

Smule Renaissance Small: Efficient General-Purpose Vocal Restoration

Yongyi Zang¹, Chris Manchester¹, David Young¹, Ivan Ivanov¹, Jeffrey Lufkin¹, Martin Vladimirov¹, PJ Solomon¹, Svetoslav Kepchelev¹, Fei Yueh Chen^{3†}, Dongting Cai^{2†}, Teodor Naydenov¹, Randal Leistikow¹

¹Smule Labs ²University of California, San Diego ³University of Rochester

[†]Work done during internship at Smule

Vocal recordings on consumer devices commonly suffer from multiple concurrent degradations: noise, reverberation, band-limiting, and clipping. We present *Smule Renaissance Small* (SRS), a compact single-stage model that performs end-to-end vocal restoration directly in the complex STFT domain. By incorporating phase-aware losses, SRS enables large analysis windows for improved frequency resolution while achieving $10.5\times$ real-time inference on iPhone 12 CPU at 48 kHz. On the DNS 5 Challenge blind set, despite no speech training, SRS outperforms a strong GAN baseline and closely matches a computationally expensive flow-matching system. To enable evaluation under realistic multi-degradation scenarios, we introduce the *Extreme Degradation Bench* (EDB): 87 singing and speech recordings captured under severe acoustic conditions. On EDB, SRS surpasses all open-source baselines on singing and matches commercial systems, while remaining competitive on speech despite no speech-specific training. We release both SRS and EDB under the MIT License.



1 Introduction

Vocal recordings captured on consumer devices often exhibit background noise, reverberation, band-limiting, and clipping. These artifacts degrade perceived quality and hinder downstream digital signal processing (DSP). While a large body of work addresses individual distortions: e.g., denoising models (Li et al., 2021; Wang et al., 2024), dereverberation models (Ernst et al., 2018; Saito et al., 2023), vocal band-width extension models (Iser and Schmidt, 2008; Wang and Wang, 2021; Li et al., 2025b), or declipping models, real-world signals typically contain several distortions simultaneously, creating train-test mismatches that limit robustness and generalization.

Recent “clean-then-synthesize” systems attempt to address compound degradations by first predicting an intermediate representation and then vocoding (Kong et al., 2020). VoiceFixer (Liu et al., 2021) predicts clean mel-spectrograms with a discriminative model and uses a GAN-trained vocoder for resynthesis; Resemble Enhance¹ similarly applies mel-domain enhancement followed by neural synthesis through latent flow matching. Although effective, two-stage designs increase computational cost, risk compounding stage-wise artifacts, and discard information when compressing to low-dimensional mel features.

We present *Smule Renaissance Small* (SRS), a compact, single-stage, end-to-end vocal restoration model that operates directly in the complex short-time Fourier transform (STFT) domain. SRS uses a band-split generator to predict complex spectrograms, coupled with a temporal-convolutional backbone augmented by SwiGLU layers to model inter-band and real-imaginary channel dependencies. An auxiliary phase-optimization loss (Li et al., 2025a) enables training with large analysis windows, improving frequency resolution and temporal efficiency while maintaining phase consistency at synthesis time. To improve robustness to real-world mixtures of artifacts, we introduce a general-purpose corruption module that stochastically perturbs both magnitude and phase in addition to targeted degradations. This design avoids mel compression, reduces opportunities for error accumulation, and yields efficient CPU inference; on an iPhone 12 CPU, SRS runs at $10.5\times$ real-time under 48 kHz.

¹<https://www.resemble.ai/introducing-resemble-enhance/>

Despite being trained without speech-specific data, SRS outperforms a stronger, GAN-trained baseline on the DNS 5 Challenge blind sets across standard metrics and is competitive with a latent flow-matching system that requires substantially more compute. A key impediment to comprehensive evaluation is the lack of realistic multi-degradation benchmarks. We therefore introduce *Extreme Degradation Bench* (EDB), a curated set of singing and speech recordings collected under diverse, severe conditions. On EDB, SRS surpasses open-source baselines and matches commercial closed-source systems on singing while exceeding a GAN-trained open-source baseline on speech. We release both SRS ² and EDB ³ under the MIT License to facilitate reproducible research.

2 Methods

2.1 Model Architecture

Input representation. Given an input mixture waveform, we compute its complex-valued Short-Time Fourier Transform (STFT) to obtain $X \in \mathbb{R}^{B \times F \times T_s \times 2}$, where the final dimension separates real and imaginary components. The generator G predicts a complex-valued estimate $\hat{X} = G(X)$ of identical shape, from which the enhanced waveform is recovered via inverse STFT: $\hat{y} = \mathcal{S}^{-1}(\hat{X})$.

