Computer Speech and Language 84 (2024) 101576

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Meta adversarial learning improves low-resource speech recognition[☆]

Yaqi Chen, Xukui Yang, Hao Zhang, Wenlin Zhang, Dan Qu [*], Cong Chen

Information Engineering University, ZhengZhou, China

ARTICLE INFO

Keywords:
Adversarial training
Meta learning
Low-resource
Speech recognition
IARPA-BABEL
OpenSLR

ABSTRACT

Low-resource automatic speech recognition is a challenging task. To resolve this issue, multilingual meta-learning learns a better model initialization from many source languages, allowing for rapid adaption to target languages. However, differences in data scales and learning difficulties vary greatly from one language to another. As a result, the model favors large-scale and simple source languages. Moreover, the shared semantic space of various languages is difficult to learn due to a lack of restrictions on multilingual pre-training. In this paper, we propose a meta adversarial learning approach to address this problem. The meta-learner will be guided to learn language-independent information by using an adversarial auxiliary objective of language identification, which makes the shared semantic space more compact and improves model generalization. Additionally, we optimize adversarial training using Wasserstein distance and temporal normalization, enabling more stable and simple training. Experiment results on IARPA BABEL and OpenSLR show a significant performance improvement. It also outperforms state-of-the-art results by a large margin in all target languages, and especially in few-shot settings. Finally, we demonstrate how our method is superior by using t-SNE visualization.

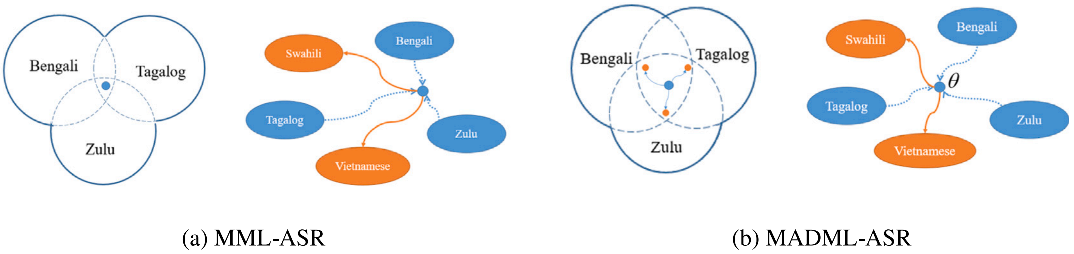
1. Introduction

Recently, end-to-end automatic speech recognition (ASR) has drawn increasing attention due to the continued success of deep neural networks. However, building an end-to-end deep ASR model often requires massive transcription data, which is hard to satisfy for two reasons. Firstly, there are about 8000 languages spoken worldwide, making it nearly impossible to collect a large amount of transcription data for each one. Secondly, it is infeasible to do so due to concerns with privacy, safety, or ethics. This is a real-world challenge that an AI system needs to handle. In recent years, there has been increasing literature on low-resource speech recognition (Pham et al., 2022; Liu et al., 2022), such as data augmentation (Park et al., 2019), transfer learning (Hu et al., 2019), multilingual transfer learning (MTL-ASR) (Hou et al., 2020), and multilingual meta learning (MML-ASR) (Hsu et al., 2020). Data augmentation is achieved by the transformation of the original data, including noise addition, speed adjustment, SpecAugment (Park et al., 2019), etc., which is applied widely in various scenarios. Transfer learning (Farooq and Hain, 2022) typically uses experience from one language with sufficient data to improve the performance of another target language that is similar. But transfer learning requires that the source language be similar to the target language, which is hard to satisfy. MTL-ASR learns general knowledge from a variety of languages to achieve better generalization in new low-resource target languages.

[☆] This work was supported by the National Natural Science Foundation of China (No. 62171470), Natural Science Foundation of Henan Province of China (No. 232300421240), and Henan Zhongyuan Science and Technology Innovation Leading Talent Project (No. 234200510019).

[*] Corresponding author.
E-mail address: qudanqudan@sina.com (D. Qu).

https://doi.org/10.1016/j.csl.2023.101576
Received 23 February 2023; Received in revised form 9 October 2023; Accepted 12 October 2023
Available online 19 October 2023



(a) MML-ASR (b) MADML-ASR

Fig. 1. Illustration: Difference between MML-ASR and MADML-ASR. On the left, the hallow blue circle denotes the semantic space of a language, and the blue dot denotes the model initialization, the orange dot denotes the adapted model initialization. On the right, the blue circles denote the source language, and the orange circles denote the target language. The dashed lines denote the multilingual pre-training, and the solid lines denote the adaptation paths from initialization θ to target languages. (a) MML-ASR: the semantic space overlapped less and model initialization was close to some languages, which have more data and are easier. (b) MADML-ASR: the semantic space overlapped more, and the model has a more balanced distance to all languages.

Learning quickly is a kind of human intelligence; e.g., children can recognize objects only from a few examples. Hence, equipping a deep model with the ability to learn new concepts from a few labeled samples is meaningful for practical usage. Meta-learning has recently been shown to be effective at transferring knowledge from previous tasks to make learning a new task easier. Meta learning is a promising paradigm for addressing the low-source problem since it builds efficient algorithms (e.g., those that need little or even no task-specific fine-tuning) that can learn the new task quickly. Inspired by this, some research has suggested multilingual meta-learning (MML-ASR). Compared to that, MTL-ASR independently trains its model on every sampled task, whereas MML-ASR is accomplished on two levels. The model learns task specific knowledge on the first level (inner loop) by training on a few training data (support set) of the task, whereas on the second level (outer loop), the model gains the knowledge between tasks by adapting to the validation data (query set) of the task. The crux lies in optimizing the generalization capability of the initialization, which is measured by the performance of the adapted model. MML-ASR develops a well-generalized model initialization. MML-ASR has previously been shown to be superior to MTL-ASR in works (Hsu et al., 2020; Xiao et al., 2021).

However, when learning from diverse source languages, MTL-ASR and MML-ASR tend to ignore some issues, resulting in unsatisfactory performance. For one thing, the model favors specific source languages due to differences in data scales and learning difficulties among various languages, both of which affect adaptation to new languages. For another, because there are no restrictions on multilingual pre-training, words and phrases with similar meanings in different languages may be far apart in the semantic space, leading to limited shared information learned by the model, as Fig. 1(a) shows.

To address the problem of unbalanced bias toward specific languages and narrow the gap between different languages in the semantic representation space, we develop a novel meta adversarial approach for multilingual low-resource speech recognition (MADML-ASR), which can boost the pre-trained model's performance and generalization. MADML-ASR incorporates the auxiliary optimization objective of language recognition into the outer loop of meta-learning, and trains it adversarially with the ASR model. The model can then develop a more compact semantic space across languages by learning a more language-independent representation, as shown in Fig. 1(b). It can also reduce the model's preference for particular languages because the model learns the shared representation features of various languages and eliminates the unique representation features of each language.

