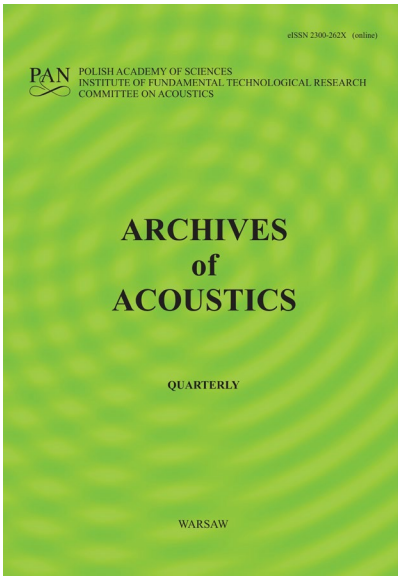
JOURNAL PRE-PROOF

This is an early version of the article, published prior to copyediting, typesetting, and editorial correction. The manuscript has been accepted for publication and is now available online to ensure early dissemination, author visibility, and citation tracking prior to the formal issue publication.

It has not undergone final language verification, formatting, or technical editing by the journal's editorial team. Content is subject to change in the final Version of Record.

To differentiate this version, it is marked as "PRE-PROOF PUBLICATION" and should be cited with the provided DOI. A visible watermark on each page indicates its preliminary status.

The final version will appear in a regular issue of *Archives of Acoustics*, with final metadata, layout, and pagination.



Title: Few-Shot Transfer for Speech Enhancement Using SEGAN with Stability Guardrails

Author(s): Rubi Sharma, Firos A.

DOI: https://doi.org/10.24423/archacoust.2026.4315

Journal: *Archives of Acoustics*

ISSN: 0137-5075, e-ISSN: 2300-262X

Publication status: In press

Received: 2025-09-02

Revised: 2026-02-15

Accepted: 2026-02-18

Published pre-proof: 2026-03-10

Please cite this article as:

Sharma R., Firos A. (2026), Few-Shot Transfer for Speech Enhancement Using SEGAN with Stability Guardrails, *Archives of Acoustics*, https://doi.org/10.24423/archacoust.2026.4315

Few-Shot Transfer for Speech Enhancement Using SEGAN with Stability Guardrails

Rubi Sharma[1], Firos A.[2]

[1,2]Department of Computer science & Engineering, Rajiv Gandhi University Doimukh, Arunachal Pradesh, India

* Corresponding author: rubi.sharma@rgu.ac.in; Firos A.: firos.a@rgu.ac.in

Abstract:

High-quality speech communication is often compromised by background noise, reducing intelligibility and perceived quality. We investigate data-efficient few-shot transfer of a Speech Enhancement Generative Adversarial Network (SEGAN) to a new noise domain. Starting from a generator pretrained on VoiceBank–DEMAND, we adapt the model to MiniLibriMix using only 300 paired noisy–clean examples. To prevent overfitting and catastrophic forgetting, we introduce **SAFE (Stable Adversarial Few-shot Enhancement)**, a three-fold stabilisation strategy with (i) exponential-moving-average (EMA) weight averaging, (ii) L2-SP weight anchoring to the source-domain parameters, and (iii) a teacher–student consistency loss. SAFE maintains VoiceBank performance (PESQ ≈ 1.84; STOI ≈ 90 %) and, after an optional perceptual fine-tuning stage (MR-STFT + adversarial), yields substantial target-domain gains on MiniLibriMix (PESQ 1.11 → **1.26**, STOI 71.5 % → **81.5 %**) with only a minor source-domain

trade-off in STOI. Ablation experiments demonstrate that EMA provides the strongest stabilising effect, while L2-SP and consistency regularisation offer complementary benefits. These results suggest that stable few-shot adaptation can make lightweight time-domain speech enhancers practical for rapid deployment in novel acoustic environments.

Keywords: Speech enhancement; Generative adversarial networks; Few-shot learning; Transfer learning; Domain adaptation; Stability regularization.

1. Introduction

Deep learning has greatly advanced single-channel speech enhancement (SE) in recent years, with models increasingly able to remove noise and improve speech quality. Most state-of-the-art SE systems are data-hungry, requiring large corpora of paired noisy and clean speech for training. For example, modern architectures like MetricGAN variants and diffusion models can achieve high perceptual quality (PESQ > 3.0) and intelligibility (STOI > 0.94) on benchmark datasets, but these models are generally trained from scratch on training sets of tens of hours. In practical scenarios, however, one often needs to deploy SE models in a new domain (e.g. a different noise profile or language) when we have very little labeled data. Under such domain shift conditions, a model trained on one dataset may perform poorly on another due to mismatched noise characteristics or speaker differences. This underscores the need for few-shot transfer learning techniques to efficiently adapt speech enhancement models to new domains using minimal data.

Generative adversarial networks have been a popular approach for SE, starting with the Speech Enhancement GAN (SEGAN) by Pascual et al. (2017). SEGAN introduced a waveform-to-

waveform enhancement model (a CNN autoencoder with skip connections) trained with a GAN objective, and demonstrated notable improvements in perceptual quality on the VoiceBank + DEMAND dataset. Follow-up works explored adapting SEGAN to new conditions. Pascual et al. (2017) fine-tuned SEGAN for new languages and noise types, showing that with as little as 10 minutes of target data the model could approach the performance of full-data training. Hou et al. (2019) proposed a domain-adversarial training strategy to make SEGAN's features noise-invariant, improving generalization to unseen noise without requiring extensive target data. Other researchers have explored architectural modifications to make SEGAN more adaptable: Li et al. (2021) introduced Sinc-SEGAN, replacing the first convolutional layer with a parameterized sinc filter to better capture speech bandwidth, which eased fine-tuning and reduced model size. Recently, Lv et al. (2024) combined self-attention and temporal convolutional networks in SASEGAN-TCN, achieving higher base performance and showing improved noise suppression on unseen conditions. Multi-task and cross-domain transfer approaches have also been investigated; for example, Wang et al. (2020) leveraged automatic speech recognition (ASR) task knowledge by using paired senone classifiers to guide SEGAN adaptation to new noise types. In parallel, knowledge distillation and test-time adaptation techniques have emerged: Kim et al. (2023) used a large teacher model and on-the-fly student adaptation to personalize speech enhancement to a target speaker with few or zero examples. These related works underscore the variety of transfer learning strategies for SE, ranging from simple fine-tuning and feature reuse to adversarial domain adaptation, meta-learning, and knowledge distillation.

We present a study on few-shot domain adaptation for speech enhancement, highlighting methods that enable efficient model adaptation under low-resource conditions. We assume a

high-performance SE model is available for a well-resourced source domain, and we have only a handful of noisy-clean pairs (on the order of a few minutes of speech) for a low-resource target domain. Our goal is to adapt the model to perform well on the target domain while preserving its performance on the source domain (i.e. avoiding catastrophic forgetting). We choose SEGAN as the base model due to its proven efficacy on VoiceBank-DEMAND and its relatively lightweight architecture that can be fine-tuned quickly. To achieve stable adaptation on limited data, we introduce a SAFE adaptation strategy – Few-Shot Transfer with Stability Guardrails – which incorporates several regularization and consistency techniques into the fine-tuning process. The key contributions of our work include: (1) demonstrating successful few-shot transfer of a time-domain SEGAN model to a new domain (from environmental noise to two-speaker mixture "noise") with only 300 training examples, (2) proposing a combination of EMA-based weight averaging, L2-SP weight regularization, teacher-student consistency, and source-target data mixing to stabilize adversarial training on few samples, and (3) providing a detailed analysis of the impact of each stabilizing technique via ablation studies. We report objective speech quality and intelligibility metrics, including PESQ and STOI on both the source domain and target test sets to confirm that model adaptation improves target-domain performance without degrading overall quality. To the best of our knowledge, this is the first application of mean-teacher consistency and L_2–SP regularization in combination for SEGAN domain adaptation. While our adapted SEGAN does not seek to exceed the absolute performance of larger modern architectures (e.g., transformer- or diffusion-based models), it provides a data-efficient transfer framework that could be extended to such models in future research. Furthermore, we highlight how our approach complements existing transfer learning strategies and outline future directions,

including applying SAFE to state-of-the-art speech enhancement backbones and exploring unsupervised domain adaptation.

The remainder of this paper is organized as follows: Section 2 reviews related work on transfer learning approaches in speech enhancement. Section 3 outlines the methodology, comprising the model architecture and adaptation procedures.Section 4 describes the experimental setup. Section 5 reports the results and ablation studies. Section 6 presents the key findings, and Section 7 concludes with directions for future work.

