

TGIF: Talker Group-Informed Familiarization of Target Speaker Extraction

Tsun-An Hsieh and Minje Kim

University of Illinois Urbana-Champaign, Siebel School of Computing and Data Science, Urbana, IL 61801, USA

Abstract—State-of-the-art target speaker extraction (TSE) systems are typically designed to generalize to any given mixing environment, necessitating a model with a large enough capacity as a generalist. Personalized speech enhancement could be a specialized solution that adapts to single-user scenarios, but it overlooks the practical need for customization in cases where only a small number of talkers are involved, e.g., TSE for a specific family. We address this gap with the proposed concept, talker group-informed familiarization (TGIF) of TSE, where the TSE system specializes in a particular group of users, which is challenging due to the inherent absence of a clean speech target. To this end, we employ a knowledge distillation approach, where a group-specific student model learns from the pseudo-clean targets generated by a large teacher model. This tailors the student model to effectively extract the target speaker from the particular talker group while maintaining computational efficiency. Experimental results demonstrate that our approach outperforms the baseline generic models by adapting to the unique speech characteristics of a given speaker group. Our newly proposed TGIF concept underscores the potential of developing specialized solutions for diverse and real-world applications, such as on-device TSE on a family-owned device.

1. INTRODUCTION

Recent advancements in speech front-end applications, such as speech enhancement (SE) [1]–[6], speech separation [7]–[10], and target speaker extraction (TSE) [11]–[14] have demonstrated substantial improvements in performance under severe noisy and reverberant acoustic environments. However, these advancements often come at the cost of increasing demand for computational resources, which poses significant challenges for on-device applications that require low compute and memory usage. The growing complexity of these models makes it difficult to deploy them efficiently on edge devices, such as intelligent speakers or smartphones, where such tasks need to run in real-time using only limited computational resources. Hence, research on reducing model complexity for the signal enhancement tasks is one of the essential challenges.

Among various task-agnostic model compression techniques for audio applications, such as low-bit quantization and pruning [15]–[20], model personalization has emerged as a promising approach, effectively reducing model size while preserving performance. Model personalization, applied to single-talker speech enhancement tasks, can be seen as a model compression technique, as it reduces the problem space from enhancing *any* speaker’s speech to a particular talker. Hence, by relieving the burden of generalization, *personalized speech enhancement* (PSE) could have gained significant complexity reduction while not sacrificing the enhancement quality. For instance, the sparse mixture of local experts (MLoE) systems [21], [22] conduct personalization by assigning mixtures to groups with similar acoustic characteristics and assigning them to local experts. Meanwhile, [23] uses zero-shot learning combined with knowledge distillation (KD) [24] to personalize SE models without requiring clean speech from the target speaker, where a pretrained large teacher model on a generic dataset guides a smaller student model to adapt effectively to unseen speakers and environments. Finally, it has also been shown that pretraining the SE model from mixtures of noisy speech and

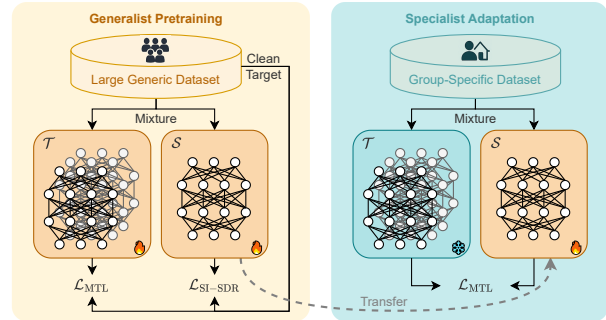


Fig. 1: An overview of the proposed knowledge distillation process. A generic dataset is used to pretrain the teacher and student generalists. Given pretrained generalists, the student model learns to adapt to a group-specific scenario with the guidance of the pretrained teacher.

additional noise is useful in reducing the amount of personal data required for PSE [25]. Note that this PSE concept is different from TSE, which conditions the speech separation model to extract only the target talker, which is sometimes also called personalized speech enhancement [26].

However, the successful model compression achieved in PSE has predominantly focused on single-talker contexts, overlooking scenarios involving multiple talkers. Meanwhile, the multi-talker tasks, TSE or speech separation, have not leveraged the fact that the problem space can also be greatly reduced in some common use cases where only a small talker group is involved in the mixture, e.g., recordings from a family-owned smart speaker.

