Learning Fast Adaptation on Cross-Accented Speech Recognition

Genta Indra Winata†, Samuel Cahywajaya†, Zihan Liu†, Xiaojiang Lin,
Andrea Madotto, Peng Xu, Pascale Fung

Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology

Abstract

Local dialects influence people to pronounce words of the same language differently from each other. The great variability and complex characteristics of accents create a major challenge for training a robust and accent-agnostic automatic speech recognition (ASR) system. In this paper, we introduce a cross-accented English speech recognition task as a benchmark for measuring the ability of the model to adapt to unseen accents using the existing CommonVoice corpus. We also propose an accent-agnostic approach that extends the model-agnostic meta-learning (MAML) algorithm for fast adaptation to unseen accents. Our approach significantly outperforms joint training in both zero-shot, few-shot, and all-shot in the mixed-region and cross-region settings in terms of word error rate.

Index Terms: speech recognition, accent-agnostic, cross-accent, meta-learning, fast adaptation

1. Introduction

Spoken languages show great variation across regions and such distinctions derive from the phonetics of local dialects and language backgrounds. Despite the high performance reported by state-of-the-art English automatic speech recognition (ASR) systems, accented speech recognition is still an unsolved real-world challenge due to the great variability of accents and their complex characteristics [1]. It is difficult for ASR models to adapt to unseen accents that have relatively distinct pronunciations and tones compared to the accents used for training the ASR models. Increasing the number of training data and exposing the model to different accents is a common solution to improving the model’s robustness to different speakers’ accents by introducing variations. However, such approaches are costly and not scalable due to the difficulties in collecting high-quality speech data with different accents. Existing data augmentation techniques such as noise injection [2] and speed perturbation [3] have been proposed to overcome the limitation on high-resource data. In this work, we explore training approaches for fast adaptation to unseen accents instead of augmenting the training data. We apply model-agnostic meta-learning (MAML) [4] to teach the model to learn new tasks faster and more efficiently, and our approach can easily be applied to few-shot learning. While a small number of previous studies have explored joint and multi-task training on multiple accent speech recognition models [5, 6, 7], none have thoroughly investigated few-shot learning on the cross-accented speech recognition task.

We introduce a cross-accented speech recognition task derived from an existing dataset, CommonVoice [8], to move toward building a robust speech recognition system. The motivation of this work is to establish a benchmark for evaluating cross-accented speech recognition. We introduce an accent-agnostic model by applying meta-learning as learning to learn method for fast accent adaptation. The trained model is able to rapidly adapt to recognize speech with unseen accents. We train our transformer [9] speech recognition model on a set of accents via meta-learning and fine-tune the trained model with a few samples of target accented speech. Experimental results show that our approach is able to quickly adapt to new accents more effectively than joint training, and interestingly, our approach is also able to handle zero-shot predictions.

2. Related Work

2.1. Meta-Learning

Meta-learning is a sub-field of machine learning that designs models for learning new tasks in a new setting with a few training examples [11, 12]. In recent work, [4] propose model-agnostic meta-learning (MAML) and show the application of meta-learning in a deep learning framework. Several meta-learning-based models have since been proposed for solving few-shot image classification [13, 14, 15] and natural language processing applications, such as text classification [16], dialogue response generation [17, 18], low-resource machine translation [10], semantic parsing [19], and sales prediction [20]. [10] makes the interesting finding that MAML is actually able to generalize the model in the low-resource machine translation task without any fine-tuning steps or when there is no information on the target accent. In speech applications, [21] introduce the practicality of applying MAML in cross-lingual speech recognition, while in another line of works, MAML has been applied to learn how to adapt respectively to
optimizing the next-step prediction on the previous characters, we train our model by applying a mask in the attention layer to avoid any probability of the outputs, we compute the softmax function of these outputs to finally calculate the logits of the outputs. To generate the encoder outputs and apply multi-head attention to its inputs, the decoder receives the encoder outputs with a learnable feature extractor module to generate input embeddings. The encoder uses input embeddings generated from the feature extractor module. Then the decoder receives the encoder outputs and applies multi-head attention to its input to finally calculate the logits of the outputs. To generate the probability of the outputs, we compute the softmax function of the logits. We apply a mask in the attention layer to avoid any information flow from future tokens, and we train our model by optimizing the next-step prediction on the previous characters and by maximizing the log probability:

$$
\max_{\theta} \sum_{i} \log P(y_i | x, y_{<i}; \theta),
$$

where $x$ is the character inputs, $y_i$ is the next predicted character, and $y_{<i}$ is the ground truth of the previous characters. In the inference time, we generate the sequence using a beam-search in an auto-regressive manner. Then we maximize the following scoring function:

$$
\eta \sum_{i} \log P(y_i | x, y_{<i}; \theta) + \gamma \sqrt{\text{wc}(y_{<i})},
$$

where $\eta$ is the parameter to control the decoding probability from the decoder, and $\gamma$ is the parameter to control the effect of the word count $\text{wc}(y_{<i})$, as suggested in [28] and [29].

### 3.2. Fast Adaptation via Meta-Learning

MAML [4] learns to quickly adapt to a new task from a number of different tasks using a gradient descent procedure, as shown in Fig. 1. In this paper, we apply MAML to effectively learn from a set of accents and quickly adapt to a new accent in the few-shot setting. We denote our Transformer ASR as $f_\theta$ parameterized by $\theta$. Our dataset consists of a set of accents $A = \{A_1, A_2, \ldots, A_n\}$, and for each accent $i$, we split the data into $A_i^{\text{val}}$ and $A_i^{\text{tra}}$, then update $\theta$ to $\theta'$ by computing gradient descent updates on $A_i^{\text{tra}}$:

$$
\theta' = \theta - \alpha \nabla_{\theta} L_{A_i^{\text{val}}} (f_\theta),
$$

where $\alpha$ is the fast adaptation learning rate. During the training, the model parameters are trained to optimize the performance of the adapted model $f(\theta')$ on unseen $A_i^{\text{val}}$. The meta-objective is defined as follows:

$$
\min_{\theta} \sum_{A_i \sim p(A)} L_{A_i^{\text{val}}} (f_\theta) = \sum_{A_i \sim p(A)} L_{A_i^{\text{val}}} (f_\theta - \alpha \nabla_{\theta} L_{A_i^{\text{tra}}} (f_\theta)),
$$

where $L_{A_i^{\text{val}}} (f_\theta)$ is the loss evaluated on $A_i^{\text{val}}$. We collect the loss $L_{A_i^{\text{val}}} (f_\theta)$ from a batch of accents and perform the meta-optimization as follows:

$$
\theta \leftarrow \theta - \beta \sum_{A_i \sim p(A)} \nabla_{\theta} L_{A_i^{\text{val}}} (f_\theta),
$$

### 3. Cross-Accented Speech Recognition

In this section, we present the architecture of our transformer-based speech recognition model and the proposed meta-learning method for fast adaptation on the cross-accented speech recognition task.

#### 3.1. Transformer Speech Recognition Model

As shown in Fig. 2, we build our model using a sequence-to-sequence transformer ASR [9, 27, 28, 29] to learn to predict graphemes from the speech input. Our model extracts audio inputs with a learnable feature extractor module to generate input embeddings. The encoder uses input embeddings generated from the feature extractor module. Then the decoder receives the encoder outputs and applies multi-head attention to its input to finally calculate the logits of the outputs. To generate the probability of the outputs, we compute the softmax function of the logits. We apply a mask in the attention layer to avoid any information flow from future tokens, and we train our model by optimizing the next-step prediction on the previous characters and by maximizing the log probability:

$$
\max_{\theta} \sum_{i} \log P(y_i | x, y_{<i}; \theta),
$$

where $x$ is the character inputs, $y_i$ is the next predicted character, and $y_{<i}$ is the ground truth of the previous characters. In the inference time, we generate the sequence using a beam-search in an auto-regressive manner. Then we maximize the following scoring function:

$$
\eta \sum_{i} \log P(y_i | x, y_{<i}; \theta) + \gamma \sqrt{\text{wc}(y_{<i})},
$$

where $\eta$ is the parameter to control the decoding probability from the decoder, and $\gamma$ is the parameter to control the effect of the word count $\text{wc}(y_{<i})$, as suggested in [28] and [29].