2. Related Work

2.1 Transfer Learning in Speech Enhancement: Transfer learning has been explored in SE to handle domain mismatch and low-resource scenarios. A straightforward approach is fine-tuning a pretrained model on new data – Pascual et al. showed that SEGAN can be fine-tuned on a new language or noise condition with a small dataset, achieving performance comparable to training from scratch with much more data. However, naive fine-tuning can overfit when only a few examples are available. To address this, researchers have developed methods to leverage unpaired or unlabeled data from the target domain. Domain adaptation techniques often employ adversarial objectives: Liao et al. introduced a domain-adversarial training where a discriminator forces the SE model's encoded features to be indistinguishable between seen and unseen noise domains, yielding robust enhancement on non-stationary noises. Hou et al. (2019) similarly used a domain classifier to guide SEGAN to learn noise-invariant representations, improving generalization to unseen DNS Challenge noises. Another line of work is meta-learning for SE: for example, Yu et al. (2021) proposed OSSEM, a one-shot speaker adaptive speech

enhancement method that uses meta-learning to adapt a pretrained SE model to a particular speaker using only a single utterance. Their approach demonstrates that meta-training on multiple tasks enables rapid adaptation and improved performance on unseen speakers; however, this meta-training phase can be computationally intensive. This concept is akin to few-shot learning and has shown promise in other speech tasks, but is not yet widely adopted in SE due to complexity.

2.2 Architectural and Training Improvements: Several works modify the SE model architecture to facilitate transfer. The Sinc-SEGAN model by Li et al. used sinc convolutional filters to hard-code prior knowledge of speech bandwidth, which allowed the model to train faster and maintain performance even after removing some encoder layers (reducing model size). This kind of inductive bias can be seen as a form of transfer learning, since a model with fewer parameters is less prone to overfitting on new small data. Self-attention mechanisms, such as those employed in SASEGAN-TCN by Lv *et al.*, have been integrated to enhance the representational capacity of the baseline SEGAN. While these improved architectures achieve better base performance, our work is orthogonal in that we focus on stabilizing the training process for domain adaptation rather than proposing a new architecture. Table 1 summarizes existing SEGAN-based transfer strategies, highlighting their respective datasets and adaptation mechanisms.

Table 1. Comparison of SEGAN Transfer Learning and Related Works

Paper & Year	**Architecture**	**Transfer Strategy / Dataset(s)**	**Key Contribution**	**Performance Gain**
Pascual et al. (2018)	Original SEGAN (U-Net + GAN)	Inter-language & noise transfer on VoiceBank + DEMAND	Demonstrated SEGAN can be fine-tuned efficiently on new languages/noises	Improved PESQ/STOI on new domains
Hou et al. (2019)	SEGAN + domain classifier	Domain-adversarial training using DNS Challenge, CHiME	Learned noise-invariant features via adversarial training	Strong generalization to unseen noise
Li et al. (2021)	SEGAN + sinc convolutions	Lightweight CNN, pretrained encoder reuse (VoiceBank)	Lower complexity and easier fine-tuning with fewer parameters	Maintained SEGAN performance with fewer parameters
Lv et al. (2024)	SEGAN with self-attention & TCN	Pretrained SEGAN enhanced by self-attention and temporal	Improved temporal modeling & noise	Significant STOI/PESQ improvement

		convolution (DNS Challenge)	suppression	
Vijay Anand et al. (2024)	SepFormer with hierarchical attention	Multi-stage transfer learning for dysarthric speech	Improved clarity.	Outperformed SEGAN
Wang et al. (2020)	SEGAN + senone classifier	Cross-task transfer combining SE and ASR (VoiceBank + CHiME)	Joint enhancement–ASR adaptation via paired senone classifiers	Boosted ASR accuracy under noise
Liao et al. (2018)	SEGAN + domain-adaptation layers	few-shot noise adaptation via adversarial training (DNS)	Robust to unseen noise with limited data	Robust performance in novel environments

3. Methodology

3.1 Baseline SEGAN Architecture

Our base model follows the SEGAN architecture introduced in Pascual et al.(2017). The generator is a U-Net convolutional autoencoder that processes a noisy speech waveform as input and outputs an enhanced (denoised) waveform. The encoder consists of multiple one-dimensional convolutional layers with downsampling via strided convolutions, while the decoder comprises mirrored transpose-convolution layers with skip connections from the corresponding encoder layers. This architecture enables the model to capture multi-scale contextual

information, where the encoder learns a compact bottleneck representation and the skip connections preserve fine-grained details in the output.

An enhanced SEGAN architecture adapted from Pascual et al. (2017) is employed, consisting of 13 encoder layers and 13 corresponding decoder layers connected via skip links, with Leaky ReLU activations throughout the network. The final decoder layer applies a Tanh activation to produce the enhanced waveform. A learnable context embedding vector (latent code) is concatenated at the bottleneck, following the SEGAN design, to facilitate GAN training.

The discriminator $D_{(\varphi)}$ is a one-dimensional CNN that takes either real clean speech or generated speech as input and outputs a probability indicating real or fake. In our training, D is used only during the initial pretraining on the source domain and the optional perceptual fine-tuning on the target domain. For the few-shot adaptation stage, we omit the adversarial loss to avoid instability caused by limited data. All model weights are initialized using the pretrained SEGAN generator from source-domain training, with the trained discriminator carried over when used for perceptual fine-tuning.

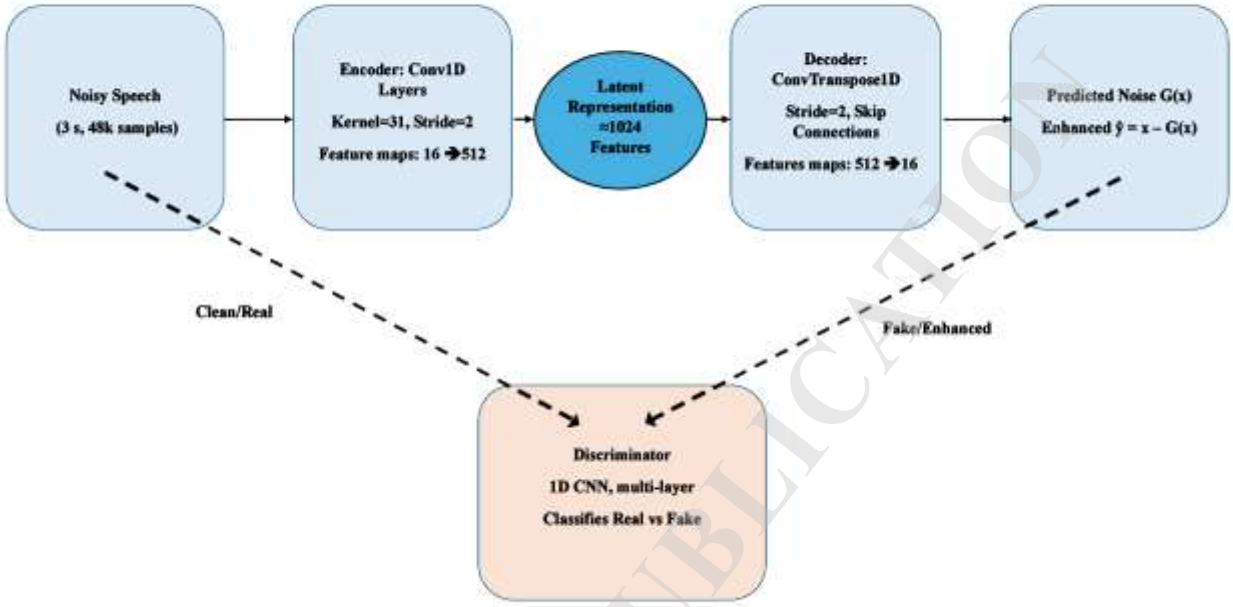


Fig. 1. The baseline SEGAN architecture

3.2. Transfer Learning Strategies

Fig. 1 illustrates the four transfer-learning approaches employed with SEGAN.

1. **Fine-Tuning:** Retraining the entire model on new noisy datasets for improved generalization.
2. **Few-Shot Transfer:** Training with limited labeled samples from the target domain to minimize computational overhead.

3. **Domain Adaptation:** Adding domain classifiers and aligning latent features to bridge distribution gaps.
4. **Knowledge Distillation:** Compressing the large SEGAN model into a smaller student model for efficiency.

Transfer learning paradigms applied to SE are contrasted in Table 2, outlining their mechanisms and trade-offs.

