TEST-TIME ADAPTATION TOWARD PERSONALIZED SPEECH ENHANCEMENT: ZERO-SHOT LEARNING WITH KNOWLEDGE DISTILLATION

*Sunwoo Kim, Minje Kim**

Indiana University, Department of Intelligent Systems Engineering, Bloomington, IN 47408, USA
kimsunw@iu.edu, minje@indiana.edu

ABSTRACT

In realistic speech enhancement settings for end-user devices, we often encounter only a few speakers and noise types that tend to reoccur in the specific acoustic environment. We propose a novel personalized speech enhancement method to adapt a compact denoising model to the test-time specificity. Our goal in this test-time adaptation is to utilize no clean speech target of the test speaker, thus fulfilling the requirement for zero-shot learning. To complement the lack of clean speech, we employ the knowledge distillation framework: we distill the more advanced denoising results from an overly large teacher model, and use them as the pseudo target to train the small student model. This zero-shot learning procedure circumvents the process of collecting users' clean speech, a process that users are reluctant to comply due to privacy concerns and technical difficulty of recording clean voice. Experiments on various test-time conditions show that the proposed personalization method can significantly improve the compact models' performance during the test time. Furthermore, since the personalized models outperform larger non-personalized baseline models, we claim that personalization achieves model compression with no loss of denoising performance. As expected, the student models underperform the state-of-the-art teacher models.

Index Terms— Speech enhancement, personalization, zero-shot learning, knowledge distillation, model compression

1. INTRODUCTION

Recent advances in deep learning-based speech enhancement (SE) models have shown superior performance to traditional machine learning and signal processing methods [1]. These models are typically with a large model capacity and trained from a large training set, so they generalize well to various test-time conditions including different speakers, noises, and signal-to-noise ratios (SNR) of the added noise. However, the growing model size and computational complexity renders them difficult to deploy onto resource-constrained devices. Hence, model compression methods have been gaining interest to facilitate the practicality of deep-learning architectures in real-time applications. Some main modes of compression such as quantization, pruning, and knowledge distillation (KD) have shown great promise in dramatically reducing the complexity [2] These *context-agnostic* compression methods do not utilize the test-time context that the model will be situated in. Hence, a loss in test-time performance is inevitable.

In some practical use cases though, e.g., a family-owned smart assistant device sitting in the living room, it suffices for the enhancement model to perform well only for the specific test environment. Hence, a *context-aware* fine-tuning method is promising, as it can turn the general-purpose SE model, *the generalist*, into a special-purpose version, *the specialist*. It can be seen as a test-time adaptation to the specific speakers and their acoustic context, overcoming the generalization losses. We call this kind of fine-tuned specialists *personalized speech enhancement* (PSE) systems.

The topic of domain adaptation has been an active area of research in computer vision and speech and audio processing. One common procedure for domain transfer is regularizing the differences between the learned representations of source and target datasets, and it has been applied for emotion, speech, and speaker recognition [3, 4]. However, these applications were provided ample target data, which cannot be assumed for usual cases. Other methods rely on few-shot adaptation in cases where a small number of ground-truth signals are available [5]. However, it can be challenging to obtain user information due to recent privacy infringement, data leakage issues and advancement in DeepFake technology rendering customers uneasy towards releasing personal information. With user compliance, the user enrollment phase can obtain trigger phrases from the users, but these recordings can be contaminated with existing background noise and might not be long enough.

In contrast to aforementioned approaches, zero-shot learning (ZSL) is a data-free solution suitable for training tasks where no additional labeled data is available [6, 7]. In the context of PSE, ZSL does not require test users' clean speech data or their home acoustic environment, while its goal is still to adapt to the test-time specificity. ZSL is an active research topic for classification tasks, where ZSL frameworks typically infer test-time labels by extracting and utilizing auxiliary information [8, 9]. Similarly, ZSL in the speech and audio classification applications extracts semantic properties or articulatory distribution to obtain labels during test-time [10, 11]. However, ZSL for speech enhancement has not been widely studied. In [12, 13], a mixture of local expert model was introduced as a ZSL solution to test-time adaptation of an SE model. It achieves the adaptation goal by employing a classifier to select the most suitable one out of pre-trained specialist models for a given noisy test signal. Although it is a valid adaptation method, it only works on a few predefined contexts, i.e., varying signal-to-noise ratio (SNR) and gender of the speaker, rather than adapting to the test-time speaker's personality or the unique context.

In this paper, we present a zero-shot learning approach to PSE, based on the KD framework [14]. As a ZSL method, it does not ask for private signals from the user, while it can still adapt to the user's speech and recording environment, thus qualifying as a PSE method. When ZSL is implemented via KD strategies, it is common to use data synthesis techniques through generative adversarial frameworks where the generator generates fake samples [15] or through KD using activation or output statistics from pre-trained

*This material is based upon work supported by the National Science Foundation under Grant No. 2046963.

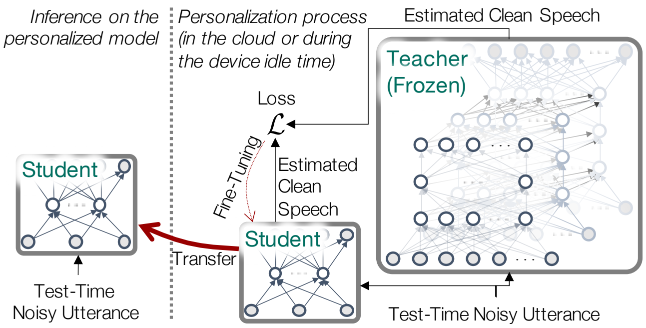


Figure 1: An overview of the proposed KD-based PSE process. The estimated clean speech by the student model is compared against the result from a larger teacher model, whose discrepancy is used to fine-tune the student model. This can be done during the test time as it does not require the clean speech target.

teacher models to synthesize pseudo samples [16, 17]. Instead of including an intermediate data synthesis step, our proposed model directly uses the teacher model's outputs as ground-truth targets to optimize the student model, where the teacher model is defined by a large generalist model trained from a large training dataset, while the student model is a relatively smaller model.

We aim at edge devices with limited computing resources, and a compact student model is used as an affordable solution. Although a small student model may not generalize well to the test environment, it can be improved via test-time adaptation without requiring ground-truth targets. The basic assumption is that the teacher model's large computational capacity guarantees the generalization goal, i.e., it works well in most test-time environments, whose excellent SE results can be considered as if they were the target clean speech from the student model's perspective. When deploying our framework, we ultimately only use the student model on the device. The teacher model is envisioned to be placed externally on a cloud server, where the actual fine-tuning operations are conducted. The student models can be frequently updated on the server side and transferred to the user device. There have been prior works using KD for training source separation models when parallel clean data is not available [18, 19] and for domain adaptation [20], but to our best knowledge, this ZSL PSE framework for model compression is novel in the field of speech enhancement.

2. THE PROPOSED ZERO-SHOT LEARNING METHOD FOR PERSONALIZED SPEECH ENHANCEMENT

We implement the proposed ZSL-based PSE via KD. Our goal is to fine-tune a compact student model during the test time, so it adapts to the unseen test speaker and environment. KD plays a key role in our ZSL framework, as its teacher model provides a *pseudo* target for the student model to learn from, while the target clean speech of the test-time noisy utterance is absent. We claim that the proposed PSE method will be helpful when the system needs to deal with the peculiarity of the test-time conditions. This kind of flexibility will be also advantageous for SE models if the system has to be frequently relocated to different test environments. Figure 1 describes the KD-based PSE process that can fine-tune the student model during the test time.

2.1. Training Teacher Speech Enhancement Models

First, we train the teacher model $\mathcal{T}(\cdot)$ using a large-scale speech corpus and noise dataset. Here, the teacher model $\mathcal{T}(\cdot)$ is defined with a large model architecture, so it can properly approximate the complex general-purpose speech denoising function. Once trained, $\mathcal{T}(\cdot)$ is frozen and *not* fine-tuned, assuming that its performance as a generalist meets the quality standard in most test cases.

The formulation of the dataset is as follows. We assume an additive signal model where the observed signal $\boldsymbol{x}$ is a mixture of a clean speech source $\boldsymbol{s}$ and noise source $\boldsymbol{n}$ of identical duration: $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$, which are all monaural. The clean speech utterances are taken from a large dataset containing many speakers, $\boldsymbol{s} \in \mathbb{G}$, and the noise recordings are similarly from a large dataset containing various noise types, $\boldsymbol{n} \in \mathbb{N}$. The objective is to denoise $\boldsymbol{x}$ and estimate waveform $\hat{\boldsymbol{s}}$ that closely approximates the target clean speech, e.g., $\boldsymbol{s} \approx \hat{\boldsymbol{s}} \leftarrow \mathcal{T}(x)$.

The optimization on $\mathcal{T}(\cdot)$ reduces the loss between the target utterance $\boldsymbol{s}$ and reconstruction $\hat{\boldsymbol{s}}$, e.g., $\arg\min_{\boldsymbol{\Theta}_\mathcal{T}} \mathcal{L}(\boldsymbol{s}||\mathcal{T}(\boldsymbol{x};\boldsymbol{\Theta}_\mathcal{T}))$, where $\boldsymbol{\Theta}_\mathcal{T}$ denotes the trainable parameters of the teacher model.

2.2. Pre-Training Student Speech Enhancement Models

Our student models $\mathcal{S}(\cdot)$ are pre-trained in a similar way to the teacher models, i.e., by updating its own model parameters $\arg\min_{\boldsymbol{\Theta}_\mathcal{S}} \mathcal{L}(\boldsymbol{s}||\mathcal{S}(\boldsymbol{x};\boldsymbol{\Theta}_\mathcal{S}))$ using the same generic datasets, $\mathbb{G}$ and $\mathbb{N}$. However, its small capacity hinders it from generalizing well to the unseen test conditions. The goal of test-time PSE is to reduce the performance gap between $\mathcal{T}(\cdot)$ and $\mathcal{S}(\cdot)$, which will be explained in detail in Sec. 2.3. Differently from $\mathcal{T}(\cdot)$, the purpose of pre-training $\mathcal{S}(\cdot)$ is to prepare the student model better than a random initialization, when it is fine-tuned.

2.3. Test-time Personalized Speech Enhancement

During the test time, we assume that the SE system is exposed to mixture signals composed of clean speech utterances from the test speaker, $\boldsymbol{s} \in \mathbb{S}$, and background noise sources, $\boldsymbol{n} \in \mathbb{M}$. Hence, in the most extreme case, our pre-trained student models can fail to generalize well to the test mixtures if those test-time speech and noise sources are not included to the generic datasets $\mathbb{G}$ and $\mathbb{N}$ used for pre-training, i.e., $\mathbb{G} \cap \mathbb{S} = \varnothing$ and $\mathbb{N} \cap \mathbb{M} = \varnothing$.

Given these assumptions, we propose a PSE framework that can adapt to a new environment without requiring test user's ground-truth clean speech samples or any other auxiliary information of the speakers and acoustic scene. Since we formulate the proposed PSE method as a fine-tuning process, we begin with a compact student model, $\mathcal{S}(\cdot)$, pre-trained in a context-agnostic manner as in Sec. 2.2. To fine-tune it, its denoising result, $\hat{\boldsymbol{s}}_\mathcal{S}$, must be compared against the target to compute the loss and perform backpropagation. However, since we assume the target is not available, we use the pseudo target computed elsewhere, i.e., using the teacher model.

This process falls in the category of the KD framework in which a student model is optimized using a teacher model's prediction [14]. In our PSE context, we employ a large pre-trained teacher model $\mathcal{T}(\cdot)$ whose predicted clean utterance serves as the target to compute the student model's loss. Both student and teacher models are initialized with pre-trained generalist SE models. During test-time, the student model is optimized as: $\arg\min_{\boldsymbol{\Theta}_\mathcal{S}} \mathcal{L}(\hat{\boldsymbol{s}}_\mathcal{T}||\mathcal{S}(\boldsymbol{x};\boldsymbol{\Theta}_\mathcal{S}))$, where $\hat{\boldsymbol{s}}_\mathcal{T}$ is the estimates of clean speech signals obtained from the teacher model and $\boldsymbol{\Theta}_\mathcal{S}$ are trainable parameters of the student model.

The teacher's estimates $\hat{\boldsymbol{s}}_{\mathcal{T}}$ are only approximations of ground-truth targets $\boldsymbol{s}$, and can contain denoising artifacts [21]. However, under a zero-shot condition, we assume having these synthesized pseudo targets is better than nothing. Hence, the performance of the fine-tuning results depends on the quality of $\hat{\boldsymbol{s}}_{\mathcal{T}}$. To this end, we employ relatively large models that surely outperform the student models on the test signals, i.e., $\mathcal{L}(\boldsymbol{s}||\hat{\boldsymbol{s}}_{\mathcal{T}}) < \mathcal{L}(\boldsymbol{s}||\hat{\boldsymbol{s}}_{\mathcal{S}})$. Thus, we hypothesize that the student will still learn from these imperfect targets and improve its test-time SE performance.

3. EXPERIMENTS

3.1. The Datasets

For pre-training SE models, we used clean speech recordings from the LibriSpeech corpus [22] and noise recordings from the MUSAN dataset [23]. We used Librispeech's `train-clean-100` and `dev-clean` subset for training and validation, which we denote as $\mathbb{G}_{\text{tr}}$ and $\mathbb{G}_{\text{va}}$ respectively. We split MUSAN's `free-sound` subset into training and validation partitions at 80:20 ratio, denoted as $\mathbb{N}_{\text{tr}}$ and $\mathbb{N}_{\text{va}}$ respectively. This exposes the generalist models to up to 251 speakers and 843 noise recordings during training. The noisy mixtures are obtained by adding the noise to speech signals at random input SNR levels uniformly chosen between -5 and 10 dB.

For fine-tuning, we used 30 speakers from Librispeech's `test-clean` and noise from the WHAM! corpus [24] whose samples are recorded in 44 different locations. From these sets, we create $K = 30$ unique test environment by assigning a unique noise location to each speaker. Given a test environment index $k \in \{1, \ldots, K\}$, we extract clean speech signals from the k-th speaker $\mathbb{S}^{(k)}$ and add noises from k-th location $\mathbb{M}^{(k)}$. For each test environment, $\mathbb{S}^{(k)}$ and $\mathbb{M}^{(k)}$ are split into separate sets: the partitions are approximately 5, 1, and 1 minutes of clean speech, which we denote by $\mathbb{S}^{(k)}_{\text{ft}}$, $\mathbb{S}^{(k)}_{\text{va}}$ and $\mathbb{S}^{(k)}_{\text{te}}$. The noise datasets are prepared similarly: $\mathbb{M}^{(k)}_{\text{ft}}$, $\mathbb{M}^{(k)}_{\text{va}}$ and $\mathbb{M}^{(k)}_{\text{te}}$. We use $\mathbb{S}^{(k)}_{\text{ft}}$ and $\mathbb{M}^{(k)}_{\text{ft}}$ to fine-tune the student model via the KD process, where the teacher model's denoising results are used as the pseudo target. $\mathbb{S}^{(k)}_{\text{va}}$ and $\mathbb{M}^{(k)}_{\text{va}}$ are mixed up to validate the student model during fine-tuning, mainly to prevent overfitting. Finally, we set aside $\mathbb{S}^{(k)}_{\text{te}}$ and $\mathbb{M}^{(k)}_{\text{te}}$ to test the final performance of the fine-tuned PSE system.

When we simulate various test conditions, the noise and speech sources are mixed under four different input SNR levels (i.e. -5 dB, 0 dB, 5 dB and 10 dB) and used them for fine-tuning, validation, and testing. All audio files are loaded at 16 kHz sampling rate and standardized to have a unit-variance.

3.2. Models

Most of our SE models are based on the uni-directional gated recurrent unit (GRU) architecture [25]. We use frequency-domain representations obtained through the short-time Fourier transform (STFT) as inputs to the SE models. STFT is with a Hann windowed frame of 1024 samples and a hop size of 256 samples. A dense layer transforms the GRU's output into the ideal ratio masks (IRM) [26]. The denoising mask is applied element-wise to the mixture spectrogram, then transformed back to the time-domain signal $\hat{\boldsymbol{s}}$ through inverse STFT. Finally, we use negative scale-invariant signal-to-noise ratio (SI-SNR) as the loss function [27].

