INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Learning Fast Adaptation on Cross-Accented Speech Recognition

Genta Indra Winata[⋆†]*, Samuel Cahyawijaya*[⋆]*, Zihan Liu*[⋆]*, Zhaojiang Lin, Andrea Madotto, Peng Xu, Pascale Fung*

Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
[†]giwinata@connect.ust.hk

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 improve a 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.

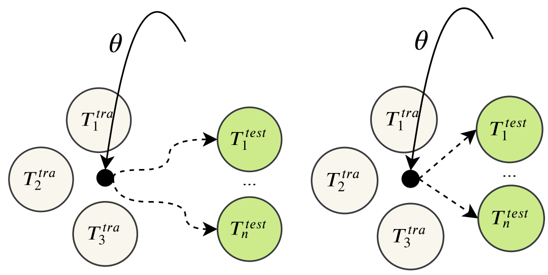

Figure 1: *Illustration modified from [10] of fine-tuning from joint training* ***(left)*** *and MAML* ***(right)****. The solid line represents the model optimization path of the initial parameters and dashed line represents the fine-tuning path. The off white circles are training accents and green circles are testing accents. MAML lets the model learn a representation that is more generalized (closer) to various tasks compared to joint training.*

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

[⋆]Equal contributions.

http://dx.doi.org/10.21437/Interspeech.2020-45

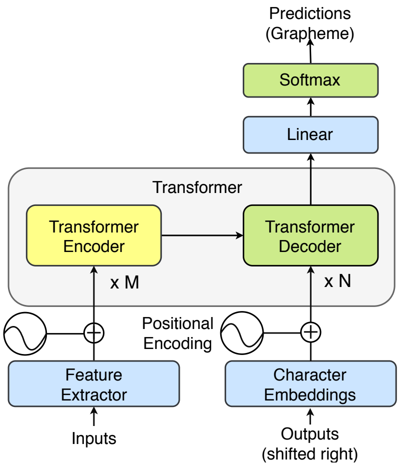

Figure 2: *Transformer ASR model architecture.*

the speaker [22, 23].

2.2. Accented Speech Recognition

Existing studies on accented speech recognition mainly focus on applying acoustic features that are accent-invariant and on adaptation methods to allow the model to accommodate accented speech. [24, 25] introduce acoustic features and an adaptation method for recognizing accented speech. Meanwhile, [6, 7] and [26] explore a multi-task architecture that jointly learns an accent classifier and an acoustic model. [7] propose a mixture of expert models to segregate accent-specific and phone-specific speech variability in a joint framework, and [5] propose an adversarial training objective to help the model to learn accent-invariant features. In this work, we explore the possibility of recognizing speech with unseen accents and extend MAML to enable fast adaptation by few-shot learning in the cross-accent setting.

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:

$$\max_{\theta} \sum_{i} \log P(y_i|x, y'_{<i}; \theta), \tag{1}$$

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:

$$\eta \sum_{i} \log P(y_i|x, \hat{y}_{<i}; \theta) + \gamma \sqrt{wc(\hat{y}_{<i})}, \tag{2}$$

where η is the parameter to control the decoding probability from the decoder, and γ is the parameter to control the effect of the word count $wc(\hat{y}_{<i})$, 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_θ parameterized by θ. Our dataset consists of a set of accents $\mathcal{A} = \{A_1, A_2, \cdots, A_n\}$, and for each accent i, we split the data into A_i^{tra} and A_i^{val}, then update θ into θ' by computing gradient descent updates on A_i^{tra}:

$$\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{A_i^{tra}}(f_\theta), \tag{3}$$

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

$$\min_{\theta} \sum_{A_i \sim p(\mathcal{A})} \mathcal{L}_{A_i^{val}}(f_{\theta'_i}) = \sum_{A_i \sim p(\mathcal{A})} \mathcal{L}_{A_i^{val}}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{A_i^{tra}}(f_\theta)}), \tag{4}$$

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

$$\theta \leftarrow \theta - \beta \sum_{A_i \sim p(\mathcal{A})} \nabla_\theta \mathcal{L}_{A_i^{val}}(f_{\theta'_i}), \tag{5}$$

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

accents	**# sample**	**duration (hr)**
Africa **(af)**	4,065	5.04
Australia **(au)**	19,625	22.86
Bermuda **(be)**	363	0.46
Canada **(ca)**	17,422	20.20
England **(en)**	58,274	64.19
Hong Kong **(hk)**	1,181	1.21
India **(in)**	23,878	29.09
Ireland **(ir)**	3,420	3.71
Malaysia **(my)**	843	1.07
New Zealand **(nz)**	6,070	7.06
Philippines **(ph)**	1,318	1.68
Scotland **(sc)**	4,376	5.08
Singapore **(sg)**	693	1.00
South Atlantic **(sa)**	212	0.23
United States **(us)**	145,692	163.89
Wales **(wa)**	1,128	1.16
Total	288,560	327.93

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

accents	**MAML**				**Joint Training**			
	`zero-shot`	`5%-shot`	`25%-shot`	`all-shot`	`zero-shot`	`5%-shot`	`25%-shot`	`all-shot`
				without pre-training				
Bermuda	33.22 $\pm$ 0.46	32.73 $\pm$ 0.47	31.85 $\pm$ 0.48	**29.90 $\pm$ 0.60**	38.92 $\pm$ 0.55	37.84 $\pm$ 0.50	36.23 $\pm$ 0.56	36.12 $\pm$ 0.65
Philippines	50.08 $\pm$ 0.56	48.22 $\pm$ 0.69	45.94 $\pm$ 0.64	**44.43 $\pm$ 0.69**	50.58 $\pm$ 0.81	49.72 $\pm$ 0.80	48.27 $\pm$ 0.85	45.47 $\pm$ 0.93
Wales	33.66 $\pm$ 0.83	33.31 $\pm$ 0.77	31.63 $\pm$ 0.86	**29.70 $\pm$ 0.87**	37.04 $\pm$ 0.68	37.43 $\pm$ 0.69	35.60 $\pm$ 0.80	32.37 $\pm$ 0.87
				with pre-training				
Bermuda	28.25 $\pm$ 0.47	28.64 $\pm$ 0.42	26.59 $\pm$ 0.43	**25.71 $\pm$ 0.43**	31.42 $\pm$ 0.57	31.43 $\pm$ 0.56	30.05 $\pm$ 0.44	27.64 $\pm$ 0.40
Philippines	40.99 $\pm$ 0.51	40.07 $\pm$ 0.52	39.06 $\pm$ 0.44	**37.48 $\pm$ 0.42**	43.17 $\pm$ 0.83	41.98 $\pm$ 0.76	40.56 $\pm$ 0.77	38.79 $\pm$ 0.69
Wales	25.91 $\pm$ 0.73	25.55 $\pm$ 0.86	23.94 $\pm$ 0.73	**23.40 $\pm$ 0.64**	29.14 $\pm$ 0.49	28.54 $\pm$ 0.52	26.70 $\pm$ 0.49	25.01 $\pm$ 0.56

where β 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:

$$\theta \leftarrow \theta - \beta \sum_{A_i \sim p(\mathcal{A})} \nabla_{\theta'_i} \mathcal{L}_{A_i^{val}}(f_{\theta'_i}). \quad (6)$$

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 for only 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 dim_{inner} of 2048, dim_{model} of 512, and dim_{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.

[1]We use CommonVoice Version 2 data (June 2019).

