Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Accelerating Diffusion-Based Speech Denoising with DDIM: A Cross-Lingual Evaluation of DiffWave on English and Arabic Speech

Abderrahmane Adjila[1, 2,*], Maamar Ahfir[3], Meriem Laouar[1], Sarra Zahouani[1]

[1]*Dept. of Mathematics and Computer Science, Lab. des Mathematiques et Sciences Appliqu ées (LMSA), Université de Ghardaia, Ghardaia, Algeria.*

[2]*Lab. d'Informatique et de Mathematiques (LIM), Amar Telidji University, Laghouat, Algeria.*

[3]*Dept. of Computer Science, Amar Telidji University, Laghouat, Algeria.*

**Corresponding Author:*

ARTICLEINFO

Received: 15 May 2026

Revised: 03 July 2026

Accepted: 11 July 2026

ABSTRACT

Speech denoising is essential for reliable human-machine interaction, yet diffusion-based generative models, despite strong reconstruction quality, remain limited by high inference latency and near-exclusive evaluation on English speech. This study adapts the DiffWave architecture to speech denoising, compares two diffusion-step choices (T=50, T=200), accelerates inference with DDIM, and evaluates cross-lingual generalization to Arabic speech. Models were trained on the SC09 dataset and evaluated on SC09, Arabic Speech Commands, and LJSpeech, using 14 noise types (6 synthetic, 8 real) at three SNR levels (-5, 0, 5 dB), and six metrics: SNR, SI-SDR, PESQ, STOI, MOS, and RTF. The results showed that T=50 outperformed T=200 on all six metrics (SI-SDR: 10.93 vs 9.94 dB) due to longer training convergence. Under the full 14-noise/3-SNR benchmark, the model achieved its best overall performance on Arabic speech (SI-SDR = 3.94 dB, success rate = 71.2%), ahead of SC09 (3.37 dB, 69.8%) and LJSpeech (1.63 dB, 59.0%). DDIM reduced sampling from 50 to 3 steps, reaching an RTF of 0.0732 (13.7× real-time) with only a -0.64 dB SI-SDR loss on Arabic, while matching or exceeding DDPM on PESQ and STOI. As a conclusion DDIM-accelerated DiffWave offers a practical balance between denoising quality and computational efficiency, and generalizes to Arabic despite training exclusively on English digits, supporting its use in real-time, cross-lingual speech enhancement applications.

Keywords: Speech Denoising; Speech Enhancement; Diffusion Models; DDPM; DDIM; DiffWave; Cross-Lingual Generalization; Real-Time Factor.

INTRODUCTION

Lorem Speech is the primary channel of human communication, yet in real-world environments it is almost always corrupted by background noise, reverberation, and electronic interference, which degrades intelligibility and undermines the performance of downstream systems such as automatic speech recognition, telecommunications, and audio streaming [1,2]. Speech denoising therefore remains a central problem in audio signal processing, aiming to recover clean speech while preserving its naturalness.

Speech enhancement research has evolved through three broad generations of methods, surveyed in detail in Section 2: classical statistical and spectral-domain estimators, deep neural network mapping/masking approaches, and, most recently, deep generative models. Among generative approaches, diffusion probabilistic models (DDPM) have emerged as a particularly promising direction: a forward process gradually corrupts clean speech with Gaussian noise, and a learned reverse process inverts this corruption step by step [3]. Denoising Diffusion Implicit Models (DDIM) subsequently introduced a non-Markovian sampling formulation that can shorten the reverse

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

trajectory dramatically, a property critical for making diffusion-based denoising tractable in real time [4].Despite this promise, two gaps limit the practical adoption of diffusion-based speech denoising. First, the iterative reverse process incurs a computational cost and inference latency that is difficult to reconcile with real-time deployment. Second, prior work has focused almost exclusively on English speech, leaving other languages—Arabic in particular—essentially unstudied in this context. This paper addresses both gaps. Building on the DiffWave architecture, originally proposed for high-fidelity audio synthesis [5], we adapt it to unconditional speech denoising, systematically compare two diffusion-step budgets (T = 50 and T = 200), accelerate inference using DDIM, and provide, to our knowledge, the first quantitative evaluation of a diffusion-based denoiser on Arabic speech.

STATE OF THE ART

2.1 Traditional Statistical and Spectral-Domain Methods

Traditional monaural speech enhancement methods rely on explicit statistical assumptions about speech and noise. Spectral-domain estimators operate on the noisy short-time Fourier transform Y(k,l) = S(k,l) + D(k,l) and apply a time-frequency gain to recover the clean spectrum [6]. Spectral subtraction directly subtracts an estimated noise spectrum from the noisy magnitude [7], while Wiener filtering derives a gain from the a priori SNR to minimize mean-square error; MMSE-based estimators (MMSE-STSA, MMSE-LSA) further refine this under Gaussian or super-Gaussian speech models [6]. A persistent challenge for all of these methods is accurate a priori SNR estimation: the widely used decision-directed approach controls musical-noise artifacts but introduces a one-frame delay bias, motivating refinements such as two-step noise reduction and harmonic regeneration noise reduction [8]. Time-domain alternatives (adaptive LMS/NLMS filtering, subspace methods such as KLT-based filtering) avoid the STFT step but are more sensitive to noise non-stationarity. Overall, traditional methods remain simple and interpretable but degrade markedly under non-stationary, colored, or speech-correlated noise, motivating the shift toward data-driven approaches [9].

2.2 Machine Learning and Deep Learning Approaches

Machine learning and, more specifically, deep learning [10] have achieved broad success across data-driven domains [11], and speech enhancement is no exception. Conventional machine learning methods introduced data-driven components while retaining much of the traditional pipeline structure: supervised non-negative matrix factorization (NMF) and dictionary/codebook methods (e.g., K-SVD) learn basis spectra for speech and noise from training data and reconstruct clean speech via sparse coding, outperforming purely heuristic filters but remaining limited by the linear nature of the decomposition and a finite dictionary size [6]. Hybrid systems such as DeepXi use a network solely to estimate the a priori SNR, which is then plugged into a classical MMSE-LSA estimator [6].

Deep neural networks subsequently became the dominant paradigm, learning direct, non-linear mappings between noisy and clean speech representations [12]. Fully connected DNNs were followed by recurrent architectures (RNN/LSTM) that better capture the temporal dynamics of speech, and convolutional networks (CNNs) that efficiently model local spectral patterns; the Convolutional Recurrent Network (CRN) combines both strengths. A further refinement came with complex-valued architectures such as the Deep Complex Convolutional Recurrent Network (DCCRN), which jointly estimate the real and imaginary STFT components and thereby recover phase information that earlier magnitude-only systems discarded, achieving state-of-the-art results in benchmarks such as the Deep Noise Suppression (DNS) Challenge. These systems are trained either to predict a time-frequency mask (Ideal Ratio Mask or Complex Ratio Mask) applied to the noisy spectrogram, or to directly map noisy input to clean output, typically under composite losses combining spectral MSE with time-domain SI-SNR or perceptual objectives such as deep feature losses [13].