While the GRU architecture for the student models is fixed with two hidden layers, we vary their hidden units from 32 to 1024 to verify the impact of PSE on the different architectural choices. Meanwhile, as for the teacher model, we use a 3×1024 GRU architecture, which is large enough to outperform the students. In addition to the large GRU architecture, we also employ ConvTasNet (CTN) [28] as an alternative teacher model. Since the CTN teacher outperforms the GRU model due to its structural advantage, we can confirm the impact of the teacher's performance on the KD-based PSE. The CTN model is configured using implementation available in Asteroid's source separation toolkit [29]. Same architecture as reported in [28] is adopted (i.e. 8 convolutional blocks and 3 repeats with global layer normalization) and trained on a single-speaker speech enhancement task. The model architectures, their respective number of parameters, and the multiplier-accumulator (MAC) operation counts are shown in Table 1. Note that CTN is not the largest model but it requires extensive MAC operations.

Table 1: Complexity of student and teacher models in MACs and number of parameters. MACs are computed given 1-second inputs.

Models		MACs (G)	Param. (M)
Student	GRU (2×32)	0.010	0.08
	GRU (2×64)	0.011	0.17
	GRU (2×128)	0.026	0.41
	GRU (2×256)	0.071	1.12
	GRU (2×512)	0.216	3.42
	GRU (2×1024)	0.729	11.55
Teacher	GRU (3×1024)	1.126	17.85
	ConvTasNet [28]	9.831	4.92

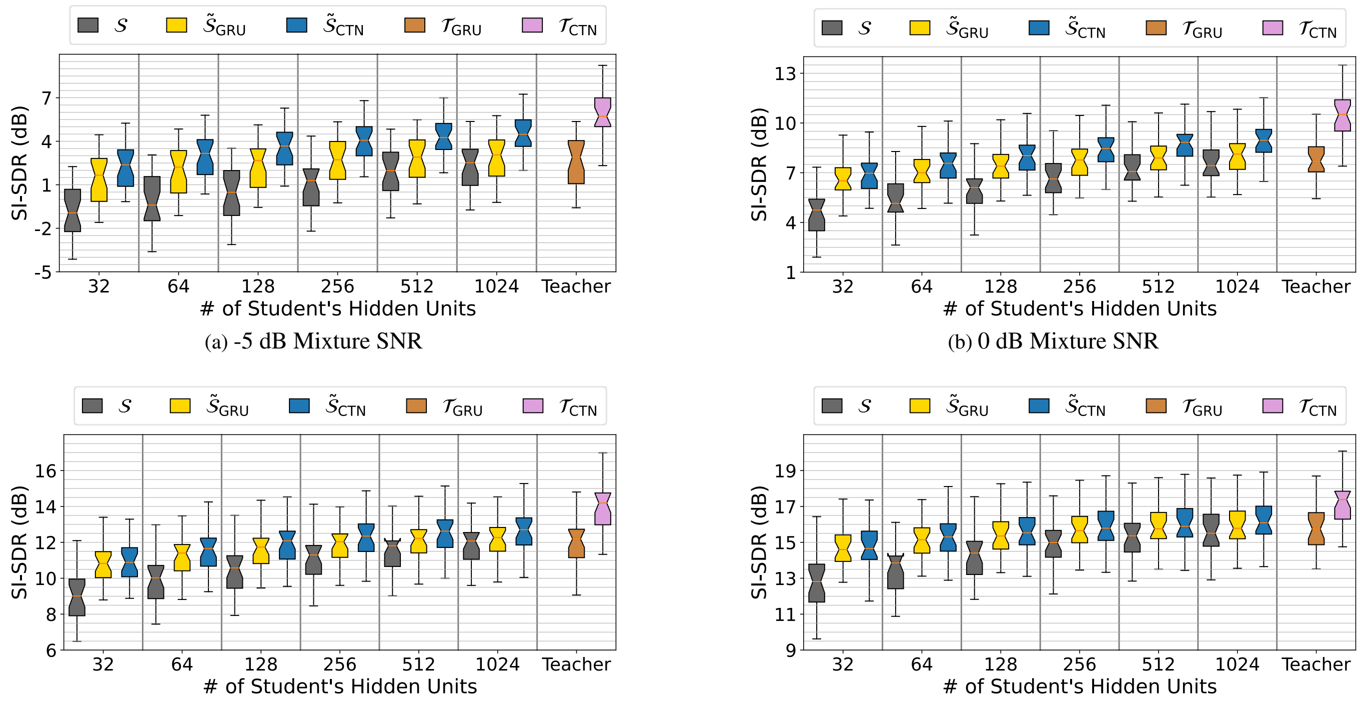
Both teacher models are pre-trained as generalist SE models using noisy mixtures generated from adding $\mathbb{G}_{\text{tr}}$ and $\mathbb{N}_{\text{tr}}$. We select the best models using early-stopping determined from validation computed using $\mathbb{G}_{\text{va}}$ and $\mathbb{N}_{\text{va}}$. During pre-training, the clean speech dataset are used as the ground-truth targets.