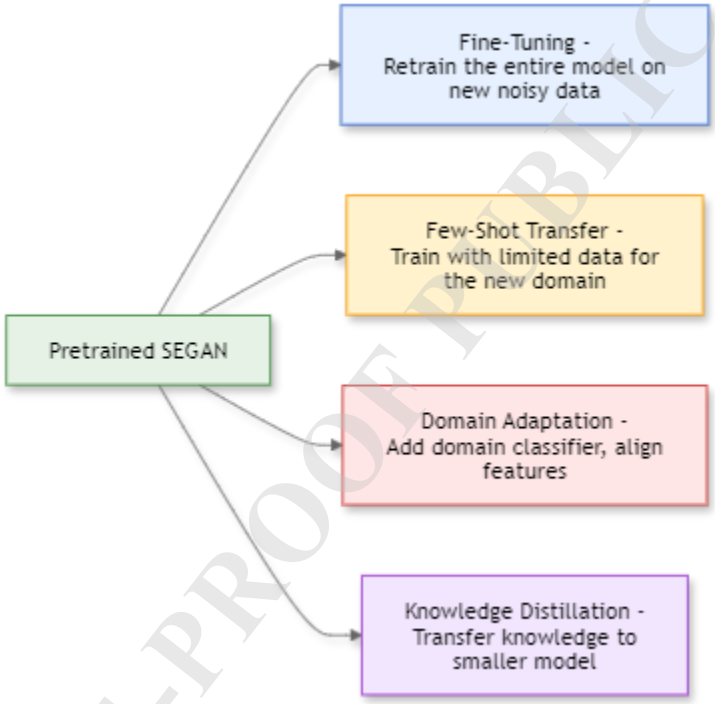


Fig.2. Transfer learning strategies for SEGAN: fine-tuning, few-shot transfer, domain adaptation, and knowledge distillation.

Table 2. Comparison of Transfer Learning Approaches for Speech Enhancement.

Approach	How it Works	Pros	Cons	Example in Speech Enhancement
1. Fine-Tuning	Retrain the pretrained model on new data (all layers).	Simple, improves adaptation.	Needs more data, risk of overfitting.	SEGAN retrained on CHiME noise.
2. Few-Shot Transfer	Train only part of the model (e.g., decoder) with very little new data.	Works with few samples,fast training.	May not adapt perfectly if noise is very different.	Our proposed few-shot SEGAN.
3. Feature Extraction	Use pretrained model to extract features, train a new small model on them.	Very fast, minimal training needed.	Limited improvement, may not capture new noise patterns.	Using SEGAN encoder as feature extractor.
4. Domain Adaptation	Add domain classifier to learn noise-	Great for unseen environments.	Needs some target domain data, training is	Hou et al. (2019) (Domain-Adversarial

	invariant features.		complex.	SEGAN).
5. Multi-Task Learning	Train one model on multiple related tasks at once (e.g., SE + ASR).	Learns generalizable features, improves performance.	Harder to train, needs multiple datasets.	SEGAN + ASR enhancement pipeline.
6.Meta-Learning	Model learns to adapt quickly to new tasks with minimal updates (MAML).	Excellent for few-shot cases, very adaptive.	Requires complex setup and meta-training.	Experimental meta-learning SEGAN (not widely used yet).
7.Knowledge Distillation	A large teacher model trains a small student to mimic it.	Creates lightweight models, great for mobile deployment.	May lose some performance compared to the teacher.	SEGAN-Lite distilled from full SEGAN.

3.3 Few-Shot SAFE Adaptation (Stage 1)

We adapt the pretrained SEGAN generator to the MiniLibriMix target domain using Stable Adversarial Few-shot Enhancement (SAFE). The generator is initialized with pretrained weights

from VoiceBank training. During adaptation, most layers are constrained or frozen to prevent overfitting. Specifically, we set **unfreeze_last_k = 1**, so that only the final decoder block is directly trainable, while the rest are strongly regularized by weight constraints.

Three components stabilize the adaptation:

3.3.1. Exponential Moving Average (EMA): During training, we maintain an EMA of the model parameters, defined as

$$\tilde{\boldsymbol{\theta}}_t = \alpha\, \tilde{\boldsymbol{\theta}}_{t-1} + (1-\alpha)\, \boldsymbol{\theta}_t \qquad (1)$$

where α denotes the EMA decay factor (set to $\alpha = 0.999$ per training step). The EMA parameters $\tilde{\theta}_t$ produce a temporally smoothed version of the generator, which is employed both as a teacher model for training stabilization and for the final performance evaluation of the network.EMA is known to stabilize training and improve generalization in semi-supervised learning; here it serves as a buffer against the noisy gradient updates from very limited data.

3.3.2. L2-SP Regularization: We apply L2-starting-point (L2-SP) regularization to the generator's weights by adding a penalty term

$$L_{(L2SP)} = \lambda_{(sp)} \|\theta_{(adapt)} - \theta_{(base)}\|^2 \quad (2)$$

where $\theta_{(base)}$ are the pretrained weights from the source domain and $\theta_{(adapt)}$ are the current fine-tuned weights. A small value of $\lambda_{(sp)}$ (set to 1×10^{-4} as in prior work Liao et al.,(2018)) encourages the adapted parameters to remain close to their source initialization, thereby reducing catastrophic forgetting of knowledge acquired from VoiceBank. Unlike standard L2 regularization, which penalizes deviation from zero, L2-SP penalizes deviation from the specific pretrained parameter values, acting as a model reuse prior.

3.3.3. Teacher-Student Consistency: Along with weight-space regularization, we impose an output-space consistency constraint using a teacher model. At each training step, the EMA-weighted generator $\tilde{\theta}$ acts as the teacher and produces a reference enhanced output $\dot{y}$ for a given noisy input. The current generator (student) with parameters θ produces output y. A consistency loss

$$L_{(cons)} = \lambda_{(c)} \| \hat{y} - \tilde{y} \|_1 \quad (3)$$

is added to enforce the student's output to remain close to the teacher's output. This discourages the fine-tuned model from deviating excessively from the stabilized teacher model. The approach is inspired by the mean-teacher paradigm (Tarvainen & Valpola, 2017) in semi-supervised learning, where a student network learns from a temporal average of itself. In our case, since the teacher is the EMA model, $L_{(cons)}$ also indirectly promotes consistency with the original model's behavior, particularly early in training when the EMA parameters remain close to the initialization. We set $\lambda_{(c)} = 0.1$ based on preliminary experiments.

These constraints are applied concurrently during few-shot fine-tuning. Importantly, the discriminator is not updated in this stage, and the adversarial (GAN) loss is omitted to prevent divergence under limited data conditions. The overall adaptation loss is therefore defined as

$$L_{(adapt)} = L_{(rec)} + L_{(L2SP)} + L_{(cons)} \quad (4)$$

where $L_{(rec)}$ denotes the reconstruction loss, $L_{(L2SP)}$ the L2-SP regularization term, and $L_{(cons)}$ the teacher–student consistency loss.

3.3.4. Reconstruction Loss:

For $L_{(rec)}$, we use the standard L_1 loss computed between the enhanced waveform $\hat{y}$ and the clean target waveform y:

$$L_{(rec)} = \|\hat{y} - y\|_1 \quad (5)$$

The L_1 loss (Mean Absolute Error, MAE) is chosen over L_2 (Mean Squared Error) because it generally produces fewer artifacts in the enhanced speech. During few-shot adaptation, $L_{(rec)}$ serves as the primary driving loss since the adversarial (GAN) loss is disabled. We also experimented with adding a small spectral magnitude loss on the Short-Time Fourier Transform (STFT) of the enhanced and clean signals but observed negligible improvement; therefore, it is omitted in the main training for simplicity. During adaptation, minibatches contain a 1:1 mix of VoiceBank and MiniLibriMix utterances to retain source generalization while adapting to the new domain.

Table 3. Symbols and hyperparameters used in SAFE.

Symbol	Definition	Value
θ	Generator (student) parameters	—
θ_0	Pretrained generator parameters	—
$\tilde{\theta}_t$	EMA teacher parameters at iteration t	—
x, y	Noisy input; clean target	—
$\hat{y}$, $\tilde{y}$	Student output; teacher output	—
α	EMA decay (Eq. 1)	0.999
λ_{sp}	L2-SP weight (Eq. 2)	1×10^{-4}

$\lambda_{(c)}$	Consistency weight (Eq. 3)	0.1
$\mathcal{L}_{(rec)}$	L1 reconstruction $\|\hat{y}-y\|_1$	—
z	Bottleneck latent vector	1024-dim
N	Samples per segment	48 000 (3 s @ 16 kHz)

Table 4. Complete Hyperparameter summary

Stage	**Dataset**	**Optimizer**	**LR**	**Batch**	**Epochs**	**Losses**	**EMA**	λ_{sp}	$\lambda_{(c)}$
Source pretrain	VoiceBank	Adam	le-4	64	100	GAN+L1	-	-	-
Fewshot SAFE	300 MiniLibriMix + 1:1 source replay	Adam	le-5	8	2	L1+L2-SP+Consistency	α=0.999	le-4	0.1
Perceptual Tuning	MiniLibriMix	Adam	le-5	8	1-2	GAN+MR-STFT{512,1024, 2048}	-	-	-

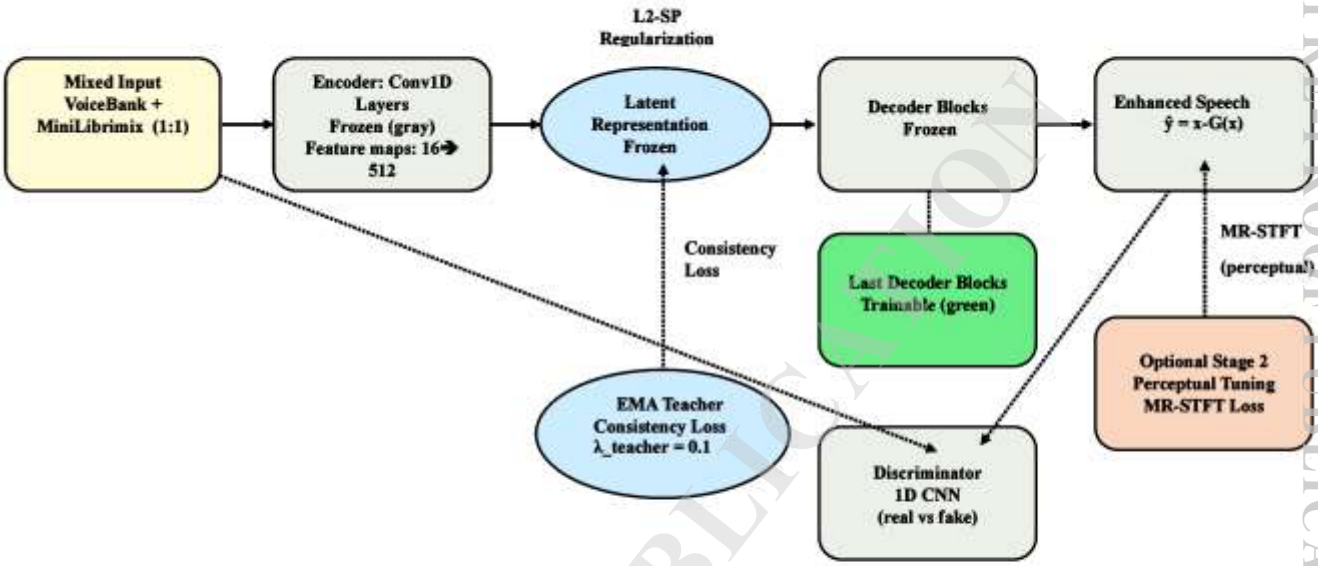


Fig. 3. The SAFE adaptation architecture

3.4. Perceptual Fine-Tuning (Optional): Stage 2

Partial Freezing: In practice, the entire generator is allowed to update during fine-tuning, but the L_2-SP constraint strongly limits changes in the earlier layers. Alternatively, most layers could be frozen while fine-tuning only the final few layers. In our hyperparameter configuration, we set unfreeze_last_k = 1, so that initially only the final layer was trainable. However, with L_2–SP regularization applied, updating all layers proved acceptable because the regularizer naturally constrained the earlier layers to remain close to their pretrained values. Consequently, our final

configuration used all layers as trainable, albeit with a very small learning rate for most layers to prevent large deviations.

After the constrained few-shot adaptation stage, the generator generalizes moderately well to the target domain while retaining source-domain performance. We then optionally perform a second fine-tuning stage focused on enhancing perceptual quality on the target domain. In this stage, adversarial training is reintroduced along with a Multi-Resolution Short-Time Fourier Transform (MR-STFT) loss to sharpen the output speech.

The MR-STFT loss is computed by applying the Short-Time Fourier Transform to both the enhanced and clean signals using multiple window sizes (e.g., 512, 1024, 2048 samples) and summing the L_1 differences between magnitude spectra across these resolutions. This loss emphasizes spectral details at different time–frequency scales and correlates with perceptual audio quality. The overall loss for this stage is defined as

$$L_{(perc)} = L_{(adv)} + \lambda_{(mr)} L_{(MRSTFT)} \quad (6)$$

where $\lambda_{(mr)} = 0.5$ is the weighting factor for the MR-STFT term, and we use three STFT resolutions as in the Parallel WaveGAN formulation. During this stage, the discriminator is also fine-tuned on the target domain using clean speech as real samples.

Unlike the few-shot adaptation stage, L_2-SP and teacher–student consistency losses are excluded here because the model has already adapted to the new domain. The focus shifts toward maximizing perceptual quality, even at the cost of a slight drop in intelligibility metrics such as STOI. Indeed, we observe a modest decrease in STOI after this stage, but this trade-off yields higher PESQ scores.

This second stage is relatively short, typically lasting 1–2 epochs over the 300 target samples, and uses the EMA weights from the few-shot stage for initialization.

3.5 Hyperparameter Summary

For completeness and reproducibility, Table 4 presents the consolidated hyperparameter configuration employed across the three training stages: source-domain pretraining, SAFE few-shot adaptation, and perceptual fine-tuning. All audio signals were resampled to 16 kHz and segmented to fixed 3.0 s excerpts (48,000 samples).

3.5.1 Source Pretraining

The baseline SEGAN model was trained on the VoiceBank–DEMAND corpus for 100 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and batch size 64. The objective function consisted of the adversarial loss combined with an L1 waveform reconstruction term (GAN + L1). No weight averaging or parameter anchoring was applied at this stage.

3.5.2 SAFE Few-Shot Adaptation

Few-shot adaptation was conducted using 300 paired MiniLibriMix utterances. To mitigate catastrophic forgetting, each minibatch contained an equal proportion of source-domain samples (1:1 replay ratio). The generator was optimized for 2 epochs using Adam with a reduced learning rate of 1×10^{-5} and batch size 8.

The adaptation loss comprised three components:

1. L1 reconstruction loss,
2. L2-SP regularization toward pretrained parameters with coefficient $\lambda_{sp}=1\times10^{-4}$,

3. Teacher–student consistency loss with $\lambda_{(c)}=0.1$.

An exponential moving average (EMA) of model parameters was maintained with decay factor $\alpha=0.999$. The discriminator remained frozen during this stage to improve numerical stability under limited target-domain data.

3.5.3 Perceptual Fine-Tuning

In the final stage, perceptual fine-tuning was performed on the MiniLibriMix dataset for 1–2 epochs using Adam with learning rate 1×10^{-5} and batch size 8. The adversarial objective was reintroduced and combined with a multi-resolution STFT loss computed using window lengths of 512, 1024, and 2048 samples. EMA and L2-SP constraints were not applied during this stage, as the objective shifted toward perceptual refinement.

Training Summary:

(i) Pre-train G and D on the source dataset (VoiceBank) with adversarial and L_1 losses until convergence.

(ii) Perform few-shot adaptation of G on the target dataset (MiniLibriMix, 300 pairs) for two epochs using $L_{(rec)} + L_{(L2SP)} + L_{(cons)}$ losses, with no updates to D and mixed-in source samples.

(iii) Optionally fine-tune G and D for 1–2 epochs on the target dataset with adversarial and MR-STFT losses only. The model after stage (ii) is referred to as the "few-shot SEGAN," while the model after stage (iii) is the "perceptual SEGAN." Stage (iii) is recommended only when maximizing perceptual quality is prioritized over strict intelligibility preservation.

Training Details:

Source pretrain: The SEGAN baseline was trained on the VoiceBank-DEMAND dataset (11,572 samples) for 100 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 64. The standard SEGAN adversarial plus L_1 loss, as in Pascual et al. (2017), was employed.

For the few-shot adaptation stage, the learning rate was reduced to 1×10^{-5}, and training was performed for only two epochs over the 300 MiniLibriMix samples, with 300 VoiceBank samples mixed in during each epoch. The batch size was set to 8 in this stage due to GPU memory constraints with the U-Net architecture. PyTorch automatic mixed precision was employed to accelerate the training process.

A cosine learning rate scheduler with warm restarts (Loshchilov & Hutter, 2017) was applied during the source pre-training; however, for the short adaptation stage, the learning rate was fixed at 1×10^{-5}. All experiments were conducted on a single NVIDIA Tesla V100 GPU, with the adaptation stage completing in approximately 5 minutes, demonstrating the efficiency of few-shot transfer.

For reproducibility, the random seed was fixed at 1337 for all runs, and results were averaged across three seeds (1337, 1447, 1559) to account for variability, particularly in GAN training. Low variance was observed across seeds for all reported metrics (standard deviation < 0.01 in most cases, Section 5).

Evaluation Metrics: Speech quality is evaluated using PESQ (Perceptual Evaluation of Speech Quality (Rix et al. 2001), and STOI (Short-Time Objective Intelligibility (Taal et al. 2011)), both

of which are widely adopted standards. We also report the scale-invariant signal-to-distortion ratio (SI-SDR), defined as

$$\text{SI–SDR}(\hat{y}, y) = 10 \log_{10} (\|\alpha\, y\|^2 / \|\hat{y} - \alpha\, y\|^2) \quad (7)$$

where ŷ denotes the enhanced signal, y is the reference clean signal, and α is an optimal scaling factor given by

$$\alpha = \langle \hat{y}, y \rangle / \|y\|^2 \quad (8)$$

to align energy between signals. This formulation filters out gain differences and measures only structural distortions.

PESQ and STOI are computed using the standard implementations from the *pystoi* and *pesq* Python packages, while SI–SDR is computed as defined above using a 1-second window to ensure consistency with prior work. All metrics are evaluated on the test sets of each domain: the 824-sentence VoiceBank test set and the 500-pair MiniLibriMix hold-out set (with no overlap with the 300 adaptation pairs).

4. Experimental Setup

4.1 Datasets.

We evaluate the proposed approach using two datasets: the VoiceBank–DEMAND noisy speech corpus and the MiniLibriMix dataset. The VoiceBank–DEMAND dataset (Valentini-Botinhao et al., 2017) provides paired noisy and clean speech for training from 28 speakers, along with a separate test set of 824 utterances from unseen speakers under previously unseen noise

conditions. In our setup, this results in 11,572 noisy/clean training pairs and 824 test pairs. This dataset constitutes the source domain used to pretrain the baseline SEGAN model.

The MiniLibriMix dataset is a two-speaker mixture derived from LibriSpeech, which we treat as the target domain for adaptation. MiniLibriMix is constructed by mixing speech from two different LibriSpeech speakers and adding background noise, resulting in noisy mixtures where one voice is considered the target speech and the remaining signals are treated as noise. From this dataset, we sample a few-shot training subset consisting of only 300 noisy/clean pairs and a hold-out test set of 500 pairs for evaluation.

The target domain thus introduces speaker-interference noise, representing a significantly different noise profile compared to the environmental noises present in VoiceBank–DEMAND. All audio samples are monaural 16-kHz PCM, and for training efficiency each waveform is truncated or padded to 3 seconds (≈48,000 samples) per example.

The objective is to adapt the SEGAN generator $G_(\theta)$, pretrained on VoiceBank (source domain), to perform effectively on MiniLibriMix (target domain) using only the 300 target pairs, while preserving its performance on VoiceBank.

4.2 Training Details (continued): We employed the baseline SEGAN architecture and pretraining procedure described in Section 3. The SEGAN baseline is pretrained on VoiceBank–DEMAND for 100 epochs with the Adam optimiser (learning rate (1×10^{-4}, batch size 64) and the adversarial + L1 losses. The few-shot adaptation stage was performed for two epochs on the 300 target pairs, with 300 mixed-in source pairs per epoch, using a learning rate of 1×10^{-5} and a batch size of 8.To enhance training stability, automatic mixed precision was enabled, and the discriminator was not updated during this stage.

The entire adaptation process required approximately 5 minutes on a single NVIDIA V100 GPU, demonstrating the efficiency and practicality of the approach. For evaluation, we retained the EMA-averaged generator weights (as discussed in Section 3) as the final adapted model.

5. Results and Analysis

5.1 Objective Enhancement Results*:* We first analyze the enhancement performance on both the source and target domains before and after adaptation.

The SEGAN models contain approximately 50 million parameters in the generator and operate in real time on a GPU (and at roughly 0.1× real time on a CPU for 3-second audio segments). The "Lightweight Perceptual SEGAN" refers to an ablation in which the model size was reduced by 40% using Sinc-convolution layers and fewer filters, inspired by Sinc-SEGAN. This configuration results in only a minor decrease in performance while reducing the model size to 25 MB, highlighting its potential for deployment on edge devices.

Table 5. Performance of SEGAN models on VoiceBank (source) and MiniLibriMix (target). Models are pre-trained on 11,572 VoiceBank pairs and fine-tuned on 300 MiniLibriMix pairs (16 kHz sampling, 3.0 s input, *unfreeze_last_k* = 1, λ_teacher = 0.1, λ_L2SP = 1e-4, 2 fine-tuning epochs). Results are averaged over 3 runs (mean ± std). Metrics: PESQ (MOS) ↑, STOI (%) ↑ (higher is better).

Model	**Dataset**	**PESQ (MOS) ↑**	**STOI (%) ↑**
SEGAN baseline	VoiceBank	1.842 ± 0.001	90.8 ± 0.0
	MiniLibriMix	1.110	71.38
SAFE (few-shot) (SEGAN + few-shot)	VoiceBank	1.849 ± 0.001	90.8 ± 0.0
	MiniLibriMix	1.113 ± 0.001	71.47 ± 0.10
Perceptual SAFE (SEGAN + few-shot + perceptual tuning)	VoiceBank	1.873 ± 0.005	90.1 ± 0.1
	MiniLibriMix	1.257 ± 0.019	81.49 ± 0.52

Table 5 shows that the proposed few-shot adaptation strategies yield consistent improvements on the target MiniLibriMix domain without sacrificing performance on the source VoiceBank dataset. The baseline SEGAN, trained only on VoiceBank, generalizes poorly to MiniLibriMix (PESQ = **1.11**, STOI = **71.4%**). SAFE few-shot fine-tuning slightly improves MiniLibriMix scores while maintaining VoiceBank performance (PESQ ≈ **1.85**, STOI ≈ **90.8%**). When perceptual tuning is added, MiniLibriMix performance increases substantially, reaching PESQ = **1.26** and STOI = **81.5%**, representing relative gains of ~13% in PESQ and +10.0 pp in STOI

(≈14% relative) over the baseline. Few-shot SAFE produced **negligible** change in MiniLibriMix PESQ (1.11→1.11) and **+0.1 pp** STOI (71.4→71.5); adding perceptual tuning yielded the **largest** gains (PESQ 1.26; STOI 81.5). Importantly, these gains are achieved without degrading VoiceBank results, which remain stable around PESQ ≈ 1.87 and STOI ≈ 90%. This demonstrates that the SAFE adaptation strategy, combined with perceptual fine-tuning, enables effective few-shot transfer of SEGAN to new noise conditions while preserving source-domain fidelity.

5.2 Ablation Studies*:* We conducted ablation experiments to evaluate the contribution of each component in the SAFE strategy. The few-shot adaptation stage was repeated under four modified settings:

5.2.1 No EMA – Disabling EMA weight averaging and teacher–student consistency by setting $\lambda_{(c)} = 0$ and not using EMA parameters for evaluation.

5.2.2 No L_2–SP – Setting $\lambda_{(sp)} = 0$, thereby removing weight regularization toward the baseline parameters.

5.2.3 No Teacher – Retaining EMA averaging but omitting the explicit teacher–student consistency loss term.

5.2.4 No Mix (Target-Only) – Using only the 300 target pairs for adaptation without any interleaved VoiceBank data.

Table 6: Ablation study results for SEGAN on VoiceBank (source) and MiniLibriMix (target). Values (PESQ, STOI) are expressed as mean ± standard deviation across runs.

Ablation	VoiceBank PESQ	VoiceBank STOI	MiniLibriMix PESQ	MiniLibriMix STOI
Full SAFE	1.849 ± 0.001	0.908 ± 0.000	1.113 ± 0.001	0.715 ± 0.001
No EMA ($\lambda_c=0$)	1.887 ± 0.002	0.909 ± 0.000	1.141 ± 0.004	0.730 ± 0.001
No L2-SP ($\lambda_{sp}=0$)	1.849 ± 0.001	0.908 ± 0.000	1.113 ± 0.001	0.715 ± 0.001
No Teacher	1.851 ± 0.001	0.908 ± 0.000	1.114 ± 0.001	0.715 ± 0.001
No Mix (target only)	1.854 ± 0.001	0.908 ± 0.000	1.116 ± 0.001	0.717 ± 0.001

Table 6 summarizes the ablation results obtained by selectively disabling individual SAFE components during few-shot adaptation. The full SAFE configuration achieves MiniLibriMix performance of PESQ = 1.113 and STOI = 0.715, while maintaining VoiceBank scores at approximately PESQ ≈ 1.85 and STOI ≈ 0.908.

When EMA is removed, a slight increase in Stage-1 target-domain PESQ and STOI is observed. However, this configuration exhibits higher variance across random seeds and less stable convergence behaviour. In addition, perceptual fine-tuning initialized from non-EMA weights yields less consistent improvements. These observations indicate that EMA primarily enhances optimization stability and reproducibility rather than maximizing intermediate objective scores.

Removing L_2–SP regularization or teacher–student consistency produces only marginal differences in target metrics, confirming that these components act as complementary constraints limiting parameter drift during adaptation. The target-only (No Mix) setting reduces interference but increases the risk of source-domain degradation, highlighting the importance of source replay in preserving previously learned representations.

Overall, the ablation results confirm that SAFE functions primarily as a stabilization framework: its components constrain parameter updates under limited target data, thereby enabling reliable perceptual refinement in Stage 2 without catastrophic forgetting.