The main contributions of this paper are:

- We propose MADML-ASR, which can build a more compact semantic space across languages and decrease the preference for particular languages since it guides the adapted model to learn more language-independent knowledge. Moreover, we adopt Wasserstein distance and temporal normalization to optimize adversarial training, making the training more stable and simple.
- We conduct extensive experiments to empirically verify the effectiveness of MADML-ASR. Results show that our method outperforms meta-learning in all target languages, and achieves significant performance in a few-shot setting. In particular, it can reduce CER from 72% to 60% by fine-tuning the model with 22 h of Vietnamese data.
- We demonstrate that our MADML-ASR can reduce the distance of similar concepts between different languages in semantic space by using t-SNE to visualize the encoder output. In addition, it can also be understood as a regularization technique to improve model generalization by adding constraints in the outer loop. We also explain the superiority of MADML-ASR compared to adversarial speech recognition.

A residual of the portion is listed: The traditional methods of low-resource speech recognition, meta learning, and domain adversarial training are described in Section 2. The prior approaches and our proposed method are presented in Section 3. Section 4 describes the experiment setting, including datasets and implementation details. Section 5 describes the performance of meta adversarial training compared to classical approaches. And a conclusion is offered in Section 6.

2. Related work

Over the past few years, there has been a lot of progress made in the area of low-resource speech recognition. Many studies attempt to alleviate the necessity of labeled data for models. These include unsupervised pre-training and semi-supervised methods to

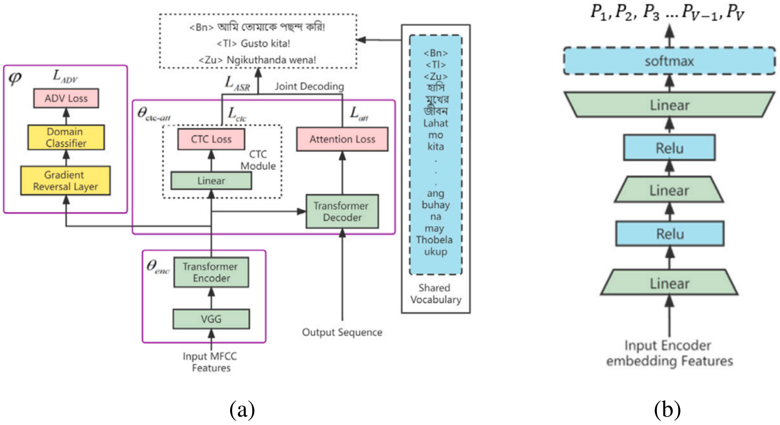


Fig. 2. Illustration: (a): Meta adversarial multilingual learning model architecture based on joint CTC-attention. (b): Language recognition classifier architecture.

utilize unlabeled data, such as wav2vec (Schneider et al., 2019), predictive coding (Chung and Glass, 2020), and self-training (Kahn et al., 2020). However, these approaches require sufficient unlabeled data and computation resources, which may be unavailable on some occasions. Other works include speech data augmentation, transfer learning, multilingual transfer learning(MTL-ASR), multilingual meta learning(MML-ASR), etc. Among them, MTL-ASR/MML-ASR learns some general knowledge from several different languages by multilingual pre-training to achieve better generalization on new low-resource target languages.

Meta-learning learns a good initialization from a large number of tasks. For each task, it learns to adapt to the validation set by fine-tuning a small amount of training data, thus improving its fast learning ability. One of the most representative algorithms is model-agnostic meta-learning algorithm (MAML) (Finn et al., 2017). MAML has achieved good performance in several low-resource domains, such as speaker adaptation (Klejch et al., 2019), accent adaptation (Winata et al., 2020), speech recognition (Hsu et al., 2020; Xiao et al., 2021), and emotion recognition (Chopra et al., 2021). In multilingual low-resource speech recognition, Hsu et al. (2020) demonstrated the superiority of MML-ASR to MTL-ASR for several source languages. Xiao et al. (2021) proposed an adversarial meta sampling algorithm to optimize the task sampling strategy, which fixed the task imbalance problem. However, training an additional sample network requires more computational resources. Hou et al. (2021a,b) developed a meta-adaptor to achieve fine-tuning effectively. Singh et al. (2022) adopted a multi-step weighted loss to improve the stability and performance of meta-learning, similar to MAML++ (Antoniou et al., 2018). What is closely related to our work is (Xiao et al., 2021), we both employ adversarial training to solve the imbalance problem, but work (Xiao et al., 2021) aims to improve the data utilization efficiency by optimizing data sampling, and we aim to learn a more compact semantic space through training process optimization.

Domain adversarial learning (Ganin et al., 2016) was proposed in the image field to find a common subspace of adversarial and clean samples to obtain a robust model. In the speech field, and it was introduced to learn invariant features, such as noise (Dalmia et al., 2018), accent (Sun et al., 2018), speaker (Saon et al., 2017), and language (Adams et al., 2019; Yi et al., 2018). Similar to GAN (Goodfellow et al., 2014), it is likely to produce the training instability problem easily for the gradient vanishing problem under the optimal discriminator (Arjovsky and Bottou, 2017). To address this problem, there are works (Arjovsky et al., 2017; Gulrajani et al., 2017; Wei et al., 2018) that use Wasserstein distance instead of Jensen–Shannon(JS) divergence to measure the loss under the optimal discriminator. What is closely related to our work is (Adams et al., 2019), which introduces the auxiliary goal of language recognition in the training process, encouraging the model to learn language-independent features. However, we use the auxiliary goal of language recognition in the outer loop of meta-learning to guide the adapted model to learn language-independent features, allowing the model to learn more generalized initializations. As shown in Fig. 1(b), the blue dots represent the pre-adaptation model, and the yellow dots represent the post-adaptation model. By constraining the post-adaptation model to be language-independent, the model learns a more generalized initialization and adapts better to the new task. In contrast, the idea of encouraging the model to learn language-independent features (Adams et al., 2019) means the model initialization (the blue dot) within the three language common regions, which is similar to what is shown in Fig. 1(a), . We also analyze the superiority of adding constraints to the outer loop compared to other approaches in Section 5.4. Moreover, we employ Wasserstein distance to measure the ASR model loss under the optimal language recognizer and adopt temporal normalization to realize more stable and easier training.

3. Methods

3.1. Multilingual learning ASR

In multilingual pre-training, network structures such as RNN (Karita et al., 2019), LAS (Toshniwal et al., 2018), Transformer (Hou et al., 2020; Zhou et al., 2018), CTC-Attention (Kim et al., 2017), and others are available. Among these works, CTC-Attention (Kim et al., 2017) is widely used in MML-ASR (Hou et al., 2020; Xiao et al., 2021) as it is highly expressive in multilingual pre-training by combining the fast alignment of CTC with the stronger learning ability of attention.

