A Simple Baseline for Domain Adaptation in End to End ASR Systems Using Synthetic Data

Raviraj Joshi
Flipkart, Bengaluru
raviraj.j@flipkart.com

Anupam Singh
Flipkart, Bengaluru
anupam.s@flipkart.com

Abstract

Automatic Speech Recognition(ASR) has been dominated by deep learning-based end-to-end speech recognition models. These approaches require large amounts of labeled data in the form of audio-text pairs. Moreover, these models are more susceptible to domain shift as compared to traditional models. It is common practice to train generic ASR models and then adapt them to target domains using comparatively smaller data sets. We consider a more extreme case of domain adaptation where text-only corpus is available. In this work, we propose a simple baseline technique for domain adaptation in end-to-end speech recognition models. We convert the text-only corpus to audio data using single speaker Text to Speech (TTS) engine. The parallel data in the target domain is then used to fine-tune the final dense layer of generic ASR models. We show that single speaker synthetic TTS data coupled with final dense layer only fine-tuning provides reasonable improvements in word error rates. We use text data from address and e-commerce search domains to show the effectiveness of our low-cost baseline approach on CTC and attention-based models.

1 Introduction

End-to-end speech recognition models simplify the speech recognition process by folding multiple components into a single model. These models directly convert the speech utterance into the spoken text (He et al., 2019). The major end-to-end architectures include CTC-based, attention-based, and transducer-based approaches (Graves et al., 2013; Graves, 2012; Chan et al., 2015). These all neural approaches are competitive in terms of performance however they require a large amount of supervised data to achieve generalization. A model trained on a single application domain doesn't work well on other target domains. Examples of such applications domains include e-commerce, voice search, medical, etc. Since it is not feasible to prepare supervised data for all the application domains, it is common to train models on large out of domain corpus followed by a small amount of in-domain finetuning (Bell et al., 2020). However, this approach still requires the availability of small labeled data. In the most basic form, the unlabelled text data from the target domain can be used to build domain-specific language models (LMs). The domain LMs are combined with the end-to-end ASR model using shallow fusion (Kannan et al., 2018; Shan et al., 2019; Meng et al., 2021). This approach has limited benefits since the main ASR model is not tuned to the target domain. Another popular technique is to prepare synthetic data using a Text to Speech (TTS) system and the target domain text data (Sim et al., 2019). This requires a sophisticated multi-speaker TTS system followed by the addition of representative noise to make the data usable. The idea is to make synthetic data as close as the real-world data. However, this approach is prone to overfitting as the synthetic data does not exactly resemble real-world noisy conditions. Different fine-tuning approaches have been explored using synthetic data to alleviate the overfitting problem.

In this work, we are concerned with domain adaptation techniques when a text-only corpus from the target domain is available (Gao et al., 2021). We present a simple baseline approach using single speaker synthetic TTS data followed by final dense layer only fine-tuning. The synthetic data is created using a single speaker TTS system which is commonly available and also easier to build in-house. The data is not subjected to any noise and is directly used to fine-tune the neural network. Although such single speaker data is easy to build it is not usable for the training of end-to-end networks. We, therefore, propose dense-only fine-tuning for effective fine-tuning. The approach solely relies on final dense layer fine-tuning

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 244 - 249
May 26, 2022 ©2022 Association for Computational Linguistics

to avoid over-fitting on single speaker and acoustic conditions. We refer to the dense layer projecting the intermediate embedding onto vocabulary space as the final dense layer. Since the acoustic encoder of the neural network is frozen, the network only learns about the linguistic characteristic of the target domain. Similar approaches have been explored in literature where only the decoder part of the neural network is fine-tuned. However, this approach is not applicable to CTC-based neural networks (Graves et al., 2006) which do not follow an encoder-decoder architecture. We present our approach in the context of CTC and Listen-Attend-Spell (LAS) based neural network architectures. For LAS-based network, we also compare dense only and decoder only fine-tuning. We consider the text from address (for delivery of e-commerce products) domain and voice search (of e-commerce products) domain (Joshi and Kannan, 2021) for fine-tuning the model trained on a generic multi-domain dataset. Although encoder only fine-tuning has been widely studied in the literature (Mimura et al., 2018), this is the first work to exploit dense-only fine-tuning which is more relevant to the CTC-based systems. Moreover, we demonstrate a way to build an ASR system for the Address domain which is not explored in the literature.

2 Related Work

Our work is at the intersection of data augmentation using the TTS system and domain adaptation. In this section, we review the recent work in these two areas. The synthetic data generated using the TTS system was used to improve the recognition of out of vocabulary (OOV) words in (Zheng et al., 2021). Both synthetic data containing OOV words and original data were used together to train the best RNN-T model. Encoder freezing and elastic weight consolidation were further shown to provide extra benefits. Similarly, (Peyser et al., 2019) used a TTS system to generate numeric training data and improve the ASR performance on the out of vocabulary numeric sequences. The importance of data augmentation over semi-supervised learning was shown in (Laptev et al., 2020). In this work, the TTS system was trained on the same supervised ASR data set and used to generate synthesized samples on a wider set. The work also highlights the importance of multi-speaker TTS systems and noise addition to build usable systems. Other data augmentation techniques like spec augment (Park et al., 2019) were shown to be complementary with TTS based augmentation in (Rossenbach et al., 2020). Effective training strategies for using synthetic data were proposed in (Fazel et al., 2021). In order to avoid catastrophic forgetting, multi-stage training was used. The encoder layers were frozen in the initial stage followed by full fine-tuning in later stages. An elastic penalty was also added to the loss function so to avoid large deviation in learned parameters.

Similar approaches have been proposed for domain adaptation as well with a bais towards fine-tuning based transfer learning approaches. An LSTM-based domain classifier was trained to select an appropriate domain adapted language model in (Liu et al., 2021). The corresponding domain-specific language model was used for second pass re-scoring. The transfer learning approaches for domain adaptation and cross-language adaptation were evaluated in (Huang et al., 2020). They compare the fine-tuning of the pre-trained QuartzNet model with the corresponding model trained from scratch. They concluded that large pre-trained models performed better than small pre-trained models and the models trained from scratch. Another form of transfer learning involves partial fine-tuning of the model instead of the entire model. The decoder only fine-tuning for domain adaptation in Listen-Attend-Spell (LAS) based model was evaluated in (Ueno et al., 2018). The model is first trained on the source domain followed by decoder only fine-tuning on the target domain. The partial fine-tuning is shown to work better than the full fine-tuning and from the models trained from scratch. An adaptation technique specific to RNN-T networks using text-only data was proposed in (Pylkkönen et al., 2021). The prediction network of RNN-T is viewed as a neural language model and is adapted using text-only corpus while keeping the encoder and joint network fixed. Another approach for adapting RNN-T network using text-only data was proposed in (Li et al., 2020). The fine-tuning of prediction and the joint network was performed using synthetic TTS domain-specific data. Partial fine-tuning was shown to work better than full fine-tuning approaches. These works mainly used RNNT-based systems and employ a multi-context multi-speaker TTS system. In this work, we use a single speaker TTS system with a focus on CTC and attention-based models. Moreover, we focus on dense only fine-tuning instead of decoder fine-tuning studied in these works.

3 Methodology

The flow of our process is depicted in Figure 1. We follow a simple pre-training and fine-tuning approach. The model is first trained on general out-of-domain data. The target domain text data is converted into audio using a single speaker TTS engine. The synthetic samples are then used to fine-tune the final dense layers of ASR models. We consider two model types i.e CTC based models and Attention-based models. The model architecture and TTS system description are provided in the following sub-sections.