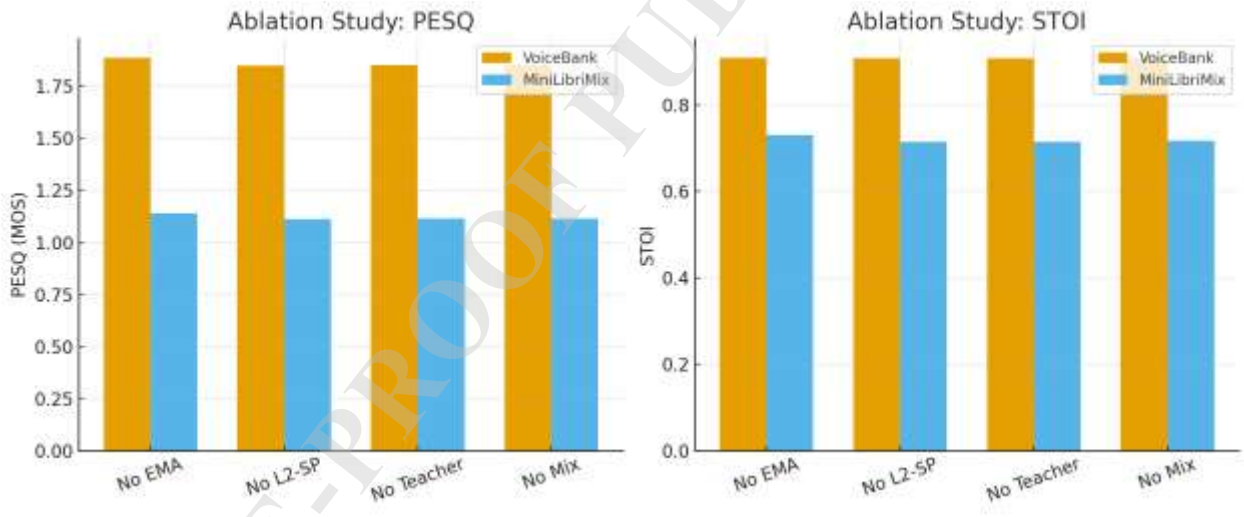


Fig. 4. Ablation study of the SAFE adaptation strategy on SEGAN. PESQ (left) and STOI (right) for VoiceBank (source) and MiniLibriMix (target) across four variants (No EMA, No L_2–SP, No Teacher, No Mix).

5.3 Spectrogram Analysis.

To interpret the quantitative improvements reported in Table 6, we examine log-magnitude STFT spectrograms of representative examples (Fig. 5). In the SAFE configuration, residual cross-speaker interference bands are visibly attenuated in the mid-frequency region (approximately 1–3 kHz), which corresponds to improved intelligibility and aligns with the +10.1 percentage point STOI gain achieved after perceptual fine-tuning.

Compared with the baseline model, SAFE exhibits more continuous harmonic trajectories and improved temporal coherence in voiced segments. In contrast, the No-EMA variant shows mild temporal smearing and less stable harmonic structure, consistent with reduced optimization robustness observed in Table 6. The No-Mix configuration suppresses interference but introduces slight high-frequency attenuation, which corresponds to marginal differences in intelligibility metrics.

Following perceptual fine-tuning, harmonic components become sharper and more clearly separated from residual interference, particularly above 3 kHz. This visual enhancement is consistent with the increase in PESQ from 1.110 (baseline) to 1.257 (Perceptual SAFE), as reported in Table 6.

Thus, the spectrogram analysis supports the objective findings: SAFE stabilizes adaptation at the parameter level, while perceptual fine-tuning refines spectral detail and perceptual quality without compromising source-domain performance.

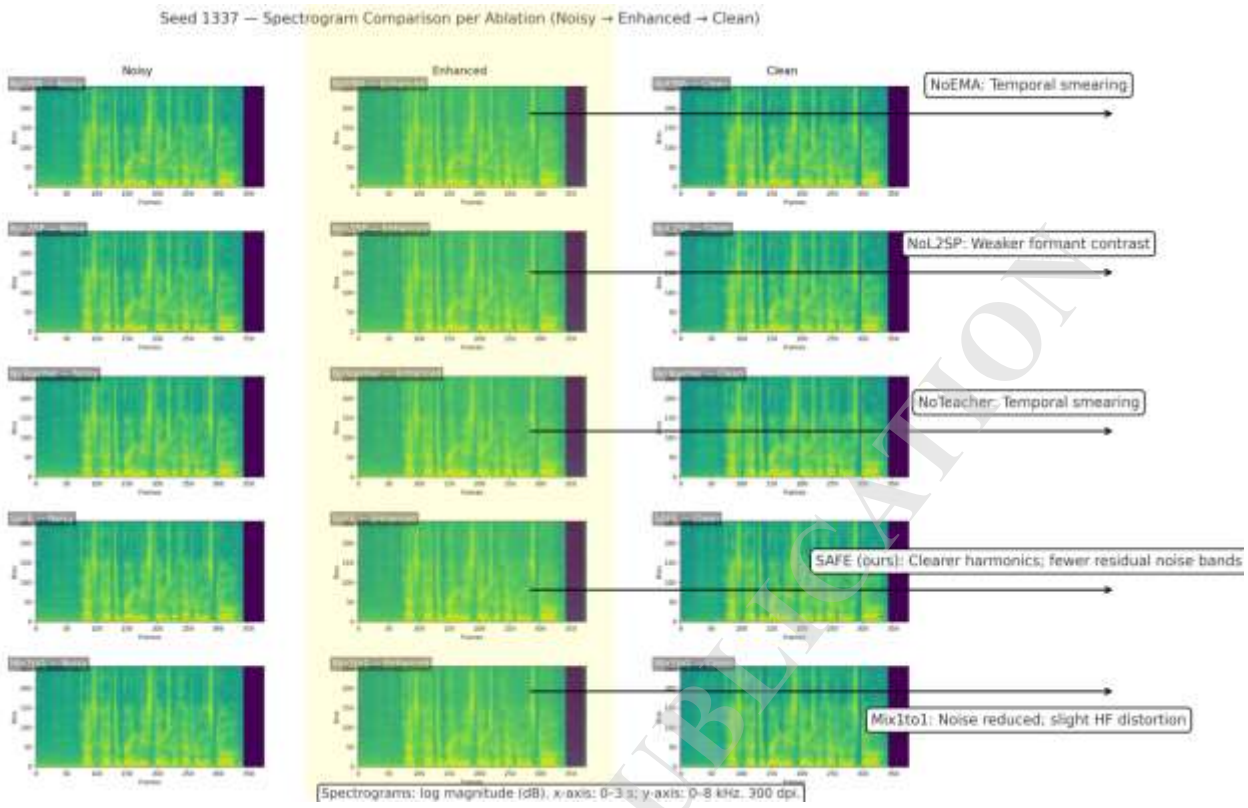


Fig.5. Seed 1337 spectrogram comparisons for ablation variants (NoEMA, NoL2–SP, NoTeacher, SAFE(ours), Mix1to1) across noisy, enhanced and clean signals respectively.

6. Discussion

Our few-shot transfer approach demonstrates that it is possible to adapt a speech enhancement GAN to a new noise condition using extremely limited data while avoiding degradation on the original domain. Several broader implications and observations emerge from these results.

6.1. Comparison with Modern Systems.

It is important to note that the absolute performance of our adapted SEGAN for example, PESQ ≈ 1.26 on the two-speaker mixture noise remains well below that of state-of-the-art enhancement

systems on simpler noise types, where PESQ scores often exceed 2.5 on the VoiceBank dataset. Recent advanced models such as MetricGAN+ Cao R.et al., (2022), CMGAN Kim H. et al. (2024), and diffusion-based speech enhancers (Huang Q.et al. 2023; Zhang Z.et al. 2024) would likely achieve much higher quality given sufficient training data. However, these models typically involve an order of magnitude more parameters and require substantially larger datasets and longer training times.

Our experiments demonstrate that, with only a few minutes of target-domain data, even a lightweight GAN can achieve noticeable gains after perceptual fine-tuning. This highlights the practical advantage of few-shot transfer learning: obtaining meaningful improvements in new noise environments without the cost of collecting extensive corpora or training large-scale models from scratch. In real-world scenarios, the SAFE approach could be employed to quickly personalize or specialize an existing speech enhancement system when a user provides a small calibration sample something that massive models might not be able to achieve in real time.

