The Journal of the Acoustical Society of America

FEBRUARY 15 2024

Zero-shot test-time adaptation via knowledge distillation for personalized speech denoising and dereverberation

Sunwoo Kim; Mrudula Athi; Guangji Shi; Minje Kim ; Trausti Kristjansson

J. Acoust. Soc. Am. 155, 1353–1367 (2024)
https://doi.org/10.1121/10.0024621

ASA ACOUSTICAL SOCIETY OF AMERICA

JASA ARTICLE

Zero-shot test-time adaptation via knowledge distillation for personalized speech denoising and dereverberation

Sunwoo Kim,[1] Mrudula Athi,[1] Guangji Shi,[1] Minje Kim,[1,2,a)] and Trausti Kristjansson[1]
[1]*Amazon Lab126, Sunnyvale, California 94089, USA*
[2]*Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

ABSTRACT:
A personalization framework to adapt compact models to test time environments and improve their speech enhancement (SE) performance in noisy and reverberant conditions is proposed. The use-cases are when the end-user device encounters only one or a few speakers and noise types that tend to reoccur in the specific acoustic environment. Hence, a small personalized model that is sufficient to handle this focused subset of the original universal SE problem is postulated. The study addresses a major data shortage issue: although the goal is to learn from a specific user's speech signals and the test time environment, the target clean speech is unavailable for model training due to privacy-related concerns and technical difficulty of recording noise and reverberation-free voice signals. The proposed zero-shot personalization method uses no clean speech target. Instead, it employs the knowledge distillation framework, where the more advanced denoising results from an overly large teacher work as pseudo targets to train a small student model. Evaluation on various test time conditions suggests that the proposed personalization approach can significantly enhance the compact student model's test time performance. Personalized models outperform larger non-personalized baseline models, demonstrating that personalization achieves model compression with no loss in dereverberation and denoising performance.
https://doi.org/10.1121/10.0024621

(Received 27 June 2023; revised 9 January 2024; accepted 11 January 2024; published online 15 February 2024)
[Editor: B. Yegnanarayana] Pages: 1353–1367

I. INTRODUCTION

Real-world speech signals are often corrupted by a varying level of interfering noise and reverberation, which can be detrimental to the performance of audio applications. Hence, speech enhancement (SE) algorithms are an essential component incorporated into the audio applications such as automatic speech recognition, diarization, voice-over-IP, and transcription (Boll, 1979; Ephraim and Malah, 1984; Gannot *et al.*, 1998). Among these potential applications, in this paper, we focus on the direct use of the enhanced speech for voice communication, i.e., improving the perceptual speech quality of the end user is our goal.

Recent advancements with deep neural networks (DNNs) for SE have shown superior performance compared to traditional machine learning and signal processing methods (Chazan *et al.*, 2017; Wang and Chen, 2018; Xu *et al.*, 2014). However, these state-of-the-art applications require significant memory and computational bandwidth, rendering them difficult for deployment onto devices for practical uses. Resource-constrained devices, such as hearing aids or wearable devices, cannot efficiently handle real-time inference tasks for the SE applications when the models are with multilayered complex architectures.

Consequently, research in model compression methods have gained interest to address the practicality of deep learning architectures for real-time applications. Common compression methods include quantization to reduce the bit-width (Bhalgat *et al.*, 2020; Jacob *et al.*, 2018), pruning to induce sparsity in model weights (Frankle and Carbin, 2018; Han *et al.*, 2016), and knowledge distillation (KD) to distill larger teacher models onto smaller student models (Hinton *et al.*, 2015). These methods, such as quantization, pruning, and KD, have shown great promise in reducing the model size and complexity while minimizing the drop in generalization performance. However, this kind of compression method can be seen as *context-agnostic* as they do not use the specificity of the test time context. Instead, they tend to seek a general-purpose compression technique that works reasonably well in various real-world test conditions. As a result, a certain level of performance drop is inevitable after compression.

In this paper, we aim at developing a *context-aware* DNN compression method for SE. We envision that a compressed model can reduce its run-time complexity without losing its performance if it focuses on a particular test environment. We contrast the proposed concept and the ordinary DNN-based SE models, which are typically designed as general-purpose frameworks with a large architecture. A DNN's large capacity is fully employed when it is trained on a large training set that consists of various speakers and noise sources, generalizing well to unseen test time conditions, e.g., different speakers, noise sources, signal-to-noise ratios (SNRs), and room acoustics. In some practical use-cases,

a)Email: minje@illinois.edu

though, it suffices for the enhancement model to perform well only for the specific test time context. For instance, a family-owned smart assistant device sitting in the living room needs to perform well only for the family members' voices and their home acoustics but not necessarily for the other situations. Compared to the general-purpose SE model, *the generalist*, our context-aware compression method can allow a model to adapt to the specific speakers and their acoustic context, overcoming the generalization losses. We call this type of context-aware SE model a *personalized* speech enhancement (PSE) system. Because the test time context contains limited variability, a small personalized model can even outperform larger and more complex universal generalist models, demonstrating personalization as a form of model compression.

The proposed PSE models achieve context-awareness by reducing the domain mismatch between the training and test datasets. The topic of domain adaptation has been an active area of research in machine learning. One common procedure for domain transfer is regularizing the differences between the learned representations of source and target datasets. It has been applied for emotion, speech, and speaker recognition (Deng *et al.*, 2014; Sun *et al.*, 2017). However, these applications rely on ample target data, which cannot be assumed if the target problem is narrowly defined as in our PSE cases. Few-shot adaptation can be a solution as it requires only a small amount of ground-truth signal (Sivaraman and Kim, 2022). However, it can be challenging to obtain ground-truth user information due to recent privacy infringement, data leakage issues. For example, the advancement in DeepFake technology increased people's concern toward releasing personal information (Rochford, 2018). Meanwhile, with user compliance, the user enrollment phase can obtain trigger phrases from the users, but there is no guarantee that these recordings are clean enough to be used as the target of PSE.

In contrast to aforementioned approaches, zero-shot learning is a solution suitable for training tasks when no additional labeled data are available (Wang *et al.*, 2019; Xian *et al.*, 2018). In the context of personalization, a zero-shot approach means that it 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. Zero-shot learning is an active research topic for classification tasks, where test time prediction is performed by computing auxiliary information (Larochelle *et al.*, 2008; Palatucci *et al.*, 2009; Romera-Paredes and Torr, 2015). Similarly, zero-shot learning in the speech and audio classification applications extracts semantic properties or articulatory distribution to obtain labels during test time (Choi *et al.*, 2019; Dauphin *et al.*, 2013; Li *et al.*, 2020).

However, zero-shot learning for SE has not been widely studied yet. In Sivaraman and Kim (2020, 2021), a mixture of local expert model is introduced as a zero-shot solution to test time adaptation of a SE model. It achieves the adaptation goal by employing an internal classifier to select the most suitable one out of predefined specialist models for a given noisy test signal, selecting a predefined specialist model for a given noisy test signal. Although it is a valid adaptation method, it only works on a few predefined contexts, i.e., varying input SNR levels, the gender of the speakers, and predefined speaker groups, rather than actively learning from the test time speaker's personality or the unique context. In Sivaraman *et al.* (2021), self-supervised learning methods are proposed to achieve PSE, where a data purification algorithm identifies clean speech frames from test time noisy speech. Although it achieves the PSE goals, it is not fully using the test time observations, which can be rare. Moreover, the data purification method does not apply to the dereverberation problem. Other works introduce a zero-shot solution for deep clustering-based speech separation models to estimate absent ground-truth labels. Provided a multichannel input, the affinity matrices were estimated through unsupervised spatial clustering or by using phase difference features (Drude *et al.*, 2019; Tzinis *et al.*, 2019). However, as a result of the sheer size and inference costs, deep clustering models are difficult to fit on small devices. In addition, these models are typically for speech separation problems rather than SE.

In this paper, we present a zero-shot learning approach to personalization for joint dereverberation and denoising based on the KD framework (Hinton *et al.*, 2015). As a zero-shot learning method, it does not ask for clean ground-truth signals from the user while it still aims at enhancing noisy reverberant mixtures. Because its goal is to train a small specialist model for a particular user's speech and recording environment, it qualifies as a personalization method. For zero-shot learning approaches implemented via student-teacher strategies, it is common to use data synthesis techniques through generative adversarial frameworks, where the generator produces fake samples (Chen *et al.*, 2019; Micaelli and Storkey, 2019; Ye *et al.*, 2020). Pseudo samples can be also synthesized using activation or output statistics of trained teacher models (Lopes *et al.*, 2017; Nayak *et al.*, 2019) among many others (Chawla *et al.*, 2021; Nayak *et al.*, 2021). Instead of including an intermediate data synthesis step, our proposed model directly uses the teacher model's outputs as if they were ground-truth targets. Previous works have successfully applied KD to develop compact SE systems (Hao *et al.*, 2020; Nakaoka *et al.*, 2021; Subramanian *et al.*, 2018; Tu *et al.*, 2019). Under the KD framework, the basic assumption is that the teacher model's large computational capacity guarantees the generalization goal. We extend this concept to a novel zero-shot learning approach for PSE. As the teacher model works well in most test time environments, we consider its excellent SE results as if they were the target clean speech from the student model's perspective. That way, we can turn any noisy and reverberant test signals into labeled training examples by passing them through the teacher model. In this process, the teacher model remains as a generalist model, whereas the student model can use the teacher's generalization power to learn from the test time input signals, fulfilling the zero-shot learning condition.

https://doi.org/10.1121/10.0024621

Using a KD learning paradigm enables us to leverage noisy unlabeled data and obtain their corresponding soft targets generated by the teacher model. In this paper, we focus on domain adaptation when we have large unlabeled target-domain dataset and assume noisy speech data of a target test time user to be more widely available as opposed to their clean labeled data. Under this assumption, we approach the PSE problem with a self-supervised method in which corresponding pseudo targets are generated from large amounts of unpaired noisy speech data (Doersch *et al.*, 2015; Manohar *et al.*, 2018; Watanabe *et al.*, 2017; Zhang *et al.*, 2020). Our experiments show that the small student models can be personalized in this way, resulting in improved performance compared to their context-agnostic counterparts. Moreover, given that these models are still small, performance improvement reconstitutes the whole KD process as a model compression method. For example, our experiments consistently show that the PSE models can compete with their larger generalist counterparts. We envision that the compact student models can work as an affordable solution in edge devices with limited computing resources.