We use the joint CTC-attention architecture (Kim et al., 2017) for multilingual speech recognition, as shown in Fig. 2. The ASR model architecture consists of three parts: the encoder and the decoder in the Transformer (Vaswani et al., 2017), and a connectionist temporal classification(CTC) module (Graves et al., 2006), which can guide the model to learn good alignments and improve the convergence speed. During training, the loss function can be expressed as a weighted sum of the decoding loss L_{att} and the CTC loss L_{ctc}:

$$L_{asr} = \lambda L_{ctc} + (1 - \lambda) L_{att} \tag{1}$$

where the hyper-parameter λ denotes the weight of CTC loss.

By using the byte pair encoding (BPE) algorithm (Sennrich et al., 2016), a shared vocabulary composed of subwords and language tokens (like "bn" and "zu") from source languages is adopted as model outputs to allow language-independent training and recognition. Additionally, a language token is added as an auxiliary language identification mark at the start of each speech sample. The language identification objective is dropped when fine-tuning the model. Because the target language has a different vocabulary from the source language, we replaced the output layer and reinitialized it to the size corresponding to the target vocabulary.

3.2. Multilingual meta-learning ASR

Model-agnostic meta-learning(MAML) learns a good model initialization from many training tasks to quickly adapt to new tasks. In this paper, we apply MAML to effectively learn from a set of languages and quickly adapt to some new target languages in few-shot settings. The dataset is a set of N languages $D_s = \{D_s^i\}_{i=1}^N$, and for each language $i(i = 1, 2, \dots, N)$, we sample tasks T_i from D_s^i and divide T_i into two subsets, the support set T_{sup}^i and the query set T_{query}^i. We denote our ASR model as f_θ parameterized by θ, and the algorithm consists of two optimization steps:

Firstly, the base learner learns every task from the initial meta-learner θ in the inner loop. Concretely, the adapted model parameters θ_i are updated from θ by performing gradient descent on the support set $T_{\mathrm{sup}}^i(i = 1, 2, \dots, N)$:

$$\theta_i = \theta - \alpha \nabla_\theta L_{T_{\mathrm{sup}}^i}(\theta) \tag{2}$$

where α is the learning rate of the inner loop, and $L_{T_{\mathrm{sup}}^i}(\theta)$ is the loss function computed by using Eq. (1).

Secondly, the meta-learner integrates the knowledge of each base learner in the outer loop. Specifically, the meta model parameters θ are updated by calculating all the task query losses using the adapted model parameters θ_i over the query set T_{query}^i:

$$\theta \leftarrow \theta - \beta \sum_i \nabla_\theta L_{T_{query}^i}(\theta_i) \tag{3}$$

where β is the learning rate of the outer loop. To sum up, let $p(T)$ be the distribution of tasks. We can formulate the meta-learning process as follows:

$$\min_\theta E_{T_i \sim p(T)} L_{T_{query}^i}(\theta - \alpha \nabla_\theta L_{T_{\mathrm{sup}}^i}(\theta)) \tag{4}$$

From Eqs. (2) and (3), we can see that the meta-learning requires the computation of the second-order derivatives of θ, which is computationally expensive. Therefore, we use the First-order MAML algorithm (FOMAML) as (Hsu et al., 2020; Winata et al., 2020), so the Eq. (3) can be reformulated as:

$$\theta \leftarrow \theta - \beta \sum_i \nabla_{\theta_i} L_{T_{query}^i}(\theta_i) \tag{5}$$

3.3. Meta adversarial multilingual learning ASR

As introduced in Section 3.2, the meta learner needs to learn a batch task sampled from a variety of languages in each meta-training iteration. Due to their sizes and difficulties, there are big variations between them in real-world situations. Hence, the model will favor a language that is simple and broad. Since there are no constraints on multilingual pre-training, words and phrases with similar meanings in several languages may be far apart in the semantic space. Hence, our goal is to address the preference issue and reduce the semantic space gap between various languages.

To solve these issues, we propose an effective meta adversarial meta-learning approach (MADML-ASR) for multilingual low-resource speech recognition. At the output of the model encoder, we add a language recognizer g_φ, which is a nonlinear neural network, with the structure depicted in Fig. 2(b). It can direct the adapted model to build language-independent representations and close the gap between various languages in the semantic space by confronting the adapted ASR model in the outer loop of meta learning. It also requires two optimization steps: an inner optimization and an outer optimization.

In the inner loop, the ASR model needs to minimize the ASR loss L_{ASR} on the support set, as Eq. (2). In this way, we can get an adapted model f_{θ_i}. And the language recognizer remains unchanged.

In the outer loop, the adapted ASR model needs to maximize the language recognition loss L_{LDV} while minimizing the ASR loss L_{ASR}, so the optimization objective takes the following forms:

$$\min_\theta L = \min_\theta (L_{ASR} - \mu L_{LDV}) = \min_\theta E_{T_i \sim p(T)} (L_{T_{query}^i}^{ASR}(\theta_i) - \mu L_{T_{query}^i}^{LDV}(\theta_i)) \tag{6}$$

where μ denotes the weight of language recognition loss L_{LDV}, and L_{ASR} is the loss function computed by using Eq. (1). Language recognition is a multi-classification task, and L_{LDV} can be computed using Eq. (16). The language recognition network is shared between different language tasks. It is worth clarifying that we use the gradient reversal layer instead of simply negating the loss as seen in Fig. 2(a).

Here, we separate the model parameters into two groups: one for the encoder and feature extraction network θ_{enc} and the other for the decoder and CTC layer $\theta_{ctc-att}$. Since only the model encoder is related to the language recognizer, the encoder parameters are updated using Eq. (7).

$$\theta_{enc} \leftarrow \theta_{enc} - \beta \sum_{i} \nabla_{\theta_i} (L^{ASR}_{T^i_{query}}(\theta_i) - \mu L^{LDV}_{T^i_{query}}(g_{\varphi}(\theta^i_{enc}))) \tag{7}$$

As the decoder and the CTC module's parameters $\theta_{ctc-att}$ have no relevance on L_{LDV}, $\theta_{ctc-att}$ are updated as follows:

$$\theta_{ctc-att} \leftarrow \theta_{ctc-att} - \beta \sum_{i} \nabla_{\theta_i} L^{ASR}_{T^i_{query}}(\theta^i_{ctc-att}) \tag{8}$$

The language recognizer needs to minimize the language recognition query loss for all tasks, whose parameters φ are updated as follows:

$$\varphi \leftarrow \varphi - \beta \sum_{i} \nabla_{\varphi} L^{LDV}_{T^i_{query}}(g_{\varphi}) \tag{9}$$

We summarize our whole algorithm as Algorithm 1.