In this paper, we propose a novel talker group-informed familiarization (TGIF) task, where a TSE system gets *familiarized* with a target talker group after the system is deployed to the unseen owner group. In this scenario, talker group-informed TSE aims to optimize model performance for a small group of speakers—an approach that remains underexplored for TSE, yet is practically significant. We see this novel concept as an extension of PSE, where the target adaptation domain grows from a single target talker to a small group of people, e.g., a family. To this end, we also propose a novel training framework specifically designed for TGIF of TSE. Our framework employs a two-stage process based on knowledge distillation (KD). In the first stage, a generalist “teacher” model is pretrained on a large, diverse dataset of noisy reverberant multi-talker speech mixtures without a notion of a speaker group. In the second stage, during test-time adaptation, the teacher model estimates the target speech source from the mixture, which, as “pseudo” targets, guides the “student” model’s training. Since the test-time adaption lacks access to the clean speech targets of the mixture, the pseudo sources are expected to serve as the proxy target, i.e., the knowledge distilled from the teacher. Meanwhile, since the talker group is a small fixed set of people, we expect that the small student model can still perform well with proper training. Unlike traditional TSE approaches that are trained to generalize to various speaker identities or PSE tailored only for one speaker, our method

adapts the student model to the unique acoustic characteristics of a talker group, including reverberation and additional noise, ensuring efficient and personalized performance without the need for clean target data.

We adopt SpEx+ [14], as a more complex and powerful teacher model to generate high-quality supervision signals. TD-SpeakerBeam [12], a more lightweight and efficient model, is optimized through KD using SpEx+’s output as pseudo targets. The models are evaluated on datasets derived from the DNS Challenge corpus [27], with experiments across various unseen speakers and acoustic settings to assess performance in realistic, talker group-specific scenarios. Our results demonstrate that the KD framework can effectively enhance the TSE performance in the proposed TGIF scenarios, and is able to outperform the teacher when there are more interfering speakers, offering a practical solution for applying *familiarized* TSE techniques to a talker group-specific setting.

2. TALKER GROUP-INFORMED FAMILIARIZATION

2.1. Problem Definition

The conventional TSE task begins with a time-domain mixture signal \mathbf{x} that consists of K speech sources, corrupted by a non-speech source \mathbf{n} and a room impulse response (RIR) kernel \mathbf{h} as follows:

$$\mathbf{x} = \mathbf{h} * \mathbf{s}_{k^*} + \sum_{k=1}^{K-1} \mathbf{s}_k + \mathbf{n}, \quad (1)$$

where \mathbf{s}_k denotes the clean speech source from the k -th speaker. The objective of the TSE model \mathcal{F} is to recover the target source signal \mathbf{s}_{k^*} from the mixture, $\mathbf{s}_{k^*} \approx \hat{\mathbf{s}}_{k^*} \leftarrow \mathcal{F}(\mathbf{x}, \mathbf{e}_{k^*})$, where the enrollment speech of the k^* -th speaker, \mathbf{e}_{k^*} , is critical to inform the system of the identity of the target speaker.

We assume that each speech source \mathbf{s}_k is spoken by a unique speaker k rather than allowing a mixture of multiple utterances spoken by the same speaker. Hence, K also indicates the number of unique talkers participating in the mixture. In this configuration, it is commonly assumed that the model is tasked with extracting any target speaker within the group, i.e., $k^* \sim \mathbb{G}$, where \mathbb{G} is a set of a very large number of speakers in the conventional definition of TSE.

Meanwhile, we can also define a *talker group-specific* TSE task, where the input signal \mathbf{x} is a mixture of a fixed set of talkers (also corrupted by noise and reverberation), i.e., $k \sim \mathbb{S}^{(l)}$, where $\mathbb{S}^{(l)}$ denotes a talker group with a group index $l \in \{1 \dots L\}$, and L defines the number of groups. Given the definition of the entire talker set \mathbb{G} and the small subset $\mathbb{S}^{(l)}$, let $|\mathbb{G}|$ be the number of total speakers in \mathbb{G} . Then, there are $\binom{|\mathbb{G}|}{K}$ combinations of unique speaker mixtures that the TSE model \mathcal{F} should generalize to, while the number of unique speaker mixtures $\binom{|\mathbb{S}^{(l)}|}{K}$ drops exponentially as $|\mathbb{S}^{(l)}| \ll |\mathbb{G}|$. As a result, \mathcal{F} learns to extract the target speaker from the group-specific mixtures more easily.