The overall generator design is similar to Li and Luo (2025):

Bandwise decomposition. We partition the frequency axis into n_{band} contiguous sub-bands with mel-spaced boundaries. The integer width bw_i of each band i satisfies $\sum_{i=1}^{n_{\text{band}}} bw_i = F$. For band i , we extract the corresponding frequency slice $X_i \in \mathbb{R}^{B \times bw_i \times T_s \times 2}$ and compute a per-frame power envelope:

$$p_i(t) = \sqrt{\sum_{f \in \text{band } i} (\Re X_{f,t})^2 + (\Im X_{f,t})^2} + \varepsilon \in \mathbb{R}^{B \times 1 \times T_s}, \quad (1)$$

which serves both for normalization and as an explicit log-power feature input to the network.

Per-band feature extraction. Within each band, we normalize X_i by dividing by $p_i(t)$ (broadcast across frequency bins and complex channels), then flatten the frequency and complex dimensions into $2bw_i$ channels. We concatenate this normalized representation with $\log p_i(t)$ to form $2bw_i + 1$ input channels per time frame. A lightweight stem consisting of RMSNorm followed by 1×1 pointwise convolution projects each band to a shared feature dimension N , yielding the initial hidden state:

$$H^{(0)} \in \mathbb{R}^{B \times n_{\text{band}} \times N \times T_s}. \quad (2)$$

Band-Sequence block (repeated L times). We design a modular processing block that couples cross-band spectral modeling with within-band temporal modeling:

- **Cross-band attention.** We reshape $H^{(\ell-1)}$ to $(B \cdot T_s, N, n_{\text{band}})$ and apply multi-head self-attention along the band dimension. Queries and keys receive rotary position encoding (RoPE (Su et al., 2024)) based on band indices to preserve relative frequency ordering. We implement projections and output mixing via 1×1 pointwise convolutions, followed by a SwiGLU feedforward layer (Shazeer, 2020), with pre-norm RMSNorm and residual connections throughout.
- **Within-band temporal modeling.** We process each band independently using a depthwise-separable 1D ConvNeXT blocks (Liu et al., 2022) stack over the temporal dimension. The convolutions employ increasing dilation rates (e.g., $\{1, d, 1\}$ where d grows with network depth and is capped at a maximum value), followed by RMSNorm, a pointwise expansion with gated linear units (GLU), and a learned Layer scale parameter γ for stable training.

The outputs of both pathways are summed with the block input via a long residual connection, maintaining the representation $H^{(\ell)} \in \mathbb{R}^{B \times n_{\text{band}} \times N \times T_s}$ across layers $\ell = 1, \dots, L$.

²<https://huggingface.co/smulelabs/Smule-Renaissance-Small>

³<https://huggingface.co/datasets/smulelabs/ExtremeDegradationBench>

Per-band synthesis and spectral reassembly. Each band employs a dedicated synthesis head comprising RMSNorm $\rightarrow 1 \times 1$ conv $\rightarrow$ SiLU $\rightarrow 1 \times 1$ conv $\rightarrow$ GLU, which maps the N -dimensional latent representation to $2bw_i$ output channels. These outputs are interpreted as real and imaginary components for each frequency bin within the band. We reshape each head’s output to $(B, bw_i, T_s, 2)$ and concatenate along the frequency axis to reconstruct the full complex spectrogram $\hat{X} \in \mathbb{R}^{B \times F \times T_s \times 2}$.

Computational complexity and design rationale. By restricting attention to the band axis rather than the frequency axis, the attention mechanism incurs cost $\mathcal{O}(B T_s n_{\text{band}}^2)$ instead of $\mathcal{O}(B T_s F^2)$. Temporal dependencies are captured through dilated depthwise convolutions with linear cost in T_s . This architectural decomposition enables effective spectral–temporal modeling with modest memory requirements, while RoPE preserves relative frequency relationships. The use of per-band synthesis heads allows for specialized reconstruction at different frequency ranges without requiring a computationally expensive global decoder.