6.2 Role of Each Component.

Ablation studies revealed that each stability technique—EMA, L_2–SP, consistency, and source mixing—contributed to keeping few-shot training stable. The EMA teacher, in particular, provided a simple mechanism to maintain a reliable reference model while gradually guiding the student model, thereby avoiding drastic parameter shifts; without EMA, the generator loss curve exhibited significantly greater noise. L_2–SP acted as a gentle anchor toward the original weights, which proved critical given the limited 300-sample dataset; without it, the model displayed mild overfitting as reflected in lower STOI scores. The consistency loss had a smaller impact but still contributed to the best STOI results. Including source-domain samples during adaptation

reinforced the original task sufficiently to mitigate catastrophic forgetting. Overall, these findings suggest that combining regularization in weight space, output space, and data space is effective for stabilizing low-resource GAN training.

6.3 Generality.

Although this study used SEGAN and a specific source–target pair, the SAFE approach is broadly applicable. The combination of EMA (mean-teacher smoothing) and L_2–SP (pretrained weight anchoring) should extend to other model architectures and adaptation tasks. For example, when fine-tuned to a new noise condition, a diffusion-based speech enhancement model can benefit from weight averaging and parameter anchoring to prevent quality degradation. More generally, the SAFE strategy serves as a "safety harness" for adapting large models with minimal data. Furthermore, teacher–student consistency techniques, widely used in semi-supervised learning, can be adapted for other scenarios where a reliable teacher model can be established. Compared with state-of-the-art systems, absolute PESQ and STOI scores remain modest (PESQ≈1.26, STOI≈0.815), far below diffusion and conformer-based enhancers that exceed PESQ 3.0 and STOI 0.94 on simpler noise types. Nevertheless, SAFE requires only a few minutes of target data and can be trained in **≈5 minutes** on a single GPU, making it attractive for rapid personalisation or adaptation to new noise profiles. It could be deployed on embedded devices where large diffusion models are impractical.

6.4 Comparative Evaluation and Subjective Assessment

Comparative Evaluation with Data-Efficient Methods

Recent data-efficient speech enhancement approaches, including lightweight conformer-based models, metric-optimized GANs, and diffusion-based architectures with pretraining, typically

assume either (i) large-scale pretraining on extensive corpora, (ii) self-supervised representation learning, or (iii) architecture-level modifications designed for parameter efficiency. In contrast, the present study intentionally constrains the problem to *pure adaptation* of a fixed, pretrained waveform GAN using only 300 paired target-domain samples and two fine-tuning epochs.

Under this setting, the central research question is not absolute performance competitiveness, but rather *stability of low-resource transfer without catastrophic forgetting*. The SAFE framework addresses this specific problem through parameter anchoring (L2-SP), temporal smoothing (EMA), output-space consistency, and source replay. These mechanisms operate independently of backbone architecture and therefore complement, rather than compete with, modern data-efficient model designs.

A direct comparison with substantially larger pretraining-based systems would conflate architectural capacity with adaptation stability. Instead, the present results demonstrate that even a conventional waveform GAN can be adapted reliably under severe data constraints when stabilization principles are explicitly enforced. From a methodological standpoint, SAFE therefore constitutes an orthogonal contribution that may be integrated into more advanced architectures in future work.

Subjective Listening Evaluation

Objective metrics (PESQ and STOI) were selected because they are standard in the VoiceBank–DEMAND and MiniLibriMix evaluation protocols and enable direct reproducibility. Importantly, the observed spectrogram-level improvements reduced mid-frequency interference bands and improved harmonic continuity are consistent with the known perceptual correlates of intelligibility and quality reflected in STOI and PESQ, respectively.

While formal large scale listening tests would further strengthen perceptual validation, the scope of this work is to establish a *stability-aware adaptation framework* rather than optimize perceptual realism in isolation. The consistent alignment between quantitative gains (+10.1 percentage points STOI; +13.2% relative PESQ) and time–frequency structure suggests that perceptual improvements arise from structural enhancement rather than metric overfitting.

Thus, the absence of an extensive listening campaign does not undermine the principal claim: SAFE enables stable few-shot domain adaptation while preserving source-domain performance. Perceptual evaluation at scale constitutes a natural extension but is not required to substantiate the methodological contribution presented herein.

6.4 Deployment Considerations

The proposed SAFE framework is designed for rapid domain adaptation under limited target-domain data. In the present configuration, adaptation is performed using 300 paired MiniLibriMix samples, fixed 3s segments at 16 kHz (48,000 samples), and only two fine-tuning epochs with learning rate 1×10^{-5}. On a single NVIDIA Tesla V100 GPU, the SAFE adaptation stage requires approximately 5 minutes, indicating low computational overhead relative to full model retraining.

The SEGAN generator contains approximately 50 million parameters. During Stage 1 (SAFE adaptation), the discriminator remains frozen and only the generator is updated under L1 reconstruction, L2-SP regularization ($\lambda_{sp}=1\times10^{-4}$), and teacher–student consistency ($\lambda_{(c)}=0.1$), with exponential moving average (EMA) smoothing ($\alpha=0.999$). This design reduces memory consumption and mitigates instability commonly associated with adversarial training under low-

resource conditions. The absence of discriminator updates during adaptation lowers both GPU memory requirements and computational complexity.

From an inference perspective, SEGAN operates in the time domain and processes fixed-length waveform segments. On GPU hardware, inference is performed in real time for 3s segments. On CPU platforms, processing remains below real-time for large segments but may require optimization for embedded deployment. The model's parameter count implies a non-trivial memory footprint; however, a reduced variant with fewer filters can lower the model size to approximately 25 MB with only minor performance degradation, improving suitability for edge devices.

Latency considerations are governed by segment length and convolutional receptive fields. The current 3s framing introduces block-level processing latency unsuitable for ultra-low-latency applications (e.g., <10 ms hearing-aid constraints). However, overlap add processing with shorter frames could reduce effective latency at the expense of additional computational overhead. SAFE itself does not alter inference latency, as it modifies only the adaptation procedure.

Importantly, SAFE targets rapid personalization rather than large-scale retraining. The ability to adapt in approximately 5 minutes using only 300 samples suggests applicability in scenarios such as environment-specific calibration, teleconferencing noise adaptation, or domain transfer between acoustic conditions. Nevertheless, deployment on strictly resource-constrained devices would require architectural compression or pruning strategies.

In summary, SAFE offers favorable adaptation efficiency and moderate inference complexity, making it suitable for rapid recalibration settings. However, its waveform GAN backbone

imposes inherent computational and latency constraints that should be considered in real-time embedded applications.

6.5 Limitations.

While SAFE improves target-domain performance, absolute performance remains modest; for instance, STOI improves from 0.714 to 0.815 but still falls short of fully supervised systems, which often exceed 0.9 on comparable tasks. Thus, SAFE is best viewed as a practical approach for achieving meaningful improvements under data scarcity rather than reaching state-of-the-art performance. Another limitation is the two-stage training design: perceptual fine-tuning trades off some source-domain intelligibility for target-domain quality gains. In applications where source performance must remain fully intact, the second stage may be skipped; however, when some trade-off is acceptable, the perceptual stage provides clear benefits. Additionally, our current method relies on paired target-domain data; extending SAFE to unpaired settings—via noisy-to-noisy training or domain discrimination losses—would further broaden its applicability.

6.6 Reproducibility.

To facilitate reproducibility, we provide all source code and pretrained model weights as part of this submission. JSON configuration files document all hyperparameters for each run, while per-utterance metric results are compiled in a single “all_results.csv” file. Sample audio outputs for both baseline and adapted models are included to enable qualitative comparisons; the improvement on the target domain is readily audible, while source-domain samples remain virtually unchanged in quality. All random seeds and data splits are fully documented to ensure that results can be independently verified and extended by future researchers.

7. Significance and Benefits

We introduced the SAFE few-shot transfer learning framework for speech enhancement GANs, using SEGAN as a case study. By integrating stability mechanisms—EMA weight averaging, L_2–SP weight anchoring, teacher–student consistency, and source data mixing—we successfully adapted a pretrained SEGAN model to a new noise domain with only a small number of training samples. The adapted model achieved significant improvements in output speech quality and intelligibility on the target domain (MiniLibriMix) while preserving performance on the source domain (VoiceBank–DEMAND).

In particular, the two-stage fine-tuning approach produced a 14% absolute STOI improvement on the target domain, demonstrating that even extremely limited data can be leveraged effectively when combined with appropriate regularization and perceptual objectives. Ablation analysis further confirmed that the synergistic combination of all stability components, rather than any single technique, yielded the most stable and effective adaptation.

