UNSUPERVISED MODEL ADAPTATION FOR END-TO-END ASR

Ganesh Sivaraman, Ricardo Casal, Matt Garland, Elie Khoury

Pindrop, Atlanta, GA, USA

ABSTRACT

End-to-end (E2E) Automatic Speech Recognition (ASR) systems are widely applied in various devices and communication domains. However, state-of-the-art ASR systems are known to underperform when there is a mismatch in the training and test domains. As a result, acoustic models deployed in production are often adapted to the target domain to improve accuracy. This paper proposes a method to perform unsupervised model adaptation for E2E ASR using first-pass transcriptions of adaptation data produced by the baseline ASR model itself. The paper proposes two transcription confidence measures that can be used to select an optimal in-domain adaptation set. Experiments were performed using the Quartznet ASR architecture on the HarperValleyBank corpus. Results show that the unsupervised adaptation technique with the confidence measure based data selection results in a 8% absolute reduction in word error rate on the HarperValleyBank test set. The proposed method can be applied to any E2E ASR system and is suitable for model adaptation on call center audio with little to no manual transcription.

***Index Terms*—** End-to-end, speech recognition, unsupervised adaptation, confidence measure, call centers, telephony audio

1. INTRODUCTION

End-to-end (E2E) ASR systems which directly map acoustic feature vectors to character symbols achieve better performance compared to the older generation of HMM-DNN ASR systems. The E2E ASR also does not require explicit alignment between speech and text for the training data. State-of-the-art E2E ASR systems are trained on several thousand hours of data spanning multiple domains [1, 2]. However even the state-of-the-art E2E ASR systems perform poorly in unseen domains like call center conversations which contain a high amount of speech variation. ASR model retraining with target domain data is an effective approach to improve the model performance in unseen domains [3, 4, 5, 6]. Retraining based model adaptation of E2E ASR systems have been explored recently with good success for multilingual ASR [7] and speaker adaptation for multichannel ASR [8]. Many studies have explored auxilliary feature based methods for unsupervised adaptation of E2E models. [9]. The auxilliary feaure based adaptation methods make use of an additional input feature vector like the i-vector, accent embedding or a domain classifier embedding. [10, 11, 12, 13]. A recent study proposed an unsupervised model adaptation technique with the transcripts obtained from the baseline model trained with supervision on a different source domain [14]. In that study, dropout layers have been used at inference time to compute a transcript uncertainty measure to choose good target domain data for unsupervised adaptation. There are several studies which propose various confidence measures for estimating the quality of ASR transcripts [15, 16, 17, 18, 14, 19].

In this paper we perform unsupervised adaptation of the pretrained Quartznet ASR model [2] using transcripts obtained from the baseline model itself. Since the baseline model performance in the target domain is known to be poor, we can expect many of its transcripts to be incorrect. Hence there is a need to assess the quality of the baseline model transcripts in order to select only target domain utterances that will produce a useful model adaptation. In section 3 of this paper we propose the use of the CTC loss and a probability ratio based confidence score (PRC score) to help assess the transcript quality for adaptation data selection. We use the transcript quality metrics for data selection and perform unsupervised and semisupervised adaptation of the E2E Quartznet ASR model on the HarperValleyBank dataset [20]. Details about the experiments and results are presented in sections 5 and 6.

2. ASR SYSTEM

All experiments in this paper have been performed using the Quartznet [2] ASR model architecture. The Quartznet architecture is a deep 1D convolutional network with time-channel separable convolutional layers. The Quartznet model is a lightweight ASR architecture with fast inference due to the separable convolutional layers. We used the pretrained Quartznet15x5Base-En model as the baseline system for all experiments in this paper. The Quartznet15x5 model consists of 15 convolutional blocks, each consisting of 5 1D time-channel separable convolutions, batch normalization, and ReLU layers. We will refer to the baseline model as the *Qnet-base* model henceforth. The *Qnet-base* model is trained on 6 different datasets consisting of both wideband and telephony speech, achieving a WER of 3.79% on Librispeech dev-clean set. Further details about the training process of the baseline model can be obtained from the Quartznet paper [2] This is a strong baseline model which is lightweight and achieves good transcription performance across a wide variety of domains including conversational speech.

3. UNSUPERVISED ADAPTATION METHODS

To improve the transcription quality of the baseline system on unseen target domains, this paper explores unsupervised model adaptation on the target domain data. In this paper, we propose the use of two metrics which can act as reliable confidence measure for the baseline model transcription - the CTC loss, and a probability ratio based confidence score.

3.1. Connectionist Temporal Classification (CTC) loss

A CTC-based E2E ASR system maps an input sequence of acoustic features denoted by $X = [X_1, X_2, ..., X_T]$ to the sequence of characters denoted by $Y = [Y_1, Y_2, ...Y_N]$. The system is trained with the CTC algorithm which assigns a probability for any output sequence Y given an input sequence X. Model training with the CTC algorithm does not require alignment between X and Y, unlike Hidden Markov Model (HMM) based ASR systems. Instead, the CTC algorithm computes the probability $p(Y|X)$ by summing over all possible alignments between X and Y. CTC-based ASR systems are trained by iteratively minimizing the CTC loss – the negative log-likelihood computed over the training dataset D as described in the following equation.

$$CTC_{loss} = \sum\nolimits_{(X,Y)\in D} - \log p(Y|X) \qquad (1)$$

The CTC loss can also be computed during inference. Assuming the decoded sequence of characters is an approximate transcript $\hat{Y}$, the CTC_{loss} $p(\hat{Y}|X)$ is an approximate measure of the baseline model's confidence in the decoded output $\hat{Y}$. The CTC loss function is already normalized for variations in duration of the input audio and length of the output character sequence. Hence the CTC loss can be used to compare the quality of transcripts across different utterances in the target domain.

We computed the CTC loss of the transcripts obtained from the Qnet-base model over the HarperValleyBank corpus. Since the HarperValleyBank dataset also has manually-produced transcripts, we could compute the WER of the Qnet-base transcripts. Figure 1 shows a scatter plot between the CTC loss and the WER on the HarperValleyBank dataset. The red line in the plot shows the median value of the CTC loss for each value of WER. The Pearson correlation coefficient between CTC loss and WER is 0.59. Hence the CTC loss with a positive correlation to WER can be used as a metric for selecting utterances with good transcript accuracy.

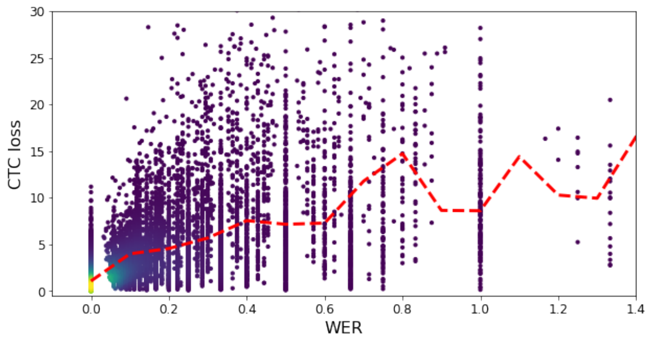

Fig. 1. Scatter plot of CTC loss and WER for the utterances from the HarperValleyBank dataset.

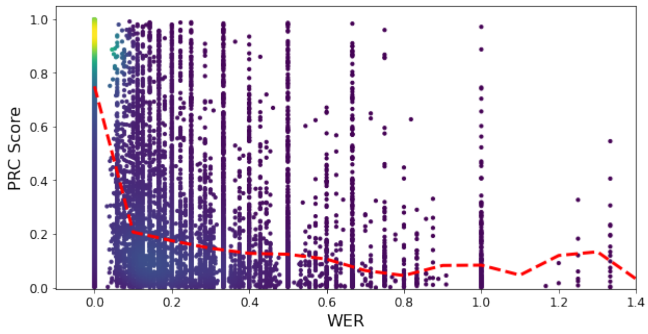

Fig. 2. Scatter plot of PRC score and WER for the utterances from the HarperValleyBank dataset.

3.2. Probability ratio based confidence score

The Quartznet ASR system predicts the posterior probability of 28 characters for every frame of audio. These frame-wise character probabilities can be decoded into a most likely sequence of characters using the beam search algorithm. The beam search algorithm, with a beam size of n, gives the n best possible character sequence candidates at each time step for a given audio file X. Let the character sequence of the i-th beam in the beam search be denoted by $\hat{Y_i}$. Each candidate sequence $\hat{Y_i}$ in the beam search has an associated conditional probability $p(\hat{Y_i}|X)$, which we denote as p_i. The beam search algorithm sorts candidate sequences in decreasing order of their conditional probability, thus $p_1 > p_2 > ... > p_n$. We define our confidence as the probability ratio confidence (PRC) score:

$$PRC = 1 - (p_2/p_1) \qquad (2)$$

The PCR score is in the range of [0, 1]. When $p_1 >> p_2$ the PRC score is close to 1, implying that the ASR system is very confident of the predicted character sequence. It is worth noting that this probability ratio based confidence measure has been defined and used before for optical character recognition [21]. This confidence score is a good candidate for assessing the transcript quality of the baseline ASR system.