Figure 1 provides an overview of the proposed KD-based PSE process. On the left, as a pretraining step, the teacher and student models are trained from a generic dataset to cover all test time variations. However, the student model's constrained capacity tends to limit its SE performance. At the center, KD-based fine-tuning learns from the target test environment: 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. The zero-shot framework enables test time adaptation.

When deploying our framework, we ultimately use only the student model on the device for the PSE inference (the rightmost part of Fig. 1). Even after being deployed, this student model can continue to be refined: the device collects more contaminated speech signals from the test scene, which are then fed to the teacher model to produce the corresponding pseudo clean speech target. The pairs of newly collected input and pseudo target signals are used to fine-tune the student model. We could carefully organize the KD fine-tuning process by keeping the teacher model on the device also and performing the KD process during the device's idle time; it is more secure because the user data stays in the device. However, the KD-based training process can be burdensome unless it is scheduled carefully. It is also possible that the teacher model and a copy of the student model are placed externally on a cloud server, where the actual fine-tuning operations are conducted. Then, the student models can be frequently updated on the server side and transferred to the user device. This cloud computing option may be more efficient, although it may be inappropriate for privacy-sensitive applications.

This paper extends our preliminary study (Kim and Kim, 2021), in which we proposed a personalization procedure for speech denoising. In addition to the previous denoising application, we extend our application to dereverberation by integrating variability in room acoustics. Reverberation introduces additional challenges as speech intelligibility is degraded when corrupted by severe reverberations and even more when combined with background noise (Han *et al.*, 2015). During test time, we assume the test time source location and room geometry are unknown and the locations of speaker and acoustic environment can change. For evaluation, we evaluate against real rooms from various settings available in public datasets. To our best knowledge, this end-to-end zero-shot personalization framework for model compression is novel in the topic of joint speech dereverberation and denoising. Our proposed framework not only demonstrates the effectiveness of personalization for the front-end denoising and dereverberation application but also illustrates the potential for using teacher's outputs as pseudo targets in a zero-shot scenario. In our experiments, we show the relationship between the amount of noisy reverberant speech samples and performances of our personalized models and draw connections to real-life scenarios where ample test time data may not be readily available. Finally, we illustrate use-cases where the teacher's estimates can be used to gauge test time performances and detect catastrophic forgetting (French, 1999) that occur from fine-tuning on specific instances (e.g., different noise sources or room conditions) and offer a simple remedy for this issue.

The rest of the paper is organized as follows. In Sec. II, we describe the student-teacher framework for test time adaptation. Experimental setups are provided in Sec. III, including the descriptions about various individual room acoustics. In Sec. IV, we provide extensive evaluation on

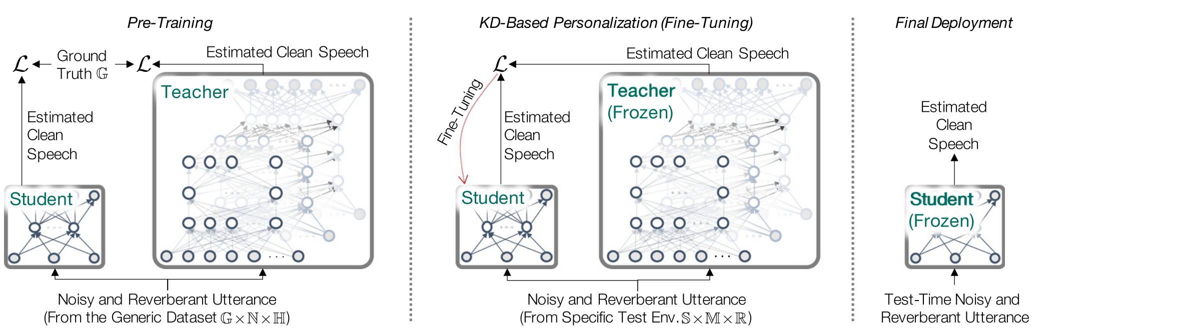


FIG. 1. (Color online) An overview of the proposed KD-based PSE process shows (left) the pretraining process for teacher and student models using generic dataset, (center) the KD-based personalization process, and (right) the student model's inference process after the personalization.

https://doi.org/10.1121/10.0024621

the effects of personalization to various unseen environments. Concluding remarks are presented in Sec. V.

II. THE PROPOSED KD-BASED ZERO-SHOT PSE ALGORITHM

Given a monaural signal recorded in a noisy and reverberant environment, we formulate the signal model as

$$\boldsymbol{y}[t] = \boldsymbol{x}[t] + \alpha\boldsymbol{n}[t] = \boldsymbol{s}[t] * \boldsymbol{h}[t] + \alpha\boldsymbol{n}[t], \quad (1)$$

where $\boldsymbol{s}$, $\boldsymbol{n}$, $\boldsymbol{h}$, and $\boldsymbol{x}$ denote speech source, background noise, room impulse response (RIR) function, and reverberant speech, respectively. The symbol "*" stands for the convolution operator. The parameter, α, controls the SNR between the reverberant speech and interfering noise source. Our goal in this study is to recover the clean anechoic version of the single-talker speech signal, $\boldsymbol{s}$, from the corresponding noisy reverberant observation, $\boldsymbol{y}$.

We propose a KD-based zero-shot PSE algorithm, which aims at joint denoising and dereverberation. Our goal is to fine-tune a compact student model after it is deployed such that it adapts to the unseen test speaker and environment continuously. In doing so, the teacher model's powerful generalization performance plays a significant role as it performs denoising and dereverberation simultaneously, a behavior that the student model attempts to learn from.

A. Training teacher SE models

First, we train the teacher model, $\mathcal{T}(\cdot)$, using a large-scale dataset consisting of dry speech sources, various noise signals, and RIR filters. Here, $\mathcal{T}(\cdot)$ is defined with a large model architecture, therefore, it can properly approximate the complex general-purpose joint speech denoising and dereverberation function. Once trained, $\mathcal{T}(\cdot)$ is frozen and *not* fine-tuned, assuming that its SE performance as a generalist meets the quality standard in most test cases. Another assumption is that it is too complex for the given test time user device to perform real-time SE inference tasks.

To train the teacher models, we use generic training datasets. The formulation of the training dataset is as follows. The clean speech utterances are taken from a large corpus containing many speakers, $\boldsymbol{s} \in \mathbb{G}$; the noise recordings are also from a large corpus containing various noise types, $\boldsymbol{n} \in \mathbb{N}$; the RIRs are similarly from a large collection recorded in various rooms, $\boldsymbol{h} \in \mathbb{H}$. We use them to synthesize the noisy and reverberant signals, $\boldsymbol{y}$, as input [Eq. (1)].

Hence, the goal of the teacher model is to jointly denoise and dereverberate $\boldsymbol{y}$ so the model can estimate the waveforms, $\hat{\boldsymbol{s}}$, that closely approximate the target clean anechoic speech, i.e., $\boldsymbol{s} \approx \hat{\boldsymbol{s}} \leftarrow \mathcal{T}(\boldsymbol{y})$. The optimization on $\mathcal{T}(\cdot)$ reduces the loss between the target utterance, $\boldsymbol{s}$, and reconstruction, $\hat{\boldsymbol{s}}$, i.e., $\mathrm{argmin}_{\boldsymbol{\Theta}_\mathcal{T}} \mathcal{L}(\boldsymbol{s}||\mathcal{T}(\boldsymbol{y};\boldsymbol{\Theta}_\mathcal{T}))$, where $\boldsymbol{\Theta}_\mathcal{T}$ denotes the trainable parameters of the teacer model. Note that the training process for the teacher model corresponds to the typical supervised learning method for general-purpose SE. Detailed model and optimization descriptions are provided in Sec. III B.

B. Pretraining student SE models

Our student models, $\mathcal{S}(\cdot)$, are pretrained in a similar way to the teacher models, i.e., by updating its own model parameters, $\mathrm{argmin}_{\boldsymbol{\Theta}_\mathcal{S}} \mathcal{L}(\boldsymbol{s}||\mathcal{S}(\boldsymbol{y};\boldsymbol{\Theta}_\mathcal{S}))$, using the same generic datasets, $\mathbb{G}$, $\mathbb{N}$, and $\mathbb{H}$. However, its small capacity hinders it from generalizing well to the unseen test conditions. Thus, we argue that further improvement is required for these student models to meet the quality requirement. We introduce the KD-based test time personalization algorithm in Sec. II C, which is designed to reduce the performance gap between $\mathcal{T}(\cdot)$ and $\mathcal{S}(\cdot)$. In this regard, the purpose of pretraining $\mathcal{S}(\cdot)$ is to prepare the student model better than a random initialization, primed for the next fine-tuning step. Further details on model and training are also given in Sec. III B.

C. Test time PSE

During the test time, we assume that the enhancement system is exposed to mixture signals composed of clean speech utterances from the test speaker, $\boldsymbol{s} \in \mathbb{S}$, background noise sources, $\boldsymbol{n} \in \mathbb{M}$, and RIRs, $\boldsymbol{h} \in \mathbb{R}$. Note that we differentiate these test sets from the training sets, i.e., $\mathbb{G} \neq \mathbb{S}$, $\mathbb{N} \neq \mathbb{M}$, and $\mathbb{H} \neq \mathbb{R}$. Meanwhile, we also assume that the noisy and reverberant speech signals defined by the combination of all speech, noise, and RIRs available in the training sets $\mathbb{G} \times \mathbb{N} \times \mathbb{H}$ mixed through Eq. (1) are representative enough to encompass the test time variations, i.e., $\mathbb{S} \times \mathbb{M} \times \mathbb{R} \subseteq \mathbb{G} \times \mathbb{N} \times \mathbb{H}$. In practice, however, there might be corner cases that even the large dataset $\mathbb{G} \times \mathbb{N} \times \mathbb{H}$ cannot successfully represent, which the proposed method could fail to adapt to. Hence, we postulate that if a teacher model is large enough, it can serve as an unbiased solution to the denoising and dereverberation problem. Meanwhile, a small student model is also in our interest if it is small enough for the resource-constrained edge device. However, it may be too biased to generalize well to the test time SE task because of its small model capacity.

Given these assumptions, we propose a personalization 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. As we formulate the proposed personalization method as a fine-tuning process, we begin with a compact student model, $\mathcal{S}(\cdot)$, pretrained in a context-agnostic manner as in Sec. II B. To fine-tune the student model, its enhancement result from dereverberation and denoising, $\hat{\boldsymbol{s}}_\mathcal{S}$, must be compared against the target to compute the loss and perform backpropagation. However, because we assume that the target is not available, we use the pseudo target computed from the teacher model.

