Towards Better Meta-Initialization with Task Augmentation for Kindergarten-aged speech Recognition

Yunzheng Zhu, Ruchao Fan, and Abeer Alwan

yunzhengzhu19@g.ucla.edu

Department of Electrical and Computer Engineering
University of California, Los Angeles, USA

UCLA Samueli Electrical & Computer Engineering

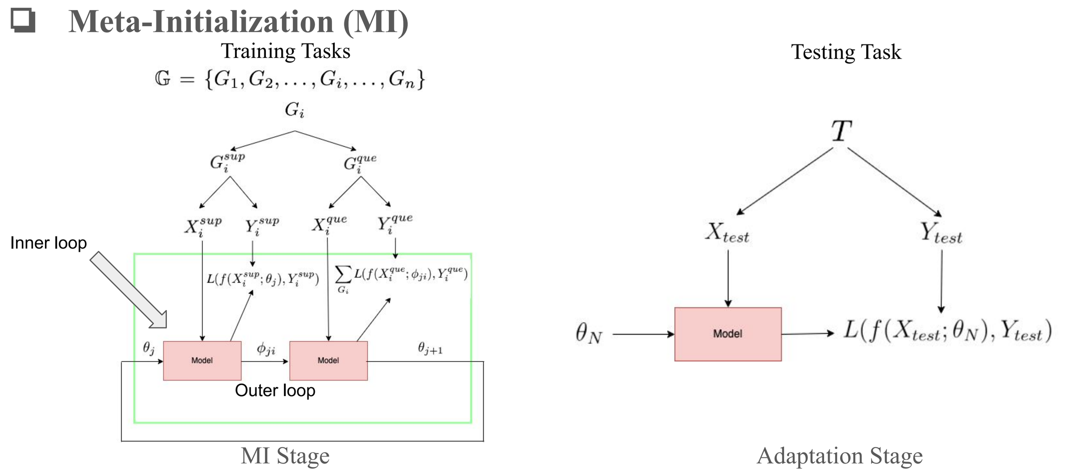
icassp 2022 Singapore

Introduction

- Child ASR is a challenging problem, in part, because of data scarcity. It is especially true for kindergarten-aged children. Data scarcity will lead to model overfitting to the training data. Thus, we need good starting points for training.
- Methods used to find a good model initialization:
  - Supervised pre-training methods
  - Unsupervised/self-supervised pre-training methods
  - **Meta-learning [1][2] to learn model initialization (MI)**
- However, meta-initialization is vulnerable to overfitting on training tasks, in terms of learner overfitting. **Task-level augmentation** is proposed by simulating new ages using time and frequency warping techniques.

Method

❑ Meta-Initialization (MI)



1) Inner loop update: $\phi_{ji} = \theta_j - \alpha \nabla_{\theta_j} L(f(X_i^{sup};\theta_j), Y_i^{sup})$
   - θ_j is the model parameters in the inner-loop at step j
   - X_i^{sup} and Y_i^{sup} are data samples and corresponding labels in the support set of task i, respectively
   - ϕ_{ji} is the model parameter updated for task i and step j
   - f is the forward computation of the model
   - L is the cross-entropy loss used in acoustic modelling
   - a is the learning rate for the inner-loop optimizer
   - ∇ is the nabla operator for computing the gradient of
2) Outer loop update: $\theta_{j+1} \leftarrow \theta_j - \beta \nabla'_{\theta_j} \boxed{\sum_{G_i} L(f(X_i^{que};\phi_{ji}), Y_i^{que}))}$ Meta-objective
   - Meta-objective function is the summation over the loss function for query set of each task, which quantifies how the adaptation behaves in the inner loop.

- Minimize this objective function with respect to θ_j to find a suitable adaptation model.
- X_i^{que} and Y_i^{que} are data samples and corresponding labels in the query set of task i, respectively
- β is the learning rate for the outer-loop optimizer, and ∇'_{θ_j} indicates that only first-order Model-Agnostic Meta-Learning (MAML) is used.
- After enough training steps, N, the final model θ_N is regards as the learned initialization for the unseen test task.

❑ Age-based Task Augmentation for MI

Two types of overfitting in MI [3]:

1. Memorization overfitting

Reason: θ_{j+1} memorizes all tasks and does not rely on support sets for inner-loop adaptation

Solution: Randomly sampling the support set and query set at each step so each sample has equal possibility of participating in either outer or inner loop update.

2. Learner overfitting:

Reason: θ_{j+1} is unable to generalize well on the test task T

Solution: Task augmentation to increase model generalization for the test tasks. We propose age-based task augmentation by simulating new tasks of children's speech using time and frequency warping techniques, such as speed perturbation and VTLP.

Experimental Setup

❑ **Database: OGI Kids Speech (Scripted)**
- Randomly split into 70% train, 8% development, and 22% test without speaker overlap for each age group (K - G10)
- Meta-learning: Nine meta-training tasks (G2-G10) (45 hours), one meta-validation task (G1) (6 hours), one meta-testing task (K) (4 hours)

❑ **Acoustic Model**
- HMM-GMM for frame-level alignment from all the meta-training tasks (G2-G10)
- HMM-DNN:
- Feature: 80-dim log-mel filterbank
- Input: 160-dim log-mel filterbank (current frame + next frame)
- Model: 4x512 BLSTM

❑ **Meta-Initialization (MI)**
- HMM-GMM & HMM-DNN: same as 2a and 2b

❑ **Augmentation**:
- Task Augmentation (MI):
  - VTLP: 3x (0.9, 1.0, 1.1)
  - SP: 3x (0.9, 1.0, 1.1)
- Data Augmentation (adaptation):
  - VTLP: 3x (0.9, 1.0, 1.1)
  - SP: 3x (0.9, 1.0, 1.1)
  - SpecAug (on-the-fly):
    - Time masking: 2 times with maximum width of 8
    - Frequency masking: 8 times with maximum width of 5

Acknowledgement

This work is supported in part by NSF.