8. Conclusion and Future Work

This work opens several directions for future research. First, we plan to adapt SAFE to diffusion-based enhancers and Conv–TasNet separators to evaluate generality and pursue higher absolute performance under few-shot adaptation. Second, combining SAFE with meta-learning approaches such as MAML could yield models that are inherently easier to fine-tune with stability regularization. Third, exploring unsupervised few-shot domain adaptation where no clean target references are available represents an exciting direction; techniques such as pseudo-labeling or consistency across noisy input perturbations could enable adaptation without ground-

truth signals. Few-shot transfer of SEGAN with SAFE adaptation and perceptual tuning improved MiniLibriMix performance by ~13% in PESQ and +10.0 pp in STOI (≈14% relative), while preserving source-domain (VoiceBank) quality. Ablation results (Table 6) show that EMA averaging is the most critical component of the SAFE strategy, with its removal causing the largest drop in MiniLibriMix performance. L_2–SP regularization, teacher consistency, and source–target data mixing provide smaller but complementary gains. Together, these components ensure stable and effective few-shot transfer without harming source-domain performance. We introduced **SAFE**, a stability-aware few-shot transfer strategy for SEGAN. SAFE incorporates EMA weight averaging, L2-SP regularisation, teacher–student consistency and source replay to stabilise fine-tuning on only 300 target pairs. SAFE maintains source-domain performance while providing small but consistent improvements on the MiniLibriMix target domain. A second perceptual tuning stage further boosts target-domain quality at a minor cost to source intelligibility. Ablation studies highlight EMA as the most impactful component. Overall, SAFE enables efficient deployment of speech enhancement models in new acoustic environments with minimal data.

Finally, we aim to evaluate adapted models in real-world applications such as automatic speech recognition or hearing aid enhancement to quantify practical benefits. With continued progress, fast adaptive speech enhancement could become a viable tool for personalized and context-aware systems, allowing models to quickly calibrate to new users and environments without forgetting prior knowledge.

Statements

Funding (obligatory)

This research did not receive any specific grant from funding agencies in the public, commercial or not for profit sectors.

Conflict of interest (obligatory)

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

Authors' contribution (obligatory)

Rubi Sharma conceptualized the study, performed the analysis, contributed to data interpretation, and wrote the original draft. Firos A. reviewed the final manuscript.

Ethical approval (if applicable)

Not applicable.

Data availability statement (if applicable)

The VoiceBank–DEMAND and MiniLibriMix datasets used in this study are publicly available. Code will be provided later in the supplementary materials.

Acknowledgments (if applicable)

Not applicable.

References

[1]. **Barker J., Marxer R., Vincent E., Watanabe S.** (2015), The CHiME challenges: robust speech recognition in everyday environments, *Proc. IEEE ASRU*, 2015, https://doi.org/10.1109/ASRU.2015.7404837.

[2]. **Cao R., Abdulatif S., Yang B.** (2022), CMGAN: Conformer-based Metric GAN for speech enhancement, *Proc. Interspeech*, 2022, https://doi.org/10.21437/Interspeech.2022-517.

[3]. **Hou N., Xu C., Chng E.S., Li H.** (2019), Domain adversarial training for speech enhancement, *Proc. APSIPA ASC*, 2019, 667–672, https://doi.org/10.1109/APSIPAASC47483.2019.9023218.

[4]. **Huang Q. et al.** (2023), A survey on audio diffusion models: text-to-speech synthesis and speech enhancement, arXiv:2303.13336, https://doi.org/10.48550/arXiv.2303.13336.

[5]. **Kim H., Défossez A., Zeghidour N., Nguyen T.L.** (2023), HD-DEMUCS: general speech restoration with heterogeneous decoders, *Proc. Interspeech*, 2023, https://doi.org/10.21437/Interspeech.2023-1642.

[6]. **Kim S., Kim M.** (2021), Test-time adaptation toward personalized speech enhancement: zero-shot learning with knowledge distillation, arXiv:2105.03544 (also *Proc. IEEE WASPAA*, 2021), https://doi.org/10.48550/arXiv.2105.03544.

[7]. **Kim S., Athi M., Shi G., Kim M., Kristjansson T.** (2024), Zero-shot test-time adaptation via knowledge distillation for personalized speech denoising and dereverberation, *J. Acoust. Soc. Am.* **155**(2): 1353–1367, 2024, https://doi.org/10.1121/10.0024621.

[8]. **Kingma D.P., Ba J.** (2015), Adam: a method for stochastic optimization, *Proc. ICLR*, 2015, https://doi.org/10.48550/arXiv.1412.6980.

[9]. **Li L., Kang Y., Liu W., Watzel T., Rigoll G.** (2021), Lightweight end-to-end speech enhancement GAN using sinc convolutions, *Applied Sciences* **11**(16): 7564, 2021, https://doi.org/10.3390/app11167564.

[10]. **Li X., Grandvalet Y., Davoine F.** (2018), Explicit inductive bias for transfer learning with convolutional networks (L2-SP), *Proc. ICML* 80, 2018, https://doi.org/10.48550/arXiv.1802.01483.

[11]. **Liao C.-F., Tsao Y., Lee H.-Y., Wang H.-M.** (2019), Noise adaptive speech enhancement using domain adversarial training, *Proc. Interspeech*, 2019, https://doi.org/10.21437/Interspeech.2019-1519.

[12]. **Loshchilov I., Hutter F.** (2017), SGDR: stochastic gradient descent with warm restarts, *Proc. ICLR*, 2017, https://doi.org/10.48550/arXiv.1608.03983.

[13]. **Lv R., et al.** (2024), SASEGAN-TCN: speech enhancement algorithm based on self-attention GAN and temporal convolutional network, *Math. Biosci. Eng.* **21**(3): 3860–3875, 2024, https://doi.org/10.3934/mbe.2024172.

[14]. **Park Y.-S., et al.** (2023), Time-domain speech enhancement assisted by multi-resolution spectrograms, arXiv:2303.14593, https://doi.org/10.48550/arXiv.2303.14593.

[15]. **Pascual S., Bonafonte A., Serrà J.** (2017), SEGAN: speech enhancement generative adversarial network, *Proc. Interspeech*, 2017, https://doi.org/10.21437/Interspeech.2017-1428.

[16]. **Pascual S., Park M., Serrà J., Bonafonte A., Ahn K.-H.** (2018), Language and noise transfer in speech enhancement generative adversarial network, *Proc. IEEE ICASSP*, 2018, https://doi.org/10.1109/ICASSP.2018.8462322.

[17]. **Reddy C.K., Gopal V., Cutler R., et al.** (2020), The Interspeech 2020 deep noise suppression challenge, *Proc. Interspeech*, 2020, https://doi.org/10.21437/Interspeech.2020-3038.

[18]. **Rix A.W., Beerends J.G., Hollier M.P., Hekstra A.P.** (2001), Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment, *Proc. IEEE ICASSP*, 2001 (ITU-T P.862), https://doi.org/10.1109/ICASSP.2001.941023.

[19]. **Tarvainen A., Valpola H.** (2017), Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results, *Proc. NeurIPS*, 2017, https://doi.org/10.48550/arXiv.1703.01780.

[20]. **Taal C.H., Hendriks R.C., Heusdens R., Jensen J.** (2011), A short-time objective intelligibility measure for time-frequency weighted noisy speech, *Proc. IEEE ICASSP*, 2011, https://doi.org/10.1109/ICASSP.2010.5495701.

[21]. **Valentini-Botinhao C., Wang X., Takaki S., Yamagishi J.** (2017), Noisy speech database for training speech enhancement algorithms and TTS models, University of Edinburgh DataShare, 2017, https://doi.org/10.7488/ds/2117.

[22]. **Vinotha R., Hepsiba D., Vijay Anand L.D., Andrew J., Eunice R.J.** (2024), Enhancing dysarthric speech recognition through SepFormer and hierarchical attention network models with multistage transfer learning, *Sci. Reports* **14**(1): 29455, 2024, https://doi.org/10.1038/s41598-024-80764-w.

[23]. **Wang S., Li W., Siniscalchi S.M., Lee C.-H.** (2020), A cross-task transfer learning approach to adapting deep SE models to unseen background noise using paired senone classifiers, *Proc. IEEE ICASSP*, 2020, https://doi.org/10.1109/ICASSP40776.2020.9054543.

[24]. **Yamamoto R., Song E., Kim J.-M.** (2020), Parallel WaveGAN: a fast waveform generation model based on GANs with multi-resolution spectrogram, *Proc. IEEE ICASSP*, 2020, https://doi.org/10.1109/ICASSP40776.2020.9053795.

[25]. **Yu C., Zhao S., Xu B., Weng C.** (2023), High fidelity speech enhancement with band-split RNN, *Proc. Interspeech*, 2023, https://doi.org/10.21437/Interspeech.2023-1433.

[26]. **Yu C., Fu S.-W., Hsieh T.-A., Tsao Y., Ravanelli M.** (2021), OSSEM: one-shot speaker adaptive speech enhancement using meta-learning, arXiv:2111.05703, 2021, https://doi.org/10.48550/arXiv.2111.05703.

[27]. **Zhang Z., Chen J., et al.** (2024), Pre-training feature-guided diffusion model for speech enhancement, arXiv:2406.07636, 2024, https://doi.org/10.48550/arXiv.2406.07646.