Algorithm 1 Meta Adversarial Training.

```
Require: learning rate $\alpha$ ,$\beta$, adversarial loss weight $\mu$.
Initialize Transformer model $f_\theta$, language recognizer $g_\varphi$
while not done do
    Sample a batch of tasks $\{T_i\}_{i=1}^n$
    For all $T_i$ do
        Sample a support set $T^i_{sup}$ and a query set $T^i_{query}$ from $T_i$
        Compute the adapted model parameters $\theta_i$ on $T^i_{sup}$ using Eq. (2) // inner loop
        Evaluate the loss of adapted model parameters $\theta_i$ on $T^i_{query}$ using Eq. (3)
    End for
    Update encoder parameters $\theta_{enc}$ using Eq. (7) // outer loop
    Update decoder and CTC modules parameters $\theta_{ctc-att}$ using Eq. (8)
    Update language recognizer parameters $\varphi$ using Eq. (9)
end while
return $\theta$
```

3.4. Optimized adversarial training

Empirical research shows that adversarial training is hard to conduct due to its instability. To address this problem, we proposed an optimized adversarial training method by adopting Wasserstein distance and temporal normalization. If we use x to represent the input features, the adversarial training process can be summarized as follows:

$$\min_{f} \max_{g} E_{x \sim p_i}[\log g(f(x))] + E_{x \sim p_{-i}}[\log(1 - g(f(x)))] \tag{10}$$

where $p_{-i} = p_0 \cup p_1 \cup \cdots \cup p_{i-1} \cup p_{i+1} \cup \cdots \cup p_{N-1}$, p_i denotes the distribution of the ith language, p_{-i} denotes the distribution of languages other than the ith language. g and f denote the language recognizer and the ASR model, respectively. We convert the language recognition task into a binary classification task using Eq. (10). We find it is similar to GAN (Goodfellow et al., 2014), whose optimization object can be expressed as:

$$\min_{G} \max_{D} V(D, G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{11}$$

where $p_{data}(x)$ and $p_z(z)$ denote the distribution of true data and fake data, respectively. D and G denote the generator and the discriminator, respectively. In this work, G represents the encoder and feature extractor, and D represents the language recognizer. For GAN, the optimization object under the best discriminator is equal to JS divergence (Lin, 1991) between $p_{data}(x)$ and $p_z(z)$.

The language recognizer can easily reach the optimal state since the language recognizer is structurally simpler than the ASR model. Our optimization function under the best language recognizer, like GAN, is equal to the JS divergence between p_i and p_{-i}. In high-dimensional space, the embedding vectors of different languages barely overlap, resulting in JS divergence to a constant and the gradient vanishing problem (Arjovsky et al., 2017). However, it is not necessary to design a special language recognizer to solve this problem, as it is more expensive and there is no guarantee that it is useful.

Table 1
Multilingual dataset statistics in terms of hours (h) languages.

IARPA BEBAL				
Source	Bengali 61.76 Zulu 62.13 Lithuanian 42.52	Tagalog 84.56 Turkish 77.18 Guarani 43.03	Target	Vietnamese 87.72 Swahili 44.39 Tamil 68.36
OpenSLR				
Source	Tamil (SLR-65) 7.08 Gujarati (SLR-78) 7.89 Basque (SLR-76) 13.86 Colombian Spanish (SLR-72) 7.58 Northern English (SLR-83) 14.60	Chilean Spanish (SLR-71) 7.15 Peruvian Spanish (SLR-73) 9.22 Galician (SLR-77) 10.31 Kannada (SLR-79) 8.68	Target	Malayalam (SLR-63) 5.51 Marathi (SLR-66) 5.7 Venezuelan Spanish (SLR-75) 4.81

To solve this problem, we borrowed the idea of Wasserstein GAN (Arjovsky et al., 2017) and measured the distance between p_i and p_{-i} using Wasserstein distance rather than JS divergence. In this case, our adversarial training object function can be described as:

$$\max_{\varphi\in\Phi} E_{x\sim p_i}[g_\varphi(f_\theta(x))] - E_{x\sim p_{-i}}[g_\varphi(f_\theta(x))] \tag{12}$$

where the language recognizer is parameterized by φ that lies in a compact space Φ (Arjovsky et al., 2017). And Φ is compact, which implies that all the functions g_φ will be K-Lipschitz.

Specifically, the pre-softmax output of the language recognizer is described as $\mathbf{X}$, $\mathbf{X} \in \Omega^{B\times T\times N}$. B, T, and N represent the batch size, the length of the feature sequence, and the number of languages, respectively. Generally, we need to average $\mathbf{X}$ over time and use the function $softmax$ to get the output probability Y':

$$\mathbf{Y}'{=}softmax(E(\mathbf{X})) \tag{13}$$

where $\mathbf{Y}' = (\mathbf{Y}'_0, \mathbf{Y}'_1, \dots \mathbf{Y}'_{B-1})$, $\mathbf{Y}'_i = (y'_{i,0}, y'_{i,1}, \dots, y'_{i,N-1})(i = 0, 1, \dots, B-1)$ with $\sum_{j=0}^{N-1} y'_{i,j} = 1$. $\mathbf{Y}_i$ represents the true label, which is a one-hot vector. Finally, the language recognition loss takes the form:

$$L_{LDV} = -\sum_{i=0}^{B-1} \mathbf{Y}_i \log \mathbf{Y}_i'^{T} = -\sum_{i=0}^{B-1}\sum_{j=0}^{N-1} y_{i,j} \log y'_{i,j} \tag{14}$$

Similar to work (Arjovsky et al., 2017), we need to cancel log at the output layer for the language recognizer to calculate the loss as Eq. (12). So in Eq. (14), log function needs to be cancelled.

This causes the unstable training due to the overflow problem in multilingual training. Therefore, we propose temporal normalization to solve this problem by applying $log\_softmax$ function to the temporal dimension, allowing the output value of each language can be maintained in a relatively stable range. It can be described as:

$$\mathbf{Z}'{=}E(log\_softmax(\mathbf{X})) \tag{15}$$

where $\mathbf{Z}' = (\mathbf{Z}'_0, \mathbf{Z}'_1, \dots \mathbf{Z}'_{B-1})$, $\mathbf{Z}'_i = (z'_{i,0}, z'_{i,1}, \dots, z'_{i,N-1})(i = 0, 1, \dots, B-1)$. In this way, the output of different languages can be unified on the same scale, which can prevent the problem of misjudgment caused by taking the output value too large in a certain period of time. Finally, the language recognition loss L_{LDV} takes the form:

$$L_{LDV} = -\sum_{i=0}^{B-1} \mathbf{Z}_i \mathbf{Z}_i'^{T} = -\sum_{i=0}^{B-1}\sum_{j=0}^{N-1} z_{i,j} z'_{i,j} \tag{16}$$

where $\mathbf{Z}_i$ represents the real label, which is a one-hot vector. This greatly increases the stability of adversarial training by eliminating the need to balance the training of the ASR model and the language recognizer.