2.2. Overview of the Knowledge Distillation Process

Regardless of its size, architecture, and performance, we call a TSE model a *generalist* if it is trained to extract the target speaker from any given mixture of speakers sampled from \mathbb{G} . Otherwise, we call it *specialist*, as it is trained to work only on a small talker group $\mathbb{S}^{(l)}$. Hence, we assume that a small specialist model is a preferred solution in resource-constrained test environments once it works well on the particular speaker group, while it is not required to generalize to other groups. However, it is challenging due to the fact that the speaker group is difficult to be defined ahead of time, leading to a lack of clean speech sources, which is essential for training a model.

Our approach to the TGIF of TSE employs the KD framework, which was originally proposed for the single-talker PSE problem in [23], [28]. As illustrated in Figure 1, KD begins with pretraining the generalists. Both the teacher and student models are trained as generalists on a generic, large-scale speaker set \mathbb{G} , where the student model is incapable of competing with the teacher’s performance.

After a generalist is trained, it is deployed to a test environment. In resource-constrained environments, smaller models are more suitable despite reduced performance, while larger models offer better generalization, but are often impractical due to their cost. In this case, the specialist model could be created by fine-tuning the student generalist using a group-specific dataset. Here, in practice, the group-specific dataset consists of mixture signals only, lacking their dry clean sources. Hence, the KD framework involves the teacher model’s output into the training process as the pseudo target. The KD process is a fine-tuning process, as the student model is already pretrained as a generalist, whose model parameters are transferred to initialize the student model before the KD process begins. This fine-tuning process adapts the student model to the unique acoustic characteristics and limited set of speakers.

2.3. Pretraining of Teacher and Student Generalists

In the generalist pretraining stage, both the teacher model \mathcal{T} and the student model \mathcal{S} are trained on a generic dataset, which originates from the generic speaker set \mathbb{G} . It is important to synthesize the mixtures with a wide variety of acoustic conditions, including different speakers, noise types, and signal-to-noise ratios (SNRs). In addition, the speakers are randomly selected without considering any specific grouping. Given this condition, the teacher model is trained to extract the target speech from the mixture with reference enrollment: $\mathbf{s}_{k^*} \approx \hat{\mathbf{s}}_{k^*} \leftarrow \mathcal{T}(\mathbf{x}, \mathbf{e}_{k^*})$. The teacher model is a high-capacity network, optimized to produce accurate source estimates $\hat{\mathbf{s}}_{k^*}$ from \mathbf{x} . Concurrently, the student model \mathcal{S} is also trained from the same dataset and procedure, although it tends to underperform the teacher due to its compressed architecture for computational efficiency.

The training of generalists is straightforward as the loss function can be defined between the predicted source $\hat{\mathbf{s}}_{k^*}$ and the ground truth source \mathbf{s}_{k^*} . The pretraining stage ensures that both models learn general knowledge for handling a variety of acoustic scenarios, providing a robust foundation for the subsequent fine-tuning task on the group-specific data. To train the generalists, we minimize the negative scale-invariant signal-to-distortion ratio (SI-SDR) [29] defined between the estimated source and the target source as follows:

$$\mathcal{L}_{\text{SI-SDR}} := -\text{SI-SDR}(\hat{\mathbf{s}}_{k^*} \parallel \mathbf{s}_{k^*}) \quad (2)$$

$$\mathcal{L}_{\text{CE}} := -\sum_{k=1}^K y_k \log(\hat{y}_k) \quad (3)$$

$$\mathcal{L}_{\text{MTL}} := \mathcal{L}_{\text{SI-SDR}} + \gamma \mathcal{L}_{\text{CE}}, \quad (4)$$

where \mathbf{y} and $\hat{\mathbf{y}}$ are the K -dimensional one-hot representations of target speaker and the prediction from the speaker embeddings, and γ serves as a regularization coefficient. Finally, the multi-task learning (MTL) loss \mathcal{L}_{MTL} is used to train the entire TSE system. Note that although we include \mathcal{L}_{CE} here, following the best-performing configuration reported in [30], the student model is trained using only a reconstruction loss $\mathcal{L}_{\text{SI-SDR}}$. We defer the training details in Sec. 3.3.

