TL;DR Speech denoising now spans a family of diffusion-flavoured designs. StoRM blends a predictive estimate with a diffusion sampler to tame hallucinations at low cost; SGMSE/SGMSE+ continue to scale score matching with variance-aware schedulers; UNIVERSE++ bakes in adversarial loss and low-rank adaptation for cross-condition robustness; few-step Schrodinger-Bridge variants target sub-10 step inference; causal diffusion architectures chase streaming deployment; and MossFormer2 remains a strong baseline when you can tolerate separation-first latency.

Why teams care in 2025

- Customer support, call centers, and meeting tooling now demand *universal* denoisers that handle noise, reverberation, codec artifacts, and far-field setups without per-domain tuning.
- Real-time AI voice agents (telephony, kiosks, wearables) force inference budgets down to single-digit diffusion steps -- or hybrids that drop to predictive streams when necessary.
- Evaluation shifted from "cleaner spectrograms" to intelligibility (STOI/SI-SDR), MOS (subjective or DNSMOS), and downstream ASR WER. Modern stacks must show gains across *all*.

The models, at a glance

Model	Key idea	Step count	Notable metrics	Where it shines
StoRM (Lemercier et al., IEEE/ACM TASLP 2023)	Predictive network provides a guided starting point for the diffusion sampler, suppressing breathing/phonation artifacts.	8-30 (configurable)	VoiceBank+DEMAND PESQ >= 2.9 with 8 steps; DNSMOS better than pure score-matching at same budget.	Low-latency deployments that still need diffusion-grade quality.
SGMSE / SGMSE+ (Richter et al., 2022; Lay & Gerkmann, 2024)	Score-based generative speech enhancement with SDEs; SGMSE+ adds a stronger UNet, variance-aware	30-60 (base), 10-20 (aggressive sampler)	Cross-dataset SI-SDR up to 11-12 dB; variance scaling trades noise suppression vs. speech distortion.	Studio/production pipelines that can batch inference and want controllable trade-offs.

	schedules, and dereverb handling.			
UNIVERSE++ (Scheibler et al., Interspeech 2024)	Hybrid universal enhancer: diffusion backbone + adversarial critic + LoRA-style adaptation for phoneme fidelity.	12-20	On DNS Challenge and WHAMR! sets, beats discriminative baselines in PESQ/STOI while preserving content.	Enterprise "single model" deployments covering noise, dereverb, compression.
Few-step Schrodinger-Bridge variants (2024-2025)	Consistency and bridge-based solvers (e.g., SE-Bridge, ICASSP/NeurIPS 2025 follow-ups) collapse 30-step samplers to ~4-8 evaluations via deterministic flows.	4-8	VoiceBank PESQ ~ 3.0 with 5 steps; MOS parity with longer samplers when paired with bridge consistency loss.	Latency-critical ASR front-ends, embedded devices.
Causal/Streaming diffusion (2024-2025 prototypes)	Chunked diffusion with causal convolutions, state caching, and look-ahead gating to keep under 40 ms algorithmic delay.	4-12 per chunk	16 kHz causal pipelines hitting DNSMOS >= 3.5 and RTF under 0.5 on laptop CPU.	Live voice agents, conferencing, cloud-to-edge streaming.
MossFormer2 (Zhao et al., ICASSP 2024)	Transformer + FSMN hybrid separation. Not diffusion, but pairs well as a front/pass for denoising or residual suppression.	Single forward pass	WSJ0-2/3mix SI-SDR > 20 dB; decent denoising when retrained on noisy mixtures.	Legacy pipelines, cascades (separate -> denoise), low-compute fallbacks.

StoRM -- diffusion guided by a predictive estimate

- **Architecture**: predictive enhancer (e.g., complex spectral mapping) plus diffusion score model. Predictive output seeds the reverse diffusion, cutting hallucinated breathing noises seen in unconditional samplers.
- **Sampling**: accepts fewer function evaluations (e.g., 8-16 vs. 100+) while maintaining MOS. Suitable for GPU and optimized CPU inference.
- **Production notes**:

    - Pair with noise-classifier gating: run predictive-only when SNR is already high, invoke diffusion only when needed.
    - Monitor "guide mismatch": if the predictive output misses entire phonemes, diffusion may overfit to the incorrect guide -- run confidence checks (entropy/energy) before sampling.

SGMSE and SGMSE+

- **Score-based diffusion** in the STFT domain; reverse process starts from noisy speech rather than pure Gaussian noise.
- **SGMSE+ upgrades**:
    - Wider UNet with cross-band attention for dereverberation.
    - Variance schedule tuning: larger variance -> stronger noise suppression but more speech smoothing; smaller variance preserves transients.
    - Cross-corpus robustness demonstrated on VoiceBank, WHAMR!, and in-the-wild recordings.
- **Ops guidance**:
    - Keep a dual-sampler setup: 30-step high quality for offline, 12-step fast mode for real-time.
    - Integrate DNSMOS or MOSNet monitoring to auto-switch step count.