This process falls in the category of the student-teacher framework in which a student model is optimized using a teacher model's prediction (Hinton *et al.*, 2015). In the context of PSE, we employ a large pretrained teacher model, $\mathcal{T}(\cdot)$,

https://doi.org/10.1121/10.0024621

whose predicted clean utterance serves as the target to compute the student model's loss. Student and teacher models are initialized with pretrained generic enhancement models as discussed in Secs. II A and II B, respectively. During the test time, the student model is optimized as $\mathrm{argmin}_{\boldsymbol{\Theta}_{\mathcal{S}}}\mathcal{L}(\hat{\boldsymbol{s}}_{\mathcal{T}}||\tilde{\mathcal{S}}(\boldsymbol{y};\boldsymbol{\Theta}_{\tilde{\mathcal{S}}}))$, where $\hat{\boldsymbol{s}}_{\mathcal{T}}$ are the estimates of clean speech signals obtained from the teacher model, and $\boldsymbol{\Theta}_{\tilde{\mathcal{S}}}$ are trainable parameters of the student model. We distinguish this fine-tuned student model, $\tilde{\mathcal{S}}(\cdot)$, from the pretrained model, $\mathcal{S}(\cdot)$, from now on.

The teacher's estimate, $\hat{\boldsymbol{s}}_{\mathcal{T}}$, is only an approximation of the ground-truth target, $\boldsymbol{s}$, and can contain artifacts from dereverberation and denoising (Xu *et al.*, 2014). However, under a zero-shot PSE setup, we assume that 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., $\mathcal{L}(\boldsymbol{s}||\hat{\boldsymbol{s}}_{\mathcal{T}})\mathcal{L}(\boldsymbol{s}||\hat{\boldsymbol{s}}_{\mathcal{S}})$. Given its large capacity, the teacher model generalizes better to unseen inputs compared to the student model. Thus, we hypothesize that the student still learns from these imperfect targets and adapts to the test environment. In our experiments on simulated signals and real-world test environments, we show that this assumed performance gap exists, guaranteeing the performance improvement by the KD process.

D. Interpretation from a manifold assumption

By training under the SE criterion, models learn to produce latent representations that are robust to corruptions in input data and useful for recovering the clean speech. Successfully learned latent representations are discriminative and can capture useful structure and variations in the input distribution as discussed in the context of denoising autoencoders (Vincent *et al.*, 2010). We interpret the process of SE using the manifold assumption: high dimensional clean speech data lie on a low dimensional manifold (Chapelle *et al.*, 2006). Data samples are mapped onto a manifold that represents a feature space that preserves the local structure of the data. The objective of our models is to learn the underlying manifold of the speech signals such that they can accurately map the noisy samples to their respective positions on the manifold of clean speech during test time. In Fig. 2, we see generic clean speech samples, $\boldsymbol{s}$ (the crosses), form a complex manifold (the thin solid line). Meanwhile, under our PSE assumption that test time environments will contain smaller subset of sources (e.g., speakers, noises, and room variation), the models would only need to learn the manifold of those subsets, which imply a simpler manifold (the thick solid line) than the generic speech's. Under this interpretation, the target speaker's corrupt examples are mapped away from the manifold of clean signals. SE models try to project the off-the-manifold examples, $\boldsymbol{y}$, back onto the manifold. The farther away $\boldsymbol{y}$ is from the manifold, the more corrupt the example is, and the model takes bigger efforts to reach the manifold. Note that the corrupted samples, $\boldsymbol{y}$, are spoken by the same target

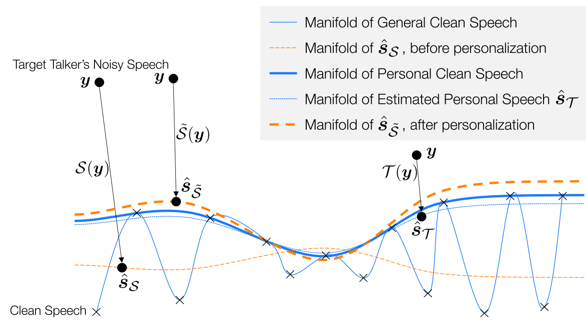


FIG. 2. (Color online) Manifold learning perspective of SE, considering small model sizes and the potential subsampling of the dataset to construct personalized dataset and model.

person and, therefore, the target manifold is the simpler manifold (thick line) than the complex manifold (the thin solid line) by all people.

We expect a large complex model, $\mathcal{T}(\boldsymbol{y})$, to better approximate the manifold given its larger architecture. Hence, its prediction of the personal clean speech forms an approximation (thin dotted line) similar to the original approximation (the thick solid line). On the contrary, smaller models are likely to learn a poor approximation, $\hat{\boldsymbol{s}}_{\mathcal{S}}$ (the thin dash). The aim of our proposed personalization framework is to distill the better manifold determined by $\hat{\boldsymbol{s}}_{\mathcal{T}}$ to the smaller student models to help better approximate the manifold. By doing so, student models fine-tuned under our framework will be able to better define and map points, $\boldsymbol{y}$, closer to the test time manifold, which is approximated by $\hat{\boldsymbol{s}}_{\tilde{\mathcal{S}}}$ (the thick dash).

III. EXPERIMENTAL SETUP

A. Datasets

Table I summarizes the datasets that were used for the experiments. For pretraining, we employed clean speech recordings from the LibriSpeech corpus (Panayotov *et al.*, 2015) and noise recordings from the MUSAN (Snyder *et al.*, 2015) and ESC50 datasets (Piczak, 2015). For RIRs, we used publicly available recordings downloaded using Kaldi scripts.[1] The RIR data sources consist of the Aachen Impulse Response Database (Jeub *et al.*, 2009), PORI

TABLE I. Corpus and notation of speech, noise, and RIR datasets used during pretraining and personalization.

	Corpus		Notation
Speech	Librispeech (Panayotov *et al.*, 2015)	`train-clean-360`	$\mathbb{G}$
		`test-clean`	$\mathbb{S}_{\mathrm{ft/va/te}}$
Noise	MUSAN (Snyder *et al.*, 2015)		$\mathbb{N}$
	ESC50 (Piczak, 2015)		$\mathbb{M}_{\mathrm{ft/va/te}}$
RIR	AIR (Jeub *et al.*, 2009)		$\mathbb{H}$
	PORI (Merimaa *et al.*, 2005)		
	RWCP (Nakamura *et al.*, 2000)		
	BUT (Szöke *et al.*, 2019)		$\mathbb{R}_{\mathrm{ft/va/te}}$
	REVERB (Kinoshita *et al.*, 2013)		

concert hall impulse responses (Merimaa *et al.*, 2005), and RWCP Sound Scene Database in Real Acoustical Environments (Nakamura *et al.*, 2000). We used Librispeech's `train-clean-360`, MUSAN's `free-sound`, and the collective RIR data for training, which we denote as $\mathbb{G}$, $\mathbb{N}$, and $\mathbb{H}$, respectively. A comprehensive summary of RIR datasets, including information on RT60, number of rooms, and microphone to loudspeaker distance, can be found in Merimaa *et al.* (2005) and Szöke *et al.* (2019). This exposes the generalist models to up to 251 speakers, 843 noise recordings, and 334 RIRs 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 or zero-shot PSE, we used 44 speakers from Librispeech's `test-clean` and noise from the ESC-50 dataset for environmental sound classification with 50 different noise types from 5 categories, consisting of animals, natural and water soundscape, nonspeech human sounds, interior-domestic sounds, and exterior-urban sounds. For RIRs, we used 5 rooms from BUT Speech@FIT Reverb Database (BUT; Szöke *et al.*, 2019) and real rooms from the Reverb 2014 Challenge (RVB) dataset (Kinoshita *et al.*, 2013) for a total of 11 rooms. Each room in the BUT dataset contains 31 microphones and 5 source positions on average. RIRs were measured for each speaker position using exponential sine sweep method. For RVB dataset, there are three types of rooms (small, medium, and large) and two types of microphone placement (near and far). RIRs are collected from two microphone angles per room. Further information on RIR datasets, along with speech and noise corpuses, can be found in Table I (Szöke *et al.*, 2019).

We synthesize $K = 44$ unique test time environments, each of which consists of a test speaker, a noise source, and a RIR configuration defined by the location of the speaker and microphone. In particular, given a test environment index, $k \in \{1, \ldots, K\}$, we sample clean utterances from the kth speaker, $\mathbb{S}^{(k)}$, convolve it with the kth room's RIR, $\mathbb{R}^{(k)}$, and add noises from the kth noise type, $\mathbb{M}^{(k)}$. For each test environment, $\mathbb{S}^{(k)}$ are split into separate sets for fine-tuning, validation, and testing: the partitions are approximately 5, 1, and 1 min of clean speech, which we denote by $\mathbb{S}^{(k)}_{\text{ft}}$, $\mathbb{S}^{(k)}_{\text{va}}$ and $\mathbb{S}^{(k)}_{\text{te}}$, respectively. The noise and RIR samples are prepared similarly and partitioned into three separate sets. We synthesize noisy and reverberant input signals by combining $\mathbb{S}^{(k)}_{\text{ft}}$, $\mathbb{M}^{(k)}_{\text{ft}}$, and $\mathbb{R}^{(k)}_{\text{ft}}$. Having them as input, the student model is fine-tuned via the KD process, where the teacher model's denoising results are used as the pseudo target. In other words, the student model for generic SE is first deployed to the device and can be personalized to the user's specificity using 5 min of noisy and reverberant recordings of the test environment. Then, $\mathbb{S}^{(k)}_{\text{va}}$, $\mathbb{M}^{(k)}_{\text{va}}$, and $\mathbb{R}^{(k)}_{\text{va}}$ are used to validate the student model during fine-tuning, mainly to prevent overfitting. Note that this does not mean that the PSE algorithm needs clean speech for validation: the validation process still relies on the teacher's estimate of clean speech as the target to compute the validation loss. Hence, early stopping is still conducted in a zero-shot manner. We report our PSE models' final performance using the test sets $\mathbb{S}^{(k)}_{\text{te}}$, $\mathbb{M}^{(k)}_{\text{te}}$, and $\mathbb{R}^{(k)}_{\text{te}}$, for which we do compute the final enhancement performance by comparing to the ground-truth clean speech signals, $\mathbb{S}^{(k)}_{\text{te}}$.