2.4. Test-Time Adaptation for Student Specialists

Test-time adaptation tailors the pretrained student network \mathcal{S} to the specific characteristics of a given talker group. By introducing a group index l , we denote a group of talkers by $\mathbb{S}^{(l)}$. Accordingly, the l -th

student model $\mathcal{S}^{(l)}$ is associated with the dataset originates from $\mathbb{S}^{(l)}$, resulting in a TSE model specialized in the acoustic conditions and speaker characteristics of the l -th specific group.

Due to the absence of ground-truth clean speech data at test time, the outputs of the generalist teacher model serve as proxy targets. The specialist learns to optimize their performance by minimizing the discrepancy between its outputs $\tilde{s}_{k^*} \leftarrow \mathcal{S}^{(l)}(\mathbf{x}, e_{k^*})$ and the corresponding proxy targets \hat{s}_{k^*} , thereby adapting effectively to the unique conditions of the talker group $\mathbb{S}^{(l)}$. In summary, the KD loss is defined as follows:

$$\mathcal{L}_{\text{MTL}} := -\text{SI-SDR}(\tilde{s}_{k^*} || \hat{s}_{k^*}) + \gamma \mathcal{L}_{\text{CE}}, \quad (5)$$

where the main difference from eq. (2) is the teacher’s estimate \hat{s}_{k^*} replaces s_{k^*} as the learning target for $\mathcal{S}^{(l)}$.

3. EXPERIMENTS

3.1. TGIF Dataset

We create our TGIF dataset based on the ICASSP 2021 Deep Noise Suppression (DNS) Challenge [27]. For generalist training, the dataset is talker-group agnostic, meaning that each mixture can contain speech from **any speaker** in \mathbb{G} . During test time, a talker group is postulated to simulate the real-world family environment, where all speakers in the mixture are drawn exclusively from **the same talker group** $\mathbb{S}^{(l)}$. The noise and reverberation conditions differ between generalist pretraining and test-time adaptation. Pretraining covers a broad range of noise types and room acoustics, while test-time adaptation is limited to household noises and a narrower range of room sizes.

3.1.1. Generic Dataset for Generalist Pretraining: We partially follow the data preparation recipe of the DNS Challenge, with modifications to incorporate multi-speaker scenarios as shown in eq. (1). Clean speech segments are sourced from LibriVox [31], and the noise clips are selected from AudioSet [32] and Freesound [33] as in the DNS Challenge.

To simulate reverberation, we sample RIRs from the OpenSLR28 dataset [34], which provides diverse acoustic profiles in room sizes and reverberation times. We apply the sampled RIR only to the target speaker’s utterance, 80% of the time, while the other sources remain dry. Each audio mixture is 10 seconds long and contains between one and five speakers, with signal-to-interference ratios (SIRs), defined between the reverberant target utterance and the *sum* of other speech utterances, uniformly sampled from the range of $[-5, 25]$ dB. While we ensure the overall SIR of the speech mixture, each interfering speech utterance’s loudness is also randomized using the same range. Background noise is then added to these reverberant speech mixtures, controlled by a signal-to-noise ratio (SNR) drawn from the same $[-5, 25]$ dB range, measured specifically between the reverberant target speech and the noise. After all, each mixture includes up to one noise source and up to four interfering speakers, possibly overlapped. In total, we generate 1,000 hours of training data and 50 hours for validation. For each training example, a 3-second enrollment is selected from a different utterance spoken by the same target speaker of the given mixture. As a result, both the mixture and enrollment signals are three seconds long as inputs to the TSE model.

3.1.2. TGIF Dataset for Test-time Adaptation: For testing, we construct 20 talker group-specific datasets using a separate set of clean speech, noise, and RIRs, sourced from different datasets: clean speech from the Voice Cloning Toolkit (VCTK) corpus [35], noise from DEMAND [36], and RIRs from OpenSLR26 [34].

