Short Term Context-based Fast Multilingual Acoustic Model for Low Resource Languages

Bipasha Sen*
Aditya Agarwal*
bise@microsoft.com
adiagar@microsoft.com
Microsoft

ABSTRACT

Multilingual automatic speech recognition (ASR) systems have led to a major step forward towards building robust ASR systems for languages with low resource availability by increasing coverage for individual languages. State of the art multilingual systems are developed with sequential networks such as recurrent neural networks (RNNs) to capture long term temporal dependencies. Training and inference in such sequential models are computationally expensive, which poses a significant challenge in terms of scalability and real-time applications. In this paper, an alternate architecture based on short term contextual temporal features learned on convolutional neural networks (CNNs) with a non-sequential discriminative network is proposed. Three low resource Indic languages, Gujarati, Tamil, and Telugu are used to ascertain that our proposed architecture trains 5.5× faster and reduces the inference time by a factor of 26 while maintaining comparable word error rates (WERs) against our baseline RNN.

KEYWORDS

Multilingual ASR, Low Resource, Fast ASR, CNN-DNN

1 INTRODUCTION

In recent years, the use of speech recognition systems has soared across the globe. However, such speech recognition systems are still largely limited to only a handful of languages. It thus becomes imperative to build robust and scalable techniques that can be used to quickly bootstrap automatic speech recognition (ASR) systems for multiple languages. Multilingual ASR system is a single entity capable of transcribing speech utterances for multiple languages with a shared phone space. Multilingual ASR systems have seen tremendous growth and have proven to be a viable solution for building robust speech recognition systems often outperforming the monolingual counterparts, especially for languages with low resource [2, 4, 8, 9, 11]. One way to build such a system is to leverage the shared phone space between languages by combining individual phone sets to build a common phone set. The shared phonetic space increases the training data for individual phonemes enabling the model to learn better acoustic representations. This is especially beneficial for low resource languages.

Several techniques have been used to build such multilingual systems. Strategies like shared hidden layers [5], bottleneck features [15], multitask learning [3] train monolingual models on multilingual data. One approach to building such monolingual models is to share the hidden representations between the languages while keeping the output layer language-specific. Such systems are operated by either placing a Language Identification (LID) front-end to switch to the corresponding monolingual model or by selecting the best hypothesis obtained by running all the monolingual models. The performance of the former approach is largely dependent on the robustness and accuracy of the front-end LID system while the latter approach requires multiple monolingual models to be operated parallelly.

Most individuals in multilingual countries like India engage in code-switching during a spoken conversation [17]. Code-switching is a phenomenon of mixing multiple languages in a single utterance. Due to the language dependency in the above-mentioned models, operating such models in a code-switched environment becomes very tricky. To employ large scale speech recognition systems in such multilingual countries, it is necessary to build language-agnostic acoustic models. This study uses the common phone set to build a computationally efficient language independent joint acoustic model capable of handling multiple languages in a single system. A parser is used to convert utf8 text format to a language-independent IT3 format [1]. The IT3 format text is then used to generate pronunciation sequences for all the words.

The discriminative networks in the acoustic models are largely built on sequential networks such as recurrent neural networks (RNNs) [7, 12, 18]. Sequential networks have the inherent property of capturing long term contextual dependencies in a sequence. Fields like natural language processing, machine translation are largely dependent on such sequential architectures. Such networks are however computationally expensive due to limited scope in parallelization. Deep neural network (DNN) on the other hand is a non-sequential architecture, providing computational efficiency but at a cost of high WERs [6]. Phonemes have very short contextual dependencies. For instance, the pronunciation of phoneme *k* is dependent on its neighboring phonemes in cat (*k*-***ae***-*t*), car (*k*-***aa***-*r*), hack (*hh*-***ae***-*k*), sky (***s***-*k*-***ay***). However, in the sentence *"this is a cat" (dh-ih-s ih-z* ***ah*** *k*-***ae***-*t)*, the pronunciation of phoneme *k* is dependent only on neighboring phones, ***ah*** and ***ae*** but is independent of all the other phonemes in the sentence [13]. CNNs have been researched in the field of monolingual ASR as a feature extraction layer to improve the WERs on DNNs while maintaining the computational efficiency by *learning* the short term contextual dependencies on the acoustic frames [6].

Recently, there have been substantial developments on end to end (E2E) acoustic modeling which combines multiple components to optimize the system as a single unit. Techniques based on attention networks [6, 7], connectionist temporal classification (CTC) [8–10] transcribe acoustic frames into phones without the need of any

*Both authors contributed equally to this research.