3.1 Model Architecture

We consider CTC and attention-based ASR models which follow the same pre-processing steps (Joshi and Kannan, 2021). The audio is segmented into 20ms chunks with an overlap of 10ms. Log-mel features are computed and provided as input to the model. Standard spec-augment is used for time and frequency masking of the spectrograms (Park et al., 2019). An 80-dimensional log mel feature is computed per time step. Three consecutive features are stacked to give a final feature of size 240. The output vocabulary size consists of sub-word units of size 5000. The sentence piece library is used to train the subword model using the generic out-of-domain data (Kudo, 2018).

The CTC-based model consists of a series of stacked LSTM layers followed by a final dense layer projecting the hidden vectors onto the vocabulary space. The LSTM consists of 700 units at all levels. A total of 12 LSTM layers are present with a final dense layer of size 700 x 5001. The final vocabulary element is reserved for the blank token. The CTC loss function is used to train the model.

The attention-based model follows a transformer LAS architecture. It consists of 10 encoder layers and 2 decoder layers. All the layers are standard transformer blocks. The internal model dimension is 512 units and the feed-forward dimension is 2048 units. Each block has 4 attention heads with 1024 units each. The final dense layer on the decoder side has a size of 512 x 5000. The same generic vocabulary is used for all the experiments. This sequence to sequence model is trained using the cross-entropy loss function.

A single speaker TTS system is used to generate the synthetic data. The system is based on Tacotron2 architecture (Shen et al., 2018) and a Clarinet (Ping et al., 2018) based vocoder. The Tacotron2 sub-system converts a sequence of phonemes to a mel-spectrogram. The generated mel-spectrogram is converted into a time-domain using a Clarinet-style vocoder. In-house single speaker studio recordings are used to train this model. Both Hindi and English queries were recorded using the voice of the same artist and the text was represented in Devnagari script. The TTS system could therefore be used to convert both English and Hindi text to audio.

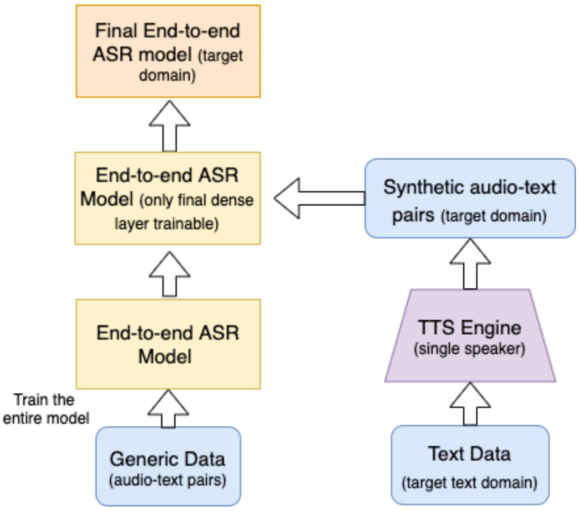
Final End-to-end ASR model (target domain)
End-to-end ASR Model (only final dense layer trainable)
Synthetic audio-text pairs (target domain)
TTS Engine (single speaker)
End-to-end ASR Model
Train the entire model
Generic Data (audio-text pairs)
Text Data (target text domain)

Figure 1: Domain Adaptation Process

3.2 Dataset Details

The train data consists of a multi-domain generic audio corpus and two domain-specific synthetic data sets. The generic data consists of crowd-sourced read speech corpus. It consists of around 4 million samples amounting to 6500 hours of data. The domain-specific data were synthetically created using a single speaker TTS engine. The two domains under consideration are the voice search domain and address domain. The search domain corresponds to the Flipkart e-commerce product search domain. The address domain corresponds to the pan India delivery address. Both the domains contain around 3 million samples which are approximately 4000 hours of data for VS domain and 5000 for the address domain. The address domain queries are longer as compared to voice search queries. All the datasets consist of queries in both English and Hindi. All the text is represented in the Devanagari script. The test data was real-world domain-specific data recorded on the Flipkart application. The test data is multi-speaker

Model	Test WER	Test WER + LM Rescoring	N-Best WER
LAS-Gen	25.31	22.18	13.71
LAS-Dense	16.25	15.55	7.6
LAS-Decoder	13.65	**13.36**	5.82
CTC-Gen	31.84	25.58	13.83
CTC-Dense	20.32	17.66	8.24

Table 1: Word Error Rate(WER) for different model variations using Voice Search Domain. The N-Best WER indicates the best WER in the top N=10 beams.