2.2 Training Setup

Input–output notation. Let $x, y \in \mathbb{R}^{B \times T}$ denote the degraded mixture and clean target waveforms, respectively. We apply STFT analysis $\mathcal{S}$ and synthesis $\mathcal{S}^{-1}$ with window size 4096 and hop size 2048 to obtain spectral representations $X = \mathcal{S}(x)$ and $Y = \mathcal{S}(y)$. The generator processes X to predict $\hat{X}$, which is converted back to the time domain as $\hat{y} = \mathcal{S}^{-1}(\hat{X})$ for comparison with y . Note that this window and hop size configuration is substantially larger than those used in most existing systems; we find that incorporating phase-aware optimization losses enables acceptable synthesis quality even with this coarse temporal resolution.

Reconstruction objective. We employ a composite reconstruction loss combining time-domain, multi-resolution spectral magnitude, and phase-aware terms:

$$\mathcal{L}_{\text{recon}} = \lambda_{\text{wav}} \|\hat{y} - y\|_1 + \lambda_{\text{spec}} \left\| |\mathcal{S}(\hat{y})| - |\mathcal{S}(y)| \right\|_1 + \lambda_{\text{omni}} \mathcal{L}_{\text{omni}}(\hat{X}, Y), \quad (3)$$

where $\mathcal{L}_{\text{omni}}$ denotes the phase-aware loss term introduced in Li et al. (2025a).

Adversarial objectives. We pair the generator G with a multi-scale discriminator D based on the Encodec architecture (Défossez et al., 2022), comprising multi-period, multi-resolution STFT, and optional multi-band discriminator branches. We modify the original implementation by replacing weight normalization with spectral normalization for improved training stability. Following the hinge loss formulation, the discriminator and adversarial losses are:

$$\mathcal{L}_D = \frac{1}{K} \sum_{k=1}^K \left(\mathbb{E} [\max(0, 1 - D_{\phi_i}^k(y))] + \mathbb{E} [\max(0, 1 + D_{\phi_i}^k(\hat{y}))] \right), \quad (4)$$

$$\mathcal{L}_{\text{adv}} = -\frac{1}{K} \sum_{k=1}^K \mathbb{E} [D_{\phi_i}^k(\hat{y})], \quad (5)$$

where K denotes the number of discriminator branches. We additionally incorporate a normalized feature-matching loss:

$$\mathcal{L}_{\text{fm}} = \frac{1}{K} \sum_{k=1}^K \frac{1}{L_k} \sum_{\ell=1}^{L_k} \frac{\|\phi^{k,\ell}(y) - \phi^{k,\ell}(\hat{y})\|_1}{\text{mean}(|\phi^{k,\ell}(y)|) + \varepsilon}, \quad (6)$$

where $\phi^{k,\ell}$ denotes the feature map at layer ℓ of discriminator k , and L_k is the number of layers in that discriminator. The total generator loss combines all objectives:

$$\mathcal{L}_G = \mathcal{L}_{\text{recon}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{fm}} \mathcal{L}_{\text{fm}}. \quad (7)$$

Optimization and regularization. We train both generator and discriminator using AdamW optimizers with shared weight decay and ϵ hyperparameters. The learning rate follows a linear warmup schedule followed by cosine decay. To ensure training stability, we apply global gradient norm clipping with a threshold of 1.0 to both networks at each optimization step.

2.3 Data

Dataset collection. Our training dataset consists of singing voice recordings collected in professional recording studios. Participants performed a standardized singing elicitation protocol ⁴, and recordings were subsequently processed by audio engineers to ensure quality. During each recording session, we captured audio from multiple microphones simultaneously; all microphone channels are utilized during training to increase data diversity.

Degradation simulation. To train the model on diverse degradation scenarios, we apply a comprehensive augmentation pipeline. We simulate frequency-dependent degradation, reverberation using parametric room models, various clipping curves, and additive environmental and instrumental noise. Additionally, we introduce stochastic perturbations in both magnitude and phase domains by randomly masking frequency bins in the magnitude spectrogram and adding noise to the phase spectrogram. These phase perturbations are computed under randomly sampled STFT parameters drawn from the following sets: window sizes $\in \{512, 1024, 2048\}$ and hop sizes $\in \{256, 512, 1024\}$. Finally, we apply time-varying gain modulation by generating random noise, applying a lowpass filter to create a smooth gain envelope, and multiplying this envelope with the audio signal to simulate realistic volume fluctuations.

3 Results

3.1 Objective Performance

	SIG	BAK	OVRL	UTMOS	Average
VoiceFixer	3.38	3.90	3.04	2.03	3.09
Resemble Enhance	3.54	3.98	3.22	2.35	3.27
SRS (Ours)	<u>3.50</u>	3.98	<u>3.18</u>	<u>2.13</u>	<u>3.20</u>

Table 1 Objective results on DNS 5 Challenge Blind Set. Best performances are in **bold** and second-best performances are underlined.

We evaluate SRS against two open-source systems on the DNS 5 Challenge (Dubey et al., 2024) blind set using predictor-based MOS metrics: DNSMOS P.835 (Reddy et al., 2022) *SIG* (speech quality), *BAK* (background intrusiveness), *OVRL* (overall quality), and *UTMOS* (Saeki et al., 2022); higher is better. Table 1 reports per-metric scores and their unweighted mean.

SRS is top-2 on all metrics, tying for the best *BAK* (3.98) and landing within 0.04 of the best system on both *SIG* (3.50 vs. 3.54) and *OVRL* (3.18 vs. 3.22). Relative to VoiceFixer, SRS improves by +0.12 *SIG* (3.50 vs. 3.38), +0.08 *BAK* (3.98 vs. 3.90), +0.14 *OVRL* (3.18 vs. 3.04), and +0.10 *UTMOS* (2.13 vs. 2.03), yielding a +0.11 gain in the average score (3.20 vs. 3.09). While Resemble Enhance attains the highest average (3.27), SRS is close behind at 3.20 ($\Delta = 0.07$) despite being a single-stage model, supporting a favorable accuracy–efficiency trade-off.

3.2 Extreme Degradation Bench (EDB)

Current vocal restoration test sets primarily consist of blind subsets from various noisy speech datasets, which typically contain limited degradation types and severity levels. To address this limitation, we propose the Extreme Degradation Bench (EDB), a benchmarking dataset comprising 87 14-second mono 48 kHz audio clips captured under diverse degradation conditions. The dataset includes samples from the UCSB Cylinder Audio Archive ⁵, as well as audio recorded in challenging acoustic environments including public transport, airports, household settings, and outdoor sports venues. EDB contains both singing and speech content across multiple languages and regions, providing comprehensive coverage of real-world degradation scenarios.

To comprehensively benchmark current system performance, we compare our proposed SRS system against the aforementioned open-source baselines and two commercial closed-source systems: Adobe Enhance V2 ⁶ and Lark

⁴Dataset manuscript is in preparation and will be released in a future publication.

⁵<https://cylinders.library.ucsb.edu/>

⁶<https://podcast.adobe.com/en/enhance>, accessed October 11, 2025 via web interface

V2⁷.

3.2.1 Subjective Ranking

We conducted a subjective ranking study in which participants evaluated 20 pairwise comparisons. For each comparison, participants heard the original degraded clip followed by outputs from two randomly selected systems, then indicated whether one system was superior or whether the systems were tied. The study ran for one week and collected responses from 34 participants. All submitted ratings were included in the analysis, even from participants who did not complete the full set of comparisons. We fitted a Bradley-Terry model to the pairwise comparison data to compute ELO scores for each system.

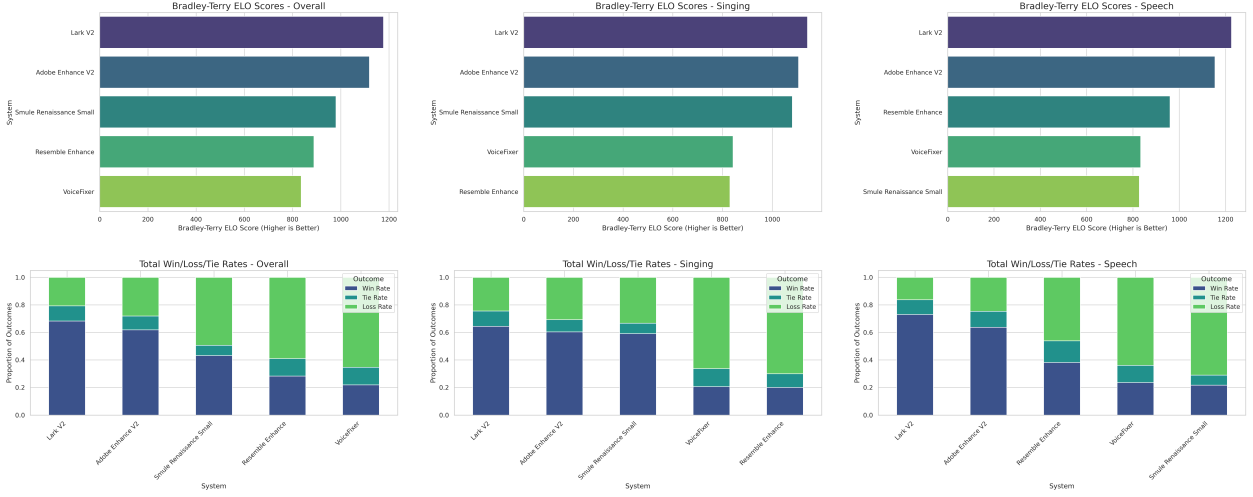


Figure 1 Overall results for all systems on Extreme Degradation Bench.

Results are shown in Figure 1. We observe that closed-source systems achieve strong performance across overall, singing, and speech categories, with Lark V2 consistently outperforming Adobe Enhance V2. Our proposed SRS system achieves the highest performance among open-source systems overall and attains performance comparable to closed-source systems on singing restoration. Notably, despite receiving no explicit training on speech data, SRS demonstrates performance on extremely degraded speech comparable to VoiceFixer, a larger GAN-based system specifically designed for speech enhancement.

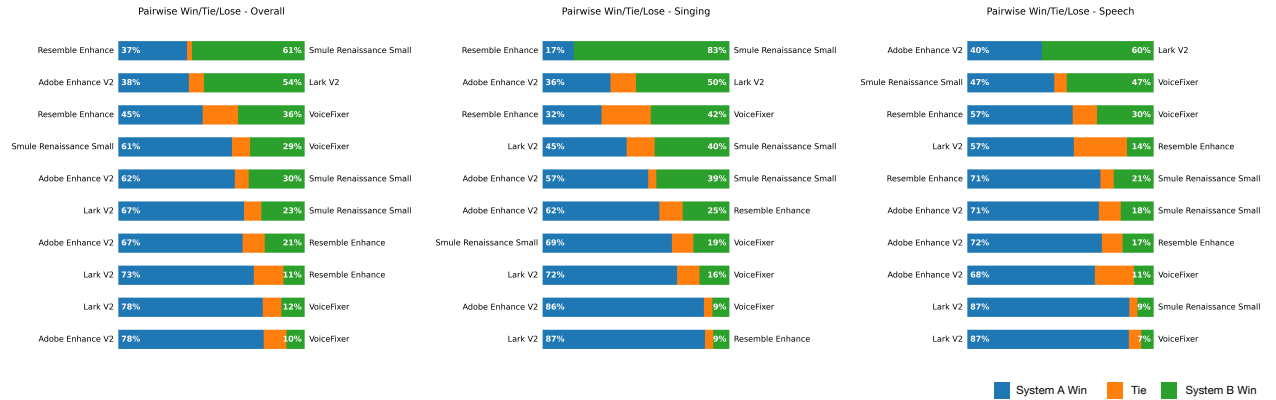


Figure 2 Pairwise win/tie/lose results for all systems on Extreme Degradation Bench.

Figure 2 presents detailed pairwise win/tie/loss rates between all systems. The results align with expected performance hierarchies: closed-source systems generally outperform open-source alternatives, with Lark V2 exhibiting a modest

⁷<https://ai-coustics.com/2025/07/29/lark-2-next-generation-reconstructive-speech-enhancement/>, accessed October 11, 2025 via official API endpoint

but consistent advantage over Adobe Enhance V2. SRS outperforms all open-source systems in both overall and singing categories. Particularly noteworthy is that SRS shows only a small performance gap compared to Lark V2 on singing restoration tasks. While SRS demonstrates relatively weaker performance on speech restoration, it remains competitive with VoiceFixer—a GAN-based network trained specifically for speech—which corroborates the overall win/loss/tie rate observations.

3.2.2 Bradley-Terry Model Validation

To verify the validity of our evaluation methodology, we assessed the goodness-of-fit of the Bradley-Terry model by computing R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) between predicted and observed pairwise comparison outcomes.

Category	R ²	MAE	RMSE
Overall	0.9540	0.0203	0.0243
Speech	0.9000	0.0298	0.0416
Singing	0.8171	0.0515	0.0591

Table 2 Goodness-of-fit metrics for Bradley-Terry model across evaluation categories.

As shown in Table 2, all categories exhibit strong model fit, with R² values exceeding 0.8 and low prediction errors, confirming that the Bradley-Terry model provides a valid and reliable framework for ranking system performance in this evaluation.