We assign up to five *fixed* members to each test group that are mixed similarly to the generic dataset along with a randomly selected

household noise source and the RIRs of a fixed room to create 10-second mixtures. We assume that there are clean enrollment segments for the individual members. For each group $\mathbb{S}^{(l)}$, the target speaker in every mixture is randomly selected from the group’s members, and the number of unique speakers K ranges from one to $|\mathbb{S}^{(l)}|$. For the talker group-specific dataset, we maintain the same SIR range used in the generic dataset, but adjust the SNR range to $[-15, 15]$ dB. This modification creates a noisier acoustic environment for us to evaluate both the generalization capabilities of the baseline models and the effectiveness of KD-based adaptation. In total, we synthesize 50 hours of mixture data: 20 hours for KD adaptation, 10 hours for validation to prevent overfitting, and 20 hours for evaluating the student specialists. Unlike generalist pretraining, test-time adaptation is tailored to household acoustic environments, using only noise types such as dishwashing, kitchen activity, and living room ambiance, within small-room reverberation settings.

3.2. Baseline Models Configurations

Our student models adopt the time-domain (TD) SpeakerBeam [30] architecture. TD-SpeakerBeam works in time domain and includes a convolutional encoder-masker-decoder structure. Speaker embeddings are fused at the 7th layer of the masker using a multiplicative conditioning mechanism.

We train two student model variants with reduced hidden dimensionality to evaluate the impact of our adaptation strategy. While the original TD-SpeakerBeam uses a hidden size of 512, our student models adopt smaller configurations with hidden size of 256 and 128. This setup allows us to examine whether KD can effectively improve test-time performance even in more compact models.

We employ SpEx+ [14] as our teacher model. SpEx+ is a also a fully TD architecture that improves upon its predecessor [13]. Different from TD-SpeakerBeam, SpEx+ consists of two parallel encoders, for the mixture and the enrollment utterance, respectively, that share weights to ensure consistency in the learned representations in early layers. These encoders are followed by a temporal convolutional network (TCN)-based separator, which predicts a mask conditioned on the target speaker’s identity. The speaker embedding, extracted from the enrollment encoder, uses a ResNet-based [37] speaker encoder, and is injected into the separator via affine transformations at multiple layers. Finally, a decoder reconstructs the waveform of the target speaker from the masked latent features.

3.3. Pretraining & KD-Based Adaptation Setup

We pretrain both the teacher and student generalists using the setup described in Section 2.3. Following [14], [30], we set the regularization coefficient γ to 0 and 0.5 respectively for student and teacher models. The generalists are trained for up to 1,000 epochs with batch sizes of 8 and 16 for \mathcal{T} and \mathcal{S} , respectively, using the Adam [38] optimizer with an initial learning rate of 0.001. A patience-based scheduler halves the learning rate if no improvement is observed for 20 validation epochs. Early stopping is employed with a patience of 120 epochs based on validation loss. The mixture and enrollment segments are cropped to 3 seconds and are resampled to 16 kHz.

For test-time adaptation, we distill the pseudo-target from the SpEx+ teacher model and use it to fine-tune the pretrained TD-SpeakerBeam student generalists. The adaptation is performed separately for each talker group from the TGIF testing set (Sec. 3.1.2). Each mixture and enrollment pair is 10 seconds long, and both are sampled at 16 kHz. The student model is initialized from the best checkpoint obtained during generalist pretraining and fine-tuned for up to 120 epochs using the Adam optimizer with a learning rate of 10^{-5} . Early stopping is not applied for adaptation.

Table 1: Mean SI-SDR and SI-SDR improvement. Values in the parenthesis show the improvements from the baseline student generalists.