Table 1: Performance comparison on the average % WER, average % WER degration, training and inference time.

				Training		Inference	
Exp	Models + Context	Avg. WER	Avg. WER deg.	Time	Speed Up	Time	Speed Up
1	lstm + mfcc	**22.63**	-	~ 4.5 days	-	780 ms	-
2	cnn + raw + $\{0, 0\}$	28.75	−6.12	7.84 hours	~ 13.5×	15ms	~ 52×
3	cnn + raw + $\{-1, +1\}$	28.05	−5.42	11.56 hours	~ 9×	29ms	~ 26×
4	cnn + raw + $\{-2, +1\}$	**24.89**	−2.26	19.08 hours	~ 5.6×	30ms	~ 26×
5	cnn + raw + $\{-1, +2\}$	27.31	−4.68	21.63 hours	~ 5×	32ms	~ 24×
6	cnn + raw + $\{-2, +2\}$	29.05	−6.42	24.42 hours	~ 4.4×	89ms	~ 8.7×

predefined alignments. Techniques based on Convolutional neural networks (CNNs) transcribe raw speech utterances directly into phones, eliminating the step of hand-crafted feature extraction [11–16]. In this paper, inspiration from the work done on CNNs in [10] and [6] is taken to build an E2E ASR system and learn short term contextual temporal features on acoustic frames using convolutional layers. Three low resource Indic languages, Gujarati, Telugu, and Tamil are used with a combined training dataset of 75 hours. WER is used as the evaluation metric for all the experiments.

The contribution of this paper is two-fold:

(1) We have shown that CNNs can be used as a viable solution for building E2E robust multilingual ASR systems.
(2) We have proposed a methodology that can be used to quickly bootstrap multilingual ASR systems and validate the compatibility of different languages in a joint multilingual setting.

2 EXPERIMENTAL DATA

The Data is a subset of Interspeech 2018's Low Resource Speech Recognition Challenge for Indian languages by Microsoft and SpeechOcean.com dataset [1] [14].

India is a country with more than 1500 recognized languages. Out of these, 30 languages have more than one million speakers and 22 languages have been accorded with the official status [16]. Such diversity in spoken languages poses a significant challenge in obtaining sizable training data to train robust monolingual systems for each of these languages. This dataset was released as an effort to explore robust multilingual systems as a means to overcome the challenge of data limitations. The data includes three Indic languages namely Gujarati, Tamil, and Telugu spoken by multiple speakers. The combined training data of all three languages is 75 hours. The test and the validation data are 5 hours per language. Text transcription along with the lexicon for the entire data is included in the dataset. State of the art systems were built on the full dataset, which has 120 hours of trainable data. The baseline DNN based system in the challenge has a WER of 27.79, 34.97, and 25.47 on Gujarati, Telugu, and Tamil respectively [14].

3 RESULTS

Table 1 presents a comparative analysis of computational time and average WER degradation of different configurations against the baseline model. The unit of the training time is kept variable to enable better readability. The inference time is calculated as T/N where T is the total inference time on the test set and N is the number of examples. All the models are trained, and inferences are drawn on single-core NVIDIA TITAN Xp GPUs in a multi-core GPU setup.

Our best model achieves a WER of 21.13, 27.80, and 25.76 on Gujarati, Telugu, and Tamil respectively with 2 left and 1 right context window (Exp 4, Table 1). The results are comparable to the baseline with an average degradation of only 2.26 WER. A computation boost of 5.6× in training time and 26× in inference time is obtained with this configuration taking only an average of 30ms to infer on an average of 5.85s long speech utterance compared to 780ms on the baseline. It is observed from Table 1 that decreasing the size of the context window boosts up the training and inference computations but at a cost of degradation in the WERs. Models trained on acoustic frames with no context (Exp 2) and large context (Exp 6) observe a significant average degradation of ~ 6.2WER. This indicates no context doesn't provide all the relevant articulation and coarticulation information for the phones while providing large contexts introduces more noise than relevancy. It is also observed that adding more context on the left (Exp 4) gives slightly better WER as compared to adding more context on the right (Exp 5) for the same context window size with comparable training and inference time speed up indicating that the left context contains more relevant information compared to the right context.