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). All speech and noise audio files are loaded at 16 kHz sampling rate and standardized to have unit-variance.

B. Models

Our student models are based on the unidirectional gated recurrent unit (GRU) architecture (Cho *et al.*, 2014). Recurrent neural networks with gating technology, such as the long short-term memory (LSTM) cell (Hochreiter and Schmidhuber, 1997) and GRUs, have been predominantly used in SE due to their ability to overcome the gradient vanishing or explosion issues during backpropagation through time (BPTT). As for GRUs, although it was first introduced as a computationally efficient alternative of LSTM for machine translation, it was quickly adopted for SE tasks as a result of their flexibility in handling continuous input sequences (e.g., audio spectra) and learning continuous latent variables (Chazan *et al.*, 2017; Luo *et al.*, 2020; Sivaraman and Kim, 2020). In the SE literature, LSTM and GRU can be combined with Convolutional Neural Network (CNN) layers for better performance (Hu *et al.*, 2020), but the hybrid architecture adds more burden to the hardware design. In this paper, we focus on the simple GRU-only architecture, which is more suitable for CPU operations, and show the PSE method's merits. However, the proposed principles should apply to other architectural choices.

We use frequency-domain representations obtained through the short-time Fourier transform (STFT) as inputs to the enhancement models. STFT is with a Hann windowed frame of 1024 samples and a hop size of 256 samples. The recurrent unit reads each STFT magnitude spectrum sequentially and updates the hidden state at each frame. For our denoising application, we apply a dense layer to map the hidden unit outputs from the GRU layer into complex ideal ratio masks (Williamson *et al.*, 2016). The denoising mask is applied element-wise to the mixture complex spectrogram, and then transformed back to the time-domain signal, $\hat{s}$, through inverse STFT. We use negative scale-invariant signal-to-noise ratio (SI-SNR) as the loss function (Le Roux *et al.*, 2019). 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 personalization on the different architectural choices.

Meanwhile, as for the teacher model, we employ two different network architectures. First, we use a 3×1024 GRU architecture, which is large enough to outperform the students. In addition, we also employ dual-path Recurrent Neural Networks (RNN) (DPRNN; Luo *et al.*, 2020) as an alternative teacher model. DPRNN was chosen because of

https://doi.org/10.1121/10.0024621

TABLE II. Complexity of student and teacher models in MACs and number of parameters. MACs are computed given 1-s inputs.

Models		MACs (G)	Parameter (M)
Student	GRU (2 × 32)	0.006	0.09
	GRU (2 × 64)	0.013	0.20
	GRU (2 × 128)	0.030	0.48
	GRU (2 × 256)	0.079	1.25
	GRU (2 × 512)	0.232	3.68
	GRU (2 × 1024)	0.762	12.08
Teacher	GRU (3 × 1024)	1.159	18.37
	DPRNN (Luo *et al.*, 2020)	15.238	3.63

its higher performance and smaller model size compared to other time-domain models such as Conv-TasNet (Luo and Mesgarani, 2019). More advanced transformer based models, such as dual-path transformer (DPTNet; Chen *et al.*, 2020), report higher performance with comparative size to DPRNN in speech separation tasks, but we found, empirically, that the DPRNN performs better for our dereverberation and denoising task.

Indeed, the DPRNN teacher outperforms the GRU teacher because of its structural advantage. Hence, we contrast the impact of the two teacher models on the PSE performance after the KD-based fine-tuning process. The DPRNN model is configured using implementation available in Asteroid's source separation toolkit (Pariente *et al.*, 2020). Same architecture as was reported in Luo *et al.* (2020) is adopted (i.e., six repeats), and we trained it with our single-speaker SE setup rather than the original speech separation task. The model architectures, their respective number of parameters, and the multiplier-accumulator (MAC) operation counts are shown in Table II. Note that DPRNN is not the largest model, but it requires extensive MAC operations.

Here, we introduce new notations to distinguish between the two teacher model architectures: $\mathcal{T}_{\text{GRU}}$ and $\mathcal{T}_{\text{DPRNN}}$. In addition, we also denote the fine-tuned student models differently from the pretrained initial model, $\mathcal{S}$, and add the subscript to indicate what it learns from: $\tilde{\mathcal{S}}_{\text{GRU}}$ and $\tilde{\mathcal{S}}_{\text{DPRNN}}$, respectively. We include a student model fine-tuned on the ground-truth oracle clean speech targets as a performance upper bound and denote it as $\tilde{\mathcal{S}}_{\text{GT}}$. Summary of notations for pretrained and fine-tuned models are listed in Table III. Note that the systems denoted with tilde, $\tilde{\mathcal{S}}_{\text{GRU}}$, $\tilde{\mathcal{S}}_{\text{DPRNN}}$, and $\tilde{\mathcal{S}}_{\text{GT}}$, represent personalized student models, whereas the generalist models, $\mathcal{S}$, $\mathcal{T}_{\text{GRU}}$, and $\mathcal{T}_{\text{DPRNN}}$, are discussed to report the performance of either the teacher models or the baseline.

TABLE III. Notations for pretrained and fine-tuned models.

Notation	Description
$\mathcal{T}_{\text{GRU}}$	The frozen GRU teacher trained from generic datasets
$\mathcal{T}_{\text{DPRNN}}$	The frozen DPRNN teacher trained from generic datasets
$\mathcal{S}$	Initial student pretrained from generic datasets
$\tilde{\mathcal{S}}_{\text{GRU}}$	Student, fine-tuned on $\mathcal{T}_{\text{GRU}}$'s test output
$\tilde{\mathcal{S}}_{\text{DPRNN}}$	Student, fine-tuned on $\mathcal{T}_{\text{DPRNN}}$'s test output
$\tilde{\mathcal{S}}_{\text{GT}}$	Student, fine-tuned on the test time ground-truth targets

The Adam optimizer (Kingma and Ba, 2015) was used with learning rate of 1×10^{-4} for pretraining and 1×10^{-5} for fine-tuning.

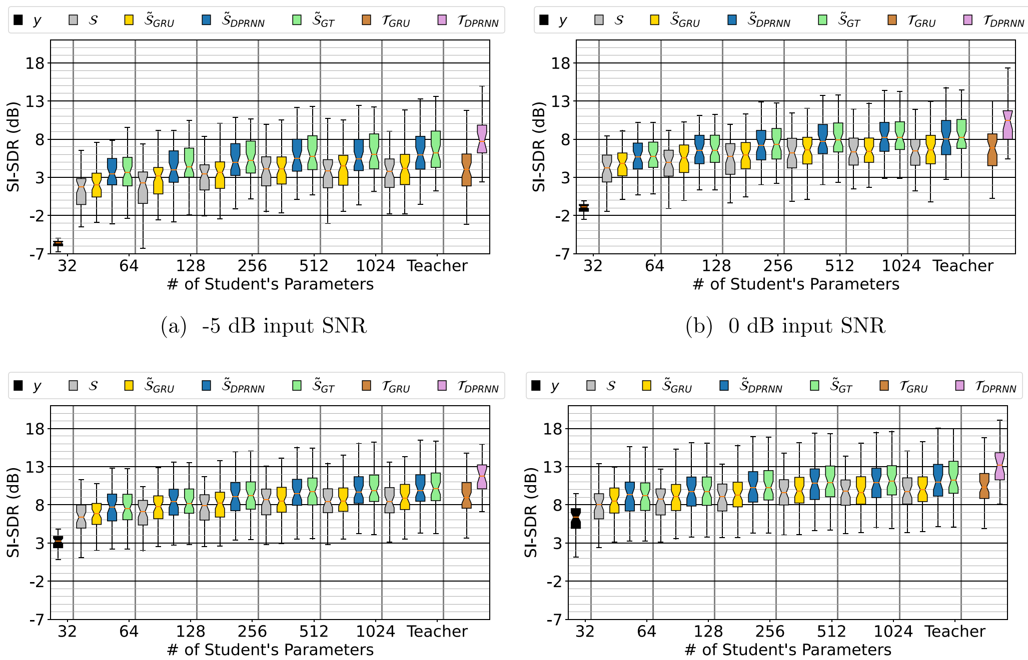
IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Objective performance analyses

The box plots in Fig. 3 show the enhancement performances of various models under $K = 44$ environments synthesized with different noise level conditions. The results are shown for pretrained and fine-tuned student models as well as for teacher models as reference. Figures 3(a)–3(d) show comprehensive denoising and dereverberation results across cases with severe (−5dB) to moderate (+10dB) test time SNR levels. We notice that the overall performance decreases due to the additive background noise. While their final input SNR values are controlled by varying the loudness of the test noise sources, $\mathbb{M}_{\text{te}}^{(k)}$, the speech sources are also degraded by the reverberation defined by the test time RIR set, $\mathbb{R}_{\text{te}}^{(k)}$. In Figs. 3(a)–3(d), we observed that our proposed personalization framework improves dereverberation and denoising performances of pretrained student models under all noise and room conditions, i.e., $\tilde{\mathcal{S}}_{\text{GRU}}$ and $\tilde{\mathcal{S}}_{\text{DPRNN}}$ results are always better than the $\mathcal{S}$ results, on average, if their model complexity is the same. From these results, we can infer that personalization helps significantly improve the joint dereverberation and denoising performances for all student architectures.

In addition, we observe that the personalized models learned from the DPRNN teacher, $\tilde{\mathcal{S}}_{\text{DPRNN}}$, always outperform their corresponding models fine-tuned using the GRU teacher, $\tilde{\mathcal{S}}_{\text{GRU}}$. $\tilde{\mathcal{S}}_{\text{GRU}}$, at times, perform similarly to $\tilde{\mathcal{T}}_{\text{GRU}}$ especially in the case of the 2×1024 and 3×1024 student and teacher models. This is because both models trained under the same complex Ideal Ratio Mask (cIRM) estimation objective, sharing similar GRU architecture, as opposed to the more advanced DPRNN's architecture and its end-to-end SE objective. The results signify the importance of the teacher model's performance. As each fine-tuned student models stem from the same pretrained GRU model, this shows that the fine-tuned performance depends on the quality of the teacher model. It is also noticeable that the structural discrepancy between the student and teacher, i.e., $\tilde{\mathcal{S}}_{\text{GRU}}$ (a GRU) and $\mathcal{T}_{\text{DPRNN}}$ (a DPRNN), is not an issue. It implies that the proposed framework can potentially employ various advanced teacher models as the deep learning research improves the state of the art in the future.