Model	Test WER	Test WER + LM Rescoring	N-Best WER
LAS-Gen	39.42	31.62	25.35
LAS-Dense	22.57	16.38	11.01
LAS-Decoder	18.96	**12.54**	8.17
CTC-Gen	31.08	22.81	19.74
CTC-Dense	22.43	15.42	12.15

Table 2: Word Error Rate(WER) for different model variations using Address Domain. The N-Best WER indicates the best WER in the top N=10 beams.

data recorded in a noisy environment and it is very different than the single speaker TTS data recorded in noise-free studio settings. The test audio data was manually transcribed by the operations team. The voice search test data consisted of 25000 examples and address test data had 7000 examples. Except for linguistic overlap, the synthetic train and real test datasets represent completely different environments, and hence improvements reported in this work are not dependent on the quality of the TTS system as long as it is a single speaker.

4 Results

In this work, we evaluate dense only fine-tuning baseline for CTC and attention-based models. The domain adaptation approach is presented on two datasets from voice search and address domain. The word error rates(WER) is used to compare the different approaches. The WER is word-level Levenshtein distance between ground truth text and output text. The results for voice search domain and address domain are shown in Table 1 and Table 2 respectively. The models are first trained on the generic multi-domain dataset and represented as CTC-Gen and LAS-Gen. These pre-trained models are then fine-tuned single speaker synthetic dataset. We show that dense only fine-tuning provides considerable improvement in accuracy while at the same time avoiding over-fitting on single speaker data. The dense-finetuned models are referred to as CTC-Dense and LAS-Dense. We also evaluate decoder-only fine-tuning for LAS models termed as LAS-Decoder. We report WER with and without external language model rescoring. A kenLM based language model is trained using text transcripts for both the domains individually. The N-Best WER is computed by picking the best beam from the top N=10 beam elements.

The results show that LAS-Dense provides around 30% relative improvement in WER over LAS-Gen for VS domain and around 50% relative improvement for the address domain. The LAS-Decoder further improves the results by 14% for VS domain and 23% for the address domain. Similarly, CTC-Dense provides an improvement of 30% and 32% for VS and address domain respectively over CTC-Gen. Note that the WER of LAS-Gen evaluated on address domain is considerably high as compared to VS domain. Moreover, this simple fine-tuning and LM-rescoring provides high improvements in WER. This shows that the text distribution of address data is very different from the initial multi-domain data. Also, the variety of named entities is very high in address data as compared to VS data. Overall we show that dense only fine-tuning can provide us a reasonable baseline for domain adaptation. For encoder-decoder architectures, decoder fine-tuning serves as a better option. This is expected as the encoder part can also be seen as the acoustic network is frozen and the decoder network which can be seen as a contextual language model is fine-tuned. For CTC-based networks, we observe that extending fine-tuning to even a single lower LSTM layer results in over-fitting and degradation in performance. Therefore for CTC networks dense only fine-tuning is the optimal approach to avoid over-fitting.

5 Conclusion

In conclusion, we demonstrate a simple baseline approach for domain adaptation using a text-only corpus from the target domain. We show that the final dense layer only fine-tuning using single speaker TTS data provides considerable improvements over the generic model. The results are shown on two different domains of voice search and address domain. For both CTC and attention-based models we show that dense-only fine-tuning is a reasonable approach for domain adaptation. Although the technique is more relevant to CTC-based models it can also be used with encoder-decoder type models. For encoder-decoder models, the decoder only fine-tuning performs better.