4. Experiment

4.1. Datasets

Our experiment is based on IARPA BABEL (Gales et al., 2014) and OpenSLR.[1] Specifically, IARPA BABEL consists of conversational telephone speech in 25 languages that was collected in various environments. The total scale of transcribed audio data varies by language and environment. There are Full Language Packs (FLP) and Limited Language Packs (LLP) for each language . We created the BABEL-3 and BABEL-6 datasets to compare with prior research (Hsu et al., 2020; Xiao et al., 2021). To construct BABEL-3, we selected three languages: Bengali (Bn), Tagalog (Tl), and Zulu (Zu). To construct BABEL-6, we selected six languages: Bengali (Bn), Tagalog (Tl), Zulu (Zu), Turkish (Tr), Lithuanian (Lt), Guarani (Gn). A pre-trained model is trained using the language's

[1] https://openslr.org/resources.php

Table 2
Character error rate (%CER) for different target languages.

	BABEL-3			BABEL-6			OpenSLR		
	Vietnamese	Swahili	Tamil	Vietnamese	Swahili	Tamil	SLR-63	SLR-66	SLR-75
MTL-ASR (Hsu et al., 2020)	57.4	48.1	65.6	59.7	48.8	65.6	/	/	/
MML-ASR (Hsu et al., 2020)	49.9	41.4	57.5	50.1	42.9	58.9	/	/	/
MML-ASR (Xiao et al., 2021)	/	/	/	45.1	36.14	50.16	/	/	/
AMS-ASR (Xiao et al., 2021)	/	/	/	43.35	32.19	48.56	/	/	/
No-Pretrain	69.06	57.31	59.08	69.06	57.31	59.08	67.61	67.84	82.57
MTL-ASR	43.84	37.63	51.54	42.03	35.47	50.84	80.76	87.13	82.11
MML-ASR	43.99	38.03	51.26	42.07	32.38	48.47	42.56	37.87	42.75
MADML-ASR	**43.23**	**35.59**	**50.05**	**40.97**	**29.83**	**46.66**	**41.62**	**35.93**	**40.58**

Table 3
Character error rate (%CER) for different proportions of target languages.

Language	Vietnamese			Swahili			Tamil		
Proportions	10%	25%	50%	10%	25%	50%	10%	25%	50%
Duration (h)	8.77	21.75	43.50	4.44	11.10	22.20	6.84	17.09	34.18
No-Pretrain	99.35	84.96	80.72	92.29	75.39	60.04	89.10	80.39	65.56
MML-ASR	79.62	72.42	58.12	76.97	62.78	54.21	75.36	70.77	65.25
MADML-ASR	**71.37**	**60.72**	**54.65**	**69.61**	**58.52**	**49.56**	**73.04**	**67.70**	**62.76**

FLP. We also selected three languages as target languages: Vietnamese (Vi), Swahili (Sw), and Tamil (Ta). We fine-tuned the model by using its FLP, and tested it on LLP(about 10 h). Table 1 describes the dataset statistics for the experimental data. It only shows the FLP for IARPA BABEL, and every language has about 10 h for LLP. Note that except for Table 2 where BABEL-6 was used, BABEL-3 was employed for the rest of the experiments. For OpenSLR, we selected 9 languages as source languages for pre-training, and fine-tuned three target languages: Malayalam (SLR-63), Marathi (SLR-66), and Venezuelan Spanish (SLR-75). We used 80% of the data for each language for training and 20% for testing.

4.2. Implementation details

We use the Kaldi (Ravanelli et al., 2019) toolkit for feature extraction to obtain 40-dimensional Mel-frequency cepstral coefficients (MFCC) features and 3-dimensional pitch features computed every 10 ms over a 25 ms window. To reduce over-fitting during training, we used spectral enhancement (Park et al., 2019) and speed perturbation (Ko et al., 2015). Transformer is a model that consists of 4 encoder blocks and 2 decoder blocks, and it uses a 6-layer VGG convolutional network for feature extraction. Each of these blocks comprises 512 hidden units, 4 attention heads, and 2048 feed-forward hidden units. It contains 64 examples for each task, of which 32 examples make up the support set and 32 examples make up the query set. The weight for CTC loss λ was set to 0.3. The weight of language recognition loss μ was set to 1. During the inference process, the best sequence is obtained using beam search. When using BPE (Sennrich et al., 2016), we set the vocabulary size as 9000. Except for the inner loop in meta-learning, we use the SGD algorithm and the learning rate is 0.0001. For the rest of model optimization, we turn to Adam (Kingma and Ba, 2015). We set the warmup steps to 12000, and k to 0.5 for IARPA BABEL. For OpenSLR, we set the warmup steps to 1000, and k to 0.5. Character error rate (CER) is used as the experiment's criterion, and the five best models are averaged to build the final model for evaluation.

5. Results

5.1. Main results

5.1.1. Fine-tuning performance in different target languages

We compared the performance of random initialization (No-Pretrain), MTL-ASR, MML-ASR, and MADML-ASR for fine-tuning target languages on IARPA BABEL and OpenSLR, and compared the results with those in works (Hsu et al., 2020; Xiao et al., 2021). As shown in Table 2, for IARPA BABEL datasets, firstly, No-Pretrain is very poor due to the lack of help from other languages. Secondly, both MTL-ASR and MML-ASR outperform No-Pretrain on all target languages, which can effectively improve the performance of low-resource languages. Our results outperform work (Hsu et al., 2020) because the joint CTC-attention structure is superior to a CTC model. And because we both use the Joint CTC-Attention structure, our baseline performance is close to work (Xiao et al., 2021). Finally, our MADML-ASR outperforms MML-ASR in all target languages, both in BABEL-3 and BABEL-6 by a large margin. Moreover, our method outperforms the meta adversarial sampling (Xiao et al., 2021).

On OpenSLR, however, we find No-Pretrain performs poorly due to the few training data for 4 h. It can be seen that MTL-ASR does not perform as well as random initialization (No-Pretrain). Experiments show that the pre-training model performs well, but the fine tuning performance is unsatisfactory, which is most likely the result of overfitting. In our analysis, this may be due to a large number of languages (9 languages) and the small data size for each language (about 10 h) on OpenSLR, making it challenging

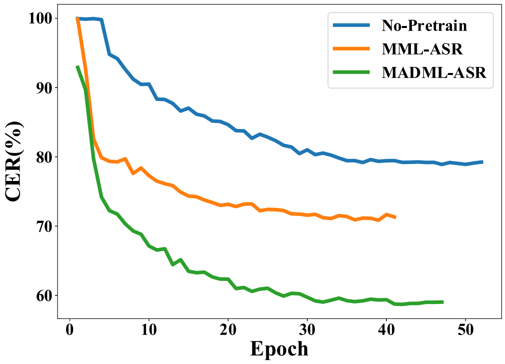


Fig. 3. Learning curves of fine-tuning three different pre-trained models by using 25% of Vietnamese data.

for the model to learn common information across multiple languages. In contrast, MML-ASR effectively improves the performance of all target languages, which shows strong generalization. And our MADML-ASR achieves the best performance across all target languages.