[2]The 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

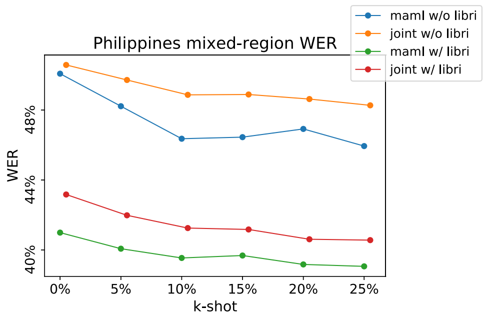

Figure 3: *Few-shot results on Philippines accent in the mixed-region setting.*

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 **af**, **au**, **ca**, **en**, **hk**, **in**, **ir**, **my**, **nz**, **sa**, **sc**, **sg**, and **us**, sampled from all regions, and we validate the model on the **ca**, **sc**, and **sa** accents and test it on the **be**, **ph**, and **wa** accents. The latter setting is to train on five accents, such as **au**, **en**, **ir**, **nz**, and **us**, from specific regions, and we validate the model on the **ca**, **sc**, and **sa** accents, and test it on the **af**, **hk**, **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. [3] We 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%-

[3]We release our code at https://github.com/audioku/cross-accent-maml-asr

Table 3: *Average Word Error Rate (% WER) with Standard Error (SE) results in the cross-region setting.*

accents	MAML				Joint Training			
	zero-shot	5%-shot	25%-shot	all-shot	zero-shot	5%-shot	25%-shot	all-shot
without pre-training								
Africa	40.38 ± 1.11	38.31 ± 1.20	36.36 ± 1.01	**34.64 ± 1.01**	41.56 ± 1.04	41.40 ± 1.08	39.34 ± 1.34	38.32 ± 1.17
Hong Kong	42.04 ± 0.74	40.20 ± 0.89	38.29 ± 0.78	**35.61 ± 0.71**	44.84 ± 0.65	44.88 ± 0.67	44.09 ± 0.66	41.28 ± 0.59
India	62.07 ± 0.90	54.60 ± 1.46	51.71 ± 1.06	**47.85 ± 1.00**	63.09 ± 0.82	56.76 ± 1.08	53.89 ± 1.00	50.73 ± 0.98
Philippines	50.06 ± 0.74	48.17 ± 0.71	47.71 ± 0.78	**45.05 ± 0.82**	53.22 ± 0.97	52.60 ± 0.99	51.64 ± 0.78	48.12 ± 0.76
Singapore	55.75 ± 0.85	55.76 ± 0.83	54.43 ± 0.68	**52.71 ± 1.06**	57.87 ± 0.64	57.21 ± 0.67	55.15 ± 0.69	53.59 ± 0.72
with pre-training								
Africa	32.63 ± 1.25	31.75 ± 1.19	31.09 ± 1.22	**29.75 ± 1.01**	34.61 ± 1.22	33.42 ± 1.18	33.12 ± 1.12	31.63 ± 1.13
Hong Kong	36.06 ± 0.56	36.04 ± 0.71	32.38 ± 0.71	**32.15 ± 0.62**	37.43 ± 0.77	36.51 ± 0.57	35.88 ± 0.51	34.18 ± 0.77
India	54.50 ± 1.41	48.73 ± 1.31	46.15 ± 1.35	**43.54 ± 1.35**	55.43 ± 1.36	50.52 ± 1.26	48.63 ± 1.32	46.58 ± 1.07
Philippines	43.73 ± 0.94	42.96 ± 1.01	40.80 ± 1.03	**40.14 ± 0.98**	45.16 ± 0.98	44.64 ± 1.04	42.38 ± 0.88	41.74 ± 0.98
Singapore	49.45 ± 0.55	48.40 ± 0.56	46.62 ± 0.62	**46.17 ± 0.67**	52.06 ± 0.71	50.48 ± 0.70	49.43 ± 0.69	47.11 ± 0.66

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 Phillippines.

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.