Model	Overall		1 speaker		2 speakers		3 speakers		4 speakers		5 speakers	
	SI-SDR	SI-SDRi	SI-SDR	SI-SDRi	SI-SDR	SI-SDRi	SI-SDR	SI-SDRi	SI-SDR	SI-SDRi	SI-SDR	SI-SDRi
\mathcal{T}	4.66	13.20	8.60	13.63	4.32	13.99	0.66	12.72	-2.03	11.54	-6.18	8.93
$\mathcal{S}_{h=128}$	1.72	10.26	6.06	11.10	0.97	10.64	-2.53	9.53	-5.51	8.06	-8.79	6.32
$\mathcal{S}_{h=256}$	2.36	10.90	6.65	11.68	1.57	11.24	-1.90	10.16	-4.50	9.07	-8.20	6.91
$\mathcal{S}_{h=128}^{\text{KD}}$	2.98 (+1.26)	11.52	6.18 (+0.12)	11.22	2.43 (+1.46)	12.10	-0.13 (+2.40)	11.92	-2.08 (+3.43)	11.49	-5.50 (+3.29)	9.61
$\mathcal{S}_{h=256}^{\text{KD}}$	3.44 (+1.08)	11.97	6.50 (-0.15)	11.54	3.03 (+1.46)	12.70	0.33 (+2.23)	12.38	-1.54 (+2.96)	12.03	-4.73 (+3.47)	10.38
$\mathcal{S}_{h=128}^{\text{KD-Oracle}}$	3.82 (+2.10)	12.36	6.63 (+0.57)	11.66	2.95 (+1.98)	12.62	1.18 (+3.71)	13.23	-0.27 (+5.24)	13.30	-2.22 (+6.57)	12.89
$\mathcal{S}_{h=256}^{\text{KD-Oracle}}$	4.42 (+2.06)	12.96	7.05 (+0.40)	12.08	3.74 (+2.17)	13.41	1.88 (+3.78)	13.93	0.57 (+5.07)	14.14	-1.78 (+6.42)	13.34

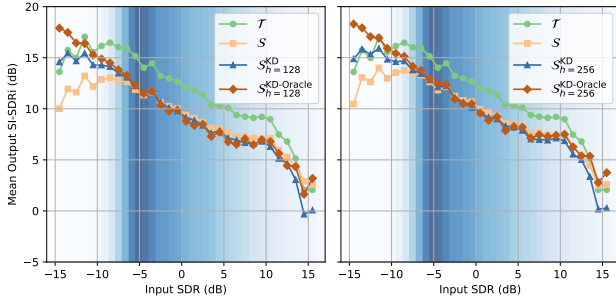


Fig. 2: SI-SDRi trends across input SDR.

To evaluate the effectiveness of adaptation under both ideal and more realistic conditions where clean references may not be available, the adaptation process considers both oracle and teacher-forcing adaptation scenarios. In the oracle setup, the student specialists are fine-tuned using the ground truth clean target speech as the supervision signal. In the teacher-forcing setup, which follows a KD paradigm, the model instead uses the output of the SpEx+ teacher as a pseudo target.

4. RESULTS AND ANALYSIS

Table 1 and Fig. 2 summarize the extraction performance for the teacher model (\mathcal{T}), the generalist student (\mathcal{S}), the student specialist fine-tuned via KD (\mathcal{S}^{KD}), and the oracle KD variant ($\mathcal{S}^{\text{KD-Oracle}}$). The results are analyzed across different numbers of interfering speakers and input SDR ranges.

4.1. Overall Performance

As shown in Table 1, overall, \mathcal{T} consistently achieves better mean SI-SDR and SI-SDRi performance (4.66 dB and 13.20 dB, respectively) than the student models. The generalist students without fine-tuning perform substantially worse, with a notable drop of 2.94 dB particularly for the smaller model. The KD student models improve the generalist baseline, by closely approaching the teacher’s performance with the gaps of 1.68 dB and 1.22 dB for $\mathcal{S}_{h=128}^{\text{KD}}$ and $\mathcal{S}_{h=256}^{\text{KD}}$, respectively, which is a 1.26 dB improvement from the small generalist student. In contrast, the larger oracle KD variant achieves further gains, reducing the SI-SDR gap relative to the teacher to 0.24 dB.