2.3 Generative Models for Speech Enhancement

Generative models depart from direct mapping/masking by learning the underlying distribution of clean speech, then sampling or reconstructing high-fidelity speech from a noisy observation. Generative Adversarial Networks (GANs) were an early and influential direction: SEGAN performs end-to-end enhancement directly on raw waveforms via an adversarial generator/discriminator pair [14], later extended by conditional GAN variants that

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

explicitly condition generation on the noisy input [15], and by MetricGAN, which replaces the standard discriminator with a critic trained to predict a target perceptual metric so the generator can directly optimize objective quality scores [16]. Waveform-domain autoregressive models such as WaveNet-based denoisers use dilated causal convolutions to model long-range temporal dependencies directly on raw samples [17].

Diffusion probabilistic models represent the most recent generative paradigm and currently define the state of the art for speech enhancement quality. DDPM defines a forward process that progressively corrupts clean speech with Gaussian noise and trains a network to predict the added noise so a reverse process can iteratively reconstruct clean speech [3]. Score-based variants apply this framework directly in the complex STFT domain, learning the gradient of the data log-density and sampling via predictor-corrector schemes such as annealed Langevin dynamics [18], with extensions that jointly perform denoising and dereverberation [19]. A central practical drawback of standard DDPM is the need for many sequential reverse steps (often 50–1000), which is incompatible with low-latency deployment. DDIM addresses this by replacing the stochastic Markovian reverse chain with a deterministic, non-Markovian sampling process capable of skipping steps while remaining consistent with the same trained model, enabling substantial inference speedups without retraining [4]. Among diffusion architectures, DiffWave [5] is notable for operating directly on raw audio waveforms via bidirectional dilated convolutions; although originally proposed for speech synthesis and unconditional generation, its authors demonstrated a zero-shot denoising capability without any dedicated denoising training—evaluated only qualitatively, with no reported objective metrics.

2.4 Positioning of This Work

Table 1 summarizes this progression. Relative to this literature, the present work makes two specific contributions. First, whereas the original DiffWave study reported only qualitative zero-shot denoising results, we provide the first quantitative, task-specific evaluation of DiffWave for speech denoising (SNR, SI-SDR, PESQ, STOI) across a systematic 14-noise-type × 3-SNR-level benchmark, and we further accelerate it with DDIM—whose speed-quality trade-off had not previously been characterized for this architecture. Second, while the diffusion-based speech-enhancement literature (score-based STFT models, DiffWave itself) has been developed and validated almost exclusively on English corpora, we provide, to our knowledge, the first quantitative evaluation of a diffusion-based denoiser on Arabic speech, directly probing whether representations learned from English digits transfer to a phonetically distinct language.

Family	Representative methods	Key idea	Main limitation
Statistical / spectral-domain	Spectral subtraction, Wiener filter, MMSE-STSA/LSA	Gain function from assumed speech/noise statistics	Degrades under non-stationary, speech-correlated noise
Conventional ML	NMF, K-SVD, DeepXi (a priori SNR)	Learned bases/dictionaries or hybrid parameter estimation	Limited capacity, poor generalization to unseen noise
Deep neural networks	DNN, CNN, RNN/LSTM, CRN, DCCRN	Learned non-linear mapping/masking, phase-aware (DCCRN)	Large data requirements; still largely discriminative
GAN-based generative	SEGAN, conditional GAN, MetricGAN	Adversarial generator/discriminator, metric-optimized critic	Training instability, model collapse risk

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Family	Representative methods	Key idea	Main limitation
Diffusion-based generative	Score-based STFT models, DiffWave, DDPM/DDIM	Iterative learned denoising of a forward noising process	High inference latency (DDPM); mainly English-only evaluation

Table 1. Summary of the methodological evolution of monaural speech denoising.

OBJECTIVES

The specific objectives of this work are to:

- Establish the first quantitative (SNR, SI-SDR, PESQ, STOI) evaluation of the DiffWave architecture for speech denoising, across 14 noise types and 3 SNR levels.
- Compare two diffusion-step configurations, T=50 and T=200, and identify the configuration that generalizes best to unseen speech.
- Accelerate inference using DDIM and quantify the resulting trade-off between denoising quality and real-time factor (RTF).
- Assess cross-lingual generalization of a model trained exclusively on English digits when evaluated on Arabic speech.
- Validate the approach on real-world recordings captured outside controlled laboratory conditions.

METHODS

4.1 Model

DiffWave [5] is a diffusion probabilistic model built on a bidirectional dilated convolutional architecture with residual connections. Its forward process corrupts a clean waveform x0 over T steps following a variance schedule βt, and a neural network εθ is trained to predict the added noise so that the reverse process can iteratively reconstruct the clean signal. Unlike the original text-to-speech application of DiffWave, this study conditions the model on noisy speech and trains it directly for denoising. Two diffusion-step choices were trained and compared: T = 50 and T = 200. Following training, DDIM [4] was applied post hoc to the T = 50 model to enable non-Markovian, deterministic sampling with as few as 3 reverse steps, without any retraining.

4.2 Datasets and Noise Protocol

Both DiffWave configurations were trained exclusively on the SC09 subset of the Speech Commands dataset [20] (spoken English digits), using additive white noise only. Trained models were then evaluated, without any further fine-tuning, on 100 unseen files from each of three speech corpora: SC09 (isolated English digits), the Arabic Speech Commands dataset [21] (isolated Arabic digits), and LJSpeech [22] (continuous English sentences). For the full benchmark, each clean file was corrupted with 14 noise types—six synthetic profiles (white, pink, brownian/running-tap, exercise-bike, dude-miaowing, doing-the-dishes) and eight real-world ambient recordings from the Arabic Speech Commands background library (battery-charger, boiler, cooking-1, cooking-2, street, walking-machine, washing-machine, water-tap)—at three SNR levels (−5, 0, and 5 dB). All audio was resampled to 16 kHz, clipped/padded to 1 second, and amplitude-normalized to [−1, 1].

4.3 Evaluation Metrics

Performance was quantified using six complementary metrics: Signal-to-Noise Ratio (SNR) and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), which measure energy-based signal fidelity; Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI), which approximate human perceptual quality and intelligibility; a Mean Opinion Score (MOS) proxy; and the Real-Time Factor (RTF, processing time divided by audio duration), which quantifies computational efficiency. A test file was labelled a “success” when the output SI-

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SDR exceeded the input SNR, i.e., when the model produced a net improvement. All experiments were run on a single NVIDIA Tesla T4 GPU (Google Colaboratory).

RESULTS

5.1 Selection of the Diffusion-Step Budget: T=50 vs T=200

Table 2 summarizes the performance of the two trained configurations on 100 unseen files corrupted with white noise at 0 dB SNR. Although T=200 provides a finer discretization of the noise schedule, T=50 outperformed it on all six metrics. This outcome is explained by training dynamics rather than architectural capacity: because each T=50 iteration is computationally cheaper, this model completed 239,598 training iterations (converging at step 165,648, loss 0.0180), whereas T=200 completed only 65,076 iterations before its lowest local loss (0.0115) at step 41,412. The lower nominal loss of T=200 therefore reflects premature, under-converged optimization rather than a genuinely better fit, while the more extensively trained T=50 network generalized better to unseen corrupted audio.

