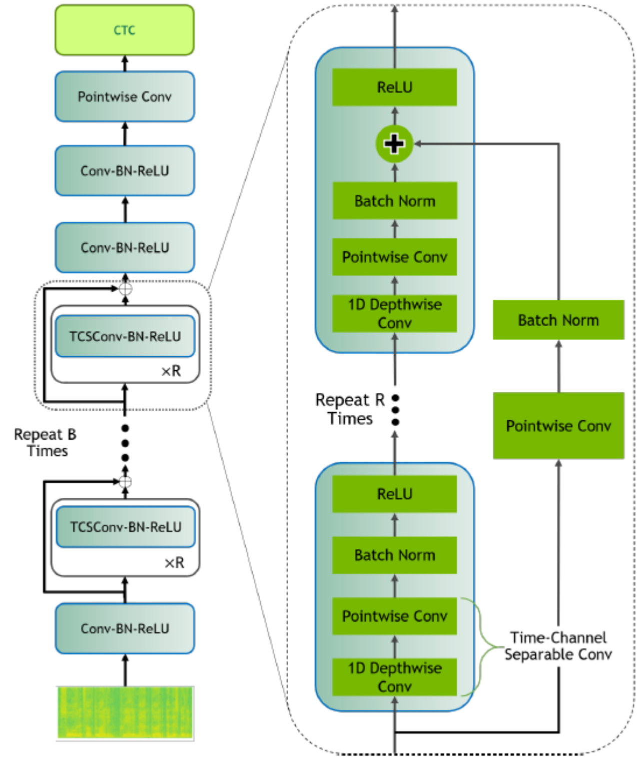

Fig. 1. QuartzNet Model Architecture. Courtesy:taken from [12]

1) Fine-tuning NeMo: NeMo pre-trained QuartzNet model is used to fine-tune on various Indian accent English data files. Fine-tuning is performed to do transfer learning on domain specific or accent-specific language datasets. Fine-tuned models are developed for 7 languages in Indic dataset. The average number of hours of training data per language is 8 hours. We performed our experiments on a single Tesla K80 GPU of 12 GB. Average training duration per language was 50 hours and fine-tuning was done for 10-20 epochs. We used a batch size of 32 or 16 depending upon the data available for training. We varied learning rate in the range of (0.0005 - 0.001) for all the experiments. NovoGrad optimizer with weight decay of 0.001 and beta values of 0.95, 0.25 was used during fine-tuning.

C. Cloud Services

Cloud-based services are very widely used for performing downstream tasks such as ASR, TTS. In this paper, along with QuartzNet model, we used a third party cloud-based service to perform experiments on Indic dataset. The third party service also shows promising results on Indic dataset.

IV. EXPERIMENTAL RESULTS

We present the performance of NeMo pre-trained and fine-tuned models on seven diverse language accents. The seven languages used in our study are Assamese, Bengali, Gujarati, Hindi, Kannada, Tamil and Telugu. Experiments are performed on both male and female. The audio files from each language accent are tested on each of the fine-tuned models to obtain the transcript. Performance is measured using WER that is calculated using the reference transcript. WER is computed using the **jiwer**[7] package from python library.

Table II depicts the WER values of the NeMo pre-trained, Nemo fine-tuned and third-party model on each of the seven datasets. It is observed from the experiments that NeMo fine-tuned model performs consistently better than the pre-trained model in case of both male and female speakers for languages like Assamese, Kannada and Bengali. It is also observerd that fine-tuning is not helping much to improve the performance for other languages.

Figure 2 demonstrates a comparison for the average WER value of male and female speakers on the three models. When performance of male speakers is considered, the average WER on NeMo pre-trained, NeMo fine-tuned and third-party models are 9.11, 9.23 and 21.28 respectively. While the WER values of female speakers are relatively better than male value. The WER value for the female speakers for the three models are 8.37, 8.00 and 19.84 respectively. This indicates that the fine-tuned models perform slightly better for female audio samples.

A. Error Analysis

Based on the above experiments, we found 3 different categories of impact of fine-tuning on language accents. We observed improvement after fine tuning for languages like Assamese, Bengali and Kannada. We observed a very less

[7] https://pypi.org/project/jiwer/

TABLE III
EXAMPLES OF UTTERANCES SPOKEN BY MALE AND FEMALE SPEAKERS AND ITS GENERATED TRANSCRIPT BY A NEMO PRE-TRAINED. **TEXT IN BOLD**: KEY-WORDS IDENTIFIED CORRECTLY, *italics*: WORDS IDENTIFIED INCORRECTLY

Spoken utterances	Generated Transcript - Male / Female
It is raining heavily throughout *Kashmir*	it as raining heavily throughout *cashmere*
I like **Jingle Bells** rhyme a lot	i like **jingle bell's** rhyme a lot
Government announced *lock down* in few of the states	Government announced *logdown* in few of the states
This **pandemic** has taken away lives of many throughout the world	This **pandemic** has taken away lives of many throughout the world
What is the use of taking medicines this disease is incurable	What is the use of taking medicines this disease is incurable