We also observe that $\tilde{\mathcal{S}}_{\text{DPRNN}}$ can catch up to the performance of $\tilde{\mathcal{S}}_{\text{GT}}$, which is only marginally better. This suggests that fine-tuning on imperfect pseudo targets generated by $\tilde{\mathcal{T}}_{\text{DPRNN}}$ has almost the same benefits as when ground-truth targets are used. Given the student models' small architecture and limited generalization capacity, our personalization procedure can fine-tune the student model to its optimal performance but by relying only on the teacher's SE results.



(a) -5 dB input SNR

(b) 0 dB input SNR

(c) 5 dB input SNR

(d) 10 dB input SNR

FIG. 3. (Color online) Comparison of joint dereverberation and denoising performances from pretrained generalists against personalized specialists under various input SNR levels. (a)–(d) demonstrate results from input SNR levels -5, 0, $+5$, and $+10$ dB, respectively. Student models are initialized as two-layered GRU generalists. Teacher models are provided as references.

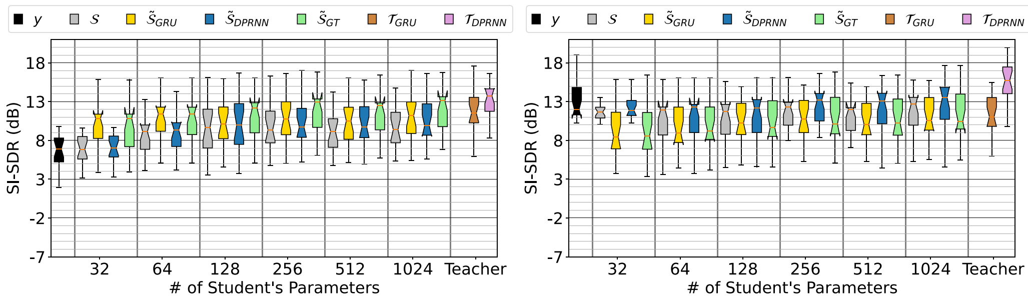
After personalization, the small student models consistently show significant improvements on their pretraining-based initialization. Hence, it verifies that our personalization framework is a model compression method if we compare the improved PSE models to those pretrained generalist models. Indeed, a smaller personalized model can compete with a large generalist, e.g., $2 \times 32\,\tilde{\mathcal{S}}_{\mathrm{DPRNN}}$ vs $2 \times 1024\,\mathcal{S}$ for -5 dB input SNR as in Fig. 3(a). According to Table II, a personalized 2×32 specialist saves 11.99 M parameters and 756 M MACs compared to a 2×1024 generalist (for 1-s inputs), which is more than 99% reduction in terms of spatial and arithmetic complexity. Therefore, even if further compression methods, e.g., eight-bit quantization, are always available to the $2 \times 1024\,\mathcal{S}$ model, it will still most likely be more complex than $2 \times 32\,\tilde{\mathcal{S}}_{\mathrm{DPRNN}}$. Furthermore, applying compression on larger models will subsequently lower their performance depending on the type and amount of compression. On the contrary, the $2 \times 32\,\tilde{\mathcal{S}}_{\mathrm{DPRNN}}$ outperforms the $2 \times 1024\,\mathcal{S}$ even after its 99% reduction of complexity. This demonstrates that our framework works as a mode of *lossless model compression*. Thus, we argue that it is more advantageous to personalize the models instead of increasing generalists' computational capacity for better generalization capabilities. In addition, as PSE shows improved performances in denoising and dereverberation tasks in various unique test environments, our personalization framework can be seen as a genuine *adaptive system* that specializes not only in each individual user but in the specific noise source and reverberant condition of the test time environment.

Figure 4 show the SE performance of various models on the reverberation-only input signals, i.e., with no additive background noise. It gives a separate view to the proposed PSE method's dereverberation performance from the joint denoising and dereverberation setup. In addition, the dereverberation results are shown separately as two cases in which one-half of the environments ($K = 22$) are with low input scale-invariant SDR (SI-SDR) [Fig. 4(a)] and the other with high input SI-SDR [Fig. 4(b)]. For the lower SI-SDR cases, the results in Fig. 4(a) show that our framework can successfully personalize to different room acoustics. Contrary to joint dereverberation and denoising results, the improvements for $\tilde{\mathcal{S}}_{\mathrm{DPRNN}}$ are minimal compared to those for $\tilde{\mathcal{S}}_{\mathrm{GRU}}$. This trend is not observed in Fig. 4(b), which reports results from upper SI-SDR cases. First, the generalist models, $\mathcal{S}$, worsen the sound quality. As the SI-SDR of the reverberant inputs were already high, the pretrained generalists must have injected artifacts that significantly decrease the quality of the signal. This is also evident in the $\tilde{\mathcal{S}}_{\mathrm{GT}}$ baselines, where the model is fine-tuned on ground-truth anechoic targets. Hence, the goal of personalization in these cases is to be able to mitigate this performance degradation of the processed signals. Indeed, $\tilde{\mathcal{S}}_{\mathrm{DPRNN}}$ show improved performances over the initial worsened estimates by the pretrained models, demonstrating the KD framework's consistent capacity to produce pseudo-labels for fine-tuning student models.

B. Subjective performance analyses on real-world test environments

Additionally, we conducted a subjective listening test with ten participants using real-world noisy reverberant

https://doi.org/10.1121/10.0024621



(a) Lower SI-SDR($\boldsymbol{y}, \boldsymbol{s}$) (b) Upper SI-SDR($\boldsymbol{y}, \boldsymbol{s}$)

FIG. 4. (Color online) Dereverberation performances of pretrained generalist and personalized specialists on reverberant inputs without additional background noise (i.e., dereverberation-only experiments). The results are two separate cases, where SI-SDR of input reverberant signals are low for half of the environments ($K = 22$) as shown in (a) and high for the other half as shown in (b).

recordings from the voiceHome-2 dataset (Bertin *et al.*, 2019). We trained ten personalized models from ten test speakers, whose noisy and reverberant samples were recorded in different rooms from different houses with varying speaker position and background noise types according to the setup of the voiceHome-2 dataset. Because the dataset was recorded in the real-world test environment, there is no clean utterance available for supervised learning, making our experiment realistic. The test was performed by asking ten listeners for their perceptual evaluation of the test sequences. Figure 5 presents their mean opinion scores (MOS). Each trial consists of an input noisy and reverberant sample, $\mathbf{y}$, the enhancement result for a small student model, $\mathcal{S}$, the output for the fine-tuned student model, $\tilde{\mathcal{S}}_{\text{DPRNN}}$, and the result for the DPRNN teacher model, $\mathcal{T}_{\text{DPRNN}}$. The student models are with 64 hidden units. From Fig. 5, we can observe that the personalized student models $\tilde{\mathcal{S}}_{\text{DPRNN}}$ outperforms the baseline student models $\tilde{\mathcal{S}}_{\text{DPRNN}}$. The statistically significant improvement aligns with the objective metrics. Whereas this dataset contains only indoor recordings, this provides additional analysis not only regarding the significance of the metrics but also the effectiveness of the PSE framework.

C. Evaluation of personalization using varying amounts of fine-tuning datasets

Our proposed personalization showed great improvements under 5 min of fine-tuning data, which is noisy and

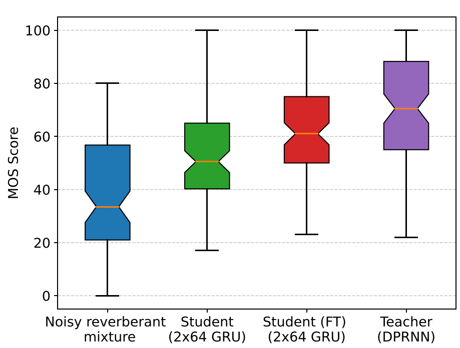


FIG. 5. (Color online) Aggregated mean opinion scores (MOS) from the listening tests. Notches represent the 95% confidence intervals.

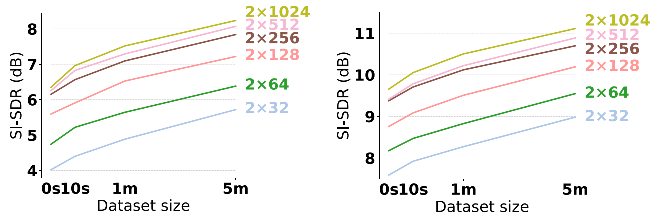
reverberant speech recorded from the same acoustic scene during test time. However, we cannot assume this amount of data to be readily available for realistic scenarios. Hence, we test our framework on varying amounts of noisy reverberant mixtures as well, i.e., 10 s and 1 min. Figure 6 shows the average SI-SDR improvements by using varying lengths of noisy input data across all K environments. We only use the DPRNN teacher's estimates for fine-tuning as we have observed its effectiveness from Sec. IV A. For brevity, we test on 0 and 10 dB input SNR cases.

As expected, the length of available datasets for fine-tuning is proportional to the test time performance of the personalized student models. This experiment illustrates a realistic use-case where, initially, 5 min of noisy data will not be directly available but rather 10 s and, later, 1 min of test time signals will be gradually collected over time in a realistic data collection scenario. From both figures in Fig. 6, we can observe that smaller personalized models can still outperform a larger generalist even with less fine-tuning data. For example, in Fig. 6(a), $2 \times 256\, \tilde{\mathcal{S}}_{\text{DPRNN}}$ personalized on only 10 s of data can outperform the largest generalist.

D. PSE models' generalization performance on unseen speakers, noises, and room RIRs

Personalization could potentially worsen the generalization performance if a model fine-tuned on a specific test environment must generalize to other unseen test environments comprised of unseen speakers, noise types, or room conditions. The performance degradation is mainly due to the *catastrophic forgetting* phenomenon (French, 1999): fine-tuning on the target test time environment changes the weights that were initially pretrained on the general-purpose training set. This can be problematic if the model is relocated or the surrounding is changed (e.g., furniture rearrangement or new additions to room, such as draping, that could alter the acoustics).