Metric	T=200	T=50
Best inference step (t)	8	5
SI-SDR (dB)	9.941 ± 2.91	10.929 ± 2.79
SNR (dB)	9.350 ± 2.20	9.497 ± 2.08
PESQ	1.389 ± 0.32	1.406 ± 0.24
STOI	0.717 ± 0.25	0.728 ± 0.25
MOS	2.235 ± 0.29	2.250 ± 0.22
Execution time (s)	0.199 ± 0.00	0.127 ± 0.01

Table 2. Performance of T=200 vs. T=50 on 100 unseen files, white noise, SNR = 0 dB.

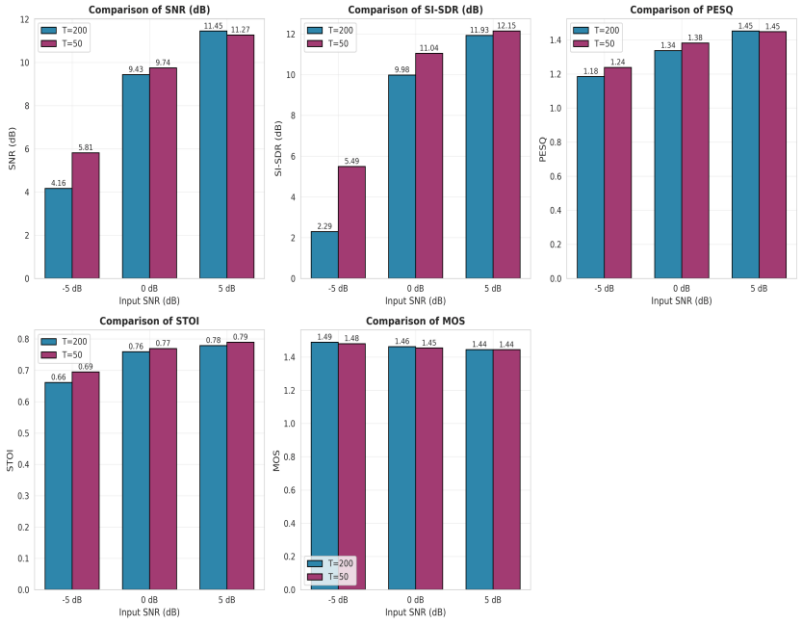


Figure 1. Performance comparison between T=200 and T=50 across SNR, SI-SDR, PESQ, and STOI.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

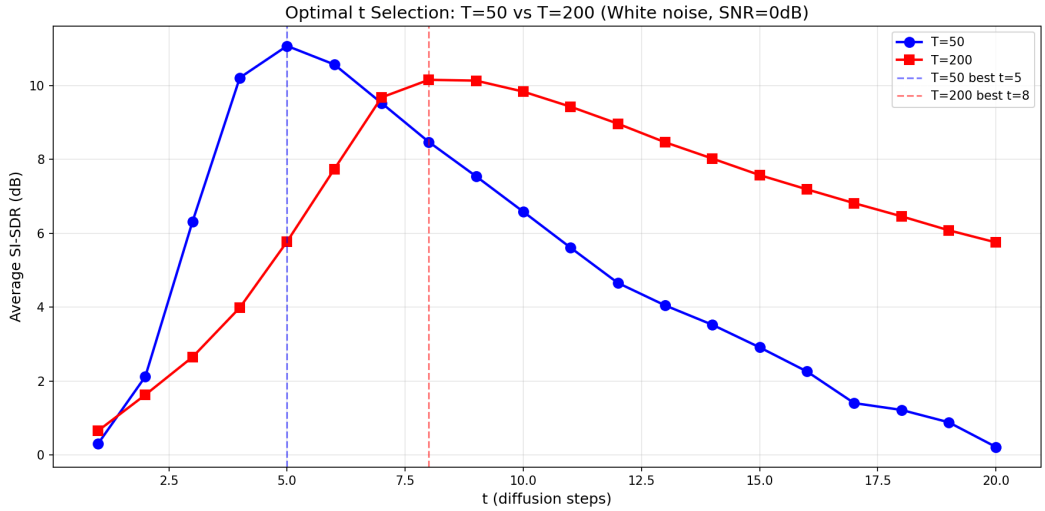


Figure 2. Optimal inference-step (t) selection for T=50 (peak at t=5) and T=200 (peak at t=8).

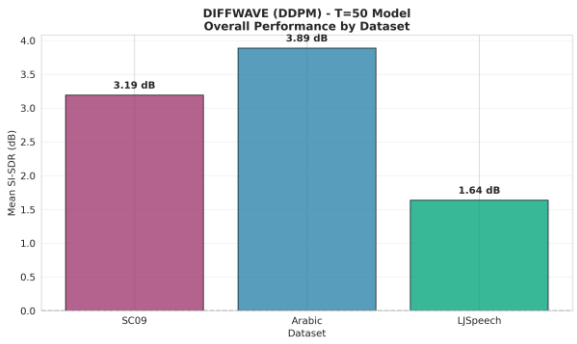
Beyond the aggregate metrics, T=50 reaches its optimal SI-SDR at an earlier inference step (t=5) than T=200 (t=8, Figure 2), consistent with its coarser noise schedule requiring fewer reverse steps to reach an equivalent noise magnitude. T=50 also processes a 1-second frame 1.57× faster (0.127 s vs 0.199 s). Given its clean sweep across all six metrics and its lower latency, T=50 with t=5 was retained as the core configuration for all subsequent experiments.

5.2 Cross-Dataset and Cross-Lingual Generalization (DDPM)

The optimal inference step (t=5) was confirmed independently on SC09, Arabic, and LJSpeech under white noise at 0 dB, all three peaking at t=5. The T=50 model, trained only on SC09 with synthetic noise, was then evaluated under the full 14-noise/3-SNR protocol. Table 3 and Figure 3 report the resulting overall performance.

Dataset	SI-SDR (dB)	PESQ	STOI	MOS	RTF	Success rate
SC09 (English digits)	3.37	1.363	0.666	2.23	0.1223	69.8%
Arabic	3.94	1.354	0.699	2.20	0.1225	71.2%
LJSpeech (English sentences)	1.63	1.263	0.742	2.12	0.1226	59.0%

Table 3. Overall DDPM (T=50) performance across datasets, 14 noise types × 3 SNR levels.



Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 3. Mean SI-SDR by dataset (DDPM, T=50).

Despite being trained exclusively on English digits, the model achieved its best overall performance on Arabic speech, both in SI-SDR (3.94 dB) and success rate (71.2%), ahead of SC09 (3.37 dB, 69.8%) and LJSpeech (1.63 dB, 59.0%). Performance scaled consistently with input SNR across all datasets (Table 4): at 5 dB, all three datasets reached a 100% success rate, while at −5 dB, continuous LJSpeech speech proved markedly harder to recover (24.8% success) than isolated digits, indicating that continuous speech is more vulnerable to severe noise than isolated-word utterances.