The student models are fine-tuned using mixtures of $\mathbb{S}_{\text{ft}}$ and $\mathbb{M}_{\text{ft}}$. Their best models are determined through validation on the set-aside validation set $\mathbb{S}_{\text{va}}$ and $\mathbb{M}_{\text{va}}$. Finally, we test the fine-tuned models on the mixture of $\mathbb{S}_{\text{te}}$ and $\mathbb{M}_{\text{te}}$, which have not been exposed to any of the pre-training and fine-tuning processes. The Adam optimizer [30] was used with learning rate of 1e-4 for pre-training and 1e-5 for fine-tuning.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

The box plots in Figure 2 show the SE performances of various models under environments synthesized from different noise level conditions. The results are shown for pre-trained and fine-tuned student models as well as the teacher models as the reference. Here, we introduce new notations to distinguish the two teacher model architectures: $\mathcal{T}_{\text{GRU}}$ and $\mathcal{T}_{\text{CTN}}$. In addition, we also denote the fine-tuned students models differently from the pre-trained initial model $\mathcal{S}$ and add the subscript to indicate what it learns from: $\tilde{\mathcal{S}}_{\text{GRU}}$ and $\tilde{\mathcal{S}}_{\text{CTN}}$, respectively. Hence, each box that represents one of the generalist models, $\mathcal{S}$, $\mathcal{T}_{\text{GRU}}$, and $\mathcal{T}_{\text{CTN}}$, is an average SI-SDR performance of the system on all 30 unique test environments. On the other hand, a box for one of the specialist architectures, $\tilde{\mathcal{S}}_{\text{GRU}}$ and $\tilde{\mathcal{S}}_{\text{CTN}}$, is an average performance of 30 different personalized models on the 30 test conditions, applied respectively.

Our proposed PSE framework improves all pre-trained student models under all noise conditions, i.e., $\tilde{\mathcal{S}}_{\text{GRU}}$ and $\tilde{\mathcal{S}}_{\text{CTN}}$ results are always better than the $\mathcal{S}$ results on average. In addition, we also observe that the personalized models learned from the CTN teacher,



(a) -5 dB Mixture SNR

(b) 0 dB Mixture SNR

(c) 5 dB Mixture SNR

(d) 10 dB Mixture SNR

Figure 2: Comparison of SE performances from pre-trained generalists against personalized specialists under various mixture SNR levels. Student models are initialized as 2-layered GRU generalists. Teacher models are provided as references.

$\tilde{\mathcal{S}}_{\text{CTN}}$, always outperform their corresponding ones fine-tuned using the GRU teacher, $\tilde{\mathcal{S}}_{\text{GRU}}$. Given that each student pair in comparison are stemmed from the same pre-trained GRU model, it showcases that the quality of the teacher model's performance is related to the performance of fine-tuning. It is also noticeable that the structural discrepancy between the student and teacher, i.e., $\tilde{\mathcal{S}}_{\text{CTN}}$ (a GRU) and $\mathcal{T}_{\text{CTN}}$ (a CTN), is not an issue.