We computed the confidence score of the transcripts obtained from the Qnet-base model over the HarperValleyBank corpus. Figure 2 shows a scatter plot between the PRC score and the WER on the HarperValleyBank dataset. The red line in the plot shows the median value of the PRC score for each value of WER. The Pearson correlation coefficient between the PRC score and the WER is -0.54. Hence the PRC score with a negative correlation with WER can be used as a metric for selecting utterances with good transcript accuracy.

4. DATASETS

For our experiments, we chose the HarperValleyBank corpus [20] which is a telephony conversational audio dataset and also no included in the baseline model training. The HarperValleyBank corpus is a spoken dialog corpus containing 23 hours of simple consumer banking conversations. The dataset contains 1,446 conversations between 59 unique speakers. The speakers follow a pre-determined script for the conversation based on a pre-assigned role of either an agent or a customer. The conversations range from 2 to 60 utterances with an average of 18 utterances per conversation. The dataset contains 735 unique words. Even though the size of the vocabulary is small, the dataset represents typical domain specific call center conversations which often use a constrained vocabulary relevant to the domain. We split the HarperValleyBank dataset into three subsets for training, development and testing. We selected 6 speakers from the dataset and designated the utterances from those speakers as the test set. In the dataset, each of the 6 speakers have played the role of the agent and the customer in different calls. Thus the test set consisted of both agent side and caller side utterances belonging to the 6 test speakers. The data from the remaining 53 speakers was used for training and development. The development set consisted of utterances from a set of 100 randomly-selected calls. The training set contained the remaining utterances, which amounted to 9.35 hours of net speech (after removing non-speech segments).

5. EXPERIMENTS AND RESULTS

In this paper we performed adaptation experiments on the Qnet-base model. We explored unsupervised and semi-supervised adaptations and compared the results with fully supervised adaptation. The adaptation of the model was performed with subsets of the train set selected using either the CTC loss or the PRC score. The baseline model was first used to decode the test set. Beam search based model decoding was performed with a beam width of 200 for all our experiments. All experiments were performed without any language model in the decoding. The baseline model without any adaptation obtained a WER of 14.33% on the development set and 20.41% on the test set.

We next adapted the baseline model in a supervised way with the manual transcripts of the complete train set. We trained the model with the novograd optimizer, and a learning rate of 1E-4 for 30 epochs and selected the best model based on the validation loss. The fully supervised model denoted by *full-sup adapt* in Table 1 and Table 2 resulted in a WER of 4.94% on the development set and 8.18% on the test set. We performed unsupervised adaptation of the baseline model using the baseline transcriptions of the complete training data. The fully unsupervised adapted model denoted by *full-unsup-adapt* resulted in a WER of 10.18% and 15.53% on the dev and test sets respectively. Next, we performed unsupervised adaptation of the baseline model by doing data selection of the train set using the CTC loss or the PRC score criteria. The objective of data selection is to exclude from the adaptation set, the utterances with poor baseline transcriptions.

5.1. CTC loss based data selection

The baseline model was used to transcribe the train set and the CTC loss was computed on all the segments of the train set conversations. The CTC loss value ranged from 0 to 100, with lower CTC loss indicating better transcript accuracy, as per Figure 1. We set an upper limit on the CTC loss and selected speech segments with CTC loss less than the limit. The selected speech segments, along with the baseline model transcripts as target labels for these segments, formed the unsupervised adaptation set. Learning rate and optimization parameters were similar to those of supervised adaptation. Validation loss was used to select the best adapted model from 30 epochs of adaptation. We perfomed the unsupervised adaptation on different adaptation subsets obtained by setting 11 different thresholds on the CTC loss function ranging from 0.2 to 50. Each experiment resulted in a new adapted Quartznet model resulting in a total of 11 adapted models. We computed the transcript WER on the development set for each of the adapted models and selected the model with the lowest WER on the development set. Table 1 shows the results of this unsupervised adaptation (unsup-adapt) for CTC loss based selection. The best unsupervised model used a CTC threshold of 0.8 and selected 3.96 hours of speech for adaptation.