Dataset	SNR = −5 dB	SNR = 0 dB	SNR = 5 dB
SC09	-2.53	3.64	8.47
Arabic	-2.13	4.29	9.50
LJSpeech	-4.15	2.28	6.77

Table 4. SI-SDR (dB) by dataset and input SNR level (DDPM, T=50).

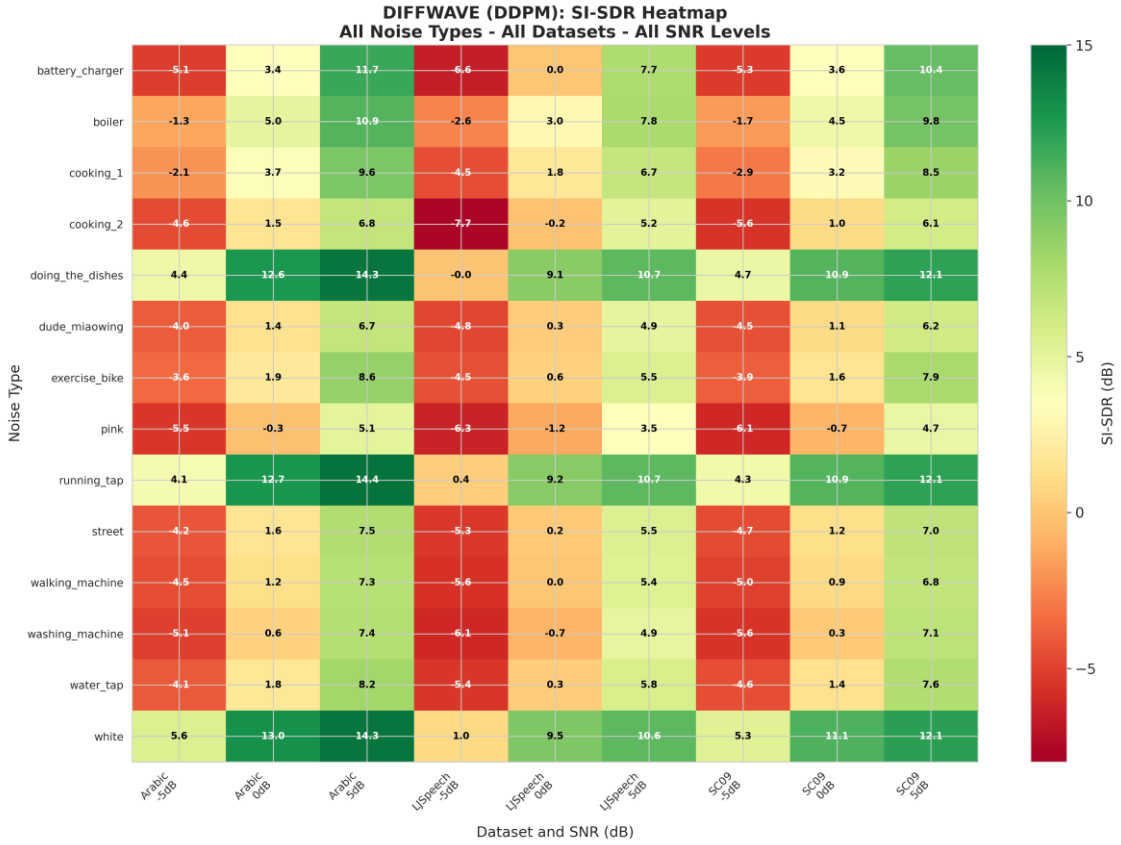


Figure 4. Complete SI-SDR heatmap across all 14 noise types, three datasets, and three SNR levels (DDPM, T=50).

Figure 5 breaks down PESQ and STOI in the same 14-noise/3-SNR benchmark. Perceptual quality (PESQ) improves monotonically with input SNR for all three datasets, while intelligibility (STOI) is already comparatively high at −5 dB and saturates near ceiling by 5 dB, indicating that the model restores intelligibility earlier, along the denoising trajectory, than it restores perceptual naturalness.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

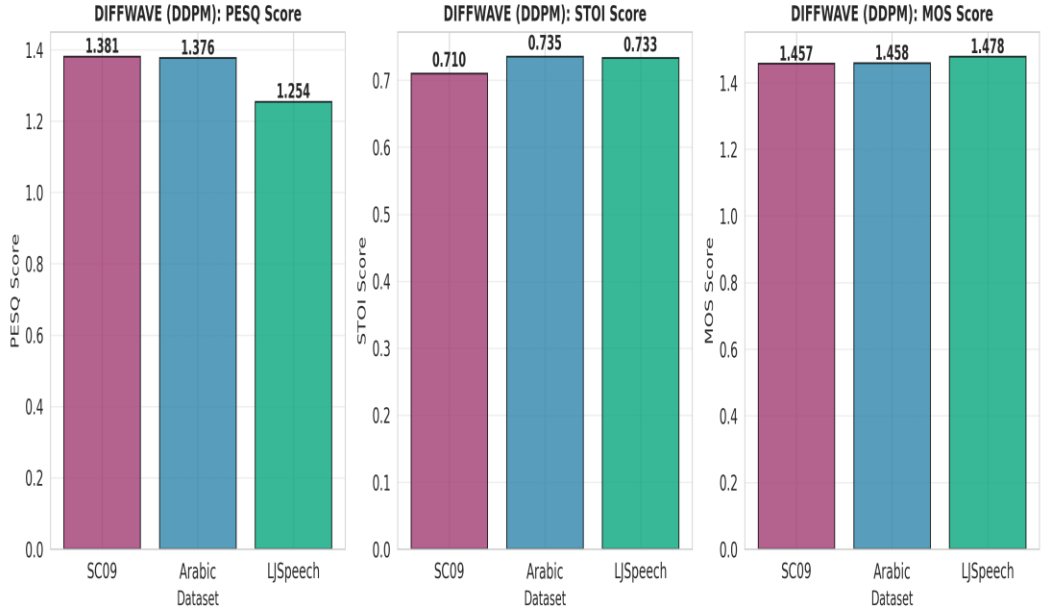


Figure 5. PESQ and STOI by dataset and input SNR level (DDPM, T=50).

Success rate, defined as the fraction of files for which output SI-SDR exceeded input SNR, follows the same dataset ranking (Figure 6): Arabic achieves the highest success rate at every SNR level, and the gap between datasets widens as SNR decreases, showing that the model's cross-lingual advantage on Arabic is most pronounced precisely in the hardest, low-SNR conditions.