The smaller student models show more significant improvements via PSE. Hence, it verifies that PSE is a model compression method, because a smaller personalized model can compete with a large generalist (e.g. 2×32 $\tilde{\mathcal{S}}_{\text{CTN}}$ vs. 2×1024 $\mathcal{S}_{\text{GRU}}$ for -5 dB mixture SNR as in Figure 2a). According to Table 1, a personalized 2×32 specialist saves 11.47M parameters and 719M MACs compared to a 2×1024 generalist (for 1-second inputs). Likewise, instead of increasing generalists' architectures for better generalization capabilities, it is more advantageous to personalize the models.

When the teacher model is better than the student by only a small margin, personalized student models are even able to outperform the relative teacher model, i.e., $\tilde{\mathcal{S}}_{\text{GRU}}$ (2×1024) vs. $\mathcal{T}_{\text{GRU}}$ (3×1024). We believe it is because of the student model's dedicated exposure to the test-time environment during finetuning.

We envision a scenario where the fine-tuning procedure can be done on the cloud, where the residing teacher model updates the small student model. To this end, the small student model needs to be transferred from the cloud server to the user device, which may not be burdensome given its small size. The cloud computing option is also convenient, as the finetuning step do not need to wait for the teacher model to denoise the test signals, which is an energy- and time-consuming process to be conducted in the small device. Likewise, frequent updates to the student does not become burdensome for the device. Since our framework is simple, we expect our framework to provide improvements under different data or loss functions, and even be applicable to other domains.

5. CONCLUSION

In this paper, we proposed a simple zero-shot learning framework that utilizes knowledge distillation to fine-tune a speech enhancement model during test-time, which we call personalization. By utilizing the teacher's estimates as the targets, which otherwise do not exist during the test time, we showed that the student model's performance greatly improves on a specific test-time speaker and the acoustic environment. Since our small personalized student model can give superior performances to large generalist models, we claim that the knowledge distillation-based fine-tuning method provides another mode of model compression that does not sacrifice performance. Our framework is flexible as it can employ heterogeneous model architectures within a teacher-student pair. Our zero-shot personalization procedure does not require any ground-truth clean speech signals from the test-time user, making it more mindful about users' privacy. Finally, we envision that PSE can be a solution to improving the model's performance on the user groups that are underrepresented in the training set. The source codes and sound examples are available at: `https://saige.sice.indiana.edu/research-projects/KD-PSE`.

6. REFERENCES

[1] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on*

Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[3] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[4] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[5] A. Sivaraman and M. Kim, "Self-Supervised Learning for Personalized Speech Enhancement," *arXiv preprint arXiv:2011.03426*, 2021.

[6] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.

[7] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[8] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *Proceedings of the AAAI National Conference on Artificial Intelligence (AAAI)*, vol. 1, no. 2, 2008, p. 3.

[9] M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[10] Y. N. Dauphin, G. Tur, D. Hakkani-Tür, and L. Heck, "Zero-shot learning for semantic utterance classification," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[11] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot learning for audio-based music classification and tagging," *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[12] A. Sivaraman and M. Kim, "Sparse Mixture of Local Experts for Efficient Speech Enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 4526–4530.

[13] ——, "Zero-shot personalized speech enhancement through speaker-informed model selection," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[15] H. Chen, *et al.*, "Data-free learning of student networks," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019, pp. 3514–3522.

[16] R. G. Lopes, S. Fenu, and T. Starner, "Data-free knowledge distillation for deep neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[17] G. K. Nayak, K. R. Mopuri, V. Shaj, R. V. Babu, and A. Chakraborty, "Zero-shot knowledge distillation in deep networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 4743–4751.

[18] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2019, pp. 695–699.

[19] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2019, pp. 81–85.

[20] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," *arXiv preprint arXiv:1708.05466*, 2017.

[21] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

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

[23] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[24] G. Wichern, *et al.*, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 1368–1372.

[25] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[26] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013, pp. 7092–7096.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[28] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[29] M. Pariente, *et al.*, "Asteroid: The PyTorch-Based Audio Source Separation Toolkit for Researchers," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 2637–2641.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.