### Table 1: Statistics of accented speech data in CommonVoice dataset sorted alphabetically.

<table>
<thead>
<tr>
<th>accents</th>
<th># sample</th>
<th>duration (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa (af)</td>
<td>4,065</td>
<td>5.04</td>
</tr>
<tr>
<td>Australia (au)</td>
<td>19,625</td>
<td>22.86</td>
</tr>
<tr>
<td>Bermuda (be)</td>
<td>363</td>
<td>0.46</td>
</tr>
<tr>
<td>Canada (ca)</td>
<td>17,422</td>
<td>20.20</td>
</tr>
<tr>
<td>England (en)</td>
<td>58,274</td>
<td>64.19</td>
</tr>
<tr>
<td>Hong Kong (hk)</td>
<td>1,181</td>
<td>1.21</td>
</tr>
<tr>
<td>India (in)</td>
<td>23,878</td>
<td>29.09</td>
</tr>
<tr>
<td>Ireland (ir)</td>
<td>3,420</td>
<td>3.71</td>
</tr>
<tr>
<td>Malaysia (my)</td>
<td>843</td>
<td>1.07</td>
</tr>
<tr>
<td>New Zealand (nz)</td>
<td>6,070</td>
<td>7.06</td>
</tr>
<tr>
<td>Philippines (ph)</td>
<td>1,318</td>
<td>1.68</td>
</tr>
<tr>
<td>Scotland (sc)</td>
<td>4,376</td>
<td>5.08</td>
</tr>
<tr>
<td>Singapore (sg)</td>
<td>693</td>
<td>1.00</td>
</tr>
<tr>
<td>South Atlantic (sa)</td>
<td>212</td>
<td>0.23</td>
</tr>
<tr>
<td>United States (us)</td>
<td>145,692</td>
<td>163.89</td>
</tr>
<tr>
<td>Wales (wa)</td>
<td>1,128</td>
<td>1.16</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>288,560</strong></td>
<td><strong>327.93</strong></td>
</tr>
</tbody>
</table>
reformulated as: an approximation MAML as in [10] and [30]. Thus, Equation 5 is achieved by a good initialization for our model. Then we can optimize our model with a small number of samples on target accents in the fine-tuning step. In this work, we use first-order approximation MAML as in [10] and [30]. Thus, Equation 5 is reformulated as:

\[ \theta \leftarrow \theta - \beta \sum_{A_i \sim p(A)} \nabla_{\theta_i} L_{A_i+1}^{\text{adapt}}(f_{\theta_i}). \]  

(6)

4. Experiments

4.1. Dataset

We use the CommonVoice dataset [8],1 a multilingual open-accented dataset collected by Mozilla. In this work, we only use the English dataset and filter out speech data with an accent label. There are 16 accents listed in the dataset, and we split the dataset into groups according to the accent label. The statistics of the English dataset are shown in Table 1. Note that the dataset is imbalanced, and some accents only have very limited data. The pre-trained models are trained on the LibriSpeech corpus [31], a 960-hour training corpus of English read speech derived from audiobooks in the LibriVox project, sampled at 16 kHz. The accents are various and unlabeled, but the majority are US English.2

4.2. Experimental Setup

We preprocess raw audio input into a spectrogram before we fetch it into our model, which utilizes a VGG model [32], a 6-layer CNN architecture, as the feature extractor. Our transformer model consists of two transformer encoder layers and four transformer decoder layers. The transformer consists of a \( \text{dim}_\text{encoder} \) of 2048, \( \text{dim}_\text{model} \) of 512, and \( \text{dim}_\text{emb} \) of 512. We use 8 heads for multi-head attention. In total, our model has around 10.2M parameters. For both the MAML and joint training models, we end the training process after 200k iterations. In the pre-training setting, we pre-train the model using the LibriSpeech dataset for 1M iterations and resume the training using the CommonVoice dataset subsequently for other 100k iterations for all approaches. During the fine-tuning step, we run ten iterations for each sample. We evaluate our model using a beam search with \( \eta = 1, \gamma = 0.1 \), and a beam size of 5. In the pre-training setting, we downsample the CommonVoice speech data to 16 kHz following the LibriSpeech dataset audio sample rate.