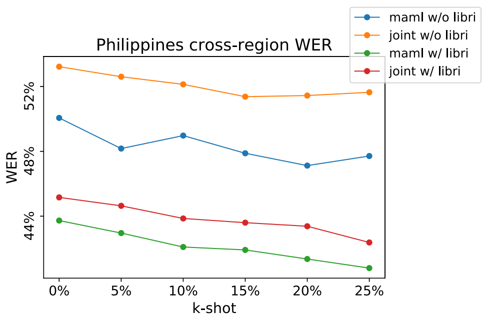

Figure 4: *Few-shot results on Philippines accent in the cross-region setting.*

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

[1] L. W. Kat and P. Fung, "Fast accent identification and accented speech recognition," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 1. IEEE, 1999, pp. 221–224.

[2] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2504–2508.

[3] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *Proc. Interspeech 2017*, pp. 949–953, 2017.

[4] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1126–1135.

[5] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4854–4858.

[6] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning." in *Interspeech*, 2018, pp. 2454–2458.

[7] A. Jain, V. P. Singh, and S. P. Rath, "A multi-accent acoustic model using mixture of experts for speech recognition," *Proc. Interspeech 2019*, pp. 779–783, 2019.

[8] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] J. Gu, Y. Wang, Y. Chen, V. O. Li, and K. Cho, "Meta-learning for low-resource neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3622–3631.

[11] J. Schmidhuber, "Learning to control fast-weight memories: An alternative to dynamic recurrent networks," *Neural Computation*, vol. 4, no. 1, pp. 131–139, 1992.

[12] S. Thrun and L. Pratt, *Learning to learn*. Springer Science & Business Media, 2012.

[13] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," *ICLR*, 2016.

[14] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[15] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, 2016, pp. 1842–1850.

[16] M. Yu, X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauro, H. Wang, and B. Zhou, "Diverse few-shot text classification with multiple metrics," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1206–1215.

[17] A. Madotto, Z. Lin, C.-S. Wu, and P. Fung, "Personalizing dialogue agents via meta-learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5454–5459. [Online]. Available: https://www.aclweb.org/anthology/P19-1542

[18] K. Qian and Z. Yu, "Domain adaptive dialog generation via meta learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2639–2649. [Online]. Available: https://www.aclweb.org/anthology/P19-1253

[19] P.-S. Huang, C. Wang, R. Singh, W.-t. Yih, and X. He, "Natural language to structured query generation via meta-learning," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 732–738.

[20] Z. Lin, A. Madotto, G. I. Winata, Z. Liu, Y. Xu, C. Gao, and P. Fung, "Learning to learn sales prediction with social media sentiment," in *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, 2019, pp. 47–53.

[21] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," *arXiv preprint arXiv:1910.12094*, 2019.

[22] O. Klejch, J. Fainberg, and P. Bell, "Learning to adapt: A meta-learning approach for speaker adaptation," *Proc. Interspeech 2018*, pp. 867–871, 2018.

[23] O. Klejch, J. Fainberg, P. Bell, and S. Renals, "Speaker adaptive training using model agnostic meta-learning," *arXiv preprint arXiv:1910.10605*, 2019.

[24] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for shanghai-accented mandarin," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[25] M. Najafian, A. DeMarco, S. Cox, and M. Russell, "Unsupervised model selection for recognition of regional accented speech," in *Fifteenth annual conference of the international speech communication association*, 2014.

[26] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition," *Proc. Interspeech 2019*, pp. 2140–2144, 2019.

[27] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[28] G. I. Winata, A. Madotto, C.-S. Wu, and P. Fung, "Code-switched language models using neural based synthetic data from parallel sentences," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 271–280.

[29] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, and P. Fung, "Lightweight and efficient end-to-end speech recognition using low-rank transformer," *arXiv preprint arXiv:1910.13923*, 2019.

[30] C. Finn, K. Xu, and S. Levine, "Probabilistic model-agnostic meta-learning," in *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. USA: Curran Associates Inc., 2018, pp. 9537–9548. [Online]. Available: http://dl.acm.org/citation.cfm?id=3327546.3327622

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.