4.2. Impact of Number of Speakers

As the number of speakers increases from two to five, all models’ performance declines, maintaining similar gaps between \mathcal{T} and the student baselines. However, the proposed TGIF framework shows more significant merit in the harsher input conditions. For example, in the single-speaker case, the fine-tuned student specialists (\mathcal{S}^{KD}) are on par with their generalist counterparts. In contrast, in the two- to five-speaker scenarios, every \mathcal{S}^{KD} model outperforms its generalist version: $\mathcal{S}_{h=128}^{\text{KD}}$ outperforms $\mathcal{S}_{h=128}$ by 1.46 dB, 2.40 dB, 3.43 dB, and 3.29 dB for two, three, four, and five speakers, respectively, showing that KD-based TGIF yields increasingly larger gains in

more complex acoustic environments. Interestingly, the TGIF student models even surpass \mathcal{T} in the four- and five-speaker conditions. This suggests that the TGIF configuration reduces the speaker encoder’s burden of discriminating thousands of diverse speakers to only a few group members, leading to improved contrasts between the speakers performance and highlighting its effectiveness for model compression.

Comparing oracle and non-oracle setups reveals that the benefit of teacher quality grows with task difficulty. The SI-SDR gaps between $\mathcal{S}_{h=128}^{\text{KD-Oracle}}$ and $\mathcal{S}_{h=128}^{\text{KD}}$ widen from 0.45 dB (one speaker) to 0.52 dB, 1.31 dB, 1.81 dB, and 3.28 dB as we move to five speakers. This trend indicates that high-quality teacher outputs become increasingly critical under heavier interference.

4.3. Performance Trends Across Input SDR

Fig. 2 visualizes model behaviors across different input signal-to-distortion ratio (SDR) values. We define SDR as the overall input signal’s quality, where the signal is defined as the dry clean target utterance, while the distortion is the difference between the reverberant input mixture and the target. The blue heatmap indicates the density of training samples across different input SDR intervals. Note that the overall SDR metric can be lower than the SIR or SNR range $[-5, 25]$ dB used for data preparation, considering additive noise and reverberation. The upper and lower plots show the mean SI-SDRi scores of various models as a function of input SDR.

At very low input SDRs, around -14 dB, where training samples are sparse, $\mathcal{S}^{\text{KD-Oracle}}$ achieves the highest SI-SDR improvement, followed by \mathcal{S}^{KD} and \mathcal{T} , while \mathcal{S} falls significantly behind. Notably, \mathcal{S}^{KD} outperforms \mathcal{T} in this region, despite relying on the teacher’s guidance during KD. This trend highlights that familiarization, whether via KD or oracle finetuning, is crucial for TSE when the speech sources are highly overlapped. Starting from approximately -4 dB, the effectiveness of KD diminishes and fails to surpass the student generalist \mathcal{S} , even with the oracle KD setup. In the high input SDR region (>13 dB), where mixtures are relatively clean, all models converge to modest SI-SDR gains (0–5 dB), leading KD versions to perform the worst.

In summary, the proposed TGIF method proves especially effective in low-quality and out-of-domain regions, where training samples are sparse and generalist models struggle to generalize.

5. CONCLUSION

This work introduces a new TGIF task, extending the single-speaker personalized speech enhancement systems to TSE for a small group of people, such as a family. By adapting a lightweight student model through KD using a teacher model’s pseudo-labels, the specialists significantly improve the performance, especially under unseen and challenging conditions. Overall, the specialists surpass their generalist counterparts by over 1 dB. Under severe conditions, e.g., multiple interfering speakers or low input SDR, the specialists even surpass the teacher, demonstrating their ability to adaptively discriminating group members for robust TSE.