We examine this behavior by challenging an already personalized student with a different unseen environment. For this experiment, we design $K = 8$ different environments with balanced speaker gender, noise class, and various room dimensions. Details of the configurations can be found in



(a) 0 dB input SNR (b) 10 dB input SNR

FIG. 6. (Color online) Relative improvements in SI-SDR from different dataset sizes for fine-tuning on various input SNR levels under severe 0 dB SNR level (a) and moderate 10 dB SNR level (b). Student models are two-layered GRU with varying number of hidden units fine-tuned on DPRNN teacher models ($\tilde{\mathcal{S}}_{\mathrm{DPRNN}}$).

Table IV. Further details on dimensions of the rooms can be found in Table VI in the Appendix. The student models are personalized to each kth room, using noisy reverberant signals generated from $\mathbb{S}_{\mathrm{ft}}^{(k)}$, $\mathbb{M}_{\mathrm{ft}}^{(k)}$, and $\mathbb{R}_{\mathrm{ft}}^{(k)}$ for K different personalized student models in total. The fine-tuned models are then evaluated on each jth room using set-aside unseen datasets $\mathbb{S}_{\mathrm{te}}^{(j)}$, $\mathbb{M}_{\mathrm{te}}^{(j)}$, and $\mathbb{R}_{\mathrm{te}}^{(j)}$ taken from the same $K = 8$ configurations, i.e., $j \in \{1, \ldots, K\}$. Thus, $k = j$ is the desired personalization setup, while $j \neq k$ represents the kth PSE model challenged to work on the jth environment. 2×64 RNN student and DPRNN teacher was used to produce Fig. 7. Additive background noise was scaled to 0 dB input SNR.

In Fig. 7, we show the relative differences between the pretrained generalist and personalized student models evaluated on all $K = 8$ environments. We report the result using SI-SDR, short-time objective intelligibility (STOI; Taal *et al.*, 2011),[2] and perceptual evaluation of speech quality (PESQ; Rix *et al.*, 2001)[3] scores for an in-depth evaluation on personalization and its following effects on other environments. We apply various metrics, SI-SDR, STOI, and PESQ, to provide a comprehensive measure of the noisy and reverberant conditions. It is not straightforward to find a metric for enhancement methods that necessarily leads to improvements for different downstream applications because different metrics capture different distortion measures. Considering automatic speech recognition (ASR) as an example, word error rate (WER) and STOI have shown a higher correlation coefficient than other objective evaluation metrics; however, under realistic conditions, there are various factors that affect the ASR performance, and there is no single metric that has been shown to necessarily lead to better WER (Chai *et al.*, 2018; Fukumori *et al.*, 2013).

TABLE IV. Descriptions of different unseen environments.

Configuration Indentification (ID)	Speaker ID (gender)	Noise	Room	RIR
A	260 (M)	Crying baby	Q301	BUT
B	1221 (F)	Rooster	E112	BUT
C	1995 (M)	Crackling fire	CR2	BUT
D	3575 (F)	Car horn	R112	BUT
E	908 (M)	Sea waves	Small-far	RVB
F	1320 (F)	Clapping	Medium-near	RVB
G	2830 (M)	Crickets	Medium-far	RVB
H	4992 (F)	Train	Large-far	RVB

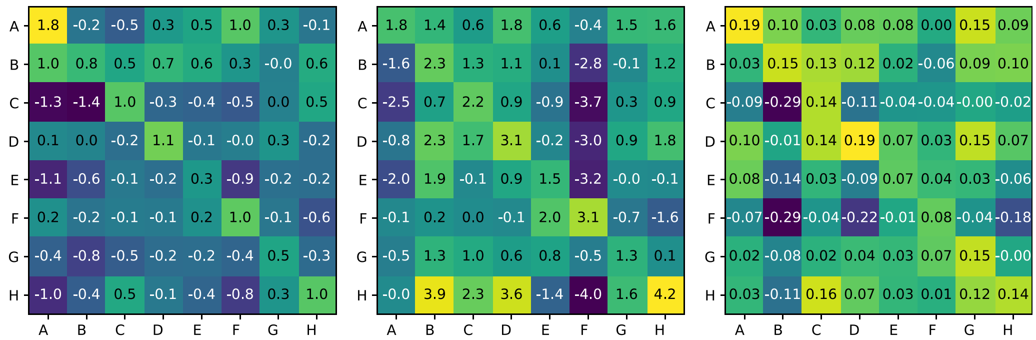
The negative values in the cells indicate performance degradation incurred from personalization. Each jth cell in the kth row corresponds to the performances of the model personalized on kth environment and evaluated on the jth environment. For example, the first row corresponds to the performances of the student model fine-tuned in environment "A" evaluated on all K configurations.

Unsurprisingly, the diagonal axes generally show highest improvements because those cells represent evaluation results of student models personalized on the same environment. This supports the main argument of our proposed framework. On the other hand, there are underperforming cases, such as models fine-tuned on "E," performing poorly on A (–1.1 dB ΔSI-SDR) and "F" (–0.9 dB ΔSI-SDR). Interestingly, the inverse relationship does not always hold. Although the model personalized on "B" generalize well to A (1.0 dB ΔSI-SDR), the model fine-tuned on A does not for B (–0.2 dB ΔSI-SDR). These results show that personalizing on one condition can incur negative effects when the model has to generalize to another environment.

Although this reveals a weakness of the proposed framework, this problem can be addressed with a simple solution by *resetting* or *readjusting* the model when sudden worsened performance is detected. However, it is not straightforward to detect such a performance drop during the test time. As a remedy, we propose to compare the PSE result to the teacher model's estimate as an indirect way to evaluate the speech quality. It is because there are no ground-truth test time data available. In Fig. 8, we show the results from a same experimental setup with Fig. 7 but with the scores computed from using the teacher's estimates as the reference, i.e., the pseudo targets. We see that the SI-SDR, STOI, and PESQ scores measured against teacher's estimates (Fig. 8) are different from those measured against the ground-truth targets (Fig. 7). However, the scores measured using teacher's outputs are still close approximates to the ground-truth metrics. This showcases another merit of using the teacher's estimates. We can reliably use these pseudo metrics to estimate a model's test time performance and decide to reset back to the pretrained generalist version or initiate a fine-tuning process to adjust the model to the new test environment.

Figures 7(b) and 7(c) also show that the relative differences in SI-SDR, STOI, and PESQ are not always correlated to one another. Cells with high relative SI-SDR improvements do not necessarily show improvements in intelligibility (e.g., model personalized on B evaluated on A). This could be a result of the loss function defined by Δ SI-SDR to optimize the student models during the fine-tuning process. A better optimization objective could be explored to prevent such differences. Nonetheless, personalization on intended environments generally shows large improvements without significant performance degradation.

https://doi.org/10.1121/10.0024621



(a) Δ SI-SDR (b) Δ STOI (c) Δ PESQ

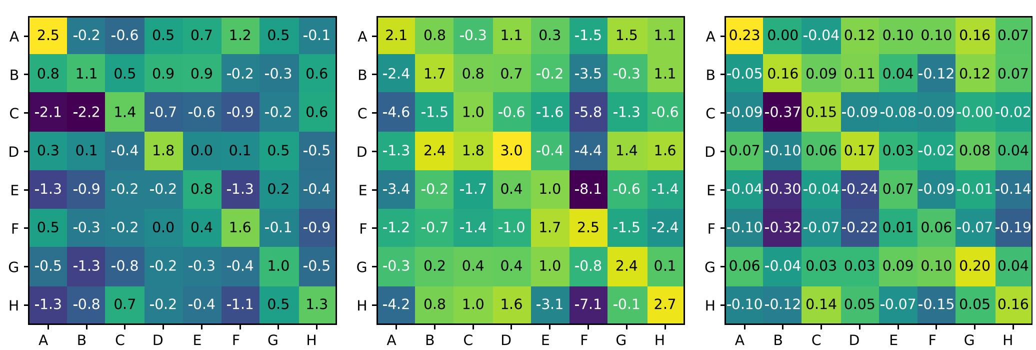
FIG. 7. (Color online) Joint denoising and dereverberation results of $2 \times 64\, \tilde{\mathcal{S}}_{\mathrm{DPRNN}}$ on different environments with 0 dB input SNR. Each of the $K = 8$ different environments are marked from "A" to "H." Each cell corresponds to the performances of the model personalized on one environment and evaluated on each environment. The objective scores, ΔSI-SDR (a), ΔSTOI (b), and ΔPESQ (c) are measured using ground-truth anechoic targets as the reference.

E. Generalization performance to unseen locations within the same room

Next, we evaluate on a scenario in which a student model is personalized on a single location of a room and tested on unseen positions within the same room. Thus far, our experiments used all RIRs, $\mathbb{R}_{\mathrm{ft}}^{(k)}$, from the kth room to construct the test time fine-tuning dataset. It was to make the PSE model robust to the variations of RIR filters, which vary vastly depending on the microphone-speaker distance, vicinity to walls or corners, occluding objects, and other factors (Shinn-Cunningham *et al.*, 2005). In this subsection, we fine-tune a student model using noisy reverberant data generated using a RIR signal from a specified location, i, within the same kth room, $\mathbb{R}_i^{(k)}$. As with other experiments, utterances from a single user, $\mathbb{S}_{\mathrm{ft}}^{(k)}$, and one noise type, $\mathbb{M}_{\mathrm{ft}}^{(k)}$, are used for fine-tuning. For evaluation, we select RIRs from unseen speaker locations within the same room, $\mathbb{R}_j^{(k)}$, where $j \neq i$. Unseen speech samples from $\mathbb{S}_{\mathrm{te}}^{(k)}$ and noise sources, $\mathbb{M}_{\mathrm{te}}^{(k)}$, are used to generate the noisy reverberant evaluation set. For brevity, we experiment on 2×64 student models and 0 dB input SNR for additive background noise.

Two rooms from the BUT were selected for this experiment: a small office (L212) and a large conference room (D105). Their speaker locations and distances from the microphone are described in Table V. Geometric information of all the other rooms can be found in Table VI in the Appendix.

Figure 9(a) shows several speaker positions from room L212 used in this experiment. We fine-tune on noisy reverberant speech from an arbitrary position A and evaluate on other locations of the room. We experiment using the same room but on multiple different speakers and noises described in Table IV. Figure 9(b) shows the average generalization results. Despite the changes in location of the test time speech source, fine-tuning on one location of a room can improve dereverberation and denoising results for unseen locations within the same room at least to some degree. The same behavior can be observed in Fig. 10 for a much larger room, D105, although the generalization power drops drastically when the unseen location is too different from that used for fine-tuning, e.g., as in F. This demonstrates that a stationary personalized device is capable of performing robust SE on a nonstationary user within the same room as opposed to suffering from drastic changes in entire room geometry (Sec. IV D).