4 CONCLUSION

In this paper, a non-sequential discriminative approach based on the features extracted by CNNs is investigated. Experimental results show that such systems vastly reduce the training and inference time while producing comparable WERs against our baseline sequential model based on RNN. This shows great potential and promise for CNN based architectures in real-time applications for achieving low latency and can also be used to quickly bootstrap multiple low resource multilingual ASR systems. It is observed that a context window of size 2 on the left and 1 on the right produces the most optimal WERs while boosting the training and inference time by 5.5 and 65× respectively. CNN based architectures can be further researched to improve and obtain better WERs compared to sequential models while speeding up the training and inference time.

[1] The dataset is available at https://msropendata.com/datasets/7230b4b1-912d-400e-be58-f84e0512985e.

ACKNOWLEDGMENTS

We would like to thank Sunayana Sitaram, Rajeev Gupta, and Sandipan Dandapat, who are researchers at Microsoft, for their valuable feedback and guidance.

REFERENCES

[1] Arun Baby, Nishanthi N. L., Anju Leela Thomas, and Hema A. Murthy. 2016. A Unified Parser for Developing Indian Language Text to Speech Synthesizers. In *Text, Speech, and Dialogue - 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings (Lecture Notes in Computer Science, Vol. 9924)*, Petr Sojka, Ales Horák, Ivan Kopecek, and Karel Pala (Eds.). Springer, 514–521. https://doi.org/10.1007/978-3-319-45510-5_59

[2] Astik Biswas, Emre Yılmaz, Febe de Wet, Ewald van der Westhuizen, and Thomas Niesler. 2019. Semi-supervised acoustic model training for five-lingual code-switched ASR. arXiv:1906.08647 [cs.CL]

[3] D. Chen and B. K. Mak. 2015. Multitask Learning of Deep Neural Networks for Low-Resource Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 7 (2015), 1172–1183.

[4] Noor Fathima, Tanvina Patel, Mahima C, and Anuroop Iyengar. 2018. TDNN-based Multilingual Speech Recognition System for Low Resource Indian Languages. 3197–3201. https://doi.org/10.21437/Interspeech.2018-2117

[5] A. Ghoshal, P. Swietojanski, and S. Renals. 2013. Multilingual training of deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7319–7323.

[6] Pavel Golik, Zoltán Tüske, Ralf Schlüter, and Hermann Ney. 2015. Convolutional neural networks for acoustic modeling of raw time signal in LVCSR. In *INTERSPEECH*.

[7] A. Graves, A. Mohamed, and G. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6645–6649.

[8] Anjuli Kannan, Arindrima Datta, Tara N. Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee. 2019. Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model. arXiv:1909.05330 [eess.AS]

[9] Hari Krishna, Krishna Gurugubelli, Vishnu Vidyadhara Raju V, and Anil Kumar Vuppala. 2018. An Exploration towards Joint Acoustic Modeling for Indian Languages: IIIT-H Submission for Low Resource Speech Recognition Challenge for Indian Languages, INTERSPEECH 2018. In *Proc. Interspeech 2018*. 3192–3196. https://doi.org/10.21437/Interspeech.2018-1584

[10] Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M. Cohen, Huyen Nguyen, and Ravi Teja Gadde. 2019. Jasper: An End-to-End Convolutional Neural Acoustic Model. arXiv:1904.03288 [eess.AS]

[11] Bhargav Pulugundla, Murali Karthick Baskar, Santosh Kesiraju, Ekaterina Egorova, Martin Karafiát, Lukas Burget, and Jan Černocký. 2018. BUT System for Low Resource Indian Language ASR. 3182–3186. https://doi.org/10.21437/Interspeech.2018-1302

[12] H. Sak, Andrew Senior, and F. Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (01 2014), 338–342.

[13] A. Senior, H. Sak, and I. Shafran. 2015. Context dependent phone models for LSTM RNN acoustic modelling. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4585–4589.

[14] Brij Srivastava, Sunayana Sitaram, Rupesh Mehta, Krishna Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. 2018. Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages. 11–14. https://doi.org/10.21437/SLTU.2018-3

[15] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky. 2012. Multilingual MLP features for low-resource LVCSR systems. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 4269–4272. https://doi.org/10.1109/ICASSP.2012.6288862

[16] Wikipedia contributors. 2020. Language — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Language&oldid=955858850 [Online; accessed 12-May-2020].

[17] Emre Yılmaz, Henk van den Heuvel, and David Van Leeuwen. 2016. Code-Switching Detection Using Multilingual DNNS.

[18] Shiyu Zhou, Yuanyuan Zhao, Shuang Xu, and Bo Xu. 2017. Multilingual Recurrent Neural Networks with Residual Learning for Low-Resource Speech Recognition. In *INTERSPEECH*.