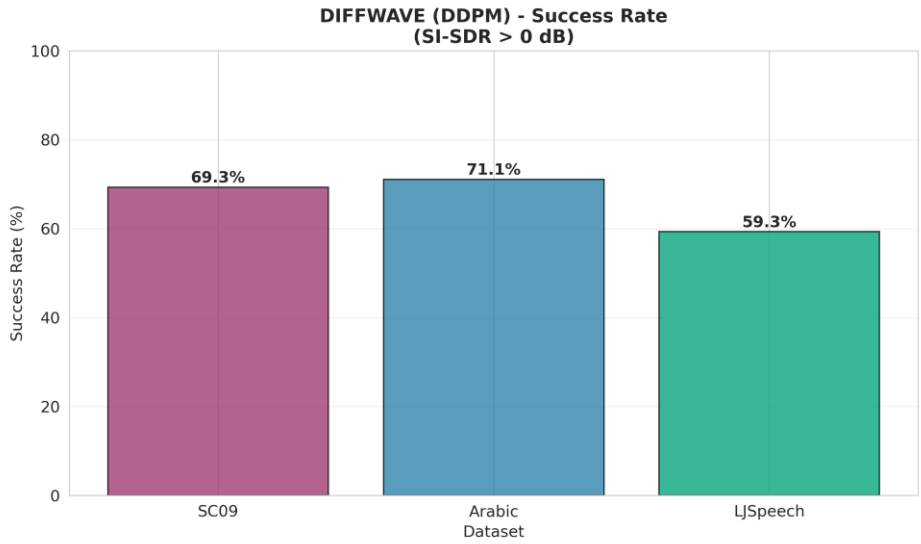


Figure 6. Success rate (output SI-SDR > input SNR) by dataset and SNR level (DDPM, T=50).

5.3 Noise-Type Sensitivity

At SNR = 0 dB, the same three noise types were consistently the easiest to remove across all datasets—white noise, running-tap, and doing-the-dishes—while pink noise, washing-machine, and walking-machine were consistently the hardest, with pink noise even yielding a slightly negative SI-SDR for SC09 (−0.6 dB) and LJSpeech (−1.2 dB) (Table 5). This pattern held for both isolated-digit and continuous-speech corpora, suggesting that the model struggles specifically with noise whose spectral envelope closely overlaps that of speech (pink, 1/f noise) or that is

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

highly non-stationary and impulsive (washing-machine cycles), rather than with any particular language or speech type.

Dataset	Best (1st)	Best (2nd)	Worst (1st)	Worst (2nd)
SC09	running_tap (11.4 dB)	white (11.4 dB)	pink (-0.6 dB)	washing_machine (0.3 dB)
Arabic	white (13.0 dB)	running_tap (12.9 dB)	pink (-0.3 dB)	washing_machine (0.5 dB)
LJSpeech	white (9.5 dB)	running_tap (9.1 dB)	pink (-1.2 dB)	washing_machine (-0.6 dB)

Table 5. Best- and worst-performing noise types by SI-SDR at SNR = 0 dB (DDPM, T=50).

5.4 DDIM Acceleration

A grid search over inference step t and number of sampling steps (20 validation files per dataset, white noise, 0 dB) identified t=5 with only 3 steps as optimal for all three datasets, reducing the reverse trajectory from 50 sequential steps to 3. Table 6 and Figure 7 report the resulting overall DDIM performance under the full 14-noise/3-SNR benchmark.

Dataset	SI-SDR (dB)	PESQ	STOI	MOS	RTF	Success rate
SC09 (English digits)	2.98	1.428	0.675	2.27	0.0732	67.1%
Arabic	3.30	1.332	0.707	2.18	0.0731	67.8%
LJSpeech (English sentences)	1.59	1.255	0.759	2.12	0.0768	58.5%

Table 6. Overall DDIM (t=5, 3 steps) performance across datasets, 14 noise types × 3 SNR levels.

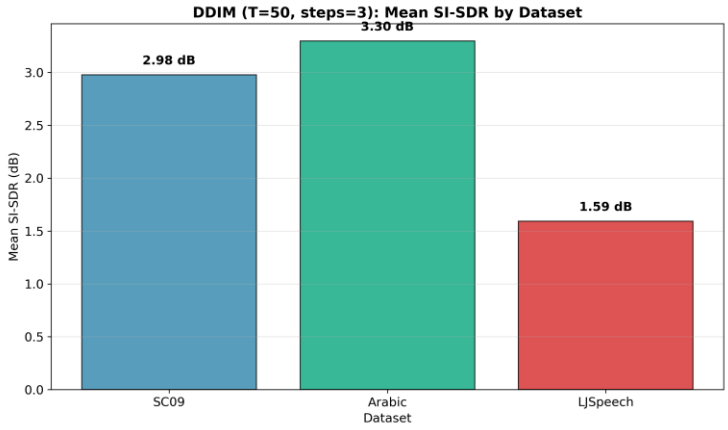


Figure 7. Mean SI-SDR by dataset under DDIM acceleration (t=5, 3 steps).

As with DDPM, Arabic remained the best-performing dataset under acceleration (SI-SDR = 3.30 dB, success rate = 67.8%), confirming that its cross-lingual robustness is not an artifact of the slower, stochastic sampling trajectory. Figure 8 shows that this ranking is preserved across all three SNR levels and in the success-rate metric (Figure 9): the relative dataset ordering established under DDPM (Arabic > SC09 > LJSpeech) is fully preserved after

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

acceleration, which indicates that DDIM compresses the sampling trajectory without altering which conditions the model handles well or poorly.

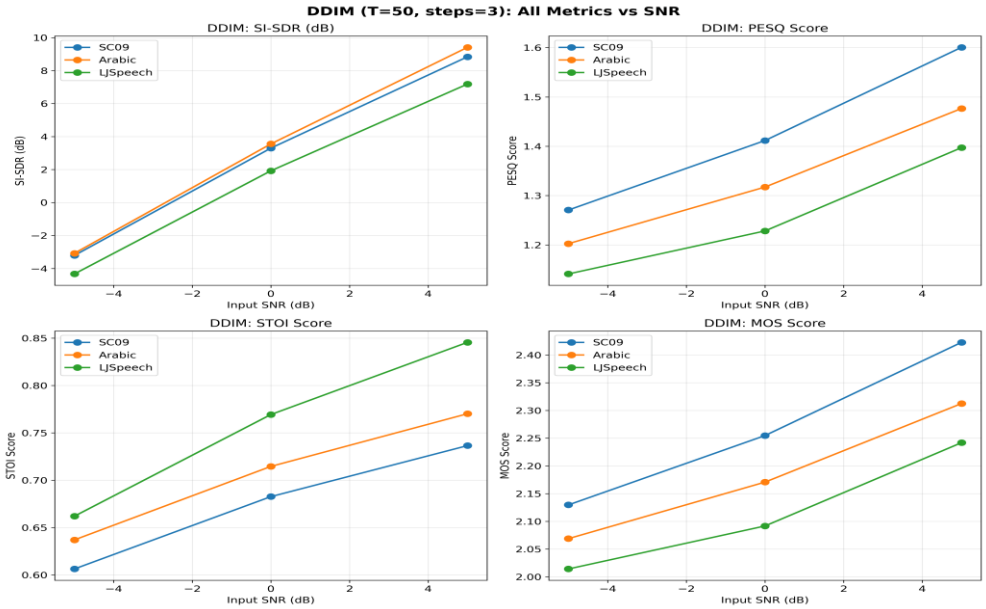


Figure 8. PESQ and STOI by dataset and input SNR level under DDIM acceleration.