1We use CommonVoice Version 2 data (June 2019).
2The LibriSpeech dataset can be downloaded at http://www.openslr.org/12/, and the list of LibriVox accents can be found at https://wiki.librivox.org/index.php/Accents_Table

Table 2: Average Word Error Rate (% WER) with Standard Error (SE) results in the mixed-region setting.

<table>
<thead>
<tr>
<th>accents</th>
<th>MAML</th>
<th>Joint Training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>zero-shot</td>
<td>5%-shot</td>
</tr>
<tr>
<td></td>
<td>without pre-training</td>
<td></td>
</tr>
<tr>
<td>Bermuda</td>
<td>33.22 ± 0.46</td>
<td>32.73 ± 0.47</td>
</tr>
<tr>
<td>Philippines</td>
<td>30.08 ± 0.56</td>
<td>48.22 ± 0.69</td>
</tr>
<tr>
<td>Wales</td>
<td>33.66 ± 0.83</td>
<td>33.31 ± 0.77</td>
</tr>
</tbody>
</table>

where \( \beta \) is the meta step size and \( f_{\theta_i} \) is the adapted network on accent \( A_i \). The meta-gradient update step is performed to achieve a good initialization for our model. Then we can optimize our model with a small number of samples on target accents in the fine-tuning step. In this work, we use first-order approximation MAML as in [10] and [30]. Thus, Equation 5 is reformulated as:

\[ \theta \leftarrow \theta - \beta \sum_{A_i \sim p(A)} \nabla_{\theta_i} L_{A_i+1}^{\text{adapt}}(f_{\theta_i}). \]

(6)

4. Experiments

4.1. Dataset

We use the CommonVoice dataset [8],1 a multilingual open-accented dataset collected by Mozilla. In this work, we only use the English dataset and filter out speech data with an accent label. There are 16 accents listed in the dataset, and we split the dataset into groups according to the accent label. The statistics of the English dataset are shown in Table 1. Note that the dataset is imbalanced, and some accents only have very limited data. The pre-trained models are trained on the LibriSpeech corpus [31], a 960-hour training corpus of English read speech derived from audiobooks in the LibriVox project, sampled at 16 kHz. The accents are various and unlabeled, but the majority are US English.2

4.2. Experimental Setup

We preprocess raw audio input into a spectrogram before we fetch it into our model, which utilizes a VGG model [32], a 6-layer CNN architecture, as the feature extractor. Our transformer model consists of two transformer encoder layers and four transformer decoder layers. The transformer consists of a \( \text{dim}_\text{encoder} \) of 2048, \( \text{dim}_\text{model} \) of 512, and \( \text{dim}_\text{emb} \) of 512. We use 8 heads for multi-head attention. In total, our model has around 10.2M parameters. For both the MAML and joint training models, we end the training process after 200k iterations. In the pre-training setting, we pre-train the model using the LibriSpeech dataset for 1M iterations and resume the training using the CommonVoice dataset subsequently for other 100k iterations for all approaches. During the fine-tuning step, we run ten iterations for each sample. We evaluate our model using a beam search with \( \eta = 1, \gamma = 0.1 \), and a beam size of 5. In the pre-training setting, we downsample the CommonVoice speech data to 16 kHz following the LibriSpeech dataset audio sample rate.

1We use CommonVoice Version 2 data (June 2019).
2The LibriSpeech dataset can be downloaded at http://www.openslr.org/12/, and the list of LibriVox accents can be found at https://wiki.librivox.org/index.php/Accents_Table

Figure 3: Few-shot results on Philippines accent in the mixed-region setting.

We train and evaluate the effectiveness of our fast adaptation method in two settings: (1) mixed-region, and (2) cross-region. The former is to train on ten accents, such as \( \text{af, au, ca, en, bk, in, ir, my, nz, sa, sc, sg, and us} \), sampled from all regions, and we validate the model on the \( \text{ca, sc, and sa} \) accents and test it on the \( \text{be, ph, and wa} \) accents. The latter setting is to train on five accents, such as \( \text{au, en, ir, nz, and us} \), from specific regions, and we validate the model on the \( \text{ca, sc, and sa} \) accents, and test it on the \( \text{af, bk, in, ph, and sg} \) accents, which come from other regions. We evaluate the model performance using the word error rate (WER) and run experiments ten times using different test folds. Each fold consists of 100 data randomly sampled from the test data. In the few-shot scenarios, we split the test accents data into training and testing sets. 75% of the data are allocated for training, and the remainder for testing.