5.1.2. Fine-tuning performance on different data scales

From Table 2, we find that the improvement of MADML-ASR is around 2%. Since we have about 44 h to 87 h of training data for Vietnamese on IARPA BABEL, we compare the performance of these methods under three settings with proportions of 10%, 25%, and 50% of target language data, and Table 3 reports our experimental results.

In the low-resource settings like 10% and 25% of training data, we can observe that MADML improves CER by about 12% than MML-ASR when fine-tuning the model with 25% (22 h) of Vietnamese data, and improves by about 8% CER when fine-tuning the model with 10% (8.7 h) of Vietnamese data. But it only improves CER by about 4% or 1% than MML-ASR with 50% (44 h) or 100% of Vietnamese data. It can be seen that the MADML-ASR is more effective under low-resource settings.

Speed of Fine-tuning. We discover that MADML-ASR converges faster than MML-ASR when adapting to new tasks. The fine-tuning process for all languages shows the same trend. Here, we only display the result of three models (No-Pretrain, MML-ASR, and MADML-ASR) with 25% of Vietnamese data, as shown in Fig. 3. It is clear that MML-ASR is effective in improving the performance of low-resource speech recognition. Moreover, MADML-ASR adapts to new tasks faster than MML-ASR. By directing the adapted model to learn language-independent features, the model learns a better initialization and can quickly adapt to novel tasks.

5.2. Ablation study

5.2.1. Impact of meta adversarial training module

We contrast the pretraining performances of multilingual transfer learning ASR (MTL-ASR), multilingual meta-learning ASR (MML-ASR), and adversarial multilingual meta-learning (MADML-ASR).

Convergence performance. As seen in Fig. 4, these are the curves of the valid CER with the number of epochs for three methods on IARPA BABEL and OpenSLR. It can be seen that MML-ASR and MTL-ASR are comparable on IARPA BABEL, but MTL-ASR is inferior to MML-ASR on OpenSLR. It can be discovered that the effectiveness of MTL-ASR falls dramatically when the number of source languages are more. Because MTL-ASR tries to learn information from many languages in turn, which is increasingly challenging as the number of languages increases. Yet, MML-ASR consistently works well no matter how many languages are used since that meta-learning incorporates the data from each task and builds a stronger initialization. Moreover, MADML-ASR consistently obtains the best performance regardless of the datasets by bridging the gap between diverse source languages and lowering the preference.

Convergence speed. As is shown in Fig. 4(a), we find that MTL-ASR converges fast in the beginning and slowly in the end, and the overall convergence speed is slower than MADML-ASR's. It shows that MADML-ASR can effectively improve the convergence speed of the model and achieve similar performance to MML-ASR with only half the epochs of MML-ASR. As is shown in Fig. 4(b), similar findings can be observed on OpenSLR, and MADML-ASR still achieves the fastest convergence while MTL-ASR performs badly.

5.2.2. Impact of optimized adversarial training

We contrast the training curves of language recognition loss L_{LDV} when using Wasserstein distance or JS divergence, as shown in Fig. 5. We can observe that without using Wasserstein distance, the loss converges to 0 quickly, making it impossible to propagate gradients and optimize the model. However, when using Wasserstein distance, the loss keeps unchanging, indicating that the model is unable to distinguish between language categories, consistent with our expectations. As Fig. 5(b) shows, it can be seen that the model performs badly without Wasserstein distance. Therefore, it is essential to use Wasserstein distance when training our model.

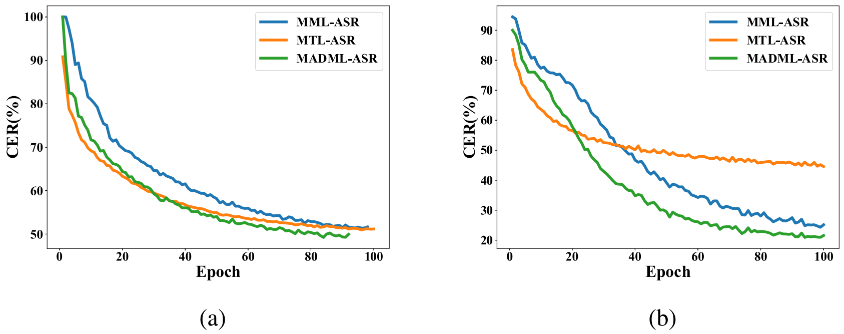
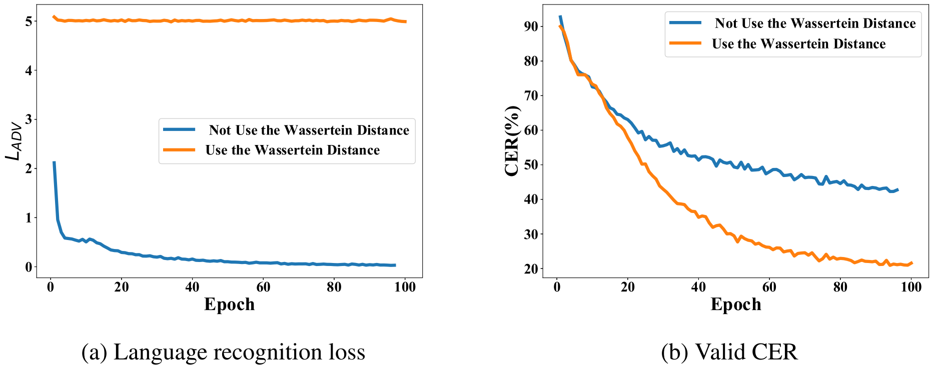


Fig. 4. Learning curves of three pre-training methods. (a): Learning curves on IARPA BABEL. (b): Learning curves on OpenSLR.



(a) Language recognition loss

(b) Valid CER

Fig. 5. Learning curves of using Wasserstein distance or JS divergence.

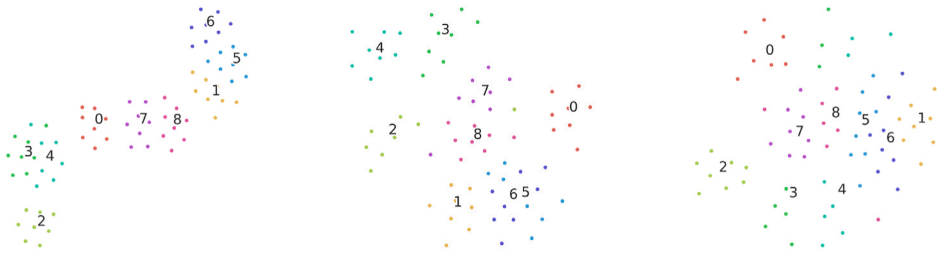
5.3. Meta adversarial training principle analysis

We use 2D t-SNE visualization for the encoder outputs of different languages, as shown in Fig. 6. The visualization results of the encoder output features for MTL-ASR, MML-ASR, and MADML-ASR are shown in Fig. 6(a) to Fig. 6(c), respectively. It is clear that MTL-ASR has a low correlation between different languages. Because MTL-ASR learns each task in turn during the learning process, it is difficult for it to learn the shared semantic space of diverse languages. The results explain the poor performance of MTL-ASR in both pre-training and fine-tuning on OpenSLR. However, MML-ASR is more evenly distributed among diverse languages, resulting in better performance than MTL-ASR. The semantic space of different languages has not overlapped, so the model has not learned very well. In contrast, some data from the second language and the third language in our MADML-ASR start to cross with other languages in the semantic space, and the distribution of all languages is denser. Therefore, it is clear that the distance and distribution among diverse languages have become closer and denser. The findings support the claim that our MADML-ASR can effectively constrain the ASR model to learn a more compact representation space between diverse languages and obtain a better initialization model.