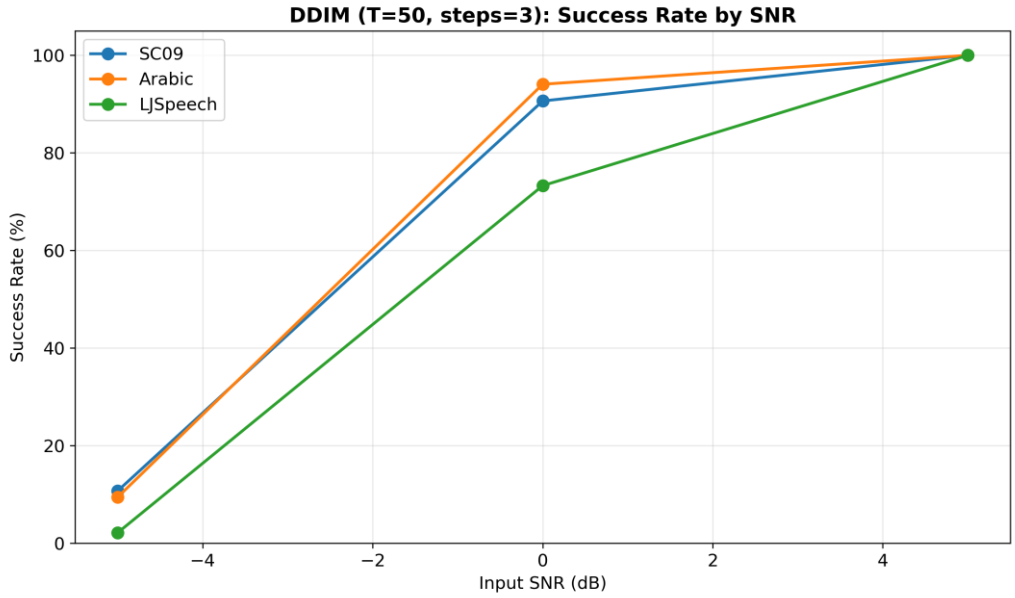


Figure 9. Success rate by dataset and SNR level under DDIM acceleration.

The average RTF dropped from 0.1225 (DDPM) to 0.0732 (DDIM), corresponding to roughly 13.7× real-time processing versus 8.2× for DDPM.

5.5 Comparative Evaluation: DDPM vs. DDIM

Table 7 juxtaposes DDPM (50 steps) and DDIM (3 steps) across all three datasets and six metrics. Figure 10 visualizes the absolute SI-SDR gap, Figure 11 the corresponding RTF reduction, and Figure 12 the success-rate comparison.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dataset	Model	SI-SDR	PESQ	STOI	MOS	RTF	Success
SC09	DDPM	3.37	1.363	0.666	2.23	0.1223	69.8%
SC09	DDIM	2.98	1.428	0.675	2.27	0.0732	67.1%
Arabic	DDPM	3.94	1.354	0.699	2.20	0.1225	71.2%
Arabic	DDIM	3.30	1.332	0.707	2.18	0.0731	67.8%
LJSpeech	DDPM	1.63	1.263	0.742	2.12	0.1226	59.0%
LJSpeech	DDIM	1.59	1.255	0.759	2.12	0.0768	58.5%

Table 7. Overall comparison of DDPM and DDIM across datasets and metrics.

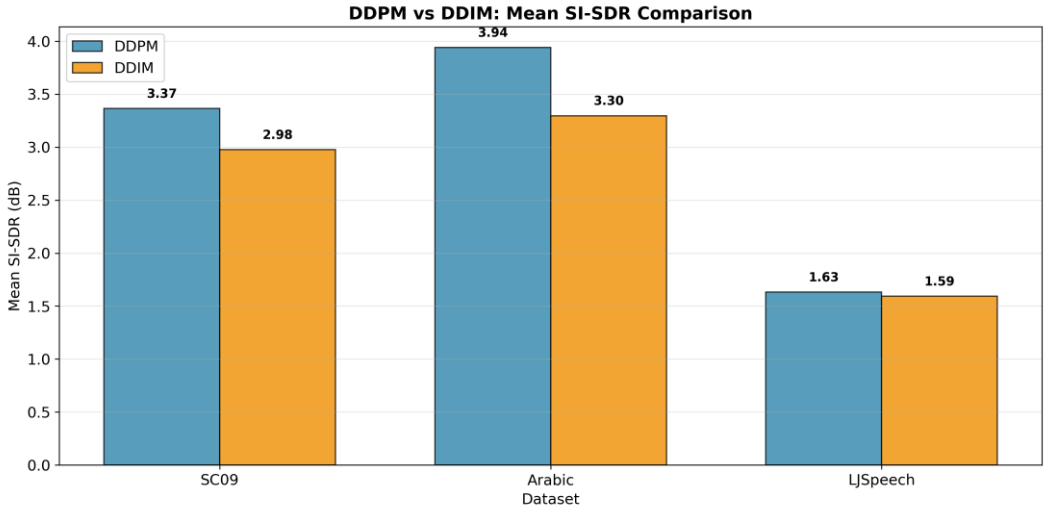


Figure 10. Absolute SI-SDR comparison between DDPM and DDIM.

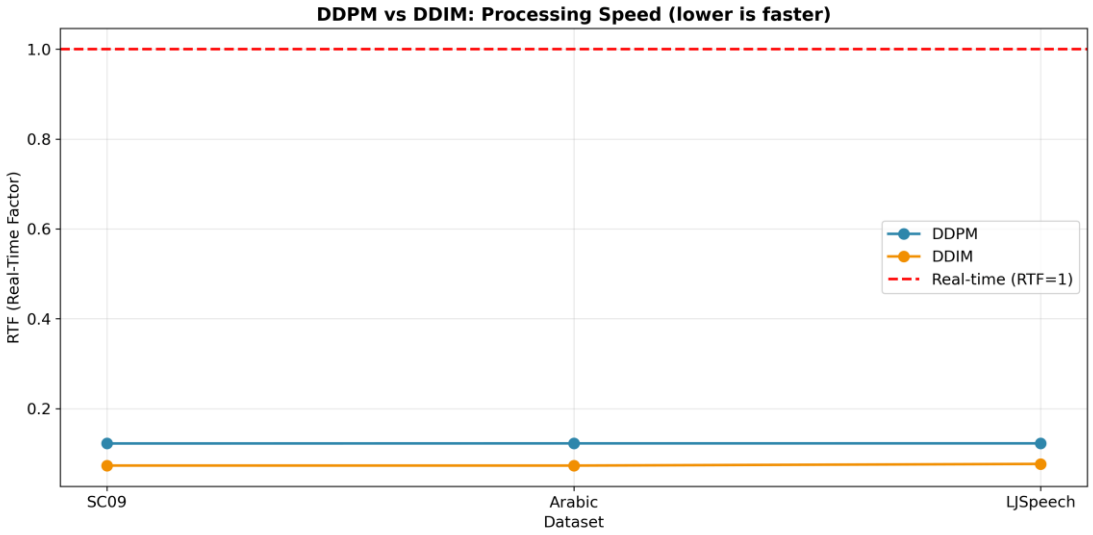


Figure 11. Real-time factor (RTF) comparison between DDPM and DDIM (lower is faster).