3We report the average and standard error of all folds in the zero-shot (0%-shot), 5%-shot, 25%-shot, and all-shot (100%-shot) settings. In addition, we also investigate the usefulness of pre-training on a large English corpus and fine-tune the model.

5. Results and Discussion

5.1. Quantitative Analysis

As shown in Table 2, MAML consistently outperforms joint training in the mixed-region setting. The approach yields up to a 4% WER margin in the zero-shot and few-shot settings. In general, for both MAML and joint training, by adding more data on fine-tuning, the WER drops at a constant rate. Using the pre-trained model on the LibriSpeech dataset significantly boosts the performance of all models by around 5% to 8% WER. In the all-shot setting, the results are similar to those in the 5%-
shot and 25%-shot settings. We observe that the WER improvement after applying the pre-trained model for the Wales accent is higher than for the Bermuda and Philippines accents since the majority of the LibriSpeech dataset is US accented speech, which is far more acoustically similar to the accent of Wales than of Bermuda or the Philippines.

5.2. Cross-region Performance

We show the cross-region performance in Table 3. As expected, the WER of the Philippines accent is slightly reduced when we remove Asian accents from the training data. Interestingly, focusing only on the results of the Philippines accent, as shown in Table 2 and Table 3, MAML on the cross-region setting yields WER performance similar to the joint training on the mixed-region setting. Based on the empirical results, we can conclude that MAML is far more accent-agnostic compared to joint training. In sum, the model trained with MAML performs better than joint training and learns more accent-invariant representations.

5.3. Effectiveness of Few-Shot Fine-tuning

We investigate the number of samples needed to start showing performance improvement after fine-tuning. We start by training the model with a very small number of samples, from one to ten, where each sample approximately consists of 4 seconds of audio. We observe that the model cannot adapt to the target accent with a minuscule amount of data. We believe that our model is not able to capture the information from a very short audio sample due to a large acoustic variation in the data. Therefore, we increase the minimum threshold to 5% of the training data, and the model starts to adapt to the target accent accordingly.

In Figure 3 and Figure 4, in general, MAML performs better than joint training in all settings. By having more target accented speech data, the model gains higher performance with a lower WER for both the mixed-accent and cross-accent settings. We observe that MAML is effectively applied to models without pre-training on the LibriSpeech dataset, and it decays much faster than joint training.

We further investigate the effectiveness of the fast adaptability of the MAML approach compared to the all-shot setting. As shown in Tables 2 and 3, the MAML approach with 25%-shot fine-tuning performs similarly or even better compared to the joint approach with all-shot fine-tuning, both in the mixed-accent and cross-accent settings. In the all-shot setting, the MAML approach can further improve the performance and outperforms the joint training approach in all experiment settings. In light of the impressive experimental results, we can infer that MAML has fast adaptability to low-resource unseen accented data.

6. Conclusions

In this paper, we introduce a cross-accented speech recognition task derived from an existing dataset, CommonVoice, and establish a new benchmark for evaluating cross-accented speech recognition in the mixed-region and cross-region scenarios. We apply a fast adaptation method via the model-agnostic meta-learning (MAML) approach to learn a robust speech recognition system to rapidly adapt to unseen accents. Based on the empirical results, MAML consistently outperforms the joint training baseline in all settings around 4% WER improvement in both the mixed-region and cross-region scenarios. Impressively, MAML leverages less data (25%-shot) and achieves comparable results to joint training with all training data (all-shot). We also further improve the performance of our model by adding pre-training on a large speech corpus.

7. Acknowledgements

This work has been partially funded by ITF/319/16FP and MRP/055/18 of the Innovation Technology Commission, The Hong Kong SAR Government, and School of Engineering Ph.D. Fellowship Award, the Hong Kong University of Science and Technology, and RDC 1718050-0 of EMOS.AI.
8. References