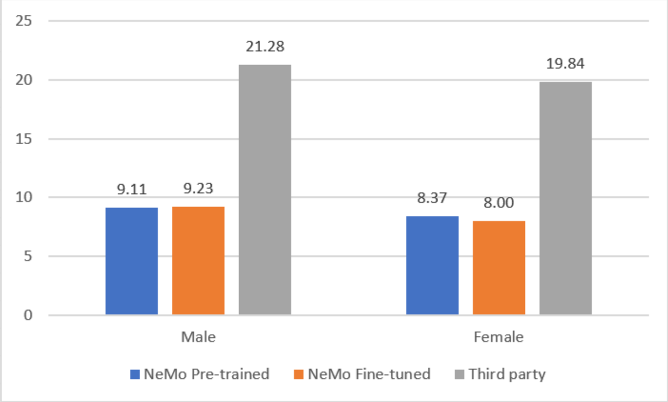

Fig. 2. Differences in WER of Male and Female speakers

degree of improvement or close to the accuracy of pre-trained for languages like Tamil and Telugu and we didn't register any improvement for languages like Hindi and Gujarati. An expert linguist manually listened to 20 audio files from each differently accented languages and based on influence of the mother tongue, accents are classified into 4 categories - Heavy (very high influence of mother tongue while speaking English), Moderate (standard influence of mother tongue accent), Faint (a very minor influence of mother tongue accent) and No-influence (no influence of mother tongue is observed while speaking English).

We observed for languages like Hindi and Gujarati, where the influence of mother tongue accent is very less - the WER is observed to be low with NeMo pre-trained model. Fine-tuning is also not recommended in such cases as it is already doing good with pre-trained model. On the other hand for languages like Tamil and Telugu, there was a faded influence of mother tongue accent and we did not observe performance improvement over pre-trained model except for Telugu female. For languages like Assamese, Bengali and Kannada, there was a prominent influence of mother tongue and and we observed a good improvement after fine-tuning. A language with an accent that has a prominent influence of mother tongue accent, is more likely to perform poor on pre-trained model and in such cases fine-tuning can be helpful to improve the performance.

B. Experiments in the Wild

The experiments shown above are performed within the train-test split of the Indic dataset. In this section, we perform experiments on a small, in-the-wild dataset that contains 20 utterances (English) spoken by one male and female non-native speakers. The male speaker has Hindi and female speaker has Telugu accent. Table III shows a sample of the spoken utterances and the generated transcript using a NeMo pre-trained model for Male speaker. The transcript generated for the female speaker is same as that of male and hence not shown in the table. The utterances were selected randomly and included some nouns like Kashmir, Jingle Bells. The male and the female speakers recorded their utterances in a noisy environment with voice recorder application in Windows machine.

It is observed that the noun ”Jingle Bells” has been recognized correctly, but the word ”Kashmir” is predicted as ”cashmere” for both male and female. Other words like pandemic, incurable are predicted correctly. The word ”lockdown” appears as ”logdown” for both speakers. The total WER computed on 40 utterances spoken by both male and female is 1.12. The WER is observed to be very less and can be computed on more in-the-wild utterances for generalization.

V. CONCLUSION AND FUTURE WORK

In this paper, an overview of the performance of the state-of-the-art convolutional model on Indian accented English is studied. The convolutional QuartzNet model has been fine-tuned for various regional accented English language and the performance of the fine-tuned model is compared against pre-trained model and a third-party cloud solution. It has been observed that the fine-tuned model outperforms pre-trained model where there is a heavy influence of mother tongue accent. This paper shows quite clearly, the influence of language accent on the performance of the pretrained ASR model. This experiment shows the suitability of using NeMo fine-tuned model for different language accents. The experimental results can be shown for all 13 language accented files and the performance can be compared against other convolutional methods for concreteness.

VI. ACKNOWLEDGEMENTS

Accenture sincerely thanks NVIDIA and team for providing early resources of NeMo for experiments presented here.

REFERENCES

[1] J. Baker, ”The DRAGON system–An overview,” in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23, no. 1, pp. 24-29, February 1975, doi: 10.1109/TASSP.1975.1162650. London, vol. A247, pp. 529–551, April 1955.

[2] Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376).

[3] Graves, A., Jaitly, N. (2014, June). Towards end-to-end speech recognition with recurrent neural networks. In International conference on machine learning (pp. 1764-1772). PMLR.

[4] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society.

[5] K. . -F. Lee, H. . -W. Hon and R. Reddy, "An overview of the SPHINX speech recognition system," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, no. 1, pp. 35-45, Jan. 1990, doi: 10.1109/29.45616.

[6] Wang, D., Wang, X., Lv, S. (2019). An overview of end-to-end automatic speech recognition. Symmetry, 11(8), 1018.

[7] Dahl, G. E., Yu, D., Deng, L., Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on audio, speech, and language processing, 20(1), 30-42.

[8] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

[9] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning (pp. 173-182). PMLR.

[10] Collobert, R., Puhrsch, C., Synnaeve, G. (2016). Wav2letter: an end-to-end convnet-based speech recognition system. arXiv preprint arXiv:1609.03193.

[11] Luo, J., Wang, J., Cheng, N., Xiao, E., Xiao, J., Kucsko, G., ... Li, J. (2021, July). Cross-Language Transfer Learning and Domain Adaptation for End-to-End Automatic Speech Recognition. In 2021 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[12] Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., ... Zhang, Y. (2020, May). Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6124-6128). IEEE.

[13] Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., ... Gadde, R. T. (2019). Jasper: An end-to-end convolutional neural acoustic model. arXiv preprint arXiv:1904.03288.

[14] Das, N., Bodapati, S., Sunkara, M., Srinivasan, S., Chau, D. H. (2021). Best of both worlds: Robust accented speech recognition with adversarial transfer learning. arXiv preprint arXiv:2103.05834.

[15] Dubey, P., Shah, B. (2022). Deep Speech Based End-to-End Automated Speech Recognition (ASR) for Indian-English Accents. arXiv preprint arXiv:2204.00977.