5.4. Why is adversarial training added in the outer loop?

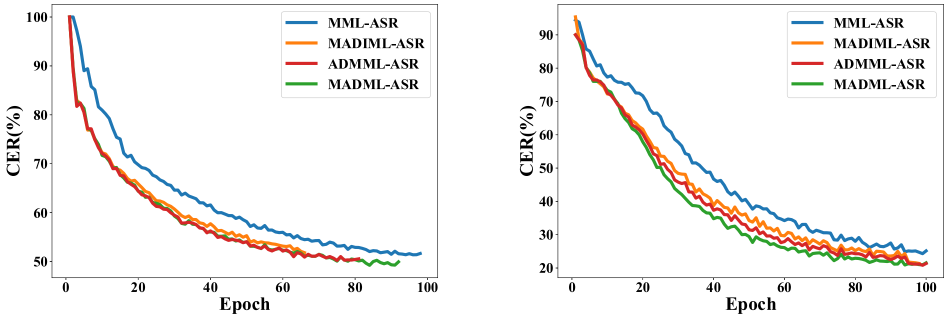
We analyzed the superiority of introducing adversarial training in the outer loop of meta-learning and compared three approaches. One is to add adversarial training to the inner loop of meta-learning, called MADIML-ASR. MADIML-ASR minimize the ASR loss and maximize the LDV loss in the inner loop, while only minimize the ASR loss in the outer loop. The second is MADML-ASR. The third is to add adversarial training to both the inner and outer loop of meta-learning, called ADMML-ASR, which minimize the ASR loss and maximize the LDV loss both in the inner loop and the outer loop. We show the pre-training performances of three methods and compare them with MML-ASR.

As is shown in Fig. 7, first, MADIML-ASR performs better than MML-ASR, but worse than MADML-ASR and ADMML-ASR. So the idea of adding adversarial training to the inner loop is effective, but not as effective as adding it to the inner and outer loop or the outer loop. Second, ADMML-ASR converges close to MADML-ASR when there are few languages involved (IARPA BABEL), but is inferior to MADML-ASR in terms of the final performance. But when there are more languages available (OpenSLR), it is inferior to MADML-ASR. This makes sense given that ADMML-ASR needs to divide its attention in order to provide adversarial training in the inner loop. The more languages there are, the more attention must be split, causing a decline in performance. Finally, we give two reasons why introducing adversarial training in the outer loop is better. First, it guides the ASR model after inner-loop adaptation to learn language-independent features, which can extend the shared semantic space as shown in Fig. 1(b). Second, it also resembles a regularization technique by adding constraints to the outer loop of meta-learning, thus improving the generalization of the model.



(a) MTL-ASR (b) MML-ASR (c) MADML-ASR

Fig. 6. T-SNE representation of encoder states corresponding to different languages on OpenSLR, and different colors mean different languages. We calculate the cluster center of every language, and mark it using 0–8. (a): The model is trained using MTL-ASR. (b): The model is trained using MML-ASR. (c): The model is trained using MADML-ASR.



(a) IARPA BEBEL (b) OpenSLR

Fig. 7. Pre-training learning curves for using MML-ASR, MADIML-ASR, ADMML-ASR, and MADML-ASR.

6. Conclusion

In this paper, we propose an meta adversarial learning approach for multilingual low-resource speech recognition, which enables the model encoder to learn more language-independent features by incorporating the auxiliary objective of language recognition into the outer loop of meta-learning and training it adversarially with the ASR model. Such an approach can broaden the shared semantic space learned by the ASR model and address the model's preference for some languages. Moreover, we propose an optimized adversarial training algorithm, which makes the training more stable and easier. Experiments results on IARPA BABEL and OpenSLR show the effectiveness of MADML-ASR. Moreover, the underlying principles of this method have also been thoroughly studied through numerous experiments. In the future, we plan to explore the integration of adversarial training algorithms with more meta-learning algorithms. And it may be applied to different applications other than speech recognition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Adams, O., Wiesner, M., Watanabe, S., Yarowsky, D., 2019. Massively multilingual adversarial speech recognition. In: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019. Vol. 1. Long and Short Papers, Association for Computational Linguistics, pp. 96–108. http://dx.doi.org/10.18653/v1/n19-1009.

Antoniou, A., Edwards, H., Storkey, A.J., 2018. How to train your MAML. ArXiv. arXiv:1810.09502.

Arjovsky, M., Bottou, L., 2017. Towards principled methods for training generative adversarial networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=Hk4_qw5xe.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. In: Proceedings of Machine Learning Research, vol. 70, PMLR, pp. 214–223, URL: http://proceedings.mlr.press/v70/arjovsky17a.html.

Chopra, S., Mathur, P., Sawhney, R., Shah, R.R., 2021. Meta-learning for low-resource speech emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, on, Canada, June 6-11, 2021. IEEE, pp. 6259–6263. http://dx.doi.org/10.1109/ICASSP39728.2021.9414373.

Chung, Y., Glass, J.R., 2020. Generative pre-training for speech with autoregressive predictive coding. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. IEEE, pp. 3497–3501. http://dx.doi.org/10.1109/ICASSP40776.2020.9054438.

Dalmia, S., Sanabria, R., Metze, F., Black, A.W., 2018. Sequence-based multi-lingual low resource speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018. IEEE, pp. 4909–4913. http://dx.doi.org/10.1109/ICASSP.2018.8461802.

Farooq, M.U., Hain, T., 2022. Investigating the impact of cross-lingual acoustic-phonetic similarities on multilingual speech recognition. http://dx.doi.org/10.48550/arXiv.2207.03390, CoRR. arXiv:2207.03390.

Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. In: Proceedings of Machine Learning Research, vol. 70, PMLR, pp. 1126–1135, URL: http://proceedings.mlr.press/v70/finn17a.html.