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

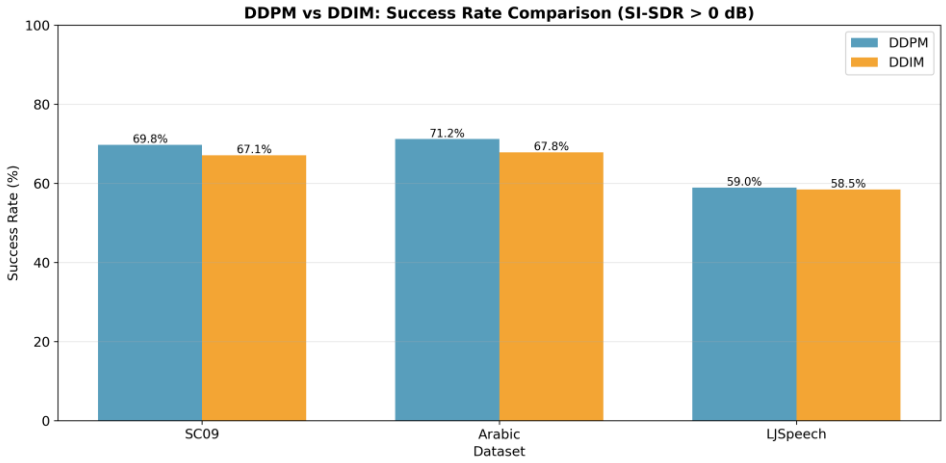


Figure 12. Success-rate comparison between DDPM and DDIM.

Table 8 quantifies the speed-quality trade-off directly: collapsing the reverse trajectory from 50 to 3 steps yields a 1.60–1.68× inference speedup, at the cost of a small SI-SDR reduction that is largest for Arabic (−0.64 dB) and smallest for LJSpeech (−0.04 dB), i.e., the quality cost of acceleration shrinks as utterances become longer and more prosodically complex. Notably, DDIM matches or slightly exceeds DDPM on PESQ for all three datasets and on STOI for all three datasets, indicating that the perceptual/intelligibility cost of acceleration is negligible even though the raw energy-based SI-SDR is somewhat lower.

Dataset	DDPM SI-SDR	DDIM SI-SDR	Δ	DDPM RTF	DDIM RTF	Speedup
SC09	3.37	2.98	-0.39	0.1223	0.0732	1.67×
Arabic	3.94	3.30	-0.64	0.1225	0.0731	1.68×
LJSpeech	1.63	1.59	-0.04	0.1226	0.0768	1.60×

Table 8. Speed-quality trade-off between DDPM and DDIM.

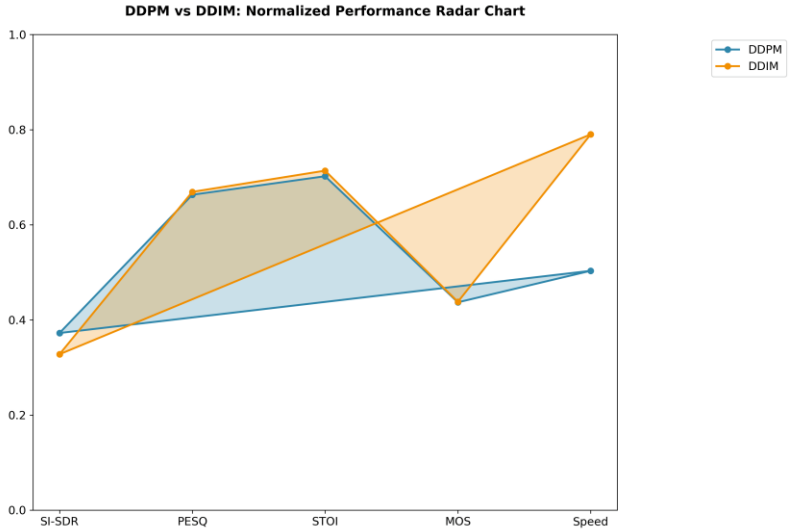


Figure 13. Holistic multi-metric radar comparison between DDPM and DDIM.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Overall, DDPM retains the absolute upper bound on signal-restoration quality (peak SI-SDR = 3.94 dB, Arabic), owing to the continuous stochastic refinement term σt z injected at each of its 50 reverse steps. DDIM, by setting this stochastic term to zero and collapsing the trajectory to 3 steps, sacrifices a small, dataset-dependent quality margin in exchange for a 1.6–1.7× inference speedup and an overall RTF close to 0.073 (13.7× real-time), which is decisive for deployment on latency-constrained or edge hardware.

5.6 Illustrative Single-Sample Analysis and Real-World Validation

To ground the aggregate statistics in a concrete example, both models were applied to a single Arabic utterance corrupted with cooking-noise at 0 dB SNR (input SNR = 0.20 dB, input SI-SDR = 0.14 dB, input PESQ = 1.095, input STOI = 0.770). DDPM (t=8) improved SNR to 2.50 dB (+2.30 dB), SI-SDR to 2.05 dB (+1.90 dB), PESQ to 1.214 (+0.119), and STOI to 0.867 (+0.097). DDIM (t=5, 3 steps) improved SNR to 2.15 dB (+1.95 dB), SI-SDR to 1.82 dB (+1.67 dB), PESQ to 1.609 (+0.514), and STOI to 0.851 (+0.080), while running roughly 3.2× faster than DDPM on the same (CPU) hardware. This single-sample comparison mirrors the aggregate pattern: DDPM edges out DDIM on energy-based metrics, whereas DDIM matches or exceeds it on the perceptual PESQ metric while being substantially faster.

Beyond the controlled, artificially corrupted benchmark described above, the trained model was additionally applied—without any further noise injection—to genuinely noisy recordings captured under real acoustic conditions (office, street, and cafeteria environments). The model produced intelligible, perceptually cleaner speech in all cases, indicating that performance gains observed under synthetic and semi-synthetic corruption transfer to authentic, uncontrolled acoustic environments.

5.7 Summary of Key Findings

Table 9 consolidates the principal quantitative findings of this study for ease of reference.