REFERENCES

- [1] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, “Universal speech enhancement with score-based diffusion,” *arXiv preprint arXiv:2206.03065*, 2022.
- [2] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, Y. Zhang, W. Han, A. Bapna, and M. Bacchiani, “Miipher: A Robust Speech Restoration Model Integrating Self-Supervised Speech and Text Representations,” in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.
- [3] H. Yang, J. Su, M. Kim, and Z. Jin, “Genhancer: High-Fidelity Speech Enhancement via Generative Modeling on Discrete Codec Tokens,” in *Proc. Interspeech*, 2024.
- [4] J.-M. Lemerrier, J. Richter, S. Welker, and T. Gerkmann, “Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, p. 2724–2737, 2023.
- [5] J. A., R. Korostik, J. Balam, and B. Ginsburg, “Schrödinger bridge for generative speech enhancement,” in *Proc. Interspeech*, 2024.
- [6] R. Scheibler, Y. Fujita, Y. Shirahata, and T. Komatsu, “Universal Score-based Speech Enhancement with High Content Preservation,” in *Proc. Interspeech*, 2024.
- [7] S. Lutati, E. Nachmani, and L. Wolf, “Separate and diffuse: Using a pretrained diffusion model for better source separation,” in *Proc. of the International Conference on Learning Representations (ICLR)*, 2024.
- [8] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, J. Yip, D. Ng, and B. Ma, “MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024.
- [9] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, “Exploring self-attention mechanisms for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2169–2180, 2023.
- [10] N. Zeghidour and D. Grangier, “Wavesplit: End-to-end speech separation by speaker clustering,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.
- [11] M. Delcroix, K. Žmolíková, K. Kinoshita, A. Ogawa, and T. Nakatani, “Single Channel Target Speaker Extraction and Recognition with Speaker Beam,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.
- [12] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, “Improving speaker discrimination of target speech extraction with time-domain speakerbeam,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.
- [13] C. Xu, W. Rao, E. S. Chng, and H. Li, “Spex: Multi-scale time domain speaker extraction network,” *Proc. of the IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, 2020.
- [14] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, “Spex+: A complete time domain speaker extraction network,” in *Proc. Interspeech*, 2020.
- [15] J.-Y. Wu, C. Yu, S.-W. Fu, C.-T. Liu, S.-Y. Chien, and Y. Tsao, “Increasing compactness of deep learning based speech enhancement models with parameter pruning and quantization techniques,” *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1887–1891, 2019.
- [16] S. Kim, M. Maity, and M. Kim, “Incremental binarization on recurrent neural networks for single-channel source separation,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.
- [17] K. Tan and D. Wang, “Towards model compression for deep learning based speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1785–1794, 2021.
- [18] E. Cohen, H. V. Habi, and A. Netzer, “Towards fully quantized neural networks for speech enhancement,” *Proc. Interspeech*, 2023.
- [19] M. Elminshawi, S. R. Chetupalli, and E. A. P. Habets, “Dynamic slimmable network for speech separation,” *IEEE Signal Processing Letters*, vol. 31, pp. 2205–2209, 2024.
- [20] R. Miccini, C. Laroche, T. Piechowiak, and L. Pezzarossa, “Scalable speech enhancement with dynamic channel pruning,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.
- [21] A. Sivaraman and M. Kim, “Zero-shot personalized speech enhancement through speaker-informed model selection,” in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.
- [22] R. E. Zezario, C.-S. Fuh, H.-M. Wang, and Y. Tsao, “Speech enhancement with zero-shot model selection,” in *Proc. of the European Signal Processing Conference (EUSIPCO)*, 2021.
- [23] S. Kim, M. Athi, G. Shi, M. Kim, and T. Kristjansson, “Zero-Shot Test-Time Adaptation Via Knowledge Distillation for Personalized Speech Denoising and Dereverberation,” *Journal of the Acoustical Society of America*, vol. 155, no. 2, pp. 1353–1367, 2024.
- [24] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [25] A. Sivaraman and M. Kim, “Efficient Personalized Speech Enhancement Through Self-Supervised Learning,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1342–1356, 2022.
- [26] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, “Personalized Speech Enhancement: New Models and Comprehensive Evaluation,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.
- [27] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, “ICASSP 2021 Deep Noise Suppression Challenge,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.
- [28] S. Kim and M. Kim, “Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation,” in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.
- [29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – half-baked or well done?” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.
- [30] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, “Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.
- [31] J. Kearns, “Librivox: Free public domain audiobooks,” *Reference Reviews*, vol. 28, no. 1, pp. 7–8, 2014.
- [32] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio Set: An Ontology and Human-Labeled Dataset for Audio Events,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.
- [33] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, “Freesound datasets: A platform for the creation of open audio datasets,” in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.
- [34] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.
- [35] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, “Superseded-ctr vctk corpus: English multi-speaker corpus for ctr voice cloning toolkit,” *URL <http://datashare.is.ed.ac.uk/handle/10283/2651>*, 2016.
- [36] J. Thiemann, N. Ito, and E. Vincent, “The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings,” *Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [38] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.