Gales, M.J.F., Knill, K.M., Ragni, A., Rath, S.P., 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. In: 4th Workshop on Spoken Language Technologies for under-Resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14-16, 2014. ISCA, pp. 16–23, URL: http://www.isca-speech.org/archive/sltu_2014/gales14_sltu.html.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.S., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17, 59:1–59:35, URL: http://jmlr.org/papers/v17/15-239.html.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 2672–2680, URL: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Graves, A., Fernández, S., Gomez, F.J., Schmidhuber, J., 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Cohen, W.W., Moore, A.W. (Eds.), Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006. In: ACM International Conference Proceeding Series, vol. 148, ACM, pp. 369–376. http://dx.doi.org/10.1145/1143844.1143891.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein GANs. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5767–5777, URL: https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

Hou, W., Dong, Y., Zhuang, B., Yang, L., Shi, J., Shinozaki, T., 2020. Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. In: Meng, H., Xu, B., Zheng, T.F. (Eds.), Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020. ISCA, pp. 1037–1041. http://dx.doi.org/10.21437/Interspeech.2020-2164.

Hou, W., Wang, Y., Gao, S., Shinozaki, T., 2021a. Meta-adapter: Efficient cross-lingual adaptation with meta-learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, on, Canada, June 6-11, 2021. IEEE, pp. 7028–7032. http://dx.doi.org/10.1109/ICASSP39728.2021.9414959.

Hou, W., Zhu, H., Wang, Y., Wang, J., Qin, T., Xu, R., Shinozaki, T., 2021b. Exploiting adapters for cross-lingual low-resource speech recognition. IEEE/ACM Tran. Audio Speech Lang. Process. 30, 317–329.

Hsu, J., Chen, Y., Lee, H., 2020. Meta learning for end-to-end low-resource speech recognition. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. IEEE, pp. 7844–7848. http://dx.doi.org/10.1109/ICASSP40776.2020.9053112.

Hu, K., Bruguier, A., Sainath, T.N., Prabhavalkar, R., Pundak, G., 2019. Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models. In: Kubin, G., Kacic, Z. (Eds.), Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019. ISCA, pp. 2155–2159. http://dx.doi.org/10.21437/Interspeech.2019-1868.

Kahn, J., Lee, A., Hannun, A.Y., 2020. Self-training for end-to-end speech recognition. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. IEEE, pp. 7084–7088. http://dx.doi.org/10.1109/ICASSP40776.2020.9054295.

Karita, S., Wang, X., Watanabe, S., Yoshimura, T., Zhang, W., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N.E.Y., Yamamoto, R., 2019. A comparative study on transformer vs RNN in speech applications. In: IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019. IEEE, pp. 449–456. http://dx.doi.org/10.1109/ASRU46091.2019.9003750.

Kim, S., Hori, T., Watanabe, S., 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, la, USA, March 5-9, 2017. IEEE, pp. 4835–4839. http://dx.doi.org/10.1109/ICASSP.2017.7953075.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1412.6980.

Klejch, O., Fainberg, J., Bell, P., Renals, S., 2019. Speaker adaptive training using model agnostic meta-learning. In: IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019. IEEE, pp. 881–888. http://dx.doi.org/10.1109/ASRU46091.2019.9003751.

Ko, T., Peddinti, V., Povey, D., Khudanpur, S., 2015. Audio augmentation for speech recognition. In: INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015. ISCA, pp. 3586–3589, URL: http://www.isca-speech.org/archive/interspeech_2015/i15_3586.html.

Lin, J., 1991. Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 37 (1), 145–151. http://dx.doi.org/10.1109/18.61115.

Liu, Q., Yang, Y., Gong, Z., Li, S., Ding, C., Minematsu, N., Huang, H., Cheng, F., Kurohashi, S., 2022. Hierarchical softmax for end-to-end low-resource multilingual speech recognition. http://dx.doi.org/10.48550/arXiv.2204.03855, CoRR. arXiv:2204.03855.

Park, D.S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In: Kubin, G., Kacic, Z. (Eds.), Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019. ISCA, pp. 2613–2617. http://dx.doi.org/10.21437/Interspeech.2019-2680.

Pham, N., Waibel, A., Niehues, J., 2022. Adaptive multilingual speech recognition with pretrained models. In: Ko, H., Hansen, J.H.L. (Eds.), Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022. ISCA, pp. 3879–3883. http://dx.doi.org/10.21437/Interspeech.2022-872.

Ravanelli, M., Parcollet, T., Bengio, Y., 2019. The pytorch-kaldi speech recognition toolkit. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019. IEEE, pp. 6465–6469. http://dx.doi.org/10.1109/ICASSP.2019.8683713.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L., Roomi, B., Hall, P., 2017. English conversational telephone speech recognition by humans and machines. In: Lacerda, F. (Ed.), Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017. ISCA, pp. 132–136, URL: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0405.html.

Schneider, S., Baevski, A., Collobert, R., Auli, M., 2019. wav2vec: Unsupervised pre-training for speech recognition. In: Kubin, G., Kacic, Z. (Eds.), Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019. ISCA, pp. 3465–3469. http://dx.doi.org/10.21437/Interspeech.2019-1873.

Sennrich, R., Haddow, B., Birch, A., 2016. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, http://dx.doi.org/10.18653/v1/p16-1162.

Singh, S., Wang, R., Hou, F., 2022. Improved meta learning for low resource speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022. IEEE, pp. 4798–4802. http://dx.doi.org/10.1109/ICASSP43922.2022.9746899.

Sun, S., Yeh, C., Hwang, M., Ostendorf, M., Xie, L., 2018. Domain adversarial training for accented speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018. IEEE, pp. 4854–4858. http://dx.doi.org/10.1109/ICASSP.2018.8462663.

Toshniwal, S., Sainath, T.N., Weiss, R.J., Li, B., Moreno, P.J., Weinstein, E., Rao, K., 2018. Multilingual speech recognition with a single end-to-end model. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018. IEEE, pp. 4904–4908. http://dx.doi.org/10.1109/ICASSP.2018.8461972.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008, URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wei, X., Gong, B., Liu, Z., Lu, W., Wang, L., 2018. Improving the improved training of wasserstein GANs: A consistency term and its dual effect. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=SJx9GQb0-.

Winata, G.I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., Fung, P., 2020. Learning fast adaptation on cross-accented speech recognition. In: Meng, H., Xu, B., Zheng, T.F. (Eds.), Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020. ISCA, pp. 1276–1280. http://dx.doi.org/10.21437/Interspeech.2020-0045.

Xiao, Y., Gong, K., Zhou, P., Zheng, G., Liang, X., Lin, L., 2021. Adversarial meta sampling for multilingual low-resource speech recognition. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, pp. 14112–14120, URL: https://ojs.aaai.org/index.php/AAAI/article/view/17661.

Yi, J., Tao, J., Wen, Z., Bai, Y., 2018. Adversarial multilingual training for low-resource speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018. IEEE, pp. 4899–4903. http://dx.doi.org/10.1109/ICASSP.2018.8461771.

Zhou, S., Xu, S., Xu, B., 2018. Multilingual end-to-end speech recognition with a single transformer on low-resource languages. CoRR. arXiv:1806.05059.