Results

Table 1: % Word error rate (WER) for Data Augmentation (Data Aug) mechanisms on baseline system, meta-initialization (MI), and the proposed task augmentation (Task Aug) mechanisms for MI with vocal tract length perturbation (VTLP) and speed perturbation (SP) on the Kindergarten-aged development and test sets. SPT stands for supervised pre-training. Raw Aug stands for augmentation within each task without creating new tasks.

Model	Data Aug Type	MI Aug Type	Dev	Test
Baseline	-	-	53.17	55.01
	SP	-	46.13	43.75
+ Data Aug	VTLP	-	45.42	46.05
	SpecAug	-	56.69	53.70
+ SPT [18]	-	-	36.27	29.06
+ MI	-	-	35.21	30.68
+ Raw Aug	-	SP	36.62	28.00
	-	VTLP	36.27	30.06
+ Task Aug	-	SP	**34.86**	**27.50**
	-	VTLP	**34.86**	29.06

- SP is better than VTLP as a method to simulate new tasks.
- Task dependent augmentation (Task Aug) outperforms augmenting the data within the original tasks (Raw Aug).
- 27.5 % WER improvement is achieved on the kindergarten test set with MI and task augmentation.

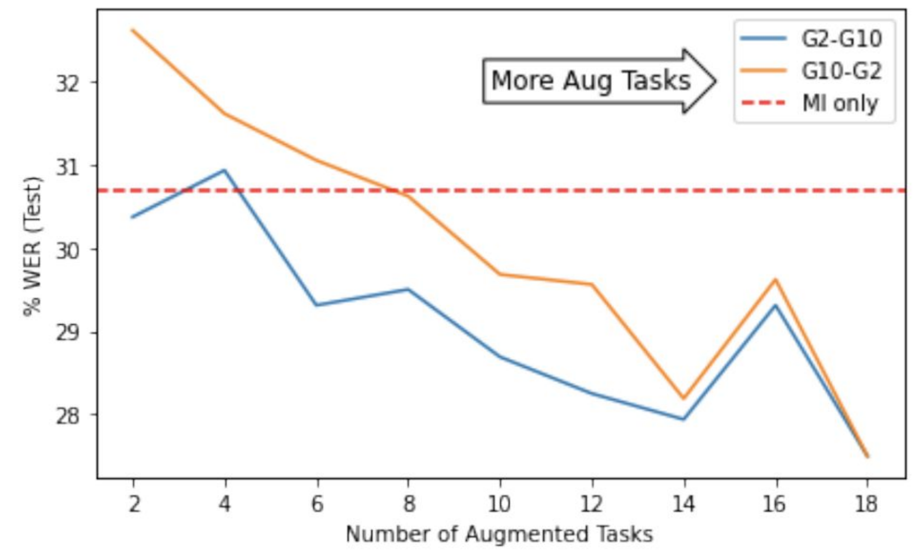


Fig. 1: % WER Results of task augmentation mechanism using speed perturbation (SP) versus the number of augmentation tasks for MI on the Kindergarten test set. The tasks are added either from G2 to G10 (in blue), or from G10 to G2 (in orange). The dashed line (in red) is MI without any task augmentation mechanism.

- To obtain insights, we added the number of tasks of SP from two directions:
  - Increasing order: G2 → G10
  - Decreasing order: G10 → G2
- Creating new tasks similar to the target task is more effective to address the learner overfitting problem.
- A 10% relative WER improvement over MI without the task augmentation.

Table 2: % Word error rate (WER) for data augmentation during the adaptation stage with SpecAug, vocal tract length perturbation (VTLP), and speed perturbation (SP) on the Kindergarten development and test sets.

Aug Type (in adaptation stage)	Dev	Test
No Aug	34.86	27.50
SpecAug	32.75	**27.01**
VTLP	**32.39**	28.13
SP	33.45	27.75

- Although all three strategies can improve the performance on the development set, only SpecAug achieves slightly better performance on the test set.

Conclusion

- To deal with the data scarcity of children's speech, particularly kindergarten-aged, meta-initialization is used to find a good starting point for training the acoustic model.
- To mitigate the overfitting in meta-initialization, particularly learner-overfitting, an age-based task augmentation mechanism is proposed to simulate new ages using time and frequency warping techniques.
- Data augmentation strategies (SP, VTLP) used in the task augmentation stage are not helpful in the adaptation stage.
- A 51% relative WER improvement over the baseline is achieved in the final system.

Future Work

- Continue to improve child ASR
- Extend the technique to other low-resource ASR tasks

References

[1] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." International Conference on Machine Learning. PMLR, 2017.

[2] Antoniou, Antreas, Harrison Edwards, and Amos Storkey. "How to train your MAML." arXiv preprint arXiv:1810.09502 (2018).

[3] Yao, Huaxiu, et al. "Improving generalization in meta-learning via task augmentation." International Conference on Machine Learning. PMLR, 2021.