References

Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski. 2020. Adaptation algorithms for neural network-based speech recognition: An overview. *IEEE Open Journal of Signal Processing*, 2:33–66.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. Synthasr: Unlocking synthetic data for speech recognition. *arXiv preprint arXiv:2106.07803*.

Changfeng Gao, Gaofeng Cheng, Runyan Yang, Han Zhu, Pengyuan Zhang, and Yonghong Yan. 2021. Pre-training transformer decoder for end-to-end asr model with unpaired text data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6543–6547. IEEE.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al. 2019. Streaming end-to-end speech recognition for mobile devices. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6381–6385. IEEE.

Jocelyn Huang, Oleksii Kuchaiev, Patrick O'Neill, Vitaly Lavrukhin, Jason Li, Adriana Flores, Georg Kucsko, and Boris Ginsburg. 2020. Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. *arXiv preprint arXiv:2005.04290*.

Raviraj Joshi and Venkateshan Kannan. 2021. Attention based end to end speech recognition for voice search in hindi and english. In *Forum for Information Retrieval Evaluation*, pages 107–113.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5828. IEEE.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. 2020. You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE.

Jinyu Li, Rui Zhao, Zhong Meng, Yanqing Liu, Wenning Wei, Sarangarajan Parthasarathy, Vadim Mazalov, Zhenghao Wang, Lei He, Sheng Zhao, et al. 2020. Developing rnn-t models surpassing high-performance hybrid models with customization capability. *arXiv preprint arXiv:2007.15188*.

Linda Liu, Yile Gu, Aditya Gourav, Ankur Gandhe, Shashank Kalmane, Denis Filimonov, Ariya Rastrow, and Ivan Bulyko. 2021. Domain-aware neural language models for speech recognition. In *ICASSP*

2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7373–7377. IEEE.

Zhong Meng, Sarangarajan Parthasarathy, Eric Sun, Yashesh Gaur, Naoyuki Kanda, Liang Lu, Xie Chen, Rui Zhao, Jinyu Li, and Yifan Gong. 2021. Internal language model estimation for domain-adaptive end-to-end speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 243–250. IEEE.

Masato Mimura, Sei Ueno, Hirofumi Inaguma, Shinsuke Sakai, and Tatsuya Kawahara. 2018. Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 477–484. IEEE.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Cal Peyser, Hao Zhang, Tara N Sainath, and Zelin Wu. 2019. Improving performance of end-to-end asr on numeric sequences. *arXiv preprint arXiv:1907.01372*.

Wei Ping, Kainan Peng, and Jitong Chen. 2018. Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv preprint arXiv:1807.07281*.

Janne Pylkkönen, Antti Ukkonen, Juho Kilpikoski, Samu Tamminen, and Hannes Heikinheimo. 2021. Fast text-only domain adaptation of rnn-transducer prediction network. *arXiv preprint arXiv:2104.11127*.

Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2020. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073. IEEE.

Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. 2019. Component fusion: Learning replaceable language model component for end-to-end speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5361–5635. IEEE.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Khe Chai Sim, Françoise Beaufays, Arnaud Benard, Dhruv Guliani, Andreas Kabel, Nikhil Khare, Tamar Lucassen, Petr Zadrazil, Harry Zhang, Leif Johnson, et al. 2019. Personalization of end-to-end speech recognition on mobile devices for named entities. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 23–30. IEEE.

Sei Ueno, Takafumi Moriya, Masato Mimura, Shinsuke Sakai, Yusuke Shinohara, Yoshikazu Yamaguchi, Yushi Aono, and Tatsuya Kawahara. 2018. Encoder transfer for attention-based acoustic-to-word speech recognition. In *INTERSPEECH*, pages 2424–2428.

Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. 2021. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678. IEEE.