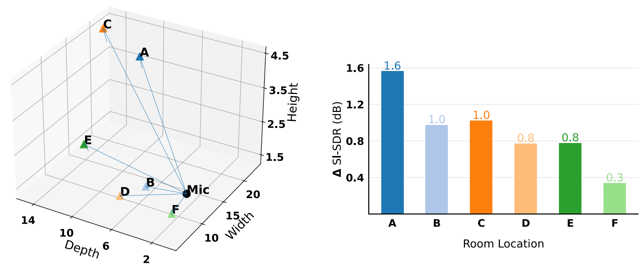


(a) Δ SI-SDR (b) Δ STOI (c) Δ PESQ

FIG. 8. (Color online) Joint denoising and dereverberation results of $2 \times 64\, \tilde{\mathcal{S}}_{\mathrm{DPRNN}}$ on different environments with 0 dB input SNR. Each of the $K = 8$ different environments are marked from A to H. Each cell corresponds to the performances of the model personalized on one environment and evaluated on each environment. In this setup, the objective scores ΔSI-SDR (a), ΔSTOI (b), and ΔPESQ (c) scores are measured using the teacher model's outputs as the reference.

TABLE V. Locations and source-mic distance of the rooms.

Room ID	Location ID	Location ($m \times m \times m$)	Source-microphone distance (m)
L212	A	$0.41 \times 1.13 \times 1.98$	4.87
	B	$6.96 \times 0.77 \times 1.98$	2.11
	C	$5.34 \times 2.48 \times 1.39$	0.90
	D	$4.72 \times 1.32 \times 1.88$	0.87
	E	$3.21 \times 1.67 \times 0.46$	2.12
D105	A	$11.93 \times 22.93 \times 3.63$	13.76
	B	$5.23 \times 9.90 \times 1.82$	4.18
	C	$14.79 \times 20.79 \times 4.47$	14.76
	D	$6.01 \times 6.06 \times 1.97$	7.85
	E	$9.65 \times 6.22 \times 3.12$	10.01
	F	$0.70 \times 5.48 \times 2.02$	7.86



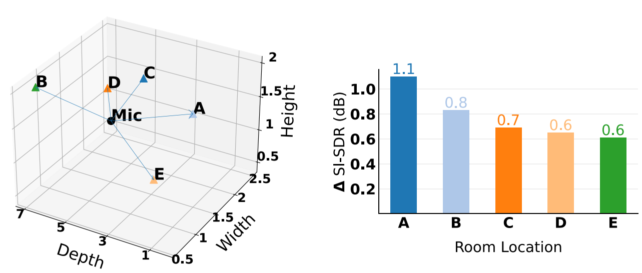
(a) Locations of speakers and microphone (b) Average Δ SI-SDR per speaker location

FIG. 10. (Color online) Denoising and dereverberation evaluation results from fine-tuning in room D105 from location *A* and generalizing to unseen locations. Locations of the microphone along with various positions within the room are shown in (a). The generalization results of the model on noisy reverberation speech from all positions in the room are provided in (b).

V. CONCLUSION

In this paper, we proposed a zero-shot KD approach to personalizing SE models for joint dereverberation and denoising. Our goal was to adapt a small model to dynamically changing test time SE environment instead of employing a large generalist model, which can be too heavy for embedded systems. In doing so, we exploited widely available noisy mixtures during test time rather than leveraging ground-truth targets or any extra information of the acoustic environment, which are rarely available in the real-world use-cases. To improve the usability of the corrupt examples found in the test scene, our framework synthesizes pseudo targets by executing a superior-quality SE routine on an overly complex teacher model. We suggest that this KD-based personalization can be performed on a regular basis or when a significant change is detected in the test time acoustic scene. As this fine-tuning task can be performed either in the cloud or when the device is idle, we envision that it is not burdensome for the device.

Evaluation results demonstrate that the student model's performance greatly improves on specific test time speakers and acoustic environments. The improvements were consistent under various noise and room conditions. Furthermore, the improvements can be observed regardless of model size or the amount of fine-tuning data available: the fine-tuned performance is dependent on the amount of data, but this does not pose a serious limitation on our framework as we can observe improvements even with minimal data. It is also noticeable that the architectural difference between the student and teacher models does not impact the personalization process. Therefore, we expect that our proposed framework can benefit from advancements in the future deep learning-based SE research. Because our small personalized student model can give superior performances to large generalist models, we claim that the KD-based fine-tuning method provides another mode of model compression that does not sacrifice performance, i.e., lossless model compression.



(a) Locations of speakers and microphone (b) Average Δ SI-SDR per source location

FIG. 9. (Color online) Denoising and dereverberation evaluation results from fine-tuning in room L212 from location *A* and generalizing to unseen locations. Locations of the microphone along with various positions within the room are shown in(a). The generalization results of the model on noisy reverberation speech from all positions in the room are provided in (b).

Although fine-tuning on specific environments can harm the generalization on other unseen scenes, the teacher's estimates can again be used to gauge the change of environments. A decision can be made to reset the student's parameter back to its pretrained value, followed by another personalization procedure for further adaptation. Also, our study shows that models personalized on one location can still show improved generalization on unseen locations within the same room, demonstrating robustness to nonstationary sources.

Major limitations of our current study for PSE comes from the dependency on the quality of the teacher model and especially the amount of fine-tuning data available. Another weakness is our experiments tested on test time environments contain only a single speaker and one unique noise source per room. Future research shall consider expanding this study to minimize the number of utterances required for the KD procedure and perform SE and separation under multi-speaker conditions.

ACKNOWLEDGMENTS

This material is based on work supported in part by the National Science Foundation under Grant No. 2046963.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Librispeech (Panayotov *et al.*, 2015),

https://doi.org/10.1121/10.0024621

TABLE VI. Descriptions of rooms from BUT.

Room ID	Size ($m \times m \times m$)	Volume (m^3)	Type
Q301	$10.7 \times 6.9 \times 2.6$	192	Office
L207	$4.6 \times 6.9 \times 3.1$	98	Office
L212	$7.5 \times 4.6 \times 3.1$	107	Office
L227	$6.2 \times 2.6 \times 14.2$	229	Stairs
R112	$4.4 \times 2.8 \times 2.6$	~40	Hotel room
CR2	$28.2 \times 11.1 \times 3.3$	1033	Conference room
E112	$28.2 \times 11.1 \times 3.3$	~900	Lecture room
D105	$17.2 \times 22.8 \times 6.9$	~2000	Lecture room
C236	$7.0 \times 4.1 \times 3.6$	102	Meeting room

MUSAN (Snyder *et al.*, 2015), ESC50 (Piczak, 2015), AIR (Jeub *et al.*, 2009), PORI (Merimaa *et al.*, 2005), RWCP (Nakamura *et al.*, 2000), BUT (Szöke *et al.*, 2019), and REVERB (Kinoshita *et al.*, 2013).

APPENDIX: DIMENSIONS OF ROOMS

TABLE VI describes the detailed geometric information about the BUT used in our experiments.

[1]See https://github.com/kaldi-asr/kaldi/blob/master/egs/aspire/s5/local/multi_condition/rirs/ (Last viewed January 26, 2024).
[2]See https://github.com/mpariente/pystoi (Last viewed December 2023).
[3]See https://github.com/vBaiCai/python-pesq (Last viewed February 16, 2023).

Bertin, N., Camberlein, E., Lebarbenchon, R., Vincent, E., Sivasankaran, S., Illina, I., and Bimbot, F. (**2019**). "Voicehome-2, an extended corpus for multichannel speech processing in real homes," Speech Commun. **106**, 68–78.

Bhalgat, Y., Lee, J., Nagel, M., Blankevoort, T., and Kwak, N. (**2020**). "LSQ+: Improving low-bit quantization through learnable offsets and better initialization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, Seattle, WA (IEEE, New York), pp. 696–697.

Boll, S. F. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. **27**, 113–120.

Chai, L., Du, J., and Lee, C.-H. (**2018**). "Acoustics-guided evaluation (age): A new measure for estimating performance of speech enhancement algorithms for robust ASR," arXiv:1811.11517.

Chapelle, O., Schölkopf, B., and Zien, A. (**2006**). *Semi-Supervised Learning* (MIT Press, Cambridge, MA).

Chawla, A., Yin, H., Molchanov, P., and Alvarez, J. (**2021**). "Data-free knowledge distillation for object detection," in *Proceeding: IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, January 3–8, 2021 (IEEE, New York), pp. 3289–3298.

Chazan, S., Goldberger, J., and Gannot, S. (**2017**). "Deep recurrent mixture of experts for speech enhancement," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY (IEEE, New York), pp. 359–363.

Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi B., Xu, C., and Tian, Q. (**2019**). "Data-free learning of student networks," in *Proceeding: IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, October 27–November 2, 2019 (IEEE, New York), pp. 3514–3522.

Chen, J., Mao, Q., and Liu, D. (**2020**). "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," arXiv:2007.13975.

Cho, K., Merriënboer, B. V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (**2014**). "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 25–29, 2014 (ACL).

Choi, J., Lee, J., Park, J., and Nam, J. (**2019**). "Zero-shot learning for audio-based music classification and tagging," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 4–8, 2019 (IEEE, New York).

Dauphin, Y. N., Tur, G., Hakkani-Tür, D., and Heck, L. (**2013**). "Zero-shot learning for semantic utterance classification," in *Proceedings of the International Conference on Learning Representations*, Banff, AB, Canada, April 14–16, 2014 (OpenReview).

Deng, J., Zhang, Z., Eyben, F., and Schuller, B. (**2014**). "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," IEEE Signal Process. Lett. **21**(9), 1068–1072.

Doersch, C., Gupta, A., and Efros, A. A. (**2015**). "Unsupervised visual representation learning by context prediction," in *Proceedings of the International Conference on Computer Vision*, Santiago, Chile, December 7–13, 2015 (IEEE, New York).

Drude, L., Hasenklever, D., and Haeb-Umbach, R. (**2019**). "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*, Brighton, UK, May 12–17, 2019 (IEEE, New York), pp. 695–699.

Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. **32**(6), 1109–1121.

Frankle, J., and Carbin, M. (**2018**). "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *Proceedings of the International Conference on Learning Representations*, Vancouver, BC, Canada, April 30–May 3, 2018 (OpenReview).