CTC loss based data selection selects utterances with good baseline transcript accuracy for adaptation. Could manual transcription efforts on poorly transcribed utterances with high CTC loss help improve the model performance? To answer this question, we selected a limited set of manual transcripts for adaptation by setting a threshold of 8 on the CTC loss and selecting utterances which had a higher loss value for limited supervision. We carefully balanced the data selection to obtain equal amounts of agent side and caller side segments limiting the selection to 1 hour of total net speech. The baseline model was adapted on this limited supervision data to obtain the lim-sup model. This model resulted in a WER of 7.63% on the development set and 12.83% on the

Table 1. Results of the model adaptation experiments on HarperValleyBank dataset performed with a CTC loss based data selection criterion. The numbers shown are WERs (%).

Adaptation type	Dev	Test	Hours of adaptation data
No adaptation	14.33%	20.41%	0
unsup adapt	6.54%	13.36%	3.96
lim-sup adapt	7.63%	12.83%	1.0
semi-sup adapt	**5.30%**	**8.89%**	3.96 (unsup) +1.0 (sup)
full-unsup adapt	10.18%	15.53%	9.35
full-sup adapt	4.94%	8.18%	9.35

Table 2. Results of the model adaptation experiments on HarperValleyBank dataset with PRC score based data selection criterion. The numbers shown are WERs (%).

Adaptation type	Dev	Test	Hours of adaptation data
No adaptation	14.33%	20.41%	0
unsup adapt	6.80%	12.13%	5.82
lim-sup adapt	6.95%	11.50%	1.0
semi-sup adapt	**4.74%**	**9.06%**	5.82 (unsup) +1.0 (sup)
full-unsup adapt	10.18%	15.53%	9.35
full-sup adapt	4.94%	8.18%	9.35

test set. The performance of the limited supervision adaptation using manual transcripts was not significantly better than the unsupervised adaptation without manual transcripts.

We also explored semi-supervised adaptation. We continued adapting the *unsup adapt* model with the 1 hour of supervised data used in the *lim-sup adapt* experiment. This resulted in the *semi-sup adapt* model with a WER of 5.30% on the development set and 8.89% on the test set. This was the best result we obtained in this paper and used only 1 hour of manually transcribed data for adaptation. Table 1 shows the results of the adaptation experiments performed using the CTC loss based data selection.

5.2. PRC score based data selection

Finally, we explored adaptation using the PRC score based data selection. The PRC score defined in section 3.2 is inversely correlated with the WER. We performed adaptation experiments similar to those for CTC based selection, setting thresholds on the PRC score to select the adaptation dataset. For unsupervised adaptation we selected utterances with high PRC score and for limited supervision we selected 1 hour of speech segments with low PRC scores. The rest of the experiment details are the same as described in section 5.1. The adaptation results of the PRC score based data selection are shown in Table 2.

6. DISCUSSION

The high WER of the baseline model on the HarperValleyBank test set indicates the degradation in the model performance under mismatched train and test conditions. The unsupervised adaptation results in around 7% absolute reduction in the WER when the adaptation is performed on a subset selected by CTC loss criterion whereas PRC score based selection is sightly better for unsupervised adaptation with 8.28% WER reduction over the baseline. Performing supervised adaptation with just 1 hour of manually transcribed audio results in a WER reduction of 7.5-8.9% over the baseline. Again in this case using PRC score selection performed slightly better than the CTC loss criterion. Semi-supervised adaptation provides the best results among all our adaptation experiments. We obtain results close to fully supervised adaptation with only 1 hour of manually transcribed speech (compared to over 9 hours of manually transcribed speech used in the fully supervised adaptation). For the semi-supervised adaptation experiment, we observe that selection using the PRC score criterion produces a lower WER on the dev set, while selection using the CTC loss produces a lower WER on the test set. We note that the utterances selected for the 1 hour of supervision data differ depending on whether the CTC loss criterion or the PRC score criterion is used, and we believe the observed difference may be due to some overfitting of the semi-supervised model with the PRC score based data selection. Future experiments on larger datasets will help us better understand the difference between the CTC loss and the PRC score based data selection for semi-supervised adaptation.

7. CONCLUSIONS

In this paper we proposed an unsupervised adaptation technique which significantly improved the performance of the ASR system on the unseen target domain. We also proposed the utilization of the CTC loss and a probability ratio based confidence measure for selecting target domain data for adaptation. Our experimental results show that both the CTC loss and the probability ratio based confidence score are effective in selecting adaptation datasets. Our experiments and results show that this unsupervised adaptation can be applied to adapt ASR models to conversational speech with a limited vocabulary setting. The experiments show significant improvement in performance without the use of language models in the decoding. Decoding with a domain adapted language model would further improve the model performance. The proposed data selection method can also be used for selecting poorly transcribed utterances for manual supervision. In the future we plan to explore additional model adaptation techniques and other metrics for assessing transcript quality.