Finding	Key figure
T=50 beats T=200 on all six metrics due to 3.7× more training iterations	SI-SDR: 10.93 dB vs 9.94 dB
Best-performing dataset overall is Arabic, despite English-only training	SI-SDR = 3.94 dB, success = 71.2%
Continuous speech (LJSpeech) is harder than isolated digits, esp. at low SNR	SI-SDR = 1.63 dB; 24.8% success at -5 dB
Easiest / hardest noise types are consistent across all datasets	Best: white, running-tap — Worst: pink, washing-machine
DDIM cuts sampling from 50 to 3 steps with minimal quality loss	RTF 0.1225 → 0.0732 (1.6–1.7× faster)
DDIM matches or beats DDPM on perceptual metrics (PESQ, STOI)	e.g., Arabic PESQ: 1.354 → 1.332; STOI: 0.699 → 0.707
Gains transfer to genuine, uncontrolled real-world recordings	Qualitatively confirmed, office/street/cafeteria

Table 9. Summary of key quantitative findings.

DISCUSSION

The results support three main conclusions. First, model capacity alone does not determine denoising performance: the finer-grained T=200 schedule was theoretically capable of a more precise reverse process, but under a fixed computational budget it received roughly one-quarter the number of gradient updates of T=50 and consequently under-converged, despite reaching a lower nominal training loss. This illustrates that, for resource-constrained

diffusion training, an easier-to-optimize, coarser schedule trained to convergence can outperform a finer schedule that is only partially trained—a trade-off likely to generalize to other diffusion-based audio tasks trained under similar GPU constraints.

Second, DDIM acceleration is an effective and low-risk trade-off for this task. Reducing the reverse trajectory by more than an order of magnitude (50→3 steps) cut the RTF by 40% and yielded a 1.6–1.7× wall-clock speed up, while the associated SI-SDR loss was modest (−0.04 to −0.64 dB depending on dataset) and, on PESQ and STOI, DDIM even outperformed DDPM. The fact that the quality cost of acceleration was smallest for the most linguistically complex corpus (LJSpeech, continuous sentences) suggests that DDIM's deterministic sampling trajectory is particularly well suited to longer, prosodically richer utterances, whereas short isolated words are more sensitive to the loss of the stochastic refinement term. This directly extends the DDIM literature [4] with the first characterization of this trade-off specifically for the DiffWave denoising architecture, complementing prior applications of DDIM to unconditional audio synthesis.

Third, and most notably, a model trained exclusively on English digits, using only synthetic white noise, generalized not only across noise types and speech continuity (isolated digits to continuous sentences) but also across languages, achieving its best performance on Arabic despite the model never encountering that language during training. This cross-lingual robustness is consistent with the hypothesis that the model learns relatively language-agnostic acoustic representations of clean speech, and that phonetic properties—Arabic's emphatic and pharyngeal consonants provide higher-energy, sharper spectral boundaries than comparable English phonemes—can make certain languages intrinsically more separable from background noise, independent of the training language. This finding directly addresses the English-centric bias identified in Section 2.4 and is practically significant: it indicates that a single model trained on a resource-rich language can plausibly be deployed for lower-resource languages without language-specific retraining, an attractive property for real-world, multilingual deployment.

Some limitations should be noted. The full 14-noise/3-SNR benchmark evaluated 100 files per dataset, and the aggregate SI-SDR values (1.6–4.0 dB) are considerably lower than the values obtained during single-condition validation (white noise only, 9–13 dB), reflecting the much greater acoustic diversity of the full benchmark rather than a weaker model; the two figures are not directly comparable. Additionally, all quantitative comparisons rely on a single base architecture (DiffWave); whether the cross-lingual and speed-quality trends reported here transfer to other diffusion backbones (e.g., score-based STFT models [18]) remains an open question for future work, as does an extension of the cross-lingual evaluation to a wider set of typologically diverse languages and to human listening tests validating the MOS proxy used here.

In summary, this work provides the first quantitative benchmark of DiffWave for speech denoising, demonstrates that a moderately-sized, fully-converged diffusion-step budget (T=50) generalizes better than a finer but under-trained one (T=200), shows that DDIM acceleration reduces inference time by 1.6–1.7× with minimal quality loss, and establishes that the resulting system generalizes to Arabic despite training exclusively on English speech, supporting its practical relevance for real-time, cross-lingual speech enhancement.

REFRENCES

[1] Hu, Y., Loizou, P.C. Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing, 2007.
[2] Vaseghi, S.V. Advanced Digital Signal Processing and Noise Reduction, 4th ed. Wiley, 2008.
[3] Ho, J., Jain, A., Abbeel, P. Denoising diffusion probabilistic models. arXiv:2006.11239, 2020.
[4] Song, J., Meng, C., Ermon, S. Denoising diffusion implicit models. arXiv:2010.02502, 2020.
[5] Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B. DiffWave: A versatile diffusion model for audio synthesis. International Conference on Learning Representations (ICLR), 2021.
[6] Zheng, C., Zhang, H., Liu, W., Luo, X., Li, A., Li, X., Moore, B.C.J. Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods. Trends in Hearing, 27, 2023.
[7] Upadhyaya, N., Karmakar, A. Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. Procedia Computer Science, 54, 574–584, 2015.

[8] Plapous, C., Marro, C., Scalart, P. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing, 14(6), 2006.
[9] Azarang, A., Kehtarnavaz, N. A review of multi-objective deep learning speech denoising methods. arXiv:2003.12108, 2020.
[10] Janiesch, C., Zschech, P., Heinrich, K. Machine learning and deep learning. Electronic Markets, 31(3), 2021.
[11] Sharifani, K., Amini, M. Machine learning and deep learning: A review of methods and applications. World Information Technology and Engineering Journal, 10(07), 2023.
[12] Ochieng, P. Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis. Artificial Intelligence Review, 56(Suppl 3), 2023.
[13] Germain, F.G., Chen, Q., Koltun, V. Speech denoising with deep feature losses. arXiv:1806.10522, 2018.
[14] Pascual, S., Bonafonte, A., Serra, J. SEGAN: Speech enhancement generative adversarial network. arXiv:1703.09452, 2017.
[15] Li, Z.-X., Dai, L.-R., Song, Y., McLoughlin, I. A conditional generative model for speech enhancement. Circuits, Systems, and Signal Processing, 37(11), 2018.
[16] Fu, S.-W., Liao, C.-F., Tsao, Y., Lin, S.-D. MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. International Conference on Machine Learning, 2031–2041, 2019.
[17] Rethage, D., Pons, J., Serra, X. A WaveNet for speech denoising. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5069–5073, 2018.
[18] Welker, S., Richter, J., Gerkmann, T. Speech enhancement with score-based generative models in the complex STFT domain. arXiv:2203.17004, 2022.
[19] Richter, J., Welker, S., Lemercier, J.-M., Lay, B., Gerkmann, T. Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 2351–2364, 2023.
[20] Warden, P. Speech Commands: A dataset for limited-vocabulary speech recognition. arXiv:1804.03209, 2018.
[21] Ghandoura, A., Hjabo, F., Al Dakkak, O. Building and benchmarking an Arabic speech commands dataset for small-footprint keyword spotting. Engineering Applications of Artificial Intelligence, 102, 104267, 2021.
[22] Ito, K., Johnson, L. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.