French, R. M. (**1999**). "Catastrophic forgetting in connectionist networks," Trends Cognit. Sci. **3**(4), 128–135.

Fukumori, T., Nakayama, M., Nishiura, T., and Yamashita, Y. (**2013**). "Estimation of speech recognition performance in noisy and reverberant environments using PESQ score and acoustic parameters," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Kaohsiung, Taiwan, October 31–November 3, 2023 (IEEE, New York), pp. 1–4.

Gannot, S., Burshtein, D., and Weinstein, E. (**1998**). "Iterative and sequential Kalman filter-based speech enhancement algorithms," IEEE Trans. Speech Audio Process. **6**, 373–385.

Han, K., Wang, Y., Wang, D., Woods, W. S., Merks, I., and Zhang, T. (**2015**). "Learning spectral mapping for speech dereverberation and denoising," IEEE/ACM Trans. Audio, Speech, Lang. Process. **23**(6), 982–992.

Han, S., Mao, H., and Dally, W. J. (**2016**). "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proceedings International Conference Learn. Representations*, San Juan, Puerto Rico, May 2–4, 2016 (OpenReview).

Hao, X., Wen, S., Su, X., Liu, Y., Gao, G., and Li, X. (**2020**). "Sub-band knowledge distillation framework for speech enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Shanghai, China, October 25–29, 2020 [International Speech Communication Association (ISCA), Baixas, Occitanie, France].

Hinton, G., Vinyals, O., and Dean, J. (**2015**). "Distilling the knowledge in a neural network," arXiv:1503.02531.

Hochreiter, S., and Schmidhuber, J. (**1997**). "Long short-term memory," Neural Comput. **9**(8), 1735–1780.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., and Xie, L. (**2020**). "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Shanghai, China, October 25–29, 2020 [International Speech Communication Association (ISCA), Baixas, Occitanie, France].

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. (**2018**). "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, June 19–21, 2018 (IEEE, New York), pp. 2704–2713.

Jeub, M., Schafer, M., and Vary, P. (**2009**). "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *IEEE International Conference on Digital Signal Processing*, Santorini, Greece, July 5–7, 2009 (IEEE, New York), pp. 1–5.

Kim, S., and Kim, M. (**2021**). "Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, NY (IEEE, New York).

Kingma, D., and Ba, J. (**2015**). "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, San Diego, CA, May 7–9, 2015 (OpenReview).

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Maas, R., Gannot, S., and Raj, B. (**2013**). "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, NY (IEEE, New York), pp. 1–4.

Larochelle, H., Erhan, D., and Bengio, Y. (**2008**). "Zero-data learning of new tasks," in *Proceedings of the AAAI National Conference on Artificial Intelligence*, Chicago, IL, July 13–17, 2008 (AAAI, Palo Alto, CA), Vol. 1, p. 3.

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (**2019**). "SDR—Half-baked or well done?," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12–17, 2019 (IEEE, New York), pp. 626–630.

Li, X., Dalmia, S., Mortensen, D. R., Li, J., Black, A. W., and Metze, F. (**2020**). "Towards zero-shot learning for automatic phonemic transcription," in *Proceedings of the AAAI National Conference on Artificial Intelligence*, New York, NY, February 7–12, 2020 (AAAI, Palo Alto, CA), Vol. 34, pp. 8261–8268.

Lopes, R. G., Fenu, S., and Starner, T. (**2017**). "Data-free knowledge distillation for deep neural networks," in *Proceedings: 31st Conference on Neural Information Processing Systems (NIPS 2017),* Long Beach, CA.

Luo, Y., Chen, Z., and Yoshioka, T. (**2020**). "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 4–8, 2020 (IEEE, New York), pp. 46–50.

Luo, Y., and Mesgarani, N. (**2019**). "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Trans. Audio. Speech. Lang. Process. **27**(8), 1256–1266.

Manohar, V., Ghahremani, P., Povey, D., and Khudanpur, S. (**2018**). "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained ASR models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, December 18–21, 2018 (IEEE, New York), pp. 250–257.

Merimaa, J., Peltonen, T., and Lokki, T. (**2005**). "Concert hall impulse responses Pori, Finland: Reference," Technical report, available at http://www.acoustics.hut.fi/projects/poririrs/.

Micaelli, P., and Storkey, A. (**2019**). "Zero-shot knowledge transfer via adversarial belief matching," in *Proceedings: Advances in Neural Information Processing Systems 32 (NeurIPS 2019).*

Nakamura, S., Hiyane, K., Asano, F., Nishiura, T., and Yamada, T. (**2000**). "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *International Conference on Language Resources and Evaluation*, Athens, Greece, May 31–June 2, 2000 (ELRA, Paris, France), pp. 965–968.

Nakaoka, S., Li, L., Inoue, S., and Makino, S. (**2021**). "Teacher-student learning for low-latency online speech enhancement using wave-u-net," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, ON, Canada, June 6–11, 2021 (IEEE, New York), pp. 661–665.

Nayak, G. K., Mopuri, K. R., and Chakraborty, A. (**2021**). "Effectiveness of arbitrary transfer sets for data-free knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, June 9–15, 2019 (IEEE, New York), pp. 1430–1438.

Nayak, G. K., Mopuri, K. R., Shaj, V., Venkatesh Babu, R., and Chakraborty, A. (**2019**). "Zero-shot knowledge distillation in deep networks," in *Proceedings of the International Conference on Machine Learning*, Brighton, UK, May 12–17, 2019, pp. 4743–4751.

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

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (**2015**). "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia, April 19–24, 2015 (IEEE, New York), pp. 5206–5210.

Pariente, M *et al.* (**2020**). "Asteroid: The PyTorch-based audio source separation toolkit for researchers," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Shanghai, China, October 25–29, 2020 [International Speech Communication Association (ISCA), Baixas, Occitanie, France], pp. 2637–2641.

Piczak, K. J. (**2015**). "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, Brisbane, Australia, October 26–30, 2015 (ACM, New York), pp. 1015–1018.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, May 7–11, 2001 (IEEE, New York), Vol. 2, pp. 749–752.

Rochford, J. (**2019**). "Accessibility and IoT/smart and connected communities," AIS Trans. Human-Comput. Interact. **11**(4), 253–263.

Romera-Paredes, B., and Torr, P. (**2015**). "An embarrassingly simple approach to zero-shot learning," in *Proceedings of the International Conference on Machine Learning*, Lille, France, July 6–11, 2015 (OpenReview). pp. 2152–2161.

Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (**2005**). "Localizing nearby sound sources in a classroom: Binaural room impulse responses," J. Acoust. Soc. Am. **117**(5), 3100–3115.

Sivaraman, A., and Kim, M. (**2020**). "Sparse mixture of local experts for efficient speech enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Shanghai, China, October 25–29, 2020 [International Speech Communication Association (ISCA), Baixas, Occitanie, France], pp. 4526–4530.

Sivaraman, A., and Kim, M. (**2021**). "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)*, Brno, Czech Republic, August 30–September 3, 2021 [International Speech Communication Association (ISCA), Baixas, Occitanie, France] (IEEE, New York).

Sivaraman, A., and Kim, M. (**2022**). "Efficient personalized speech enhancement through self-supervised learning," IEEE J. Sel. Top. Signal Process. **16**(6), 1342–1356.

Sivaraman, A., Kim, S., and Kim, M. (**2021**). "Personalized speech enhancement through self-supervised data augmentation and purification," in *Proceedings of the Annual Conference of the International Speech Comm. Association (Interspeech)*, Brno, Czech Republic, August 3–September 3, 2021 [International Speech Communication Association (ISCA), Baixas, Occitanie, France].

Snyder, D., Chen, G., and Povey, D. (**2015**). "MUSAN: A music, speech, and noise corpus," arXiv:1510.08484.

Subramanian, A. S., Chen, S.-J., and Watanabe, S. (**2018**). "Student-teacher learning for BLSTM mask-based speech enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Hyderabad, India, September 2–6, 2018 [International Speech Communication Association (ISCA), Baixas, Occitanie, France], pp. 3249–3253.

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

Szöke, I., Skacel, M., Rosner, L., Paliesek, J., and Cernocky, J. (**2019**). "Building and evaluation of a real room impulse response dataset," IEEE J. Sel. Top. Signal Process. **13**(4), 863–876.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. Audio. Speech. Lang. Process. **19**(7), 2125–2136.

Tu, Y.-H., Du, J., and Lee, C.-H. (**2019**). "Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition," IEEE/ACM Trans. Audio. Speech. Lang. Process. **27**(12), 2080–2091.

https://doi.org/10.1121/10.0024621

Tzinis, E., Venkataramani, S., and Smaragdis, P. (**2019**). "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*, Brighton, UK, May 12–17, 2019 (IEEE, New York), pp. 81–85.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (**2010**). "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," J. Mach. Learn. Res. **11**, 3371–3408.

Wang, D. L., and Chen, J. (**2018**). "Supervised speech separation based on deep learning: An overview," IEEE/ACM Trans. Audio, Speech, Language Proc. **26**(10), 1702–1726.

Wang, W., Zheng, V. W., Yu, H., and Miao, C. (**2019**). "A survey of zero-shot learning: Settings, methods, and applications," ACM Trans. Intell. Syst. Technol. **10**(2), 1–37.

Watanabe, S., Hori, T., Roux, J. L., and Hershey, J. R. (**2017**). "Student-teacher network learning with enhanced features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, LA, March 5–9, 2017 (IEEE, New York), pp. 5275–5279.

Williamson, D., Wang, Y., and Wang, D. (**2016**). "Complex ratio masking for monaural speech separation," IEEE/ACM Trans. Audio. Speech. Lang. Process. **24**(3), 483–492.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (**2018**). "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2251–2265.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (**2014**). "An experimental study on speech enhancement based on deep neural networks," IEEE Signal Process. Lett. **21**(1), 65–68.

Ye, J., Ji, Y., Wang, X., Gao, X., and Song, M. (**2020**). "Data-free knowledge amalgamation via group-stack dual-GAN," in *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, Seattle, WA, June 14–19, 2020 (IEEE, New York), pp. 12516–12525.

Zhang, Z., Song, Y., Zhang, J., McLoughlin, I. V., and Dai, L. (**2020**). "Semi-supervised end-to-end ASR via teacher-student learning with conditional posterior distribution," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Shanghai, China, October 25–29, 2020 [International Speech Communication Association (ISCA), Baixas, Occitanie, France], pp. 3580–3584.