8. REFERENCES

[1] Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel, "Toward Cross-Domain Speech Recognition with End-to-End Models," *arXiv preprint arxiv:2003.04194*, 2020.

[2] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang, "Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions," in *ICASSP*, may 2020, vol. 2020-May, pp. 6124–6128.

[3] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*, oct 2013, pp. 7893–7897.

[4] Roberto Gemello, Franco Mana, Stefano Scanzio, Pietro Laface, and Renato De Mori, "Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training," in *ICASSP*, 2006, vol. 1.

[5] Hank Liao, "Speaker adaptation of context dependent deep neural networks," in *ICASSP*, oct 2013, pp. 7947–7951.

[6] Pawel Swietojanski and Steve Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," *IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, pp. 171–176, apr 2014.

[7] Sibo Tong, Philip N. Garner, and Hervé Bourlard, "Cross-lingual adaptation of a CTC-based multilingual acoustic model," *Speech Communication*, vol. 104, 2018.

[8] Tsubasa Ochiai, Shinji Watanabe, Shigeru Katagiri, Takaaki Hori, and John Hershey, "Speaker Adaptation for Multichannel End-to-End Speech Recognition," in *ICASSP*, 2018, pp. 6707–6711.

[9] M. A. Tugtekin Turan, Emmanuel Vincent, and Denis Jouvet, "Achieving multi-accent ASR via unsupervised acoustic model adaptation," in *Proceedings of INTERSPEECH*, 2020, vol. 2020-Octob, pp. 1286–1290.

[10] Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, Christian Huemmer, and Tomohiro Nakatani, "Context adaptive neural network based acoustic models for rapid adaptation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 5, 2018.

[11] Markus Müller, Sebastian Stüker, and Alex Waibel, "Language adaptive multilingual CTC speech recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10458 LNAI.

[12] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 2013.

[13] Marc Delcroix, Shinji Watanabe, Atsunori Ogawa, Shigeki Karita, and Tomohiro Nakatani, "Auxiliary feature based adaptation of end-to-end ASR systems," in *INTERSPEECH*, 2018, vol. 2018-Septe, pp. 2444–2448.

[14] Sameer Khurana, Niko Moritz, Takaaki Hori, and Jonathan Le Roux, "Unsupervised Domain Adaptation for Speech Recognition via Uncertainty Driven Self-Training," in *ICASSP*, 2021, pp. 6553–6557.

[15] Ahmed Ali and Steve Renals, "Word error rate estimation without asr output: E-WER2," in *INTERSPEECH*, 2020, vol. 2020-Octob, pp. 616–620.

[16] Kaustubh Kalgaonkar, Chaojun Liu, Yifan Gong, and Kaisheng Yao, "Estimating confidence scores on ASR results using recurrent neural networks," in *ICASSP*, aug 2015, vol. 2015-Augus, pp. 4999–5003.

[17] David Qiu, Qiujia Li, Yanzhang He, Yu Zhang, Bo Li, Liangliang Cao, Rohit Prabhavalkar, Deepti Bhatia, Wei Li, Ke Hu, Tara N. Sainath, and Ian McGraw, "Learning word-level confidence for subword end-to-end ASR," in *ICASSP*, 2021, vol. 2021-June, pp. 6393–6397.

[18] Apoorv Vyas, Pranay Dighe, Sibo Tong, and Herve Bourlard, "Analyzing Uncertainties in Speech Recognition Using Dropout," in *ICASSP*, 2019, vol. 2019-May.

[19] Ankur Kumar, Sachin Singh, Dhananjaya Gowda, Abhinav Garg, Shatrughan Singh, and Chanwoo Kim, "Utterance confidence measure for end-to-end speech recognition with applications to distributed speech recognition scenarios," in *INTERSPEECH*, 2020, vol. 2020-Octob, pp. 4357–4361.

[20] Mike Wu, Jonathan Nafziger, Anthony Scodary, and Andrew Maas, "Harpervalleybank: A domain-specific spoken dialog corpus," 2021.

[21] Noam Mor and Lior Wolf, "Confidence Prediction for Lexicon-Free OCR," in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, may 2018, vol. 2018-Janua, pp. 218–225.