3.3 Mobile Device Performance

To evaluate the practical deployment of our system, we benchmark inference performance on consumer iOS devices using a 10-second audio input at 48 kHz sampling rate. We test on two devices representing different generations: iPhone 12 (released 2020, mid-tier) and iPhone 14 Pro (released 2022, flagship), measuring latency on both CPU and GPU accelerators⁸.

Device	Compute	Median (s)	P90 (s)	Mean (s)
iPhone 12	GPU	1.057	1.540	1.384
iPhone 12	CPU	0.948	0.970	0.952
iPhone 14 Pro	GPU	0.631	0.795	0.780
iPhone 14 Pro	CPU	0.727	0.739	0.731

Table 3 Inference latency on mobile devices for processing 10 seconds of audio. Median, 90th percentile (P90), and mean values computed over repeated runs.

Results are presented in Table 3. On the iPhone 12, CPU execution achieves faster and more consistent performance (0.948s median) compared to GPU (1.057s median). We attribute the slower GPU performance to cold start overhead between CPU and GPU on this older device architecture. Nevertheless, even on this five-year-old consumer device, the system achieves approximately 10.5× real-time processing speed⁹.

On the more recent iPhone 14 Pro, both accelerators demonstrate improved performance. The GPU achieves a median latency of 0.631s, corresponding to 15.8× real-time speed, while the CPU maintains competitive performance at 0.727s (13.7× real-time). The improved GPU performance on iPhone 14 Pro suggests that newer device architectures have reduced memory transfer bottlenecks, making GPU acceleration more effective.

These results demonstrate that our system achieves practical real-time performance across multiple generations of consumer mobile hardware, validating its deployment viability for on-device audio restoration applications.

⁸Benchmarks conducted using Xcode’s built-in profiling tools for MLPackage deployment.

⁹Note that the system is not causal, and therefore cannot actually be used in real-time.

4 Conclusion

We presented Smule Renaissance Small (SRS), a compact single-stage vocal restoration system operating directly in the complex STFT domain. By combining band-split architecture with phase-aware losses, SRS achieves effective restoration with $10.5\times$ real-time inference on iPhone 12 CPU, demonstrating practical viability for on-device deployment. Our evaluations show that SRS outperforms strong GAN baselines and approaches more expensive flow-matching systems on DNS 5 Challenge—despite no speech training. On the Extreme Degradation Bench, SRS surpasses all open-source alternatives on singing and matches commercial systems, while remaining competitive on speech without speech-specific training. We release both SRS and EDB under the MIT License to facilitate reproducible research in robust vocal restoration.

References

- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.
- Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Mehrsa Golestaneh, et al. Icacss 2023 deep noise suppression challenge. *IEEE Open Journal of Signal Processing*, 5:725–737, 2024.
- Ori Ernst, Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger. Speech dereverberation using fully convolutional networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 390–394. IEEE, 2018.
- Bernd Iser and Gerhard Schmidt. Bandwidth extension of telephony speech. In *Speech and Audio Processing in Adverse Environments*, pages 135–184. Springer, 2008.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.
- Andong Li, Wenzhe Liu, Chengshi Zheng, Cunhang Fan, and Xiaodong Li. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1829–1843, 2021.
- Andong Li, Tong Lei, Zhihang Sun, Rilin Chen, Erwei Yin, Xiaodong Li, and Chengshi Zheng. Learning neural vocoder from range-null space decomposition. *arXiv preprint arXiv:2507.20731*, 2025a.
- Chang Li, Zehua Chen, Liyuan Wang, and Jun Zhu. Audio super-resolution with latent bridge models. *arXiv preprint arXiv:2509.17609*, 2025b.
- Kai Li and Yi Luo. Apollo: Band-sequence modeling for high-quality audio restoration. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
- Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. Voicefixer: Toward general speech restoration with neural vocoder. *arXiv preprint arXiv:2109.13731*, 2021.
- Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.
- Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 886–890. IEEE, 2022.
- Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.
- Koichi Saito, Naoki Murata, Toshimitsu Uesaka, Chieh-Hsin Lai, Yuhta Takida, Takao Fukui, and Yuki Mitsufuji. Unsupervised vocal dereverberation with diffusion-based generative models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- Heming Wang and DeLiang Wang. Towards robust speech super-resolution. *IEEE/ACM transactions on audio, speech, and language processing*, 29:2058–2066, 2021.
- Ju-Chiang Wang, Wei-Tsung Lu, and Jitong Chen. Mel-roformer for vocal separation and vocal melody transcription. *arXiv preprint arXiv:2409.04702*, 2024.