UNIVERSE++

- **Decoupled feature extractor + diffusion**: adversarial loss stabilizes high-frequency detail while diffusion handles coarse structure.
- **LoRA-style adaptation** allows per-customer fine-tuning without re-training the base model.
- **Phoneme fidelity loss**: ensures enhanced speech stays aligned for ASR/TTS.
- **Deployment tips**:
    - Maintain a library of low-rank adapters (e.g., meeting rooms, vehicles). Swap adapters dynamically based on environment classifiers.
    - Expect higher GPU memory use (extra critic). For CPU batches, freeze the critic during inference and prune LoRA ranks.

Few-step Schrodinger-Bridge denoisers

- **Consistency models + bridges** (e.g., SE-Bridge) learn deterministic flow matching between clean and noisy distributions.

- 2025 iterations leverage **Schrodinger bridges with amortised solvers**, landing 4-8 inference steps without adversarial training.
- **Strengths**: near-diffusion quality with autoregressive speeds; robust to step mis-specification.
- **Considerations**:
    - Sensitive to forward noise model mismatch -- keep an online noise estimator (e.g., non-stationary SNR tracker) to adjust bridge endpoints.
    - For far-field reverberation, combine with a dereverb pre-filter or train with multi-condition noise to avoid residual tails.

Causal diffusion for streaming

- Research prototypes (ICASSP & Interspeech 2024/2025) reorganize diffusion as **causal convolutional blocks with recurrent state cache** and limited look-ahead (no more than 20 ms).
- **Techniques in play**:
    - Parallelizable causal convolutions instead of global UNet skip connections.
    - Frame-wise conditioning using noise decoupling (predictive front-end + diffusion refinement per chunk).
    - Curriculum training with progressively shorter context windows to reduce drift.
- **Rolling out in production**:
    - Budget CPU-friendly kernels: depthwise separable or low-rank convs to hit RTF no greater than 0.3.
    - Combine with voice activity detection to skip diffusion on silence segments.
    - Provide fallback to predictive-only mode during CPU spikes.

MossFormer2 as a complementary tool

- **Hybrid stack**: MossFormer2 inserts FSMN-style memory into transformer blocks, covering both long/short dependencies.
- **Why mention it in denoising?** When trained on noisy mixtures, MossFormer2 can output a "mostly clean" track quickly. Use it to pre-condition diffusion models or as a fast approximate cleanup where diffusion is overkill.
- **Limitations**: Not state-of-the-art on heavy babble noise; struggles with non-stationary backgrounds compared to diffusion models.

Putting it together -- choosing a stack

1. **Latency budget under 50 ms**: Start with StoRM (8 steps) or a few-step bridge model. Add predictive bypass for high-SNR frames.
2. **All-rounder desktop or cloud**: UNIVERSE++ with adapter bank; fall back to SGMSE+ fast sampler when critic has not been specialised.
3. **Streaming voice agent**: Causal diffusion + VAD gating; optionally run MossFormer2 front pass to stabilise speech before diffusion.
4. **Batch studio cleanup**: Full SGMSE+ or UNIVERSE++ 30-step sampler for maximum perceived quality.

Evaluation checklist

- Track DNSMOS, STOI, SI-SDR, and downstream ASR WER. Diffusion models can boost MOS but hurt ASR if phoneme timing drifts.
- Measure computational load: record real-time factor across CPU and GPU targets; monitor GPU memory when loading adapters.
- Include hard cases: clipped inputs, codec artifacts (VoIP), non-stationary backgrounds (sirens), and reverberant rooms.
- Run user studies when targeting customer-facing audio; generative models can sound "too clean" compared to human expectations.

Practical deployment tips

- Use **noise classifiers** to dispatch between predictive-only, hybrid (StoRM/bridge), and full diffusion.
- Cache intermediate embeddings for streaming use-cases (e.g., StoRM predictive output or MossFormer2 latent) to warm-start subsequent chunks.
- Keep fine-tuning loops lightweight: prefer LoRA/adapter-based updates (as in UNIVERSE++) over full retraining.
- Budget observability: log per-frame confidence, diffusion step count, and VAD decisions for postmortems.

References

- StoRM (GitHub)
- StoRM (arXiv)
- SGMSE (GitHub)
- SGMSE (arXiv)
- open-universe (GitHub)
- UNIVERSE++ (arXiv)

- MossFormer2 (GitHub)
- MossFormer2 (arXiv)

CTA: