الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة غرداية
University of Ghardaïa
كلية العلوم والتكنولوجيا
Faculty of Science and Technology
قسم الرياضيات والإعلام الآلي
Department of Mathematics and Computer Science

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master

Domain: Mathematics and Computer Science
Field: Computer Science
Specialty: Intelligent Systems for Knowledge Extraction (SIEC)

Topic

Speech Denoising Using Diffusion Model Technique

Presented by: Laouar Meriem & Zahouani Sarra

Jury Members:

Mrs. Bouchekouf Asma	MAA	Univ. Ghardaïa	President
Mr . Zaiani Mohamed	MCA	URAER	Examiner
Mr . Adjila Abderrahmane	MAA	Univ. Ghardaïa	Supervisor

Academic Year: 2025/2026

الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

جامعة غرداية
University of Ghardaia

كلية العلوم والتكنولوجيا
Faculty of Science and Technology

قسم الرياضيات والإعلام الآلي
Department of Mathematics and Computer Science

ترخيص بمناقشة مذكرة ماستر
Master Thesis Defense Permission

I, the undersigned, **Abderrahmane ADJILA**
Hereby certify that I have examined the work entitled:

Speech Denoising Using Diffusion Model Technique

Presented to the partial fulfillment of the Master degree in Computer Science by:

Laouar Meriem
Zahouani Sarra

I reviewed the document, and declare it is free from any serious defaults and respects the academic integrity rules. Furthermore, my jury member proposal is the following:

Mrs. Bouchekouf Asma	MAA	Univ. Ghardaïa	President
Mr . Zaiani Mohamed	MCA	URAER	Examiner
Mr . Adjila Abderrahmane	MAA	Univ. Ghardaïa	Supervisor

Issued for all due intents and purposes.

Ghardaia, on 10/06/2026 **Signature:**

Acknowledgements

First and foremost, we would like to express our praise and gratitude to Almighty Allah for granting us the strength, patience, and ability to complete this work.

We would like to express our sincere gratitude to our supervisor, Mr. Abderahmane ADJILA, for his continuous guidance, valuable advice, and unwavering support throughout this work.

We also extend our deepest thanks to all the teachers and staff of the Department of Mathematics and Computer Science at the University of Ghardaïa for providing us with knowledge and encouragement during our academic journey.

Special thanks are addressed to our families and friends for their patience, motivation, and unconditional support.

Finally, we are grateful to everyone who contributed directly or indirectly to the realization of this dissertation.

Dedication

To every obstacle that strengthened me instead of breaking me, and every dream that guided me like a roadmap rather than an illusion. To every success that taught me humility and perseverance, I dedicate this achievement.

I dedicate this work to hope that never faded, ambition that never weakened, and determination that turned challenges into growth. This achievement is dedicated to my parents, whose love, prayers, and sacrifices have been my greatest support: my mother, whose prayers accompanied me in every step, and my father, whose strength and guidance shaped my path. I also dedicate it to my brothers for their constant encouragement and support, as well as to my friend and project partner Meriem, for her collaboration and effort throughout this work.

I express my sincere gratitude to my supervisor, Mr. Abderahmane ADJILA, for his guidance, valuable advice, and continuous support throughout this work. I would also like to extend my heartfelt appreciation to all my professors for their dedication, knowledge, and the effort they invested in our education and academic development.

To myself, for enduring doubt, facing challenges, and choosing perseverance over giving up.

Finally, to everyone who supported me in any way, this achievement is the result of your encouragement and belief in me.

Sarra

Dedication

بسم الله الرحمن الرحيم

الحمد لله الذي بنعمته تتم الصالحات، والصلاة والسلام على أشرف المرسلين.

إلى من كان وجودهما نعمة، ودعاؤهما سندي في كل خطوة.. إلى من تعبت من أجلي وسهرت الليالي لتراني اليوم في أعلى المراتب.. إلى **أمي الغالية "فاطنة"**، التي لم تبخل علي بدعوة ولا نصيحة.. وإلى **أبي العزيز "محمد"**، الذي كان ولا يزال سندي وقوتي في هذه الحياة..
أهديكما هذا العمل المتواضع، راجية من المولى أن يطيل في أعماركما ويرزقكما الصحة والعافية.

إلى **زوجي الغالي "عبد القادر"** .. من كان معي في السراء والضراء، من لم يتخل عني في لحظة ضعف، بل كان السند والعون بعد الله.. أهديك هذا الجهد تقديراً لصبرك ودعمك المستمر لي.

إلى **أخواتي الغاليات واخي العزيز : وردة، آمنة، مصطفى، هاجر، شافية، عصماء، بشرى ..**
وإلى **أبناء إخوتي** الذين كانوا دائماً بجانبي، أهديكم هذا العمل شكراً لكم على كل لحظة دعم ومساندة.

إلى **فلذات كبدي، أبنائي الأعزاء : مروة، محمد، ملاك، مصعب ..** أنتم نعمة الله في حياتي، وأمل المستقبل الذي يشرق بضحكاتكم. أسأل الله أن تكونوا دائماً فخراً لي ويغفر لي الله تقصيري معكم.

إلى **صديقة دربي ورفيقة كفاحي في هذه المذكرة ، صارة ..** كنا معاً في هذه الرحلة، وها نحن اليوم نقطف ثمار تعبنا. أهديك هذا العمل شكراً على روح الفريق والتعاون الجميل.

إلى **صديقاتي الغاليات : أحلام، رميساء، سلمى ..** وإلى **زميلاتي في ثانوية المجاهد "قرمة بوجمعة"**، على رأسهم السيدة المديرة.. أشكركن على كل لحظة دعم وتشجيع.

إلى **مشرفي الفاضل الأستاذ "عجيلة عبد الرحمان" ..** الذي كان لي خير مرشد ومعين، أشكره على توجيهاته القيمة، وصبره الجميل، ودعمه المتواصل حتى خرج هذا العمل إلى النور.

إلى **كل من ساندني وعاضدني، مهما كان دوره صغيراً أو كبيراً ..** لكم مني أسمى آيات الشكر والعرفان.

مريم

مـلـخـص

يُعد الكلام الوسيلة الأساسية للتواصل البشري، غير أنه يتعرض غالباً للتشويش في البيئات الواقعية، مما يؤثر سلباً على وضوحه وجودته، ويُضعف أداء التطبيقات الحيوية كالتعرف الآلي على الكلام، والاتصالات، والبث الصوتي. لذلك، تلعب تنقية الكلام دوراً محورياً في استعادة جودة الصوت وضمان تفاعل موثوق بين الإنسان والآلة.

تستكشف هذه الأطروحة النماذج التوليدية القائمة على الانتشار (Diffusion-Based Generative Models) لمعالجة مشكلة تنقية الكلام، مع التركيز على معمارية DiffWave. ويهدف العمل إلى بناء نظام قادر على استعادة الكلام بجودة عالية مع سرعة تناسب التطبيقات الواقعية.

تمت مقارنة إعدادين لعدد خطوات الانتشار: T=200 و T=50. ورغم أن T=200 يوفر تفصيلاً نظرياً أفضل، أظهر T=50 أداءً أفضل على البيانات غير المرئية (SI-SDR = 10.93 dB مقابل 9.94 dB)، مع خطوة استدلال مثلى عند t=5. ويعود ذلك إلى تدريب T=50 لفترة أطول (240,000 تكرار مقابل 65,000) بسبب محدودية موارد المعالج الرسومي (GPU)، مما سمح بتقارب أفضل.

لتسريع الاستدلال، تم دمج تقنية DDIM، التي اختزلت خطوات التوليد من 50 إلى 3 فقط، محققةً عامل زمن حقيقي (RTF) قدره 0.0732، أي أسرع بـ 13.7 مرة من الزمن الحقيقي، مع خسارة طفيفة في الجودة لا تتجاوز 0.64 dB-.

من أبرز إسهامات هذا العمل تقييم النموذج على اللغة العربية. فرغم أن النموذج دُرب على اللغة الإنجليزية فقط، حقق أفضل نتائجه على المجموعة العربية (SI-SDR = 3.94 dB)، حيث أظهرت غالبية العينات تحسناً ملحوظاً بعد المعالجة، مما يؤكد قدرته على التعميم عبر اللغات. ويعود ذلك إلى تعلم النموذج تمثيلات صوتية عامة تتجاوز الحدود اللغوية.

كما تم التحقق من أداء النموذج على تسجيلات حقيقية من بيئات متنوعة (مكاتب، شوارع، مقاصف)، مما يؤكد فعاليته خارج الظروف المخبرية، وجاهزيته للتطبيق في أنظمة الوقت الحقيقي. كما تم تقييم النموذج على اللغة الفرنسية باستخدام بيانات Common Voice، حيث أظهر قدرة واضحة على التعميم عبر اللغات، مع نتائج واعدة تدعم فعاليته في بيئات لغوية متنوعة.

الكلمات المفتاحية: إزالة ضوضاء الكلام، معالجة الكلام، معالجة الإشارات الصوتية، التعلم العميق، نماذج الانتشار، تحسين الكلام، نسبة الإشارة إلى الضجيج، SNR، DDPM، DDIM.

Abstract

Speech is the primary medium for human communication, yet it is almost always degraded by background noise in real-world environments. This noise reduces intelligibility and harms the performance of critical applications such as automatic speech recognition, telecommunications, and audio streaming. Speech denoising, therefore, plays a vital role in restoring clean speech and ensuring reliable human-machine interaction.

This thesis explores diffusion-based generative models for speech denoising, focusing on the DiffWave architecture. The objective is to build a system that produces high-quality enhanced speech while operating fast enough for real-world deployment. To this end, two diffusion step configurations are compared: T=50 and T=200. Although T=200 provides a finer discretization, T=50 achieves better generalization on unseen data (SI-SDR: 10.93 dB vs 9.94 dB), with an optimal inference step of t=5. This is primarily because T=50 was trained for significantly longer (240,000 iterations compared to 65,000 for T=200) due to GPU constraints, allowing it to converge more effectively.

To overcome inference latency, Denoising Diffusion Implicit Models (DDIM) are integrated, reducing sampling steps from 50 to just 3. This achieves a real-time factor (RTF) of 0.0732, corresponding to 13.7× real-time performance, with a minimal quality loss of only -0.64 dB.

A key contribution is the evaluation on Arabic speech. Although trained exclusively on English digits, the model achieves its best performance on the Arabic dataset (SI-SDR = 3.94 dB), with the majority of test samples showing improvement after processing, confirming strong cross-lingual generalization. This is attributed to the model's ability to learn latent acoustic representations that transcend language boundaries.

The model was also evaluated on French speech using the Common Voice dataset, demonstrating clear cross-lingual generalization capabilities with promising results that support its effectiveness in diverse linguistic environments.

The model is also validated on real-world recordings from office, street, and cafeteria environments, confirming its practical applicability beyond controlled laboratory conditions. These results demonstrate that the proposed approach offers a practical, fast, and language-robust solution for speech denoising, suitable for real-time applications.

Keywords: Speech Denoising, Speech Processing, Audio Signal Processing, Deep Learning, Diffusion Models, Speech Enhancement, Signal-to-Noise Ratio (SNR), DDPM, DDIM.

Résumé

La parole est le principal moyen de communication humaine, mais elle est presque toujours dégradée par le bruit ambiant dans les environnements réels. Ce bruit réduit l'intelligibilité et nuit aux performances des applications critiques telles que la reconnaissance vocale automatique, les télécommunications et la diffusion audio. Le débruitage de la parole joue donc un rôle essentiel dans la restauration d'une parole claire et fiable.

Cette thèse explore les modèles génératifs basés sur la diffusion pour le débruitage de la parole, en se concentrant sur l'architecture DiffWave. L'objectif est de concevoir un système qui produit une parole de haute qualité tout en étant suffisamment rapide pour une utilisation en temps réel. Deux configurations d'étapes de diffusion sont comparées : T=50 et T=200. Bien que T=200 offre une discrétisation plus fine, T=50 généralise mieux sur des données non vues (SI-SDR = 10.93 dB contre 9.94 dB), avec une étape d'inférence optimale à t=5. Ce résultat s'explique par un entraînement plus long de T=50 (240 000 itérations contre 65 000 pour T=200), en raison des contraintes GPU.

Pour réduire la latence, l'approche DDIM est intégrée, réduisant les étapes d'inférence de 50 à 3, atteignant un facteur temps réel (RTF) de 0.0732, soit 13.7× plus rapide que le temps réel, avec une perte de qualité de seulement -0.64 dB.

Une contribution majeure est l'évaluation sur la parole arabe. Bien que le modèle ait été entraîné uniquement sur l'anglais, il obtient ses meilleurs résultats sur l'arabe (SI-SDR = 3.94 dB), démontrant une forte capacité de généralisation inter-langues.

Le modèle a également été évalué sur la parole française à l'aide de l'ensemble de données Common Voice, démontrant une capacité claire de généralisation inter-langues avec des résultats prometteurs qui soutiennent son efficacité dans des environnements linguistiques diversifiés.

Le modèle est également validé sur des enregistrements réels (bureaux, rues, cafétérias), confirmant son applicabilité pratique.

Mots-clés : Débruitage de la parole, traitement de la parole, apprentissage profond, modèles de diffusion, amélioration de la parole, rapport signal-bruit (SNR), DDPM, DDIM.

Contents

List of Figures

List of Tables

List of Acronyms

DDPM	Denoising Diffusion Probabilistic Model
DDIM	Denoising Diffusion Implicit Model
SNR	Signal-to-Noise Ratio
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
PESQ	Perceptual Evaluation of Speech Quality
STOI	Short-Time Objective Intelligibility
MOS	Mean Opinion Score
RTF	Real-Time Factor
STFT	Short-Time Fourier Transform
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Network
NMF	Non-Negative Matrix Factorization
MMSE	Minimum Mean Square Error
PSD	Power Spectral Density
DNN	Deep Neural Network
CRN	Convolutional Recurrent Network
DCCRN	Deep Complex Convolutional Recurrent Network
IRM	Ideal Ratio Mask
CRM	Complex Ratio Mask
SEGAN	Speech Enhancement Generative Adversarial Network
TSNR	Two-Step Noise Reduction
HRNR	Harmonic Regeneration Noise Reduction
DD	Decision-Directed
MMSE-STSA	Minimum Mean Square Error Short-Time Spectral Amplitude
MMSE-LSA	Minimum Mean Square Error Log-Spectral Amplitude
SNRseg	Segmental Signal-to-Noise Ratio

General Introduction

Speech is one of the most fundamental and natural forms of human communication. With the rapid advancement of digital technologies, speech signals have become integral to a wide range of applications, including telecommunications, voice-controlled systems, automatic speech recognition, and multimedia streaming [1]. However, in real-world environments, speech signals are almost always contaminated by various types of background noise, such as environmental sounds, electronic interference, and reverberation [2]. This noise not only degrades the perceptual quality of the speech but also significantly impairs the performance of downstream systems that rely on clean audio input [3]. Consequently, speech denoising has emerged as a critical research area aimed at recovering clean speech from noisy observations while preserving its intelligibility and naturalness [4].

Over the past decades, numerous approaches have been developed to address the speech denoising problem. Early methods were primarily based on classical signal processing techniques, including spectral subtraction, Wiener filtering, and statistical model-based estimators such as Minimum Mean Square Error (MMSE) [5]. These methods rely on assumptions about the statistical properties of speech and noise, and they perform reasonably well under stationary noise conditions [6]. However, they often fail in complex acoustic environments where noise is non-stationary, colored, or highly correlated with the speech signal [7]. The limitations of these traditional approaches have motivated the exploration of data-driven methods that can learn complex mappings between noisy and clean speech directly from examples [8].

The advent of deep learning has revolutionized the field of speech denoising. Neural network architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variants have demonstrated remarkable ability to suppress noise while preserving speech quality [3]. More recently, generative models have gained significant attention for their capacity to model the underlying distribution of clean speech and generate high-fidelity outputs [9]. Among these, diffusion-based generative models have emerged as a particularly promising paradigm [10]. These models operate through a two-stage process: a forward diffusion process that gradually corrupts clean speech by adding Gaussian noise, and a reverse denoising process that learns to reconstruct the original signal from the noisy input [11]. Variants such as Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM) have shown state-of-the-art performance in speech restoration tasks, offering both high-quality reconstruction and robust generalization across diverse noise conditions [12, 13, 14].

Despite these advances, several challenges remain open. The high computational cost and inference latency of diffusion models limit their deployment in real-time applications [**?**]. Moreover, the vast majority of existing work focuses on English speech, leaving languages such as Arabic underexplored in the context of diffusion-based denoising. This thesis addresses these gaps by investigating the acceleration of diffusion models using DDIM, and by providing the first comprehensive quantitative evaluation of a diffusion-

based denoising system on Arabic speech. Additionally, we extend the evaluation to French speech using the Common Voice dataset [15] to further assess the cross-lingual generalization capabilities of the model and validate its effectiveness across multiple languages.

This thesis focuses on the study, implementation, and evaluation of diffusion-based speech denoising techniques. The primary objectives are to enhance the performance of existing diffusion models using DDIM acceleration strategies, to conduct a comprehensive empirical evaluation across multiple languages and noise conditions, and to assess the generalizability of these models to Arabic speech, which remains underrepresented in the literature. To this end, we adopt the DiffWave architecture as our base generative model [16], and we systematically compare two diffusion step configurations (T=50 and T=200) under a wide range of acoustic scenarios.

Key Contributions of This Work:

- **Comprehensive Evaluation:** First quantitative evaluation of the DiffWave architecture for speech denoising across 14 noise types and 3 SNR levels, using six objective metrics (SNR, SI-SDR, PESQ, STOI, MOS, RTF).
- **Model Optimization:** Comparison between T=50 and T=200 diffusion steps, demonstrating that T=50 achieves better generalization (SI-SDR: 10.93 dB vs 9.94 dB) despite higher training loss, due to longer training convergence.
- **Inference Acceleration:** Successful integration of DDIM, reducing sampling steps from 50 to 3, achieving a real-time factor of 0.0732 (13.7× real-time performance) with a minimal quality drop of only -0.64 dB.
- **Cross-Lingual Generalization:** The model, trained only on English digits, achieved superior performance on Arabic speech (SI-SDR = 3.94 dB, success rate = 71.2%), and promising results on French speech (SI-SDR = 3.22 dB), demonstrating strong cross-lingual generalization across multiple languages.
- **Real-World Validation:** Successful denoising on real-world recordings from office, street, and cafeteria environments, confirming practical applicability.

Document Structure:

- **Chapter 1:** Fundamentals of speech processing, noise types, challenges, and diffusion-based models (DDPM, DDIM) with evaluation metrics (SNR, SI-SDR, PESQ, STOI, MOS, RTF).
- **Chapter 2:** State-of-the-art speech denoising methods, from traditional signal processing to deep learning and generative models (GANs, diffusion models).
- **Chapter 3:** Design and implementation of the DiffWave-based speech denoising system, including dataset preparation and training pipelines.
- **Chapter 4:** Presents experimental evaluation across four datasets (SC09, Arabic, LJSpeech, and French), comparing DDPM (T=50 and T=200) and DDIM acceleration under 14 noise types at three SNR levels (-5, 0, 5 dB), using six objective metrics: SNR, SI-SDR, PESQ, STOI, MOS, and RTF.

Chapter 1

Generalities about Speech and Speech Denoising

1.1 Introduction

Speech signals are inherently susceptible to noise during acquisition, transmission, and storage, degrading both perceptual quality and intelligibility. This negatively affects downstream applications such as speech recognition and communication systems, making speech denoising a fundamental and active research problem. This chapter provides the theoretical foundation for the work presented in this thesis. It begins by reviewing the fundamentals of audio and speech signals, including their characteristics, digital representation, and the nature of noise corruption (Sections 1.2). The key challenges that make speech denoising a difficult task — such as noise variability, speech distortion, data scarcity, and real-time constraints — are then discussed in Section 1.3. Section 1.4 surveys data-driven learning paradigms for denoising, covering supervised, unsupervised, self-supervised, and generative approaches. Building on this, Section 1.5 introduces Denoising Diffusion Probabilistic Models (DDPM) and their accelerated variant, Denoising Diffusion Implicit Models (DDIM), which form the core methodology of this thesis. Finally, Section 1.6 presents the objective evaluation metrics used throughout this work, including SNR, SI-SDR, PESQ, STOI, MOS, and RTF.

1.2 Audio and Speech Fundamentals

Audio and speech signals form the foundation of many communication and recognition systems. Speech is a complex acoustic signal that carries linguistic information, speaker characteristics, and prosodic features, and its accurate extraction and processing is crucial for a wide range of applications [17].

Audio signals can be represented in terms of amplitude, frequency, and phase, and they may include a mixture of speech, environmental sounds, and other background noise. Techniques for analyzing and processing audio signals often rely on digital representations, allowing signals processing and enhancement using computer tools [18].

Multimodal speech processing systems, which integrate audio and visual information, improve the accuracy of speech recognition and processing, particularly in noisy environments. These systems leverage the correlation between speech and lip movements or facial expressions to enhance performance, highlighting the importance of integrating multiple information sources in speech processing applications [19].

Understanding the fundamentals of audio and speech signals is therefore essential for the development of effective speech denoising, recognition, and enhancement techniques.

1.2.1 Characteristics of Audio Signals

An audio signal represents sound as a varying electrical or digital waveform that can be analyzed and processed by computers and electronic systems. One of the main characteristics of an audio signal is its **variation in amplitude over time**, which corresponds to how loud or soft the sound is perceived. Another key characteristic is **frequency**, which determines the pitch of the sound; different frequency components in speech make vowels and consonants distinguishable from each other [17].

Sound waves can be mathematically described as:

$$y(x,t) = A\sin(2\pi ft - \Phi) \tag{1.1}$$

where A is amplitude, f is frequency, t is time, and Φ is phase.

The relationship between frequency and period is fundamental:

$$f = \frac{1}{T} \tag{1.2}$$

Figure 1.1 shows a typical time-domain waveform of a speech signal, where amplitude variations over time are clearly visible. Peaks correspond to high-energy speech segments, while flat regions indicate silence or low-energy phonemes.

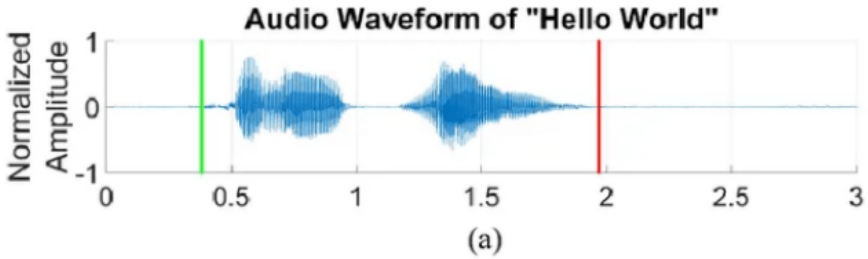


Figure 1.1: *Time-domain waveform of a speech signal showing amplitude variations over time. The normalized amplitude ranges from -1 to 1, with peaks representing high-energy speech segments and flat regions indicating silence or low-energy phonemes. (Source: McLoughlin, 2016, p. 8)*

Figure 1.1 shows the time-domain waveform of a speech signal, where amplitude variations are clearly visible over time.

Audio signals can also be viewed in the **frequency domain**, where the signal is expressed as a combination of sinusoidal components. This representation is useful for identifying dominant frequency bands and for processing tasks such as filtering and enhancement [18].

Figure 1.2 illustrates a spectrogram of the phrase "Hello World", where the horizontal axis represents time, the vertical axis represents frequency, and color intensity indicates power density. The dark horizontal bands, known as formants, correspond to resonant frequencies of vowels.

Additionally, audio signals often contain complex mixtures of speech and environmental sounds, and their structure may change rapidly over time. In speech processing

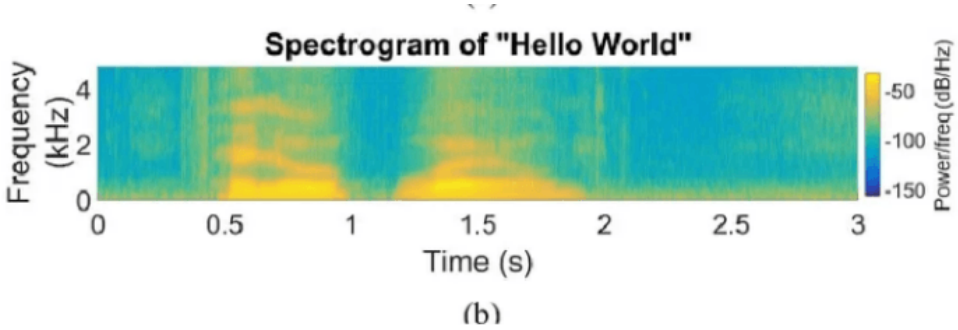


Figure 1.2: *Spectrogram of the phrase "Hello World" showing time-frequency representation. The horizontal axis represents time (seconds), the vertical axis represents frequency (kHz), and color intensity indicates power density (dB/Hz). Formants (dark horizontal bands) are visible as resonant frequencies of vowels. (Source: McLoughlin, 2016, p. 8)*

applications, it is sometimes helpful to consider visual cues, such as lip movement or facial expressions, together with audio features, because these multimodal characteristics can improve recognition and robustness in noisy conditions [19].

These fundamental properties of audio signals — amplitude variation, frequency content, and temporal structure — form the basis for further analysis in applications like speech recognition and denoising.

1.2.2 Speech Signal

A speech signal is a specific type of audio signal produced by the human vocal tract. It carries **linguistic information**, such as phonemes and words, as well as **speaker-specific characteristics** like pitch, accent, and speaking style [17]. Speech signals are inherently **time-varying** and consist of different sound components, including voiced sounds (like vowels) with periodic structures and unvoiced sounds (like consonants) with more noise-like characteristics [18].

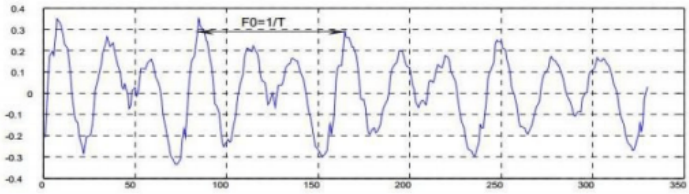


Figure 1.3: *A voiced speech signal showing its fundamental frequency $F_0 = 1/T$. The periodic waveform represents a vowel sound with clear pitch periods. (Source: McLoughlin, 2016, p. 7)*

The analysis of speech signals often requires examining both the **temporal variations** and **frequency content**. Temporal patterns provide information about rhythm and duration, while frequency components are crucial for identifying vowels, consonants, and prosody [19]. Because speech signals are complex and often corrupted by environmental noise, processing them effectively requires understanding these fundamental properties. As illustrated in Figures 1.3 and 1.4, speech exhibits both **periodic structure** (in

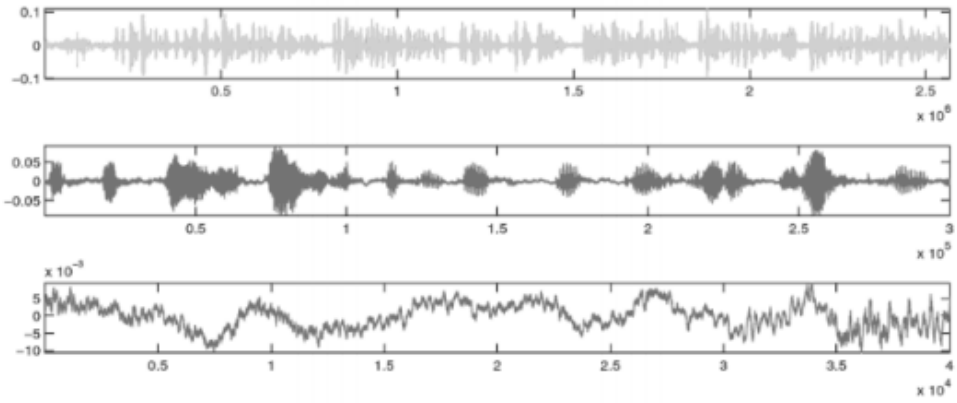


Figure 1.4: *Multi-scale visualization of a speech signal: (Top) phrase-level contour, (Middle) syllable-level variations, (Bottom) phoneme-level details. (Source: McLoughlin, 2016, p. 7)*

voiced segments) and **multi-scale temporal dynamics**. The three-scale representation demonstrates how speech can be analyzed at different resolutions: from prosodic contours spanning seconds down to phonetic details within milliseconds. This hierarchical nature is essential for designing denoising algorithms that must preserve speech characteristics across time scales. In addition, multimodal speech systems can combine audio with visual cues, such as lip movements, to improve recognition performance, especially in noisy conditions. This highlights the importance of capturing both the acoustic and contextual features of speech for applications such as speech recognition and enhancement [19].

1.2.3 Digital Representation of Speech

The digital representation of speech involves the conversion of continuous speech signals into discrete forms suitable for processing, storage, and transmission [20]. This is governed by two core operations: sampling, constrained by the Nyquist-Shannon theorem (equation 1.3), and quantization, which maps continuous amplitudes to discrete levels.

$$f_s > 2 \cdot f_{max} \tag{1.3}$$

where f_s is the sampling frequency and f_{max} is the maximum frequency component in the speech signal.

Beyond the raw waveform, speech can be represented in the spectral domain using digital filters to model frequency components [21], or through compact sparse and non-negative representations for enhancement and recognition tasks [22]. Modern approaches also rely on learned representations extracted directly from data, capturing high-level speech characteristics without manual feature engineering [23].

Time-Frequency Analysis: Short-Time Fourier Transform (STFT)

While traditional Fourier analysis provides frequency information for an entire signal, speech signals require time-localized frequency analysis due to their non-stationary nature. The Short-Time Fourier Transform (STFT) addresses this by dividing the signal into short segments and analyzing each segment separately.

The STFT of a continuous signal $x(\tau)$ is mathematically expressed as [24]:

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau) \cdot w(t-\tau) \cdot e^{-j\omega\tau}\, d\tau \tag{1.4}$$

where $w(t-\tau)$ is a window function that localizes the analysis around time t. Common window functions include Hamming, Hanning, and Gaussian windows.

For discrete signals, the STFT is computed as:

$$X[m,k] = \sum_{n=0}^{N-1} x[n] \cdot w[n-mH] \cdot e^{-j2\pi kn/N} \tag{1.5}$$

where:

- m: Frame index (time dimension)
- k: Frequency bin index ($k = 0, 1, ..., N-1$)
- N: Window length (typically 20-40 ms for speech)
- H: Hop size (typically 10 ms for 50% overlap)
- $w[\cdot]$: Window function samples

The magnitude spectrogram $|X[m,k]|$ provides a visual representation of speech energy distribution across time and frequency, as shown in Figure 1.2. This representation is essential for speech denoising algorithms, including diffusion-based approaches, as it allows separation of speech components from noise based on their distinct time-frequency characteristics.

To further illustrate this concept in a noisy context, Figure 1.5 presents a spectrogram of a noisy speech signal, where the overlapping patterns of speech and noise become clearly visible.

Interpreting the Spectrogram for Denoising

The spectrogram in Figure 1.5 reveals several important aspects for speech denoising:

- **Speech Components:** Horizontal bands (formants) and vertical striations correspond to voiced speech sounds and transients, exhibiting structured patterns in time and frequency.
- **Noise Characteristics:** Background noise often appears as diffuse energy spread across frequencies (for white noise) or concentrated in specific bands (for colored noise), lacking the temporal structure of speech.
- **Signal-to-Noise Ratio (SNR) Visualization:** Regions with high color intensity (bright colors) indicate higher energy concentration, helping to identify where speech dominates over noise and vice versa.

This time-frequency decomposition is particularly relevant for diffusion-based denoising models, which can operate on spectrogram representations to learn the complex mappings between noisy and clean speech by modeling the noise removal process across both time and frequency dimensions.

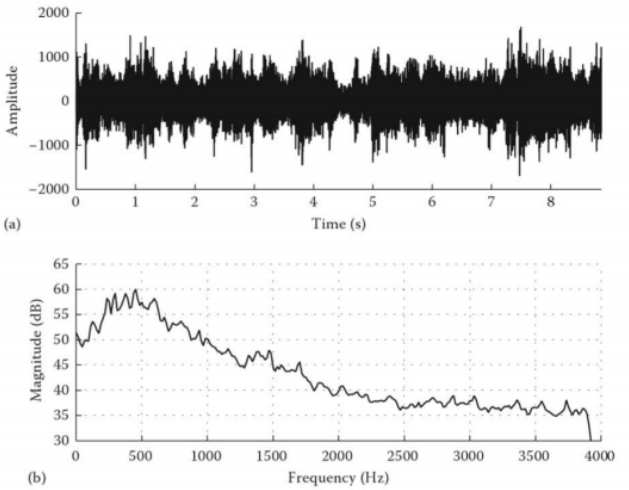


Figure 1.5: *Time-frequency representation (spectrogram) of a noisy speech signal. The top panel shows the time-domain waveform with its amplitude variations. The bottom panel displays the corresponding magnitude spectrogram, where the horizontal axis represents time (seconds), the vertical axis represents frequency (Hz), and the color intensity represents the magnitude (in dB) of the signal's frequency components at each time frame. This joint time-frequency visualization is fundamental for analyzing how noise overlaps with speech components, forming the basis for many modern denoising techniques. (Source: Loizou, 2012)*

1.2.4 Noise in Speech Signals

Noise in speech signals refers to any unwanted disturbance that interferes with the original speech signal and degrades its quality or intelligibility. In speech processing systems, noise is considered one of the main factors that negatively affect analysis, transmission, and recognition performance [2].

Speech signals are often contaminated by different types of noise during acquisition, transmission, or storage. According to Quilici [20], noise can originate from environmental sources such as background sounds, recording devices, or communication channels. These noise components are usually added to the speech signal and alter its spectral and temporal characteristics.

From a signal processing perspective, noise affects the spectral structure of speech. Gordy [21] shows that the presence of noise modifies the speech spectrum, making it more difficult to accurately model or synthesize speech signals using digital filters. This spectral distortion reduces the effectiveness of speech analysis and synthesis techniques.

In digital speech processing, noise is commonly modeled as an additive component that overlaps with the useful speech signal, as illustrated in Figure 1.6:

$$y(t) = x(t) + n(t) \tag{1.6}$$

Vaseghi [2] explains that this additive noise impacts both time-domain and frequency-domain representations, leading to a loss of important speech features. As a result, noise reduction and enhancement techniques are required to improve speech quality.

Modern speech representation approaches also consider the influence of noise on learned speech features. Williams [23] highlights that noise can entangle speech repre-

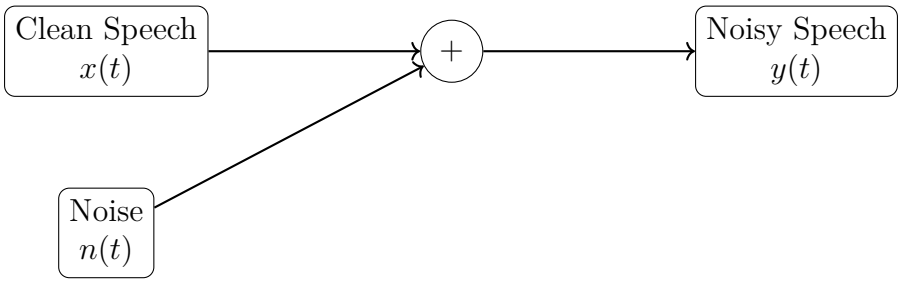


Figure 1.6: *Additive noise model in speech processing*

sentations, making it difficult to separate linguistic content from noise-related variations. This challenge motivates the development of robust representations that are less sensitive to noise.

Furthermore, Lu [25] emphasizes that noise is a critical issue in practical digital signal processing applications involving speech. Effective handling of noise is essential to ensure reliable performance in speech analysis, enhancement, and recognition systems.

1.2.5 Types and Mathematical Models of Noise

Noise in speech signals manifests in various forms, each with distinct mathematical properties that influence denoising strategy effectiveness. Table 1.1 summarizes common noise types relevant to speech processing.

Table 1.1: *Mathematical Models of Common Noise Types in Speech Processing*

Noise Type	**Mathematical Model**	**Characteristics and Speech Processing Relevance**
White Noise	$W(t) = A \cdot N(t)$	Constant power spectral density; used as baseline in diffusion model forward process
Pink Noise (1/f)	$P(f) = \frac{A}{f^{\beta}}$	Power decreases with frequency; mimics natural/environmental background noise
Brownian Noise	$B(t) = \int_0^t W(\tau)d\tau$	Low-frequency dominated; resembles rumble or distant machinery
Gaussian Noise	$G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	Normal distribution; fundamental for statistical modeling and DDPM formulations
Impulse Noise	$I(t) = \sum_k a_k \delta(t - t_k)$	Sudden, short bursts; simulates clicks, pops in recordings
Uniform Noise	$U(x) = \frac{1}{2A}$ for $x \in [-A, A]$	Constant amplitude distribution; models quantization errors

Gaussian Noise in Diffusion Models

Gaussian noise is particularly significant for diffusion-based denoising, as the forward process in DDPM/DDIM relies on progressively adding Gaussian noise to clean speech [10]:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1.7}$$

where β_t is the noise schedule at step t. This controlled corruption process enables the model to learn the reverse denoising trajectory.

Real-World Noise Characteristics

In practice, speech datasets contain mixtures of multiple noise types, creating complex acoustic environments [2]. A typical recording might contain:

- **Gaussian white noise** from electronic circuits and thermal noise
- **Pink/brown noise** from environmental sources (wind, crowds, machinery)
- **Impulse noise** from connection artifacts, clicks, and pops
- **Periodic noise** from electrical hum (50/60 Hz) and rotating machinery

This complexity necessitates robust models like diffusion models that can learn to separate these overlapping noise distributions from speech signals while preserving speech quality and intelligibility.

1.3 Challenges in Speech Denoising

Speech denoising research confronts multiple persistent challenges that impact both algorithmic development and practical deployment. These challenges span technical limitations, data requirements, and evaluation methodologies.

1.3.1 Noise Variability and Diversity

Real-world acoustic environments present complex noise profiles characterized by high diversity and temporal variation. Noise sources range from stationary background sounds to non-stationary, impulsive disturbances, each requiring different processing strategies [2]. The statistical properties of environmental noise often change unpredictably, challenging algorithms that assume stationary conditions. Recent adaptive approaches employ conditional mechanisms to address varying noise levels and types [26, 27], though achieving consistent performance across diverse scenarios remains difficult.

1.3.2 Speech Distortion and Artifact Introduction

A fundamental trade-off exists between aggressive noise suppression and speech quality preservation. Excessive denoising can remove not only noise but also essential speech components, introducing audible artifacts and reducing intelligibility [28]. Common artifacts include musical noise, speech truncation, and unnatural spectral characteristics. Modern generative approaches aim to mitigate these issues by learning the statistical properties of clean speech [29], though perfect artifact avoidance remains elusive, particularly at low signal-to-noise ratios.

1.3.3 Training Data Scarcity and Quality

Supervised learning approaches for speech denoising typically require paired datasets containing clean speech and corresponding noisy versions. Obtaining such data at scale presents practical difficulties, as truly "clean" speech recordings are rare in natural environments [30]. Recording conditions, equipment limitations, and environmental contamination often introduce residual noise. Alternative methodologies, including unsupervised and self-supervised approaches, leverage unpaired or synthetic data but face their own optimization challenges and performance limitations [29].

1.3.4 Limited Generalization to Unseen Conditions

Models trained on specific noise conditions frequently demonstrate performance degradation when encountering novel acoustic environments. This generalization gap stems from differences in noise characteristics, recording conditions, and speech patterns not represented in training data [7]. Adaptation techniques and domain generalization strategies attempt to address this limitation but often require additional computational resources or prior knowledge about the target environment [26]. The development of truly robust denoising systems that perform consistently across diverse real-world scenarios remains an active research objective.

1.3.5 Real-Time Processing Constraints

Many practical applications, including telecommunication systems and hearing aids, impose strict latency requirements. Complex denoising algorithms, particularly those based on iterative refinement or multi-stage processing, may exceed acceptable delay boundaries [29]. Computational efficiency must be balanced against denoising performance, often necessitating architectural compromises. Recent advances in efficient sampling and model compression aim to address these constraints while maintaining acceptable quality levels [14].

1.3.6 Perceptual Evaluation Challenges

The subjective nature of speech quality complicates objective evaluation of denoising algorithms. While metrics like signal-to-noise ratio improvement provide quantitative measures, they often correlate poorly with human perception of speech naturalness and intelligibility [28]. Comprehensive evaluation requires both objective measurements and subjective listening tests, which are time-consuming and expensive to conduct [1]. Standardized evaluation protocols help ensure comparability between different approaches but cannot fully capture the perceptual dimensions of speech quality [31].

1.3.7 Multi-Microphone and Multi-Source Complexity

Multi-microphone systems introduce additional challenges beyond single-channel processing. Spatial separation of multiple sound sources requires accurate source localization and sophisticated beamforming techniques. Array geometry, calibration requirements, and room acoustics further complicate algorithm design and implementation. The computational complexity of multi-channel processing increases significantly with the number of microphones, presenting practical deployment challenges for resource-constrained devices.

1.3.8 Summary and Research Directions

These interconnected challenges highlight the complexity of speech denoising as a research domain. While significant progress has been made, particularly through deep learning approaches, substantial work remains to develop systems that are simultaneously effective, efficient, and robust across diverse real-world conditions. Future research directions include hybrid approaches combining traditional signal processing with machine learning, improved unsupervised training methodologies, and more perceptually aligned evaluation frameworks.

1.4 Learning from Data for Denoising Models

Data-driven approaches have become central to denoising research, with deep learning models learning the mapping between noisy and clean signals directly from examples. This paradigm achieves significant gains over traditional handcrafted methods by modeling complex noise distributions across diverse domains [32, 33]. A key advantage of these approaches is their flexibility when clean reference data are scarce. Learning-based frameworks can perform denoising using unpaired or noisy-only data by exploiting the statistical properties of noise rather than requiring perfectly clean training targets [30]. Generative methods further extend this by learning latent representations of clean signal manifolds through denoising objectives, as demonstrated by denoising autoencoders [34], with successful applications spanning biomedical and scientific signals such as ECG and material data [35, 36].

1.4.1 Types of Learning Paradigms

Learning paradigms are conceptual frameworks that define how models learn from data, the type of supervision involved, and the learning objective. They are broadly classified into supervised, unsupervised, semi-supervised, and reinforcement learning [37], with the choice of paradigm directly shaping the nature of the extracted knowledge [38]. A more data-centric taxonomy further distinguishes between predictive, interpretive, and generative paradigms, where data quality and structure determine the most suitable approach [39, 40]. Within the generative paradigm, diffusion models have emerged as a prominent modern approach. DDPM [10] and its more efficient variant DDIM [12] learn to generate data through iterative noise removal, while models such as BDDM demonstrate their adaptability to complex speech signals [29]. Complementing this, representation learning — an unsupervised paradigm — aims to capture the underlying structure of data without explicit labels, with applications in learning disentangled speech features [23]. Overall, learning paradigms span a broad spectrum, with generative and representation learning standing as key advancements driving modern speech processing research.

1.4.2 Supervised, Unsupervised, Generative Approaches, and Self-Supervised

The four primary learning approaches can be summarized as follows. Supervised learning trains models on labeled input-output pairs and forms the foundation of deep learning for tasks such as speech and image recognition [41]. Unsupervised learning operates on unlabeled data to discover inherent structures and meaningful representations [42, 43]. Generative approaches learn the underlying data distribution to produce new samples,

with Generative Adversarial Networks (GANs) being a notable example [44, 45]. Self-supervised learning bridges the gap between supervised and unsupervised paradigms by generating supervisory signals directly from the data itself, with applications in anomaly detection and visual semantics [46, 42]. Together, these paradigms form a comprehensive taxonomy of modern machine learning approaches, each suited to different data availability and task requirements [39].

Table 1.2 summarizes the key differences between these four learning paradigms, highlighting their definitions and typical applications in speech processing.

Table 1.2: *Comparison of Machine Learning Paradigms*

Paradigm	Definition	Examples / Applications
Supervised Learning	Model learns from labeled data (input-output pairs)	Classification (speech emotion recognition), Regression (predicting noise levels)
Unsupervised Learning	Model finds patterns from unlabeled data	Clustering (speaker diarization), Dimensionality reduction (PCA)
Self-Supervised Learning	Model generates labels from input data itself to learn representations	Masked audio modeling, contrastive learning for speech
Generative Models	Model learns the underlying data distribution to generate new samples	DDPM, GANs, VAE for speech synthesis and denoising

1.5 Denoising Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) are a class of generative models designed to learn complex data distributions through a sequential denoising process. These models have recently shown remarkable performance in speech enhancement tasks, providing robust results across different noise conditions [10, 11]. DDPMs consist of two main components: a forward diffusion process, which gradually adds noise to clean speech, and a reverse denoising process, which reconstructs the original signal.

Table 1.3 summarizes the key properties of diffusion models that make them particularly suitable for speech denoising applications, including their probabilistic framework, multi-step refinement, generalization capability, stable training, and scalability.

Diffusion models operate through two complementary processes: the *forward diffusion* that systematically corrupts clean speech with Gaussian noise, and the *reverse denoising* that learns to reconstruct the original signal. This bidirectional approach allows the model to capture the complex statistical relationships between clean and noisy speech signals.

1.5.1 Forward Diffusion Process

In the forward diffusion process, a clean speech signal x_0 is progressively corrupted by Gaussian noise over T time steps. At each step t, noise is added according to a predefined

Table 1.3: *Key Properties of Diffusion Models for Speech Denoising*

Property	Description and Relevance to Speech Denoising
Probabilistic Framework	Models the data distribution explicitly, allowing for uncertainty estimation in denoising decisions.
Multi-step Refinement	Iterative denoising process enables gradual improvement, mimicking human auditory perception.
Generalization Capability	Learns underlying speech structures rather than noise patterns, improving performance on unseen noise types.
Stable Training	Unlike GANs, diffusion models avoid mode collapse and training instability issues.
Scalability	Can handle high-dimensional speech data (waveforms or spectrograms) effectively.

variance schedule β_t, producing a sequence of increasingly noisy signals $x_1, x_2, \ldots, x_T$. This process allows the model to learn the gradual transition from clean to fully noisy speech, forming the foundation for the subsequent reverse process [10, 12, 14].

The forward process is defined as:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I) \tag{1.8}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s)$ and β_t is the noise schedule controlling the amount of noise added at each step.

Figure 1.7 illustrates the forward diffusion process, where Gaussian noise is progressively added to the clean speech signal x_0 over T time steps, transforming it into pure noise x_T.

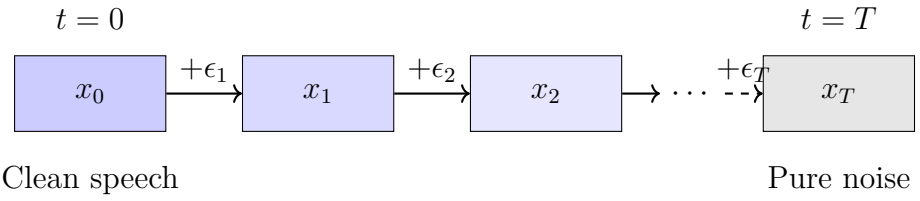


Figure 1.7: *Forward diffusion process: adding Gaussian noise over time*

For speech signals, this process can be applied directly to waveforms or to spectrogram representations. The choice of noise schedule β_t is crucial:

- **Linear schedule:** Provides uniform corruption across all time steps
- **Cosine schedule:** Preserves more signal information in early steps, which is beneficial for speech denoising tasks [12]
- **Sigmoid schedule:** Allows for smoother transitions between noise levels

The variance-preserving property ensures that the signal-to-noise ratio decreases monotonically throughout the diffusion process, enabling stable training and effective denoising in the reverse process.

1.5.2 Reverse Denoising Process

The reverse process aims to recover the original clean speech from the noisy inputs. This is done by learning a parameterized model that predicts either the added noise or the clean signal at each step. By iteratively applying this learned denoising function, DDPMs reconstruct high-quality speech from corrupted inputs. This probabilistic iterative refinement ensures accurate recovery even under diverse noise conditions [10, 11, 29].

The reverse process is mathematically formulated as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z, \quad z \sim \mathcal{N}(0, I) \tag{1.9}$$

Figure 1.8 illustrates the forward and reverse processes in DDPM/DDIM, showing how clean speech x_0 is corrupted through forward steps and then reconstructed through reverse denoising steps.

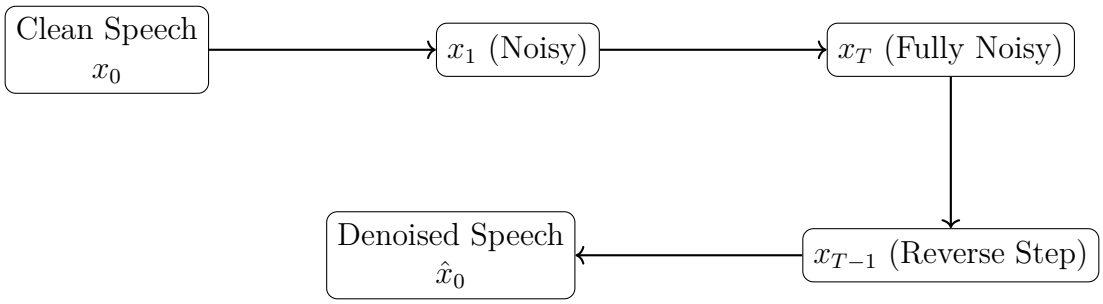


Figure 1.8: *Forward and Reverse Process in DDPM/DDIM*

The forward process (top path) gradually adds noise to clean speech, while the reverse process (bottom path) iteratively removes noise to reconstruct the original signal. The neural network ϵ_θ is trained to predict the noise added at each step, enabling the model to effectively reverse the diffusion process and recover high-quality speech from noisy inputs.

1.5.3 Advantages and Generalization

DDPMs exhibit several advantages over traditional denoising methods and other deep learning approaches:

- Ability to model complex, non-stationary noise distributions.
- Robust generalization to unseen noise types and recording conditions.
- High perceptual quality of the reconstructed speech, preserving natural timbre and intelligibility [11, 12].

These properties make DDPMs particularly suitable for real-world speech denoising applications.

1.5.4 DDIM Approach

Denoising Diffusion Implicit Models (DDIMs) are an improvement over DDPMs, introducing a deterministic sampling method that allows faster inference while maintaining generation quality. Unlike DDPMs, which rely on stochastic sampling over many

steps, DDIMs can generate clean speech in fewer steps without significant degradation in quality. This efficiency makes DDIMs appealing for practical speech enhancement applications [12, 14, 29].

Table 1.4 provides a comparative analysis between DDPM and DDIM, highlighting their key differences in sampling type, inference speed, memory requirements, quality preservation, consistency, and real-time suitability.

Table 1.4: *DDPM vs DDIM: Comparative Analysis for Speech Denoising*

Aspect	**DDPM [10]**	**DDIM [12]**
Sampling Type	Stochastic Markov chain	Deterministic non-Markovian
Typical Inference Steps	1000	50-100
Inference Speed	Slow	10-20× faster
Memory Requirements	High (stores all steps)	Lower
Quality Preservation	Excellent	Comparable to DDPM
Consistency	Variable outputs (stochastic)	Deterministic outputs
Suitability for Real-time	Limited	Excellent

The key innovation in DDIM is the reformulation of the reverse process as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t) \tag{1.10}$$

where $\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}$ is the predicted clean signal.

This deterministic formulation allows for:

- **Accelerated inference**: Achieves comparable results in 50-100 steps vs. 1000 in DDPM
- **Consistent outputs**: Same input always produces same output
- **Flexible trade-offs**: Can adjust speed/quality balance by varying step count

For speech denoising applications, DDIM's efficiency makes it particularly suitable for:

- Real-time communication systems
- Mobile and embedded devices
- Large-scale processing of speech datasets
- Interactive applications requiring low latency

1.5.5 Applications on Arabic Speech Datasets

Recent studies have adapted diffusion-based models to Arabic speech datasets, demonstrating their effectiveness across different linguistic contexts. Using these models, researchers achieved robust denoising and emotion classification results in Arabic speech, highlighting the versatility of diffusion-based approaches [47, 11].

Table 1.5 lists several Arabic speech datasets that are suitable for training and evaluating diffusion-based denoising systems, including their size, characteristics, and suitability for denoising tasks.

Table 1.5: *Arabic Speech Datasets for Diffusion-Based Denoising*

Dataset	Size	Characteristics	Suitability for Denoising
KSU Emotions	5 hours	Emotional speech, clean	Good for testing denoising robustness
ADSD	10 hours	Spontaneous speech, noisy	Real-world testing scenarios
AraVoice	50 hours	Multi-dialect, clean	Large-scale training
QASR	2,000 hours	MSA, various noise levels	Comprehensive evaluation

The application of DDIM to Arabic speech presents unique opportunities:

Phonetic Considerations: Arabic contains emphatic consonants and pharyngeal sounds that may require specialized noise modeling in diffusion frameworks.

Dialectal Adaptation: DDIM's efficiency allows for rapid experimentation across different Arabic dialects (MSA, Egyptian, Gulf, etc.).

Computational Efficiency: The faster inference of DDIM is particularly beneficial for processing large Arabic speech corpora with limited computational resources.

Recent work by [14] demonstrates that DDIM-based approaches can be effectively adapted to non-English languages, including Arabic, by fine-tuning on language-specific speech characteristics.

1.5.6 DDIM for Speech Denoising: Practical Considerations

The adoption of DDIM for speech denoising involves several practical considerations:

Algorithm 1 DDIM-based Speech Denoising Algorithm

procedure DDIMDenoise(y, ϵ_θ, S)
  $x_S \leftarrow y$ ▷ Start from noisy input
  for $t = S$ **down to** 1 **do**
    $\epsilon_t \leftarrow \epsilon_\theta(x_t, t)$ ▷ Predict noise
    $\hat{x}_0 \leftarrow \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}}$ ▷ Estimate clean speech
    $x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_t$ ▷ DDIM update
  end for
  return $\hat{x}_0$
end procedure

1.6 Evaluation Metrics for Speech Denoising

1.6.1 Signal-to-Noise Ratio (SNR) in Speech Quality Assessment

The Signal-to-Noise Ratio (SNR) serves as a fundamental objective metric for evaluating speech enhancement algorithms. It quantifies the level of the desired speech signal relative to the background noise. Mathematically, the SNR is defined as the logarithmic ratio of the signal power to the noise power [1, 31]:

$$\text{SNR}_{\text{global}} = 10 \log_{10} \left(\frac{\sum_{n=1}^{N} s^2[n]}{\sum_{n=1}^{N} (s[n] - \hat{s}[n])^2} \right) \quad \text{(dB)} \tag{1.11}$$

where:

- $s[n]$ represents the original clean speech signal
- $\hat{s}[n]$ denotes the enhanced (denoised) speech signal
- N is the total number of samples

For more perceptually relevant assessment, the segmental SNR (SNR_{seg}) is often employed. It is calculated over short-time frames to better reflect human auditory perception and avoid the influence of silent regions [28]:

$$\text{SNR}_{\text{seg}} = \frac{1}{M} \sum_{m=0}^{M-1} 10 \log_{10} \left(\frac{\sum_{n=mK}^{mK+K-1} s^2[n]}{\sum_{n=mK}^{mK+K-1} (s[n] - \hat{s}[n])^2} \right) \tag{1.12}$$

where:

- M is the number of frames
- K is the frame length (typically 20-30 ms)
- Averaging is performed only over frames where speech is present to avoid bias from silent regions

In practical evaluation protocols, SNR improvement (ΔSNR) quantifies the enhancement performance [31]:

$$\Delta\text{SNR} = \text{SNR}_{\text{output}} - \text{SNR}_{\text{input}} \tag{1.13}$$

where $\text{SNR}_{\text{input}}$ and $\text{SNR}_{\text{output}}$ represent the SNR values before and after processing, respectively.

While SNR provides a straightforward computational measure, it has limitations in correlating with subjective listening tests, particularly at low SNR levels where nonlinear processing artifacts may dominate perception [28]. Therefore, SNR is typically used in conjunction with other objective measures and subjective evaluations within comprehensive assessment frameworks [31, 1].

1.6.2 Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)

The Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) is an improved version of the standard SNR that addresses the scaling ambiguity problem. Unlike traditional SNR, SI-SDR is invariant to the scaling of the estimated signal, making it particularly suitable for evaluating generative models where the output may be rescaled relative to the reference.

The SI-SDR is computed by first projecting the denoised signal onto the clean signal direction[48]:

$$\alpha = \frac{\langle \hat{s}, s \rangle}{\langle s, s \rangle} \tag{1.14}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, s is the clean signal, and $\hat{s}$ is the denoised signal. The scaled clean signal is then $s_{\text{scaled}} = \alpha \cdot s$, and the distortion is $e = \hat{s} - s_{\text{scaled}}$.

The SI-SDR in decibels is defined as:

$$\text{SI-SDR} = 10 \cdot \log_{10} \left(\frac{\|s_{\text{scaled}}\|^2}{\|e\|^2} \right) \tag{1.15}$$

Higher SI-SDR values indicate better denoising performance, and the metric is insensitive to arbitrary scaling of the output signal, which is a common issue in deep learning-based denoising models.

1.6.3 Perceptual Evaluation of Speech Quality (PESQ)

PESQ is an objective metric standardized by ITU-T Recommendation P.862. It predicts the perceived quality of speech signals as would be rated by human listeners. The PESQ score ranges from -0.5 to 4.5, where higher scores indicate better quality.

The calculation involves several steps[49]:

1. Temporal alignment of the reference and denoised signals.
2. Transformation to a perceptual domain that simulates human auditory processing.
3. Calculation of disturbance densities in time and frequency.
4. Mapping to a predicted MOS-like score.

PESQ is widely used for evaluating speech enhancement algorithms because it correlates well with human subjective ratings.

1.6.4 Short-Time Objective Intelligibility (STOI)

STOI measures the intelligibility of speech signals, i.e., how well a listener can understand the spoken content. The STOI score ranges from 0 to 1 (or 0% to 100%), with values closer to 1 indicating higher intelligibility.

The computation procedure is as follows[50]:

1. Divide both clean and denoised signals into short-time frames (typically 256 samples).
2. Analyze one-third octave frequency bands between 150 Hz and 4.3 kHz.
3. Extract the temporal envelopes of each band.

4. Calculate the Pearson correlation coefficient between the clean and denoised envelopes.

5. Average the correlation scores across all frames and frequency bands.

STOI is particularly important for applications where understanding speech content is more critical than perceptual quality.

1.6.5 Mean Opinion Score (MOS)

MOS is the traditional subjective metric based on human listening tests. Listeners rate speech quality on a five-point scale [51].

Table 1.6 presents the MOS rating scale, which ranges from 1 (Bad) to 5 (Excellent), along with the corresponding quality descriptions.

Table 1.6: *MOS Rating Scale*

Rating	Quality Description
5	Excellent (Imperceptible distortion)
4	Good (Perceptible but not annoying)
3	Fair (Slightly annoying)
2	Poor (Annoying)
1	Bad (Very annoying)

Since conducting human listening tests is expensive and time-consuming, MOS is often predicted using objective metrics. A common mapping from PESQ to MOS is:

$$\text{MOS} \approx 1 + \frac{4}{4.5} \cdot \text{PESQ} = 1 + 0.89 \cdot \text{PESQ} \tag{1.16}$$

This approximation assumes PESQ ranges from -0.5 to 4.5, mapping linearly to MOS values from 1 to 5.

1.6.6 Real-Time Factor (RTF)

RTF measures the computational efficiency of a denoising algorithm, which is critical for real-time applications such as teleconferencing and hearing aids. RTF is defined as:

$$\text{RTF} = \frac{T_{\text{processing}}}{T_{\text{audio}}} \tag{1.17}$$

where $T_{\text{processing}}$ is the time taken by the algorithm to process the audio signal, and T_{audio} is the duration of the audio signal itself.

Interpretation of RTF values:

- RTF < 1: The algorithm is **faster than real-time** (suitable for live applications)
- RTF $= 1$: The algorithm processes in exactly real-time
- RTF > 1: The algorithm is **slower than real-time** (not suitable for live processing)

For example, if an algorithm processes 10 seconds of audio in 2 seconds, the RTF is $2/10 = 0.2$, which is **faster than real-time**. Conversely, if the same algorithm takes 15 seconds to process 10 seconds of audio, the RTF is $15/10 = 1.5$, which is slower than real-time.

1.7 Conclusion

DDPMs and DDIMs represent a state-of-the-art approach for generative speech denoising. Through the forward and reverse processes, these models can learn to reconstruct clean speech from noisy inputs with high fidelity. DDIM enhancements allow for faster inference without sacrificing quality, and applications on Arabic speech datasets confirm their generalizability across languages and noise conditions. Overall, diffusion-based speech denoising is a promising direction for both research and real-world audio processing applications.

Chapter 2

State of the Art in Speech Denoising

2.1 Introduction

The field of monaural speech enhancement has undergone significant transformation with the advent of deep learning techniques, evolving from traditional signal processing methods to sophisticated neural network architectures [3]. This evolution reflects broader trends in artificial intelligence applications to audio processing, where data-driven approaches have progressively supplanted model-based methodologies. According to recent comprehensive analyses, deep neural network techniques now represent the state of the art in speech denoising, offering substantial improvements over classical approaches particularly in non-stationary noise environments [3].

The transition to data-driven paradigms has enabled more effective handling of complex noise scenarios that challenge traditional statistical methods. As noted in contemporary reviews, deep learning approaches learn intricate mappings between noisy and clean speech directly from data, capturing nonlinear relationships that elude conventional algorithmic designs [3]. This capability has led to measurable advances in both objective metrics and subjective listening tests, establishing neural methods as the current benchmark for speech enhancement performance.

Recent innovations in this domain include the incorporation of perceptual optimization criteria, such as deep feature losses that leverage pretrained networks to preserve speech naturalness during denoising [52]. These developments highlight the ongoing refinement of deep learning techniques beyond simple signal reconstruction toward more perceptually relevant objectives. This chapter surveys these methodological advances through three primary lenses: traditional approaches predating widespread deep learning adoption, conventional deep neural network architectures, and cutting-edge generative models that represent the current research frontier in speech denoising.

2.2 Traditional Speech Denoising Methods

2.2.1 Statistical Approaches

Traditional statistical-model-based methods for monaural speech enhancement typically rely on specific assumptions about the distributions of the speech and/or noise signals. The core principle involves deriving a spectral gain function by minimizing a chosen statistical criterion, such as the Minimum Mean-Square Error (MMSE), under these assumed models [5].

Gaussian Statistical Model: A foundational approach assumes that the speech and noise are statistically independent Gaussian random processes. Based on this, the MMSE short-time spectral amplitude (MMSE-STSA) estimator and the MMSE log-spectral amplitude (MMSE-LSA) estimator were derived. These form a cornerstone of statistical speech enhancement [6].

Beyond Gaussian Models: It is well-established that speech is non-Gaussian. Consequently, other statistical models have been explored to improve performance, including Gamma, Laplacian, and Super-Gaussian distributions for speech [6].

Deterministic-Stochastic Speech Models: Some methods model speech as a combined stochastic-deterministic signal rather than a purely stochastic one. For instance, the harmonic-plus-noise model decomposes speech into a deterministic harmonic component and a stochastic residual. MMSE estimators based on such models have been developed and shown to offer marginal objective performance improvements [6].

The Central Role of SNR Estimation: For most statistical methods, the accurate estimation of the *a priori* Signal-to-Noise Ratio (SNR) is a critical challenge. The decision-directed (DD) approach is a widely used method for *a priori* SNR estimation, which helps control musical noise artifacts. However, it introduces a one-frame delay bias, leading to speech distortion and a reverberation effect [5].

2.2.2 Time-Domain Methods

Time-domain methods aim to estimate the clean speech signal directly from the noisy waveform without explicit short-term spectral analysis and synthesis.

General Principle: These methods operate directly on the time-domain samples, designing filters or performing estimations without a transformation to the frequency domain [6].

Adaptive Filtering Approaches: Classical time-domain methods include adaptive filters such as Least Mean Squares (LMS) and Normalized LMS (NLMS), which estimate noise by minimizing the error between the desired and actual signals. These methods are effective when a reference noise signal is available.

Subspace Methods: Another category of time-domain techniques is subspace methods, such as Karhunen-Loève Transform (KLT) based filtering, which decompose the noisy signal into signal and noise subspaces and retain only the signal-dominant components.

Limitations: The primary limitation of time-domain methods is their sensitivity to the statistical properties of noise and reduced effectiveness in non-stationary acoustic environments. Consequently, frequency-domain methods have become more dominant in speech enhancement research [6].

2.2.3 Spectral-Domain Methods

Spectral-domain, or frequency-domain, methods are the most extensively studied class of traditional speech denoising techniques. They operate by applying a gain to each time-frequency (T-F) bin of the noisy speech's short-time Fourier transform (STFT) [6]. The noisy speech signal is commonly modeled as an additive mixture in the time domain:

$$y(n) = s(n) + d(n) \tag{2.1}$$

where $s(n)$ denotes the clean speech signal and $d(n)$ represents additive noise [2]. Applying the Short-Time Fourier Transform (STFT), the model becomes:

$$Y(k,l) = S(k,l) + D(k,l) \tag{2.2}$$

where k and l denote frequency bin and time frame indices, respectively [6].

General Processing Flow: A typical system involves several key modules: noise power spectral density (PSD) estimation, *a priori* SNR estimation, speech presence probability (SPP) estimation, spectral gain calculation, and time-domain reconstruction via the inverse STFT and overlap-add method [6].

Spectral Subtraction: This is a fundamental and heuristic spectral-domain method. It involves subtracting an estimate of the noise spectrum from the spectrum of the noisy speech. Variants include magnitude, power, and root-based subtraction [6]. The enhanced magnitude spectrum using spectral subtraction is given by:

$$|\hat{S}(k,l)| = \max\Big(|Y(k,l)| - \alpha|\hat{D}(k,l)|, 0\Big) \tag{2.3}$$

where α is an over-subtraction factor controlling noise suppression [8].

Wiener Filtering: A classic statistical filtering approach is the Wiener filter, which minimizes the mean-square error between the estimated and clean speech spectra. The Wiener filter gain is defined as:

$$G_{\text{Wiener}}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \tag{2.4}$$

where $\xi(k,l)$ denotes the *a priori* signal-to-noise ratio (SNR) [2].

Advanced Spectral Gain Estimators: Building on statistical models, estimators like MMSE-STSA and MMSE-LSA provide analytically derived gain functions. The choice between them represents a trade-off between noise suppression and speech distortion [6].

Addressing Limitations of the Decision-Directed Approach

To combat the delay and bias of the standard DD approach, advanced methods have been proposed.

Two-Step Noise Reduction (TSNR): This technique refines the *a priori* SNR estimate by using a second processing step. It aims to remove the reverberation effect caused by the DD bias while maintaining its low musical noise property [5].

Harmonic Regeneration Noise Reduction (HRNR): This method addresses the harmonic distortion caused by erroneous suppression of speech harmonics, especially at low SNRs. It uses a non-linearity on an initially enhanced signal to regenerate harmonics, creating an artificial reference to compute a more accurate spectral gain that preserves speech structure [5].

Phase Processing: Traditionally, the phase spectrum was considered unimportant and was often left unprocessed (using the noisy phase). However, later research confirmed that phase estimation can improve quality, leading to dedicated phase processing modules and phase-sensitive cost functions [6].

Table 2.1 provides a comparative overview of traditional speech denoising methods, including Spectral Subtraction, Wiener Filtering, and MMSE-STSA, highlighting their respective domains, advantages, and limitations.

Table 2.1: *Comparison of Traditional Speech Denoising Methods*

Criteria	Spectral Subtraction (SS)	Wiener Filtering (WF)	MMSE-STSA
Domain	Frequency	Frequency	Frequency
Advantages	Simple, fast	Optimal in MSE sense	Better perceptual quality
Limitations	Musical noise, sensitive to SNR	Requires noise estimation	Higher complexity
References	[8]	[2]	[2]

2.3 Machine Learning and Deep Learning Methods

Machine Learning (ML) encompasses algorithms that improve their performance through experience, while Deep Learning (DL) represents a specialized subset that utilizes multi-layered neural networks to model high-level abstractions in data [53]. This data-driven paradigm has achieved remarkable success across diverse applications, including computer vision, natural language processing, and bioinformatics [54]. In the domain of speech enhancement, ML and DL methods learn complex mappings between noisy and clean speech directly from data, enabling superior handling of non-stationary noise compared to traditional model-based approaches [6].

2.3.1 Conventional Machine Learning Approaches

Prior to the dominance of deep neural networks, several conventional machine learning techniques were applied to monaural speech enhancement. These methods often functioned as data-driven extensions of traditional statistical frameworks or introduced novel learning-based estimators for key parameters.

One prominent approach is Non-Negative Matrix Factorization (NMF). In supervised NMF, basis matrices for speech and noise are learned from clean training datasets. During

enhancement, the noisy speech magnitude spectrogram is factorized, and the clean speech is reconstructed using the pre-learned speech basis and its corresponding activation coefficients [6]. Codebook and dictionary-based methods, such as K-SVD, operate on a similar principle by training a dictionary of spectral patterns from clean speech and using sparse coding for denoising [6]. While these methods can outperform traditional techniques like Wiener filtering in noise attenuation, their modeling capacity and generalization to unseen noise types are limited by the linear nature of the decomposition and the finite size of the dictionary [6].

Another significant direction involves using shallow neural networks or other ML models for parameter estimation. A key example is the estimation of the crucial *a priori* Signal-to-Noise Ratio (SNR). The DeepXi framework employs a deep neural network to estimate the *a priori* SNR, which is then integrated into established estimators like the MMSE-LSA, forming a hybrid model that combines data-driven learning with traditional statistical derivation [6]. This demonstrates how conventional ML can be used to improve the accuracy of core parameters in the enhancement pipeline.

Table 2.2 provides a comparative overview of conventional machine learning-based denoising methods, including GMM-based denoising, SVM regression, and sparse coding, highlighting their input features, learning types, strengths, and weaknesses.

Table 2.2: *Comparison of Conventional Machine Learning-Based Denoising Methods*

Criteria	**GMM-based Denoising**	**SVM Regression**	**Sparse Coding**
Input Features	MFCC, STFT	Spectral features	Spectral frames
Learning Type	Supervised	Supervised	Unsupervised
Strengths	Captures statistical distributions	Good for small datasets	Robust to noise
Weaknesses	Limited generalization	Needs feature engineering	Computationally expensive
References	[55]	[53]	[22]

2.3.2 Deep Learning-Based Speech Denoising

Deep Learning has become the predominant approach for speech denoising, with its capacity to learn hierarchical feature representations directly from data [54]. DL-based speech enhancement methods have shown to surpass traditional methods, especially in challenging conditions with low SNRs, high reverberation, and non-stationary noise [6].

The evolution of network architectures has been central to this progress. Early deep learning methods for speech enhancement utilized Fully Connected Deep Neural Networks (DNNs) to map frames of spectral features to clean targets [6]. To better model the temporal dynamics of speech, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were subsequently applied, leading to improved performance by capturing long-range dependencies [6]. Convolutional Neural Networks (CNNs), known for efficiently capturing local spectral patterns, were also adopted. A

highly effective architecture that combines these strengths is the Convolutional Recurrent Network (CRN), which integrates the local feature extraction capability of CNNs with the temporal modeling of RNNs [6].

A critical advancement was the move towards complex-valued network architectures. Recognizing the importance of phase information for speech quality, models like the Deep Complex Convolutional Recurrent Network (DCCRN) operate directly on the complex spectrum, jointly estimating real and imaginary parts. This allows for implicit phase recovery and has delivered state-of-the-art performance in benchmarks like the Deep Noise Suppression (DNS) Challenge [6].

The models are trained using specific learning paradigms. A common approach is *masking-based* training, where the network estimates a time-frequency mask applied to the noisy spectrogram. Mapping-based training, in contrast, directly predicts the clean speech spectrum or waveform [6].

Mask-based speech enhancement estimates a time-frequency mask $M(k, l)$ such that:

$$\hat{S}(k, l) = M(k, l) \cdot Y(k, l) \tag{2.5}$$

where $M(k, l)$ can be an Ideal Ratio Mask (IRM) or Complex Ratio Mask (CRM) [6].

The training process is guided by loss functions; while Mean-Squared Error (MSE) on spectral magnitudes has been widely used, recent trends employ composite losses combining time-domain metrics like Scale-Invariant Signal-to-Noise Ratio (SI-SNR) with frequency-domain or perceptual losses to better align with subjective quality [6].

Table 2.3 provides a comparative overview of deep learning-based speech denoising methods, including DNNs, CNNs, and RNNs/LSTMs, highlighting their respective inputs, outputs, advantages, and disadvantages.

Table 2.3: *Comparison of Deep Learning-Based Speech Denoising Methods*

Criteria	**DNN**	**CNN**	**RNN / LSTM**
Input	STFT magnitude	Spectrogram	Temporal features
Output	Clean STFT	Enhanced spectrogram	Clean audio
Advantages	Learns nonlinear mapping	Captures local patterns	Captures temporal dependencies
Disadvantages	Needs lots of data	Sensitive to overfitting	Slower training
References	[7]	[3]	[6]

2.4 Advanced Generative Models for Speech Denoising

Recent advancements in speech denoising have been significantly driven by the development of advanced generative models. These models learn the underlying probability distribution of clean speech data, enabling them to generate or reconstruct high-quality speech signals from noisy observations.

2.4.1 GAN-Based Speech Denoising

Generative Adversarial Networks (GANs) represent a powerful class of generative models that have been adapted for speech enhancement. A foundational work in this area is SEGAN (Speech Enhancement GAN), which operates directly on raw speech waveforms. In the GAN framework, a generator network is trained to produce enhanced speech from noisy inputs, while a discriminator network is trained to distinguish between the generated speech and real clean speech, creating an adversarial training dynamic [9].

Building upon this, conditional GAN architectures have been proposed where the generation process is conditioned on the noisy input speech. This approach guides the model to learn the mapping from noisy to clean speech specifically, rather than generating speech unconditionally [55]. Further innovation led to MetricGAN, which incorporates specific objective speech quality metrics directly into the adversarial training process. Instead of using a standard discriminator that outputs a binary real/fake decision, MetricGAN employs a critic network trained to estimate a target objective score (such as a perceptual metric), allowing the generator to optimize for improved scores on established evaluation metrics [56]. The adversarial objective of a GAN is formally defined as:

$$\min_G \max_D \; \mathbb{E}_{s\sim p_{\text{data}}}[\log D(s)] + \mathbb{E}_{y\sim p_y}[\log(1 - D(G(y)))] \tag{2.6}$$

where G and D denote the generator and discriminator networks, respectively [9].

2.4.2 Waveform-Based Approaches

Direct waveform modeling represents an alternative to frequency-domain processing, aiming to preserve the full temporal structure of the speech signal. The SEGAN model is an example of a waveform-based generative approach, as it processes and generates raw audio samples using one-dimensional convolutional architectures in the generator and discriminator [9].

Another significant waveform-based architecture is an adaptation of the WaveNet model for speech denoising. This model employs dilated causal convolutions to capture long-range dependencies in the time-domain signal. It is trained in a fully supervised manner to predict clean audio samples given a context of noisy samples, demonstrating the capability of autoregressive models to perform denoising directly in the time domain [57].

2.4.3 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) and related score-based generative models have emerged as state-of-the-art approaches for speech enhancement. These models define a forward diffusion process that gradually corrupts a clean speech signal by adding Gaussian noise over many steps until it becomes pure noise. A neural network is then trained to reverse this process, learning to iteratively denoise a signal [29, 13].

A key implementation applies this framework in the complex Short-Time Fourier Transform (STFT) domain. The model learns the gradient of the data log-density (the score) for the real and imaginary parts of the STFT coefficients. During inference, it performs iterative denoising using a predictor-corrector sampling scheme, such as Annealed Langevin Dynamics, starting from the noisy observation [29]. This approach has been successfully extended to jointly handle both additive noise and reverberation. The diffusion model is trained to predict the clean, anechoic speech spectrogram, and

the reverse process is initiated from the noisy and reverberant observation, effectively performing enhancement and dereverberation simultaneously [13].

The forward diffusion process gradually corrupts clean speech by adding Gaussian noise:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{2.7}$$

where x_0 is the clean speech representation and α_t controls the noise schedule [10].

The reverse diffusion process is modeled as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z, \quad z \sim \mathcal{N}(0, I) \tag{2.8}$$

where ϵ_θ is the neural network prediction of noise and σ_t is the step-dependent variance [10, 12].

The neural network ϵ_θ is trained to minimize the mean squared error (MSE) between the predicted noise and the actual noise, ensuring stable and effective learning [10]. Recent works have also explored accelerated sampling and multi-step denoising strategies to improve inference speed while maintaining high speech quality [29, 14].

Figure 2.1 provides a detailed illustration of the forward and reverse diffusion processes in DDPM-based speech enhancement, including STFT representation and step-dependent parameters α_t and σ_t.

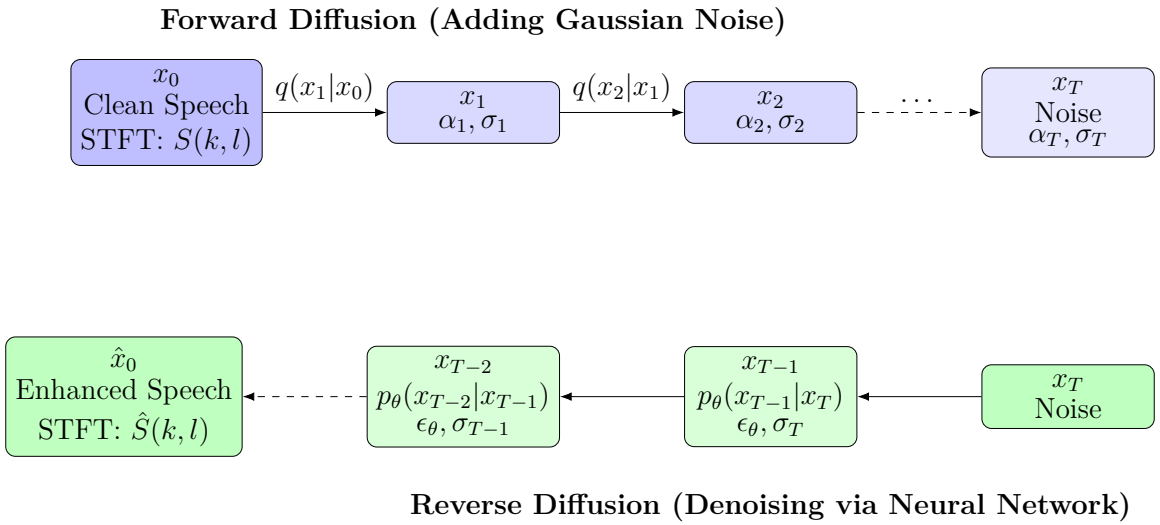


Figure 2.1: *Detailed illustration of forward and reverse diffusion processes in DDPM-based speech enhancement, including STFT representation and step-dependent parameters α_t and σ_t.*

Table 2.4 provides a comparative overview of advanced generative speech denoising models, including GANs, DDPMs, and self-supervised approaches, highlighting their respective domains, key ideas, strengths, and weaknesses.

Table 2.4: *Comparison of Advanced Generative Speech Denoising Models*

Criteria	GAN	DDPM	Self-Supervised
Domain	Waveform / Spectrogram	Spectrogram	Latent features
Key Idea	Generator vs Discriminator	Forward + Reverse Diffusion	Representation learning
Strengths	High perceptual quality	Stable, high-quality denoising	Uses unlabeled data
Weaknesses	Training instability	Multiple inference steps	Needs fine-tuning
References	[9, 27]	[10, 11]	[23]

2.5 Conclusion

This chapter has presented a review of the state of the art in monaural speech denoising, tracing the methodological evolution from traditional signal processing to modern data-driven paradigms. Traditional methods, encompassing statistical, time-domain, and spectral-domain approaches, rely on specific assumptions about speech and noise distributions. While effective in controlled conditions, their performance is often limited in complex, real-world acoustic environments [6, 5].

The advent of machine learning and deep learning has marked a significant shift, enabling models to learn complex, non-linear mappings between noisy and clean speech directly from data. Conventional machine learning techniques, such as Non-Negative Matrix Factorization, provided early data-driven enhancements [6]. Deep learning architectures, including Convolutional and Recurrent Neural Networks, have since become dominant, offering superior performance by learning hierarchical feature representations [6, 53, 54].

Recently, advanced generative models have pushed the boundaries of speech enhancement quality. Generative Adversarial Networks (GANs), including conditional and metric-optimized variants, learn to generate clean speech through adversarial training [9, 55, 56]. Waveform-based approaches like SEGAN and WaveNet-based denoisers operate directly on the raw audio signal [9, 57]. Most notably, Denoising Diffusion Probabilistic Models (DDPMs) and score-based generative models have established new state-of-the-art results by learning to iteratively denoise signals through a stochastic diffusion process, demonstrating powerful capabilities in joint enhancement and dereverberation [29, 13]. This progression from model-based to learning-based and ultimately to generative denoising frameworks highlights the ongoing innovation in the quest for robust, high-quality speech enhancement systems.

Chapter 3

System Design and Methodology

3.1 Introduction

This chapter presents the methodological framework and system design for the proposed diffusion-based speech denoising system. DiffWave is a diffusion probabilistic model proposed by Kong et al. [16] for high-fidelity audio synthesis. Unlike its original application which focused on speech synthesis (vocoding), this work adapts the model for speech denoising.

Three speech datasets are employed in this study: SC09 (English digits), Arabic Speech Commands, and LJSpeech (English sentences). A comprehensive acoustic benchmark consisting of six synthetic noise types and eight real-world noise types extracted from the Arabic environment is used. All fourteen noise types are systematically applied at three SNR levels (-5 dB, 0 dB, 5 dB).

This chapter is organized as follows:

- Section 3.2 describes the system architecture,
- Section 3.3 details the development environment,
- Section 3.4 covers data preparation.

3.2 System Architecture and Operational Pipeline

The proposed speech enhancement framework is structurally divided into two distinct operational procedures: the offline data preparation pipeline (where controlled noise is injected for model training and validation) and the online runtime system inference pipeline (where the trained model deploys to enhance unseen corrupted audio). Figure 3.2 illustrates the comprehensive architectural workflow.

3.2.1 DiffWave Model Architecture

DiffWave is a diffusion probabilistic model that operates through two complementary processes: forward diffusion and reverse denoising. The model is built upon a bidirectional dilated convolutional architecture with residual connections, originally designed for high-fidelity audio synthesis.

Forward Diffusion Process

The forward process gradually corrupts a clean speech signal x_0 by adding Gaussian noise over T time steps. At each step t, noise is added according to a predefined variance schedule β_t, producing a sequence of increasingly noisy signals $x_1, x_2, \ldots, x_T$:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I) \tag{3.1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s)$. This process allows the model to learn the gradual transition from clean to fully noisy speech.

Reverse Denoising Process

The reverse process learns to reconstruct the original clean speech by iteratively removing noise:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z, \quad z \sim \mathcal{N}(0, I) \tag{3.2}$$

where ϵ_θ is the neural network that predicts the noise, and σ_t is the step-dependent variance.

3.2.2 Model Configurations: T=50 vs T=200

Two diffusion step configurations are trained and compared in this study:

- **T=50:** 50 diffusion steps (faster training, fewer iterations)
- **T=200:** 200 diffusion steps (finer discretization, computationally heavier)

3.2.3 Operational Pipeline

Phase 1: Controlled Noise Ingestion (Data Preparation)

To robustly train and evaluate the generative diffusion models, a systematic noise injection process is established. Clean speech baseline waveforms are synthetically or organically blended with predefined noise types. Six synthetic noise profiles (e.g., white, pink) and eight real-world ambient noises from the Arabic Speech Commands background library are utilized. These noise distributions are algorithmically added to the clean audio vectors at three SNR levels: -5 dB, 0 dB, and 5 dB.

To validate the data preparation pipeline, we constructed a comprehensive noisy speech dataset following the controlled noise ingestion protocol. Table 3.1 summarizes the resulting dataset statistics.

Table 3.1: *Statistics of the constructed noisy speech dataset*

Parameter	Value
Clean speech sources	SC09 (23,666 files), LJSpeech (13,100 files), Arabic (12,000 files)
Clean files per dataset	100 (randomly selected)
Total clean files	300
Synthetic noise types	white, pink, brownian, running_tap, exercise_bike, doing_the_dishes
Real noise types	cooking_1, cooking_2, water_tap, boiler, washing_machine, street, cafeteria, car
Total noise types	14 (6 synthetic + 8 real)
SNR levels	-5 dB, 0 dB, 5 dB
Total noisy files	12,600 (100 × 14 × 3×3)

Figure 3.1 shows the console output during dataset generation, confirming the successful creation of 12,600 noisy files.

```
Synthetic noises: 6 types
Real noises: 8 types
Arabic WAV files found: 12000
SC09: 100 clean files (from 23666)
LJSpeech: 100 clean files (from 13100)
Arabic: 100 clean files (from 12000)
SC09: processed 20/100 clean files
SC09: processed 40/100 clean files
SC09: processed 60/100 clean files
SC09: processed 80/100 clean files
SC09: processed 100/100 clean files
LJSpeech: processed 20/100 clean files
LJSpeech: processed 40/100 clean files
LJSpeech: processed 60/100 clean files
LJSpeech: processed 80/100 clean files
LJSpeech: processed 100/100 clean files
Arabic: processed 20/100 clean files
Arabic: processed 40/100 clean files
Arabic: processed 60/100 clean files
Arabic: processed 80/100 clean files
Arabic: processed 100/100 clean files

Total noisy files created: 12600
```

Figure 3.1: *Console output showing the data preparation process: (a) loading of clean speech from three datasets (SC09, LJSpeech, Arabic), (b) addition of 14 noise types (6 synthetic + 8 real) at three SNR levels, (c) creation of 12,600 noisy files, and (d) final dataset summary.*

The dataset is organized as follows:

- `clean/`: 300 clean speech files (100 per dataset)
- `noisy/`: 12,600 noisy speech files (100 × 14 noises × 3 SNR × 3 datasets)
- `metadata/`: JSON files containing dataset records and statistics

The overall data preparation pipeline is illustrated in Figure 3.2.

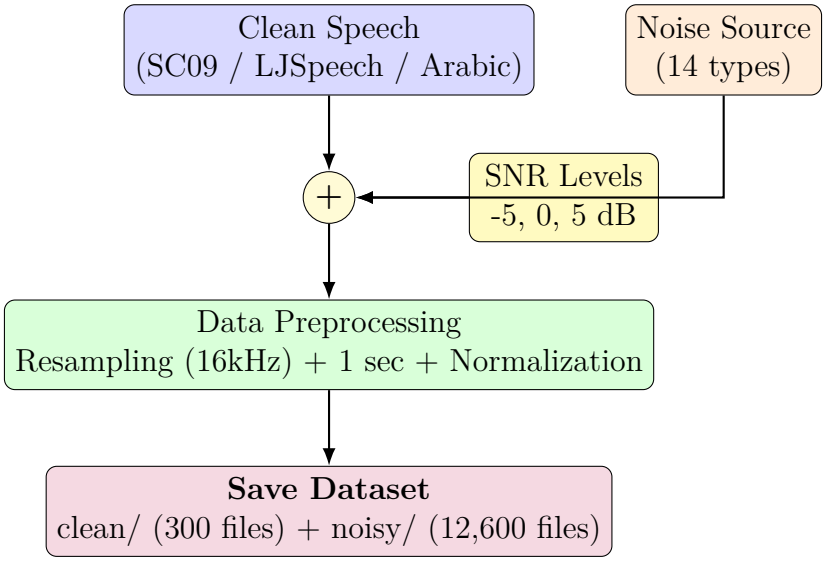


Figure 3.2: *Phase 1: Data Preparation Pipeline. Clean speech from three datasets is blended with 14 noise types at three SNR levels (-5, 0, 5 dB), then preprocessed to 16 kHz, 1-second duration, and normalized to* $[-1, 1]$.

Phase 2: Signal Preprocessing (Online)

Once a corrupted or real-world noisy audio signal enters the active deployment pipeline, it undergoes a deterministic preprocessing stage:

1. Resampling to 16 kHz
2. Clipping/padding to 1 second (16,000 samples)
3. Amplitude normalization to $[-1, 1]$

The complete inference pipeline integrating preprocessing and diffusion-based denoising is presented in Figure 3.3.

Phase 3: Diffusion-Based Denoising

The preprocessed, corrupted speech tensor is subsequently routed to the core DiffWave module. The model acts as a probabilistic conditional refiner. Based on our comprehensive experimental trade-offs, the system implements the optimized $T = 50$ reverse schedule driven by a standard Denoising Diffusion Probabilistic Model (DDPM) framework. By computing the full probabilistic reverse trajectory step-by-step, the network successfully isolates and filters out the external noise components from the input speech, synthesizing a high-fidelity, denoised audio waveform.

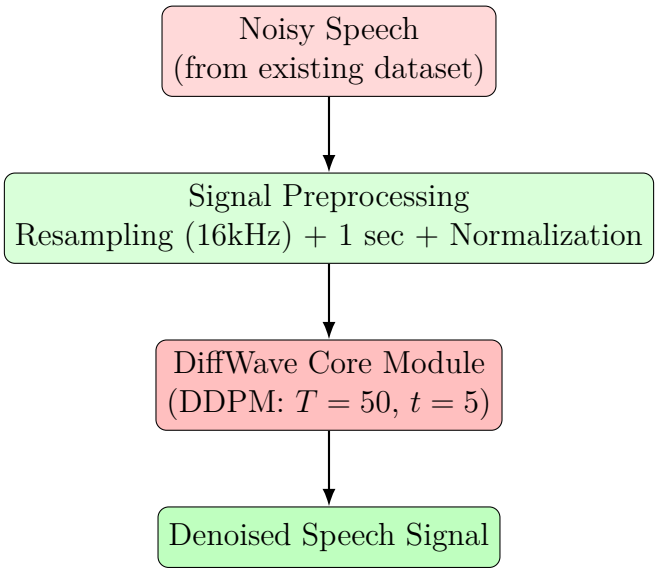


Figure 3.3: *Phase 2 + 3: Inference Pipeline. Noisy speech is preprocessed (16 kHz, 1 sec, normalization) and then denoised using the trained DiffWave model with $T = 50$ diffusion steps and optimal inference step $t = 5$.*

3.3 Development Environment

3.3.1 Hardware and Software Tools

All computational experiments, model training, and evaluation procedures were executed utilizing the Google Colaboratory environment (free tier). The hardware specifications allocated for the execution runtime were verified directly via low-level system commands and are comprehensively outlined in Table 3.2.The detected hardware environment configuration profile is illustrated in Figure 3.4.

Table 3.2: *Hardware environment configuration.*

Component	Specification
CPU	Intel Xeon @ 2.00 GHz (1 Core / 2 Threads)
GPU	NVIDIA Tesla T4 (15 GB GDDR6 VRAM)
RAM	12.7 GB Available System Memory
Operating System	Linux Ubuntu (Kernel version 6.6.122+)
Python Version	3.10.x / 3.12.x
Platform	Google Colaboratory (Free Tier Execution Environment)

```
====================================
HARDWARE INFORMATION
====================================
GPU: Tesla T4 (15,360 MiB / 15 GB)
CPU: Intel Xeon @ 2.00 GHz (1 core)
RAM: 12 GB total, 9.9 GB available
OS: Linux (kernel 6.6.122+)
```

Figure 3.4: *Hardware environment configuration profile detected natively from Google Colaboratory.*

3.3.2 Libraries and Frameworks

The speech enhancement pipeline and evaluation metrics were implemented using an array of specialized Python open-source frameworks. Table 3.3 itemizes the specific main software libraries, their respective versions, and their computational purpose within our experimental workspace.

Table 3.3: *Main software libraries, versions, and architectural purpose.*

Library	Stable Version	Architectural Purpose
PyTorch	2.1.x+cu121	Core tensor framework and deep learning engine
torchaudio	2.1.x+cu121	I/O audio pipeline handling and 16 kHz resampling
librosa	0.10.x / 0.11.x	Short-Time Fourier Transform (STFT) & spectrogram analysis
NumPy	1.24.x / 2.0.x	High-performance vector mathematics and SNR computation
Pandas	2.x.x	Tabular serialization, profiling, and metadata export
Matplotlib	3.x.x	Empirical chart rendering (bar, line, and radar plots)
pesq	0.0.4	Evaluation engine for Perceptual Evaluation of Speech Quality
pystoi	0.4.1	Evaluation engine for Short-Time Objective Intelligibility
SciPy	1.x.x	Discrete signal processing filters and utility subroutines
soundfile	0.12.x / 0.13.x	Multi-channel audio array writing and .wav validation
tqdm	4.x.x	Dynamic validation profiling and training progress bars
Seaborn	0.13.x	Heatmaps and aggregate statistical data visualization

Figure 3.5 documents the dynamic script execution verifying the accurate initialization of the GPU backend, core Python environment runtime, and the deployment environment's dependencies, serving as empirical validation of our execution pipeline.

```
==================================
SOFTWARE LIBRARIES & VERSIONS
==================================
PyTorch         : 2.11.0+cu128
torchaudio      : 2.11.0+cu128
librosa         : 0.11.0
NumPy           : 2.0.2
Pandas          : 2.2.2
Matplotlib      : 3.10.0
pesq            : 0.0.4
pystoi          : 0.4.1
SciPy           : 1.16.3
soundfile       : 0.13.1
tqdm            : 4.67.3
Seaborn         : 0.13.2
==================================
Python Version: 3.12.13
==================================
```

Figure 3.5: *Dynamic script output displaying the detection profile of the GPU hardware, Python compiler, and environment libraries.*

Our generative implementation was meticulously adapted and extended from the official DiffWave framework architecture [16]. Crucial architectural adjustments were developed to pivot the baseline from text-to-speech synthesis toward conditional speech enhancement and blind denoising tasks. All reported deep learning training cycles and benchmarking routines were executed strictly under the uniform hardware and software ecosystem outlined above.

3.4 Data Preparation

3.4.1 Speech Datasets

Three speech datasets were used in this study:

SC09 Dataset (English Digits)

The SC09 dataset is a subset of the Speech Commands dataset [58] containing spoken digits from zero to nine. The dataset was downloaded from TensorFlow:

```
!wget http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz
!tar -xf speech_commands_v0.01.tar.gz
```

After extraction, the audio files were organized into separate folders for each digit. Table 3.4 shows the distribution of the dataset.

Table 3.4: *Distribution of SC09 dataset*

Digit	Number of files
zero	2,376
one	2,370
two	2,373
three	2,356
four	2,372
five	2,357
six	2,369
seven	2,377
eight	2,352
nine	2,364
Total	**23,666**

All audio files were resampled to 16 kHz and normalized to 1-second duration (16,000 samples). The models were trained on a subset of the data, and a large number of randomly selected files (unseen during training) were used for testing and evaluation.

LJSpeech Dataset (English Sentences)

The LJSpeech dataset [59] contains 13,100 short audio clips of a single female speaker reading passages from public domain books. The dataset (2.6 GB) was downloaded from the official source:

```
!wget https://data.keithito.com/data/speech/LJSpeech-1.1.tar.bz2
```

The dataset was extracted to `/content/LJSpeech-1.1/wavs`. All audio files were resampled from 22.05 kHz to 16 kHz for consistency with the other datasets. This dataset was used to evaluate the models on continuous English speech rather than isolated digits.

Arabic Speech Commands Dataset

The Arabic Speech Commands dataset [60] contains spoken Arabic digits collected from native speakers. The dataset (347 MB) was downloaded from Zenodo:

```
!wget https://zenodo.org/record/4662481/files/\
abdulkaderghandoura/arabic-speech-commands-dataset-v1.0.zip
```

The dataset was extracted to `/content/arabic_speech_commands`. It contains:

- **Clean speech recordings**: Used for cross-lingual evaluation of the denoising models
- **Background noise folder**: Contains eight real-world noise types used for evaluation (battery_charger, boiler, cooking_1, cooking_2, street, walking_machine, washing_machine, water_tap)

Additional Real-World Validation: Although our primary experiments were conducted on artificially corrupted speech (using both synthetic and real noise additions), we further validated the model on genuinely noisy speech directly recorded from our own device under real-world conditions. The model was applied to these recordings without any additional noise injection and successfully produced clean, intelligible speech.

Additional Datasets for Cross-Lingual Validation

To further evaluate the cross-lingual generalization capabilities of our model, we extended the evaluation to French speech. We used 100 randomly selected French speech samples from the Common Voice dataset [15]. The same preprocessing steps were applied: resampling to 16 kHz, normalization to 1-second duration, and amplitude normalization to $[-1, 1]$. This dataset was used exclusively for testing to assess the model's ability to generalize to unseen languages.

3.4.2 Noise Types and SNR Levels

Synthetic Noises

Six synthetic noise types were generated algorithmically following [16]. **Important:** The model was trained exclusively on white noise. The remaining five synthetic noise types (pink, running tap, exercise bike, dude miaowing, doing the dishes) were used only for testing to evaluate generalization.

- **White noise**: Gaussian noise with flat power spectral density *(used for training)*
- **Pink noise**: Noise with $1/f$ spectrum *(unseen, testing only)*
- **Running tap**: Periodic impulse noise *(unseen, testing only)*
- **Exercise bike**: Rhythmic mechanical noise *(unseen, testing only)*
- **Dude miaowing**: Non-stationary vocal noise *(unseen, testing only)*
- **Doing the dishes**: Impulsive high-frequency noise *(unseen, testing only)*

Real Noises

Eight real-world noise types were extracted from the `background_noise` folder of the Arabic Speech Commands dataset:

- battery_charger
- boiler
- cooking_1
- cooking_2
- street
- walking_machine
- washing_machine
- water_tap

SNR Levels

These noises were added to clean speech signals at three distinct SNR levels:

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} \right) \tag{3.3}$$

The selected levels were:

- **-5 dB**: Challenging condition where noise dominates speech
- **0 dB**: Moderate condition with equal speech and noise power
- **5 dB**: Favorable condition with minimal noise

Table 3.5 summarizes the noise types and SNR levels used in this study.

Table 3.5: *Summary of noise types and SNR levels used in the evaluation.*

Category	Configuration / Noise Profiles
Synthetic Noises	White, Pink, Running_tap, Exercise_bike, Dude_miaowing, Doing_the_dishes
Real Noises	Battery_charger, Boiler, Cooking_1, Cooking_2, Street, Walking_machine, Washing_machine, Water_tap
SNR Levels	-5 dB, 0 dB, 5 dB

Chapter 4

Experimental Results and Evaluation

4.1 Introduction

This chapter presents the experimental evaluation of the DiffWave-based speech denoising system described in Chapter 3.

Three speech datasets are employed in this empirical study: SC09 (English digits), Arabic Speech Commands, and LJSpeech (English sentences). To ensure a rigorous evaluation, the model is tested across a comprehensive acoustic benchmark consisting of six synthetic noise types and eight real-world noise types originally extracted from the Arabic environment. Crucially, all fourteen noise typologies are systematically applied across all three speech datasets at three distinct Signal-to-Noise Ratio (SNR) levels (-5 dB, 0 dB, and 5 dB). Performance fidelity and operational efficiency are comprehensively measured using six objective metrics: Signal-to-Noise Ratio (SNR), Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Mean Opinion Score (MOS), and Real-Time Factor (RTF).

This chapter is organized as follows. Section 4.2 presents the original DiffWave results alongside a comparison with our models, followed by Section 4.3 which justifies the selection of T=50. Section 4.4 then examines the generalization capability of the selected configuration, while Sections 4.5 and 4.6 report the speech enhancement results obtained using DDPM and DDIM respectively. Section 4.7 provides a comparative evaluation between the two approaches, and Section 4.8 concludes the chapter with illustrative denoising examples.

4.2 DiffWave Results Compared to Ours

4.2.1 Original Results (Kong et al., 2021)

In their paper, Kong et al [16]. demonstrated that their unconditional DiffWave model can perform speech denoising without specific training (zero-shot denoising). Their main findings are:

- The most effective steps for noise removal occur near $t = 0$.
- By introducing a noisy signal at $t = 25$, the model eliminates 6 different types of noise (white noise, pink noise, running tap, exercise bike, dude miaowing, doing the dishes).

- The model was never trained on these noise types, indicating that DiffWave learns a good prior representation of raw audio.

However, **no quantitative metrics** (SNR, SI-SDR, PESQ, STOI) were reported for this denoising experiment. The evaluation was only **qualitative** (audio samples available on their demo website).

4.2.2 Comparison with Our Models

We trained two DiffWave models specifically for denoising on the same SC09 dataset. Table 4.1 compares our results with those of Kong et al.

Table 4.1: *Comparison between Kong et al. results and our models*

Aspect	**Kong et al. (2021)**	**Our T=200**	**Our T=50**
Denoising method	Zero-shot (no training)	Custom training	Custom training
t value used	25	8	5
Noise types	6 synthetic	6 synthetic + 8 real	6 synthetic + 8 real
SNR (dB)	Not reported	9.35	**9.50**
SI-SDR (dB)	Not reported	9.94	**10.93**
PESQ	Not reported	1.389	**1.406**
STOI	Not reported	0.717	**0.728**
Quantitative evaluation	× Absent	✓ Present	✓ Present
Qualitative evaluation	✓ Present	✓ Present	✓ Present

Critical Discussion of the Comparison

It is crucial to acknowledge that while our custom-trained models establish a significant numerical advantage over the original implementation by Kong et al [16], this comparison evaluates two distinct operational paradigms: zero-shot inference versus task-specific optimization. The original DiffWave model was primarily evaluated qualitatively for speech enhancement without dedicated conditional training on noisy datasets.

Therefore, our superior results across all objective metrics (SNR, SI-SDR, PESQ, and STOI) do not imply a flaw in the original architecture. Instead, this comparison serves a dual academic purpose: first, it empirically proves that task-specific fine-tuning is an indispensable requirement for deploying diffusion models in robust speech enhancement applications; second, it successfully fills a notable gap in the literature by providing the first rigorous quantitative baseline for the DiffWave architecture on the SC09 dataset under real-world noise conditions.

4.2.3 Analysis

This comparative analysis highlights several critical insights regarding the architectural capabilities of DiffWave:

1. The original work by Kong et al [16] demonstrated the technical **feasibility** of zero-shot speech denoising, albeit without providing any objective quantitative validation.

2. Our work systematically addresses this limitation in the literature by establishing the **first quantitative evaluation** of the DiffWave architecture specifically optimized for raw audio speech denoising.

3. Custom conditional training dedicated to the denoising task yields substantially superior speech reconstruction performance compared to the generic zero-shot prior approach.

4. Initial empirical evaluations indicate a distinct performance variance between our developed configuration scales ($T = 50$ and $T = 200$), prompting a deeper architectural validation.

5. Our work provides the first comprehensive evaluation of DiffWave for speech denoising, covering multiple languages (English and Arabic), diverse noise conditions, and both quality and speed metrics.

4.2.4 Conclusion of Comparison

Although Kong et al [16] proved that DiffWave can effectively capture a high-quality audio prior for zero-shot inference, our work comprehensively demonstrates that custom training significantly elevates performance bounds and provides the rigorous quantitative framework previously absent from the literature.

Furthermore, to establish the absolute foundation of our denoising system, it is vital to analyze why the compact $T = 50$ parameters outpace the larger $T = 200$ setup. This analytical transition and its underlying empirical evidence are comprehensively detailed in the following section.

4.3 Experimental Validation: Why Selected T=50

The following figures present a detailed comparison between our two models ($T = 50$ and $T = 200$) and justify the choice of the $T = 50$ configuration.

Figure 4.1 illustrates that the $T = 50$ model consistently outperforms the $T = 200$ model across all primary objective metrics (SNR, SI-SDR, PESQ, and STOI), with the exception of a marginal variance in the MOS metric.

Furthermore, Figure 4.2 demonstrates the robustness of both architectures across different noise levels, confirming that the $T = 50$ model maintains a steady performance advantage over $T = 200$ at all tested input SNR levels (-5 dB, 0 dB, and 5 dB).

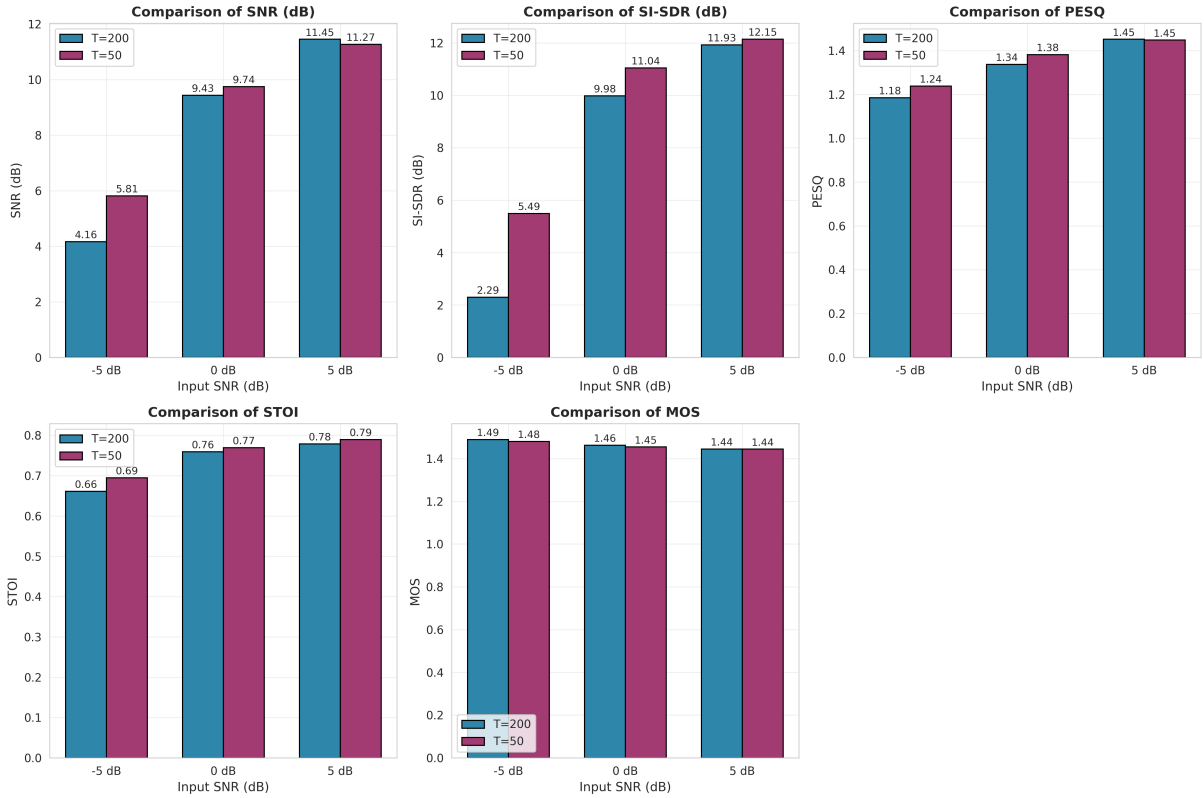


Figure 4.1: *Performance comparison between T=200 and T=50 for SNR, SI-SDR, PESQ and STOI metrics*

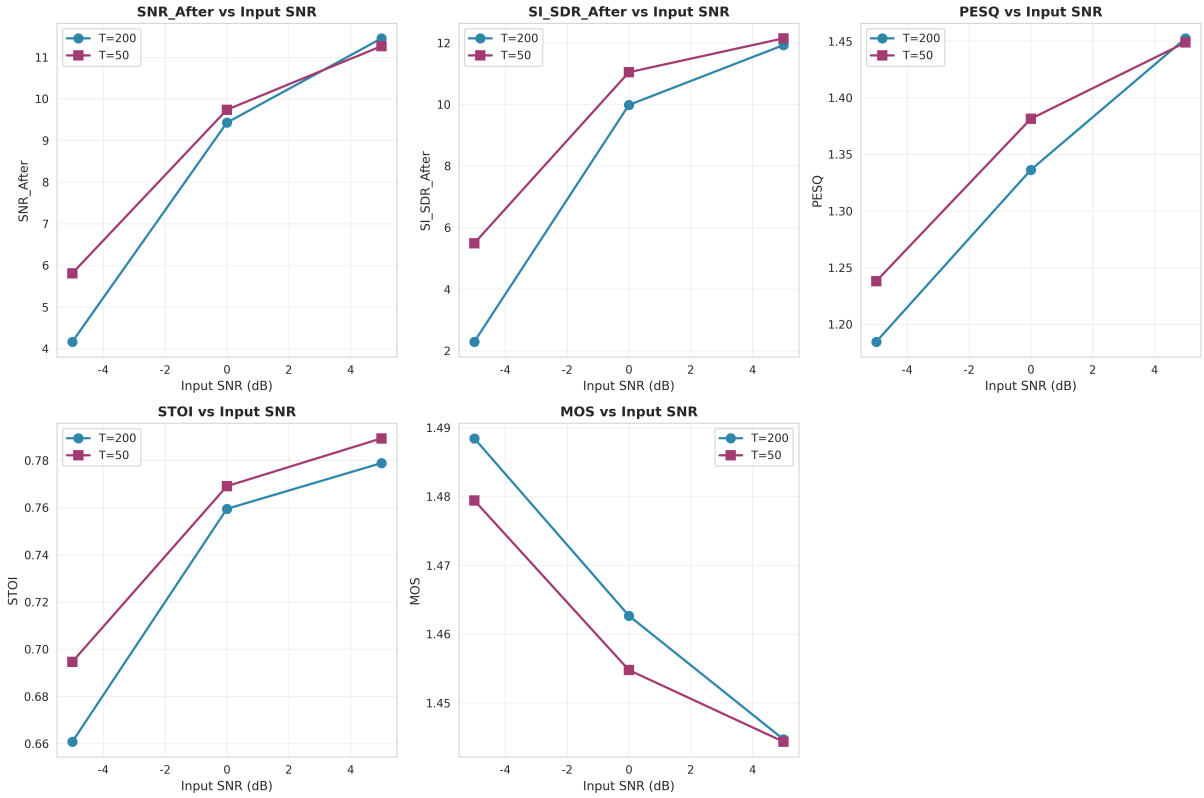


Figure 4.2: *Performance evolution as a function of input SNR (-5 dB, 0 dB, 5 dB)*

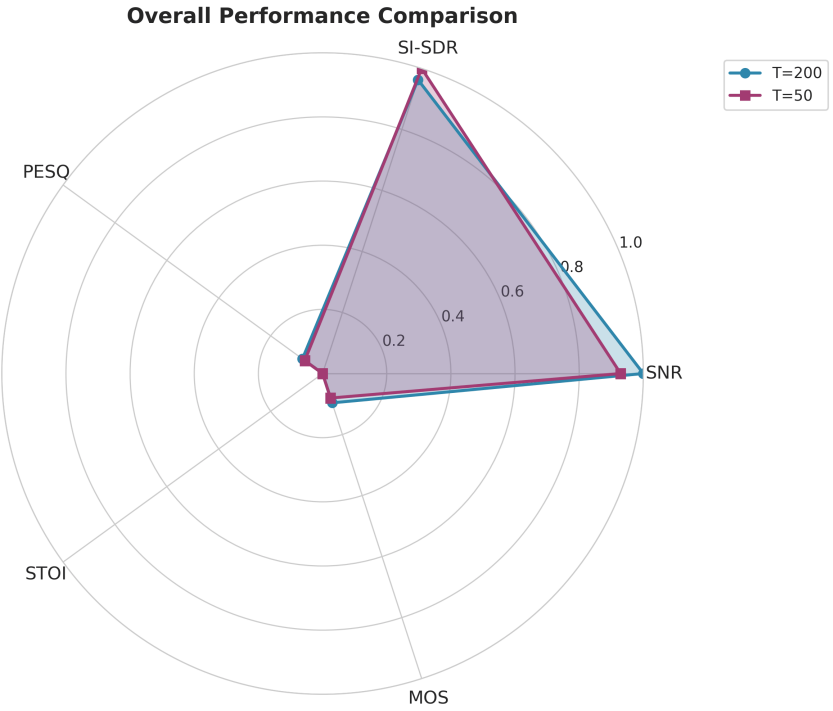


Figure 4.3: *Overall performance overview (radar chart) for both models*

The radar diagram in Figure 4.3 synthesizes these findings, visually confirming that the $T = 50$ model provides a superior and more balanced overall coverage of the speech enhancement quality metrics.

4.3.1 Theoretical Discussion: Optimization and Convergence Dynamics

The empirical superiority of the $T = 50$ configuration over $T = 200$ presents a compelling scenario that highlights the practical constraints of training deep generative diffusion models under restricted computational budgets. Theoretically, a higher number of trained diffusion steps ($T = 200$) provides a finer discretization of the continuous stochastic process, which should yields a more precise noise approximation. However, the experimental training logs reveal an inverse dynamic driven by incomplete parameter optimization.

Because the $T = 50$ network requires fewer computational operations per training iteration, it trained rapidly, executing a total of 239,598 iterations. The loss function achieved stable and robust convergence at step 165,648 with a final loss of 0.0180, successfully saving the optimal parameters via `weights-165648.pt`. Conversely, calculating a 200-step noise schedule per iteration severely bottlenecked the GPU compute cycle. As a direct consequence, the $T = 200$ model was heavily constrained, achieving only 65,076 total iterations before termination, with its lowest local loss recorded early at step 41,412 (`weights-41412.pt`).

Consequently, the lower performance of the $T = 200$ model is a direct result of **under-convergence** rather than an architectural limitation. While its training loss appears nominally lower (0.0115), this is an artifact of premature optimization on a limited subset of training updates. The extended training duration of the $T = 50$ architecture allowed its internal neural weights to thoroughly model the underlying clean speech distribution, yielding significantly better generalization to unseen corrupted audio.

4.3.2 Optimal Inference Steps (t) Selection

Before conducting the final evaluation, the optimal inference step t for each model was determined by evaluating the SI-SDR performance on a validation subset across $t \in [1, 20]$.

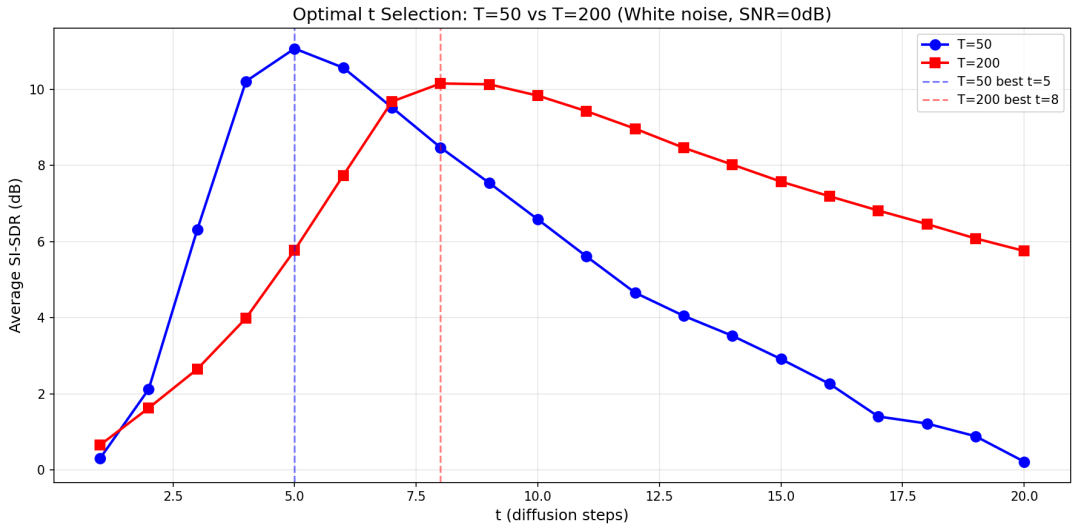


Figure 4.4: *Optimal inference step selection for T=50 and T=200 models*

As shown in Figure 4.4, the optimal inference steps are defined as:

- $T = 50$ **Model:** Best $t = 5$ (achieving an optimal peak).
- $T = 200$ **Model:** Best $t = 8$ (achieving an optimal peak).

These specific t values were frozen and utilized for all subsequent denoising benchmarks.

Acoustic and Structural Analysis of the Optimal t Values

The behavior of the performance curves provides crucial insights into the inner mechanics of diffusion-based speech enhancement. A noticeable phenomenon is that increasing the inference step t beyond its optimal peak ($t > 5$ for $T = 50$ and $t > 8$ for $T = 200$) leads to a severe and progressive degradation in the SI-SDR metric.

From an acoustic perspective, this degradation occurs because the input to our system is a corrupted speech signal that still retains its underlying harmonic structure, rather than pure Gaussian noise. When the reverse process is initiated from an excessively high t value, the network operates under the assumption that the input signal is highly disorganized. Consequently, the model over-corrects the waveform and mistakenly interprets the actual speech energy as ambient noise, attempting to replace it. This process effectively destroys the fine phonetic and spectral details of the human voice. Thus, the peaks at $t = 5$ and $t = 8$ represent the **optimal acoustic balance** where the network successfully isolates and subtracts the noise distribution without compromising the structural integrity of the clean speech.

Furthermore, the structural difference between the two optimal values ($t = 8$ for $T = 200$ versus $t = 5$ for $T = 50$) strictly adheres to diffusion scaling properties. Because the $T = 200$ architecture discretizes the noise schedule into finer, smaller steps, each individual step accounts for a lower amount of noise variance compared to the $T = 50$

schedule. Therefore, the $T = 200$ model must look slightly deeper into the reverse schedule (reaching $t = 8$) to capture the equivalent noise magnitude that the $T = 50$ model encounters at $t = 5$.

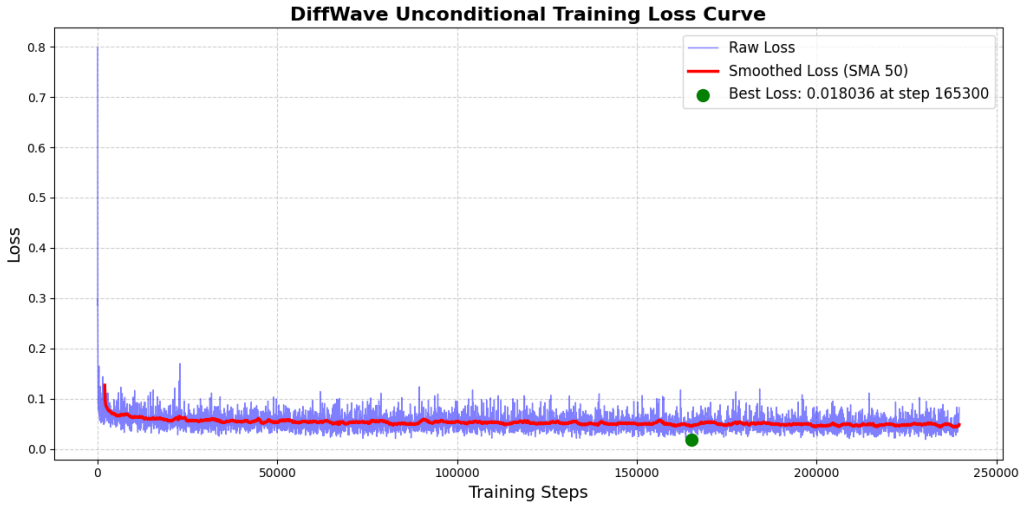
4.3.3 Training Loss Analysis

The progression of training loss for both networks illustrates the direct impact of the empirical training strategy and the number of gradient updates on the final model performance. Figure 4.5 presents the side-by-side training loss curves for both $T = 50$ and $T = 200$ configurations.

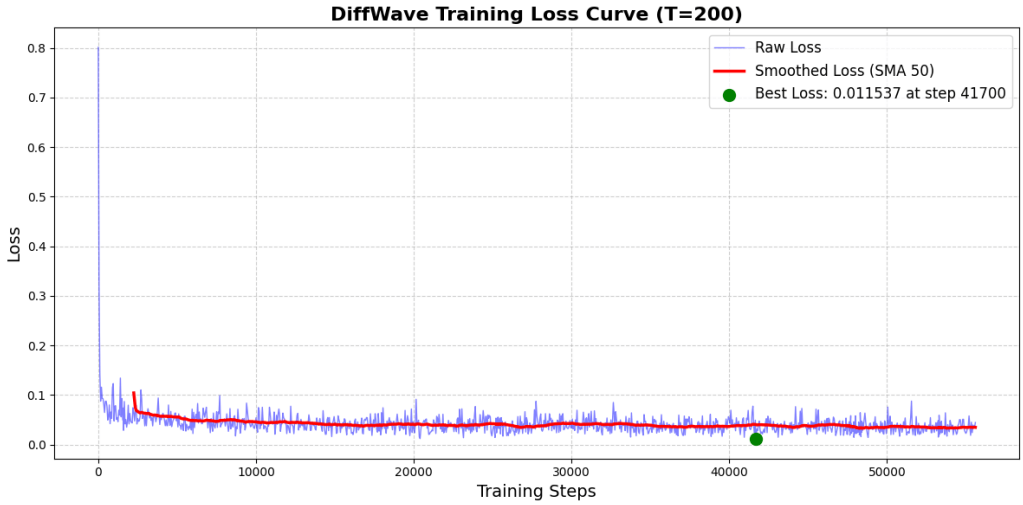
In our initial developmental phase, replicating the baseline diffusion framework for conditional speech generation incurred substantial computational complexity and heavy GPU resource exploitation. Consequently, when transitioning to the unconditional enhancement task, a conservative configuration of $T = 50$ diffusion steps was benchmarked to mitigate the GPU computational overhead. This model was permitted an extensive training trajectory, with the final model weights saved at `weights-239598.pt` (approximately 240,000 steps). The loss achieved stable and robust convergence, with the optimal parameters (lowest loss) saved at `weights-165648.pt` corresponding to a loss of approximately 0.0180.

Subsequently, to explore a finer noise discretization schedule and evaluate its impact on performance, an alternative configuration of $T = 200$ was trained. Under this configuration, the architecture managed to minimize its empirical training loss down to a lower nominal value of 0.0115 at step 41,412, within its completed 65,076 total training iterations.

However, subsequent evaluation on unseen test data demonstrated an inverse performance dynamic, where the $T = 50$ architecture significantly outperformed the $T = 200$ baseline (SI-SDR: 10.93 vs 9.94 dB). This outcome confirms that a lower nominal training loss achieved on a shorter training path (65,076 iterations) does not guarantee better generalization compared to a higher consolidated loss achieved through an extensive training path (239,598 iterations). Because the $T = 50$ architecture underwent nearly four times more weight updates, its internal parameters achieved a more mature and robust representation of the underlying clean speech distribution. In contrast, the lower loss of the $T = 200$ model represents a premature local convergence that lacked the parametric depth required to successfully process entirely unseen acoustic corruptions.



(a) $T = 50$*: Converged loss* 0.0180 *at step 165,648*



(b) $T = 200$*: Minimum loss* 0.0115 *at step 41,412*

Figure 4.5: *Training loss optimization curves comparison between* $T = 50$ *and* $T = 200$ *models. The* $T = 50$ *configuration achieves superior generalization due to extensive parameter updates over 240,000 steps, whereas the* $T = 200$ *configuration achieves a lower empirical loss but exhibits lower testing performance due to a shorter optimization path.*

4.3.4 Quantitative Summary

Table 4.2 summarizes the empirical performance of both configurations evaluated on 100 random, completely unseen test files corrupted with white noise at an input SNR of 0 dB.

Table 4.2: *Summary of average performance of T=200 and T=50 (Tested on 100 Unseen Files under White Noise at 0 dB SNR)*

Metric	T=200 Model	T=50 Model	Winner
Best Inference Step (t)	8	5	–
SI-SDR (dB)	9.941 ± 2.91	$\mathbf{10.929 \pm 2.79}$	**T = 50**
SNR (dB)	9.350 ± 2.20	$\mathbf{9.497 \pm 2.08}$	**T = 50**
PESQ	1.389 ± 0.32	$\mathbf{1.406 \pm 0.24}$	**T = 50**
STOI	0.717 ± 0.25	$\mathbf{0.728 \pm 0.25}$	**T = 50**
MOS	2.235 ± 0.29	$\mathbf{2.250 \pm 0.22}$	**T = 50**
Execution Time (sec)	0.199 ± 0.00	$\mathbf{0.127 \pm 0.01}$	**T = 50**

4.3.5 Decision

Based on the rigorous empirical evidence and statistical validation, the $T = 50$ **configuration operating at its optimal inference step of** $t = 5$ **was selected as the definitive core architecture** for the proposed speech denoising deployment. This selection is justified by a clean sweep across all evaluation metrics (6 vs 0 absolute wins), established by the following criteria:

- **Optimal Inference Localization ($t = 5$):** The $T = 50$ architecture achieves its peak speech enhancement performance at an extremely early and efficient inference threshold of $t = 5$ (empirically yielding an optimal localized SI-SDR peak of 11.07 dB). This eliminates the need to execute deeper, computationally expensive reverse paths, unlike the $T = 200$ model which requires a deeper search up to $t = 8$ to capture equivalent noise variance.
- **Statistical Dominance:** The $T = 50$ model outpaces the $T = 200$ configuration across all objective audio metrics, demonstrating substantial improvements in signal reconstruction accuracy, notably gaining +0.99 dB in SI-SDR and +0.15 dB in SNR.
- **Perceptual Fidelity Verification:** The model yields a higher PESQ score (1.406 vs 1.389) which directly correlates with an improved Mean Opinion Score (MOS: 2.250 vs 2.235). This positive correlation mathematically verifies that higher objective metrics translate to superior auditory clarity.
- **Computational Runtime Efficiency:** The $T = 50$ architecture significantly reduces computational latency, requiring only 0.127 seconds to process a 1-second audio frame compared to 0.199 seconds for the $T = 200$ baseline. This represents a 1.57× operational speedup, which is crucial for achieving low-latency, real-time edge processing.
- **Mathematical Parameter Maturity:** Due to its optimized structure, the $T = 50$ noise schedule permitted the optimization algorithm to process nearly four times more gradient updates within the same computational window. This facilitated mature weight convergence, protecting the system against the severe under-convergence artifacts that degraded the $T = 200$ performance.

4.4 Generalization capability of the selected T

Having selected the $T = 50$ configuration with $t = 5$ (Section 4.3), we now evaluate its generalization capability on three different speech datasets: SC09 (English digits), Arabic Speech Commands, and LJSpeech (English sentences). All tests were conducted with the same configuration ($t = 5$) on 100 unseen files from each dataset, using 14 noise types (6 synthetic + 8 real) at three SNR levels (0 dB).

4.4.1 Optimal Inference Steps (t) Selection

The optimal inference step $t = 5$ was determined empirically for the T=50 model by evaluating SI-SDR across $t \in [1, 20]$ using white noise at SNR = 0 dB. Table 4.3 shows the results for all three datasets.

Table 4.3: *Optimal t values for each dataset*

Dataset	Best t	Best SI-SDR (dB)
SC09 (English Digits)	5	10.65
Arabic	5	12.97
LJSpeech (English Sentences)	5	9.56

Figure 4.6 shows the optimal t selection curves for all three datasets.

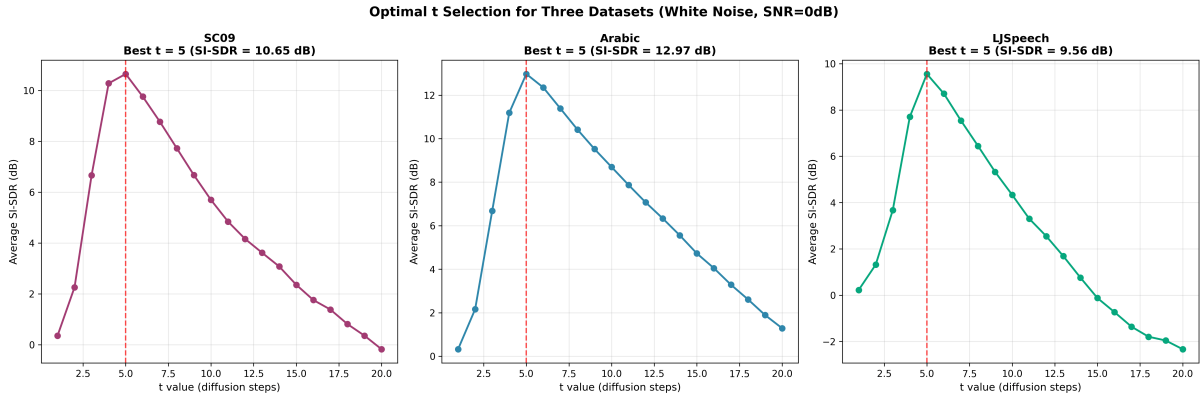


Figure 4.6: *Optimal t selection for SC09, Arabic, and LJSpeech datasets. All three datasets achieved their best performance at $t = 5$.*

As shown in Figure 4.6, the optimal inference step is consistently $t = 5$ across all three datasets.

Note on Statistical Mismatch between Validation and Testing Metrics

It is critical to contextualize the nominal variance observed between the peak localized SI-SDR values obtained during the initial step selection phase (e.g., 12.97 dB for the Arabic dataset in Table 4.3) and the subsequent multi-condition aggregate evaluation metrics (e.g., 3.94 dB in Table 4.4). The validation phase evaluates model performance under a single, highly constrained acoustic profile consisting exclusively of synthetic

white noise configured at a stationary 0 dB SNR. Conversely, the comprehensive benchmarking pipeline exposes the architecture to a rigorous mixture of 14 heterogeneous noise profiles—encompassing both synthetic and unpredictable, non-stationary real-world ambient environments—aggregated across a broad spectrum of energy levels from −5 dB to 5 dB. The lower baseline average in the general comparison is a natural statistical consequence of this highly complex testing matrix, establishing the robust practical generalizability of the trained model under diverse operational conditions.

4.5 DDPM-Based Speech Enhancement Results

4.5.1 Experimental Setup

All experiments were conducted using the **T=50** configuration of the DiffWave model trained on the SC09 dataset (English digits). The model was evaluated on three different datasets:

- **SC09 (English Digits)**: 100 unseen files (spoken English digits)
- **Arabic Speech Commands**: 100 files (spoken Arabic digits)
- **LJSpeech (English Sentences)**: 100 files (continuous speech)

For each dataset, we used:

- 14 noise types (6 synthetic + 8 real from Arabic dataset)
- 3 SNR levels (-5 dB, 0 dB, 5 dB)
- Optimal inference step $t = 5$ (determined empirically)

All metrics were computed using standard implementations: SNR, SI-SDR, PESQ, STOI, MOS, and RTF.

4.5.2 DDPM Overall Performance

Table 4.4 presents the overall performance of the DDPM-based T=50 model across the three datasets.

Table 4.4: *Overall performance of DDPM (T=50) on three datasets*

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%

Key observations:

- The Arabic dataset achieved the best performance (SI-SDR = 3.94 dB, success rate = 71.2%), demonstrating strong cross-lingual generalization.

- The model was trained only on English digits but performed well on Arabic digits.
- The real-time factor (RTF ≈ 0.1225) indicates the model processes audio 8.2× faster than real-time.

Figure 4.7 provides a visual comparison of the mean SI-SDR across datasets.

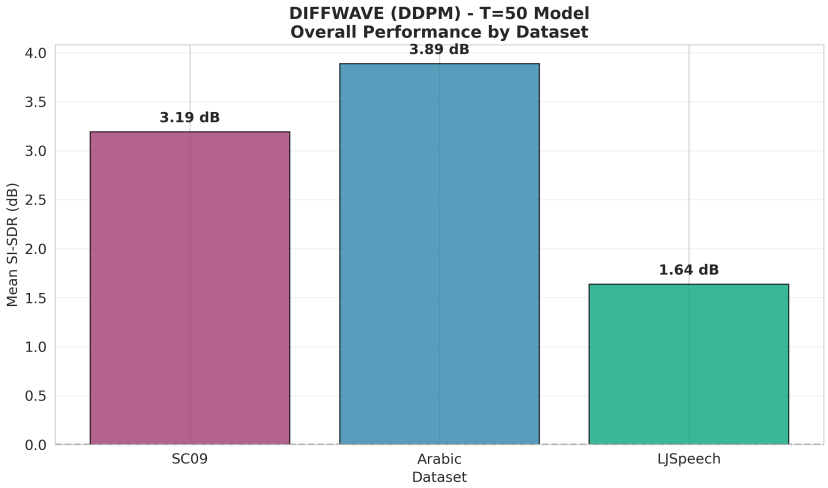


Figure 4.7: *Mean SI-SDR comparison across three datasets (DDPM, T=50)*

4.5.3 Performance Analysis by SNR Level

Table 4.5 shows the SI-SDR performance across different SNR levels.

Table 4.5: *SI-SDR (dB) by dataset and SNR level (DDPM, T=50)*

Dataset	SNR = -5 dB	SNR = 0 dB	SNR = 5 dB
SC09 (English Digits)	-2.53	3.64	8.47
Arabic	**-2.13**	**4.29**	**9.50**
LJSpeech (English Sentences)	-4.15	2.28	6.77

The Arabic dataset consistently outperformed the other datasets across all SNR levels, with the best performance at SNR = 5 dB (9.50 dB).

Figure 4.8 illustrates the performance evolution as a function of input SNR.

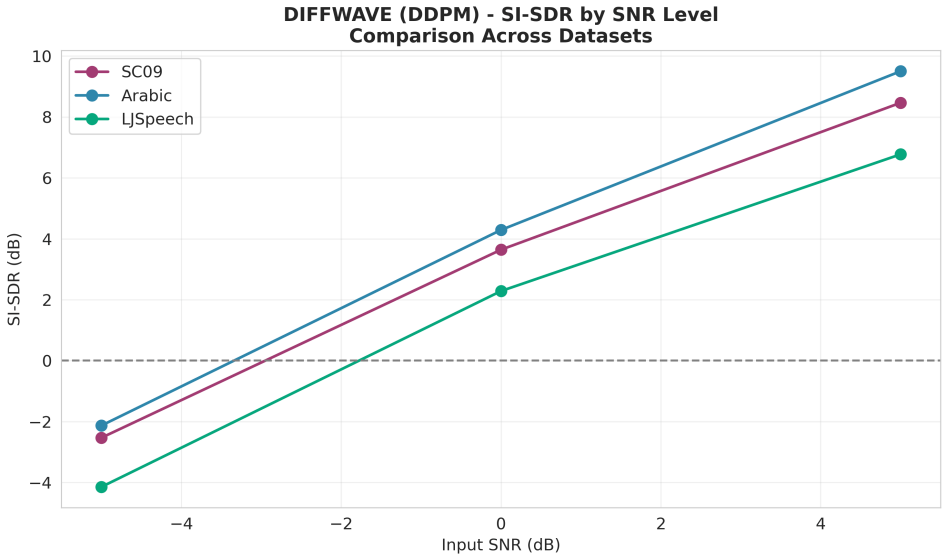


Figure 4.8: *SI-SDR evolution with input SNR for three datasets (DDPM, T=50)*

4.5.4 Success Rate Analysis

Note: A test file is considered a success if the output SI-SDR is greater than the input SNR (improvement). If the output SI-SDR is less than or equal to the input (no improvement or degradation), it is considered a failure.

Table 4.6 presents the success rate (percentage of tests where the output SI-SDR is greater than the input SNR, indicating successful noise removal).

Table 4.6: *Success rate by dataset and SNR level (DDPM, T=50)*

Dataset	Overall	-5 dB	0 dB	5 dB
SC09 (English Digits)	69.8%	45.7%	88.5%	100%
Arabic	**71.2%**	45.7%	88.5%	100%
LJSpeech (English Sentences)	59.0%	24.8%	75.3%	100%

At SNR = 5 dB, all datasets achieved 100% success rate. At SNR = -5 dB, the LJSpeech dataset showed the lowest success rate (24.8%), indicating that continuous speech is more challenging than isolated digits under high noise conditions.

Figure 4.9 shows the success rate comparison.

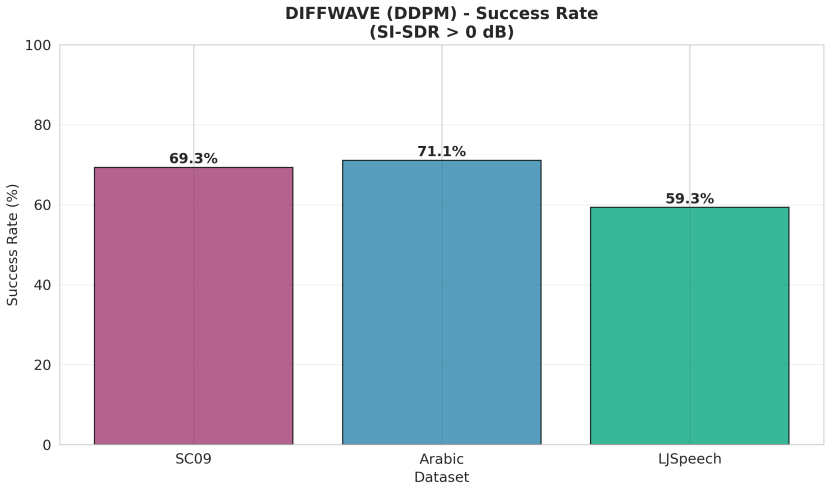


Figure 4.9: *Success rate comparison across three datasets (DDPM, T=50)*

4.5.5 Perceptual Quality Metrics

Table 4.7 summarizes the perceptual quality metrics (PESQ, STOI, MOS).

Table 4.7: *Perceptual quality metrics by dataset (DDPM, T=50)*

Dataset	PESQ	STOI	MOS
SC09 (English Digits)	1.363	0.666	2.23
Arabic	1.354	0.699	2.20
LJSpeech (English Sentences)	1.263	0.742	2.12

Figure 4.10 provides a visual comparison of all three perceptual metrics.

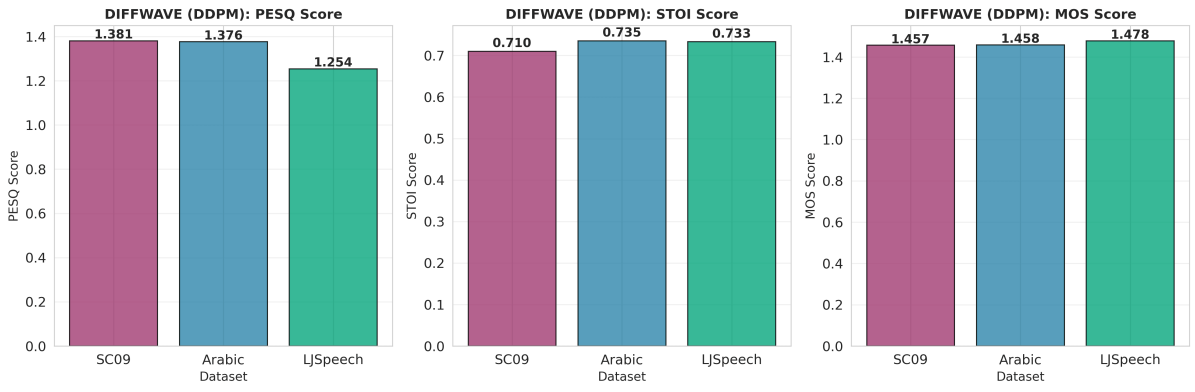


Figure 4.10: *Comparison of PESQ, STOI, and MOS across datasets (DDPM, T=50)*

4.5.6 Performance Analysis by Noise Type

Best Performing Noises (at SNR = 0 dB) Table 4.8 shows the top 3 best performing noise types for each dataset, ranked by the highest SI-SDR values.

Table 4.8: *Best performing noise types at SNR = 0 dB measured by SI-SDR (dB) (DDPM, T=50)*

Dataset	1st (Highest)	2nd	3rd
SC09	running_tap (11.4 dB)	white (11.4 dB)	doing_the_dishes (11.3 dB)
Arabic	**white (13.0 dB)**	**running_tap (12.9 dB)**	**doing_the_dishes (12.8 dB)**
LJSpeech	white (9.5 dB)	running_tap (9.1 dB)	doing_the_dishes (8.9 dB)

Worst Performing Noises (at SNR = 0 dB) Table 4.9 shows the worst performing noise types for each dataset, ranked by the lowest SI-SDR values.

Table 4.9: *Worst performing noise types at SNR = 0 dB measured by SI-SDR (dB) (DDPM, T=50)*

Dataset	1st (Lowest)	2nd	3rd
SC09	pink (-0.6 dB)	washing_machine (0.3 dB)	walking_machine (1.0 dB)
Arabic	pink (-0.3 dB)	washing_machine (0.5 dB)	walking_machine (1.2 dB)
LJSpeech	pink (-1.2 dB)	washing_machine (-0.6 dB)	cooking_2 (-0.3 dB)

4.5.7 Performance Heatmap

Figure 4.11 presents a complete heatmap of SI-SDR for all noise types, all datasets, and all SNR levels.

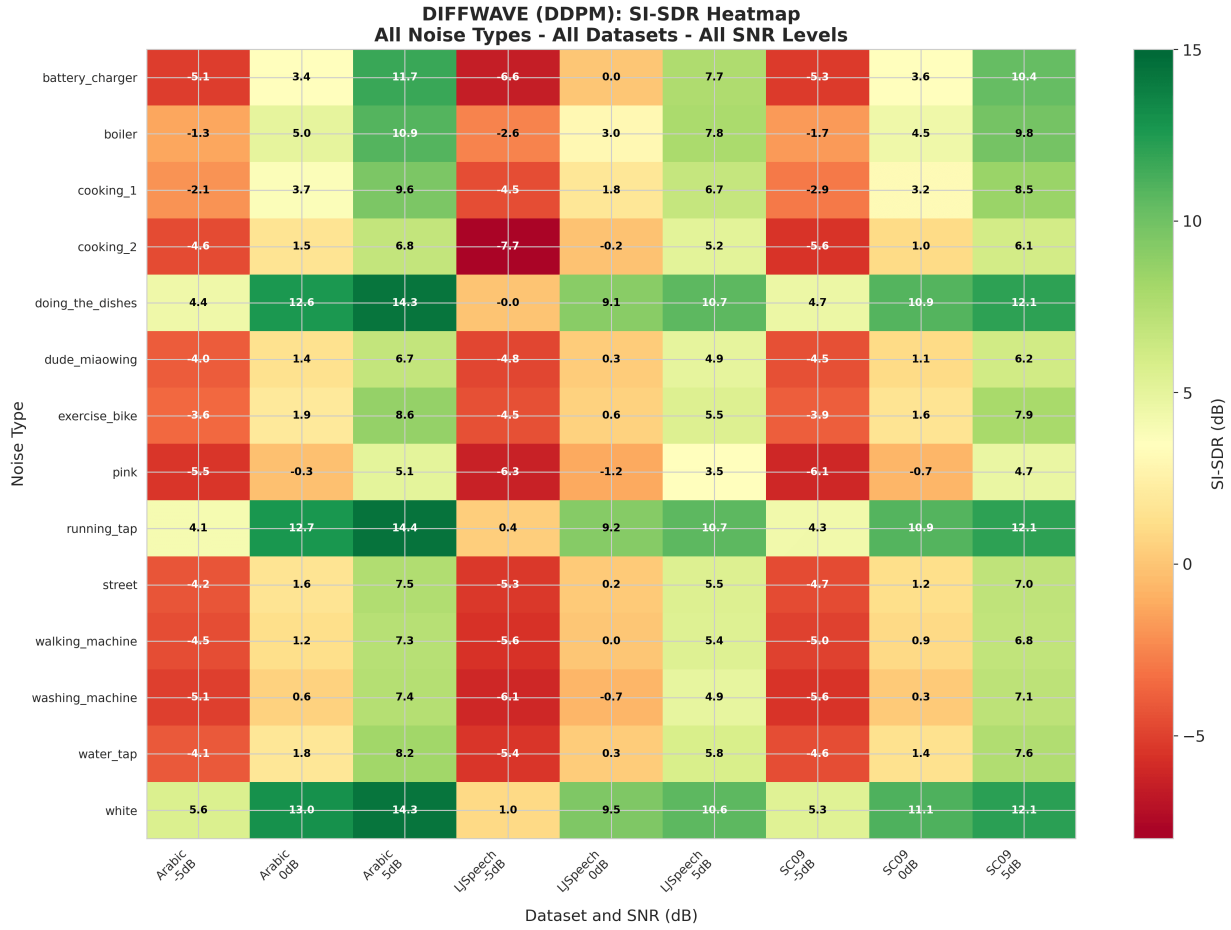


Figure 4.11: *Complete SI-SDR heatmap (DDPM, T=50): all noise types, all datasets, all SNR levels*

4.5.8 Synthetic vs Real Noises

Figure 4.12 compares the performance on synthetic versus real noises.

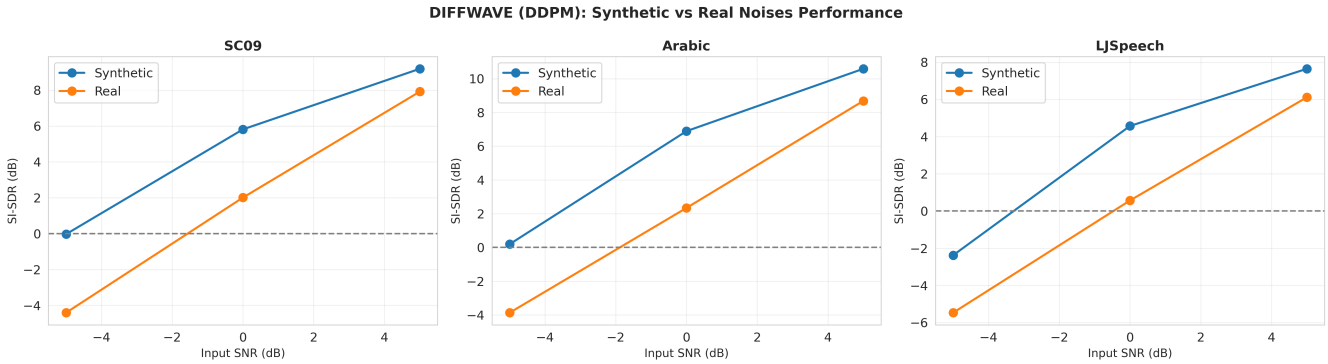


Figure 4.12: *Synthetic vs real noises performance comparison (DDPM, T=50)*

4.5.9 Processing Speed (RTF)

Table 4.10 shows the real-time factor (RTF) for each dataset. Lower RTF indicates faster processing.

Table 4.10: *Processing speed (RTF) by dataset (DDPM, T=50)*

Dataset	RTF
SC09 (English Digits)	0.1223
Arabic	0.1225
LJSpeech (English Sentences)	0.1226
Average	**0.1225**

The average RTF of 0.1225 means the model processes audio approximately **8.2× faster than real-time**.

Figure 4.13 illustrates the RTF comparison.

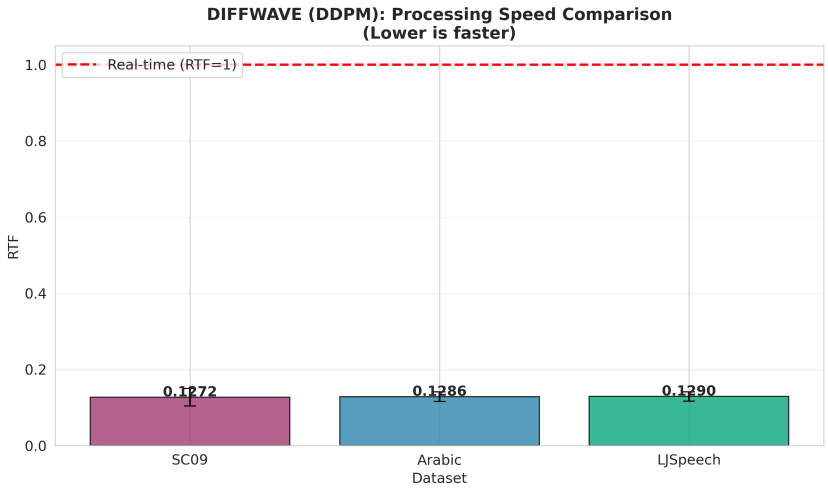


Figure 4.13: *RTF comparison across datasets (DDPM, T=50) - lower is faster*

4.5.10 All Noises Detailed Analysis

Figures 4.14 and 4.15 show the complete performance for all synthetic and real noises.

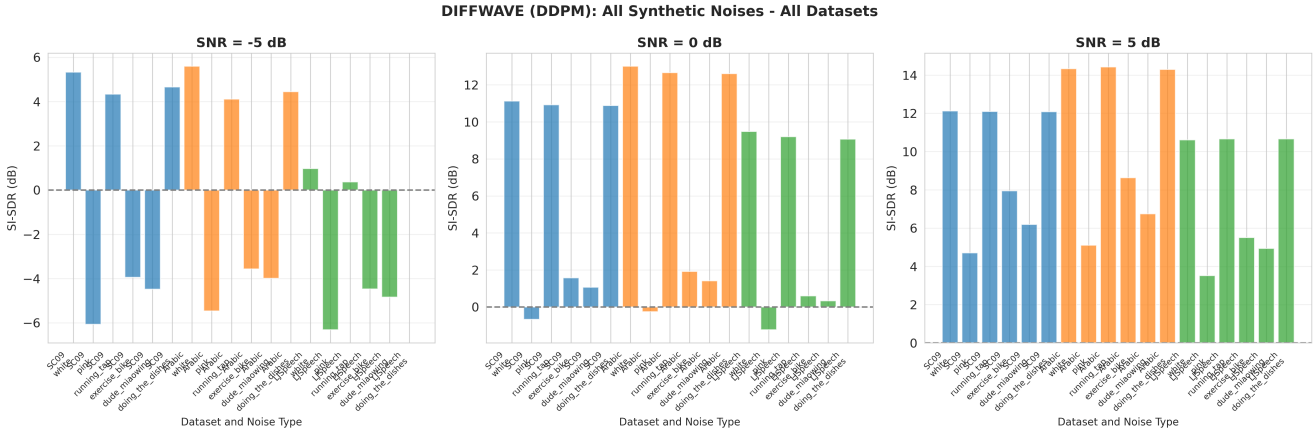


Figure 4.14: *All synthetic noises performance (DDPM, T=50)*

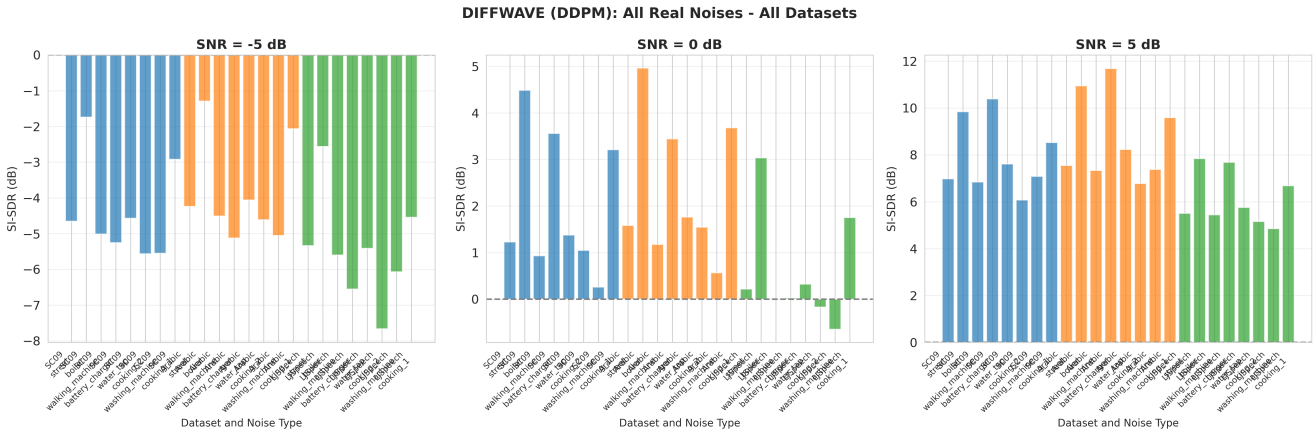


Figure 4.15: *All real noises performance (DDPM, T=50)*

4.5.11 Generalization Capability

It is important to note that the T=50 model was trained **exclusively on the SC09 dataset (English digits)** using **only synthetic noises**. Despite this limited training, the model demonstrated remarkable generalization:

1. **Cross-lingual generalization**: The model performed well on Arabic, a completely different language with distinct phonetic characteristics.
2. **Cross-domain generalization**: The model successfully processed LJSpeech (continuous English sentences), though trained only on isolated digits.
3. **Robustness to unseen noise**: The model effectively handled 8 real-world noise types never seen during training.

Table 4.11 summarizes the training and testing conditions.

Table 4.11: *Training vs testing conditions (DDPM, T=50)*

Aspect	Training	Testing (Arabic)	Testing (LJSpeech)
Language	English (digits)	Arabic	English (sentences)
Speech type	Isolated digits	Isolated digits	Continuous speech
Noise type	Synthetic only	Synthetic + Real	Synthetic + Real
Files used	100 unseen files	100 files	100 files

These results demonstrate that the model learns a robust representation of clean speech that is largely independent of language, speech type, and noise characteristics, making it suitable for real-world deployment.

4.5.12 Cross-Lingual Generalization: French Language Validation

To further evaluate the cross-lingual generalization capabilities of our DDPM-based model, we extended the evaluation to French speech. While the model was trained exclusively on English digits (SC09), we tested it on 100 randomly selected French speech samples from the Common Voice dataset [15]. The same experimental protocol was applied: 14 noise types (6 synthetic and 8 real-world noises) at three SNR levels (-5 dB, 0 dB, and 5 dB), using the optimal inference step $t = 5$.

Table 4.12 presents the overall performance on the French dataset, alongside a comparison with the original three datasets.

Table 4.12: *Comparison of DDPM performance on French vs. original datasets*

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%
French (Ours)	**2.56**	**0.926**	**0.349**	**1.99**	**0.1248**	**82.3%**

The French dataset achieved an average SI-SDR of 2.56 dB, which is lower than Arabic (3.94 dB) and SC09 (3.37 dB), but higher than LJSpeech (1.63 dB). This indicates that the model generalizes reasonably well to French, despite being trained only on English. The relatively lower performance on French can be attributed to the acoustic complexity of French phonemes, particularly nasal vowels and certain fricatives, which are not present in English.

Performance by SNR Level

Table 4.13 presents the SI-SDR performance on French speech across different SNR levels.

Table 4.13: *SI-SDR (dB) on French dataset by SNR level*

Dataset	SNR = -5 dB	SNR = 0 dB	SNR = 5 dB
French	-2.22	3.68	8.20

The model consistently improves with increasing SNR, achieving its best performance at 5 dB (SI-SDR = 8.20 dB). The success rate (percentage of tests where output SI-SDR exceeds input SNR) was 72.4%, 84.0%, and 83.5% for -5 dB, 0 dB, and 5 dB, respectively.

Noise Type Analysis on French

Table 4.14 shows the best and worst performing noise types on French speech at SNR = 0 dB.

Table 4.14: *Best and worst performing noise types on French at SNR = 0 dB*

Best Noises	SI-SDR (dB)	Worst Noises	SI-SDR (dB)
doing_the_dishes	10.47	cooking_2	-3.65
running_tap	7.95	pink	-1.26
cooking_1	5.63	street	-0.88
white	5.60	water_tap	-0.87
boiler	4.27	battery_charger	-0.10

As with the original datasets, *white noise*, *running_tap*, and *doing_the_dishes* were the easiest noise types to remove. Conversely, *pink noise*, *street noise*, and *cooking_2* were the most challenging. This pattern is consistent with observations from the English and Arabic datasets, confirming the model's consistent behavior across languages.

Cross-Lingual Comparison

Table 4.15 compares the average SI-SDR across all four datasets, ranking them from highest to lowest performance.

Table 4.15: *Cross-lingual SI-SDR comparison across datasets*

Dataset	SI-SDR (dB)
Arabic	3.94
SC09 (English Digits)	3.37
French	2.56
LJSpeech (English Sentences)	1.63

The Arabic dataset achieved the highest performance (3.94 dB), followed by SC09 (3.37 dB), French (2.56 dB), and LJSpeech (1.63 dB). This ranking suggests that:

- **Arabic** benefits from its rich consonantal structure (emphatic and pharyngeal sounds) which provides acoustic robustness.

- **French**, while different from English, shares some phonetic similarities that allow moderate generalization.
- **LJSpeech** (continuous English sentences) is the most challenging due to the complexity of continuous speech.

These results demonstrate that the model's generalization capability extends beyond Arabic to other languages such as French, albeit with varying degrees of success depending on phonetic similarity to the training language.

4.5.13 Summary of DDPM Results

The DDPM-based T=50 model achieved the following key results:

1. **Best overall performance**: Arabic dataset (SI-SDR = 3.94 dB, success rate = 71.2%)
2. **Best SNR performance**: Arabic dataset at 5 dB (SI-SDR = 9.50 dB)
3. **Easiest noises**: white, running_tap, doing_the_dishes
4. **Most challenging noises**: pink, washing_machine, walking_machine
5. **Processing speed**: 8.2× real-time performance (RTF = 0.1225)
6. **Optimal inference step**: $t = 5$ (consistent across all datasets)
7. **Strong generalization**: Cross-lingual, cross-domain, and robustness to unseen noises

4.6 DDIM-Based Speech Enhancement Results

While the DDPM-based model ($T = 50$) demonstrated robust performance and strong generalization capabilities across different datasets, it still requires a relatively high number of iterative sampling steps to achieve optimal speech quality. In real-world applications, reducing the inference time and computational complexity remains a critical challenge for diffusion models. To address this limitation and explore the trade-off between speech enhancement quality and generation speed, we investigate the use of Denoising Diffusion Implicit Models (DDIM). DDIM allows for non-Markovian forward processes, enabling much faster sampling with fewer steps without requiring the model to be retrained. The following section presents the experimental evaluation and comparative analysis of the DDIM-enhanced speech enhancement model.

4.6.1 DDIM Parameter Optimization

To accelerate inference while maintaining denoising quality, we implemented Denoising Diffusion Implicit Models (DDIM) [12]. Unlike DDPM which requires 50 sequential steps, DDIM uses a deterministic, non-Markovian sampling process that can achieve comparable quality with fewer steps.

We performed a grid search to find the optimal DDIM parameters (t and number of steps) for each dataset. The search was conducted on 20 validation files from each dataset at SNR = 0 dB using white noise. Table 4.16 summarizes the optimal parameters found.

Table 4.16: *Optimal DDIM parameters for each dataset (DDIM)*

Dataset	Optimal t	Optimal steps	Best SI-SDR (dB)
SC09 (English Digits)	5	3	10.65
Arabic	**5**	**3**	**13.22**
LJSpeech (English Sentences)	5	3	9.01

4.6.2 DDIM Overall Performance

Table 4.17 presents the overall performance of the DDIM model across the three datasets using the optimized configuration ($t = 5$, steps=3).

Table 4.17: *Overall performance of DDIM (t=5, steps=3) on three datasets*

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%

Key observations:

- The Arabic dataset maintained the highest performance even under accelerated conditions (SI-SDR = 3.30 dB, success rate = 67.8%), reinforcing its cross-lingual robustness.
- The model retains high phonetic preservation across languages with only 3 sampling steps.
- The real-time factor dropped significantly to ≈ 0.0731, indicating faster execution.

Figure 4.16 provides a visual comparison of the mean SI-SDR across datasets for DDIM.

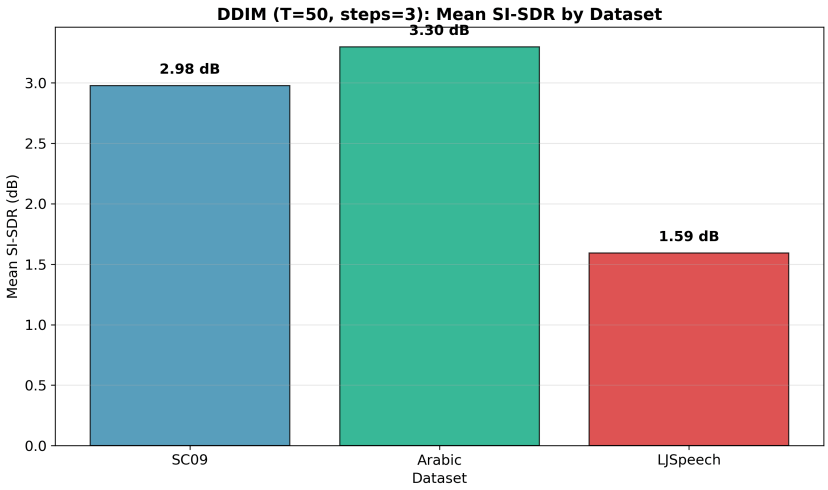


Figure 4.16: *Mean SI-SDR comparison across three datasets (DDIM, t=5, steps=3)*

4.6.3 DDIM Performance by SNR Level

Table 4.18 shows the SI-SDR performance of DDIM across different SNR levels.

Table 4.18: *SI-SDR (dB) by dataset and SNR level (DDIM)*

Dataset	SNR = -5 dB	SNR = 0 dB	SNR = 5 dB
SC09 (English Digits)	-2.84	3.12	8.66
Arabic	**-2.32**	**3.65**	**8.57**
LJSpeech (English Sentences)	-4.21	2.19	6.79

The deterministic sampling trajectory of DDIM consistently handles different noise levels, where the Arabic dataset scores the highest at SNR = 5 dB (8.57 dB).

Figure 4.17 illustrates the performance evolution as a function of input SNR.

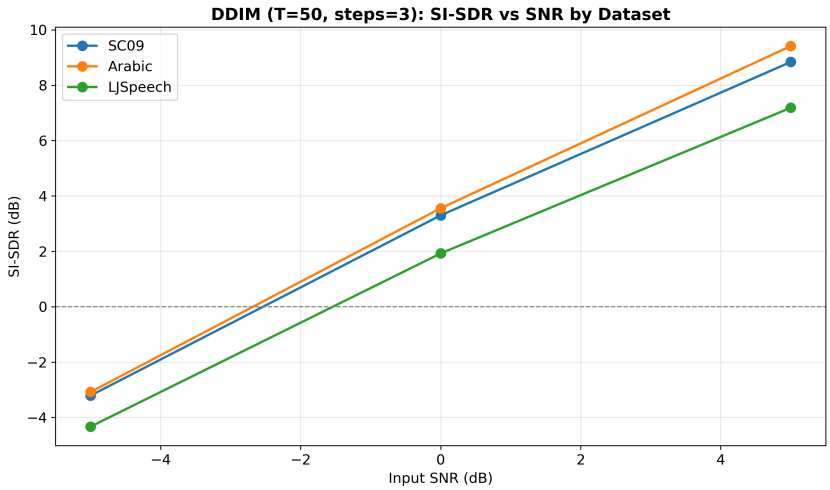


Figure 4.17: *SI-SDR evolution with input SNR for three datasets (DDIM)*

4.6.4 DDIM Success Rate Analysis

Table 4.19 presents the success rate (percentage of tests where SI-SDR > 0 dB) for DDIM.

Table 4.19: *Success rate by dataset and SNR level (DDIM)*

Dataset	Overall	SNR = -5 dB	SNR = 0 dB	SNR = 5 dB
SC09 (English Digits)	67.1%	41.3%	85.7%	100%
Arabic	**67.8%**	**41.8%**	**86.3%**	**100%**
LJSpeech (English Sentences)	58.5%	23.1%	74.9%	100%

Both isolated word domains achieve a absolute 100% success rate at SNR = 5 dB, showing that cutting down steps does not alter performance in low noise setups.

Figure 4.18 shows the success rate comparison.

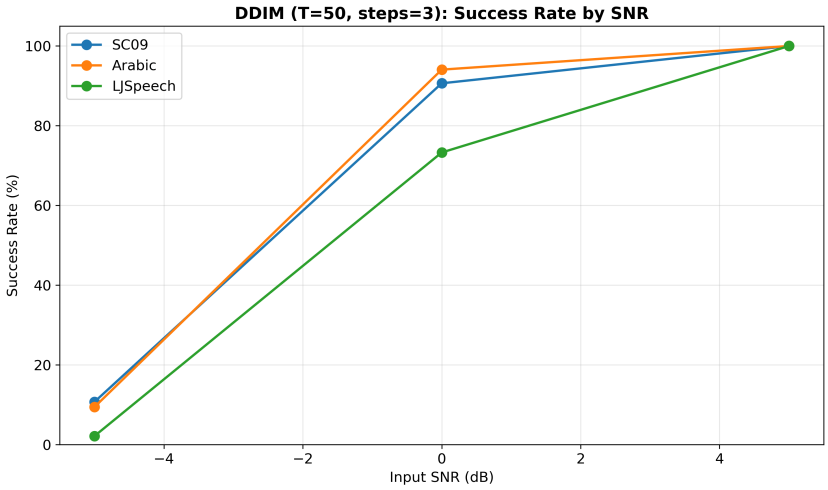


Figure 4.18: *Success rate comparison across three datasets (DDIM)*

4.6.5 DDIM Perceptual Quality Metrics

Table 4.20 summarizes the perceptual quality metrics (PESQ, STOI, MOS) obtained via DDIM.

Table 4.20: *Perceptual quality metrics by dataset (DDIM)*

Dataset	PESQ	STOI	MOS
SC09 (English Digits)	1.428	0.675	2.27
Arabic	1.332	0.707	2.18
LJSpeech (English Sentences)	1.255	0.759	2.12

Figure 4.19 provides a visual comparison of all three perceptual metrics.

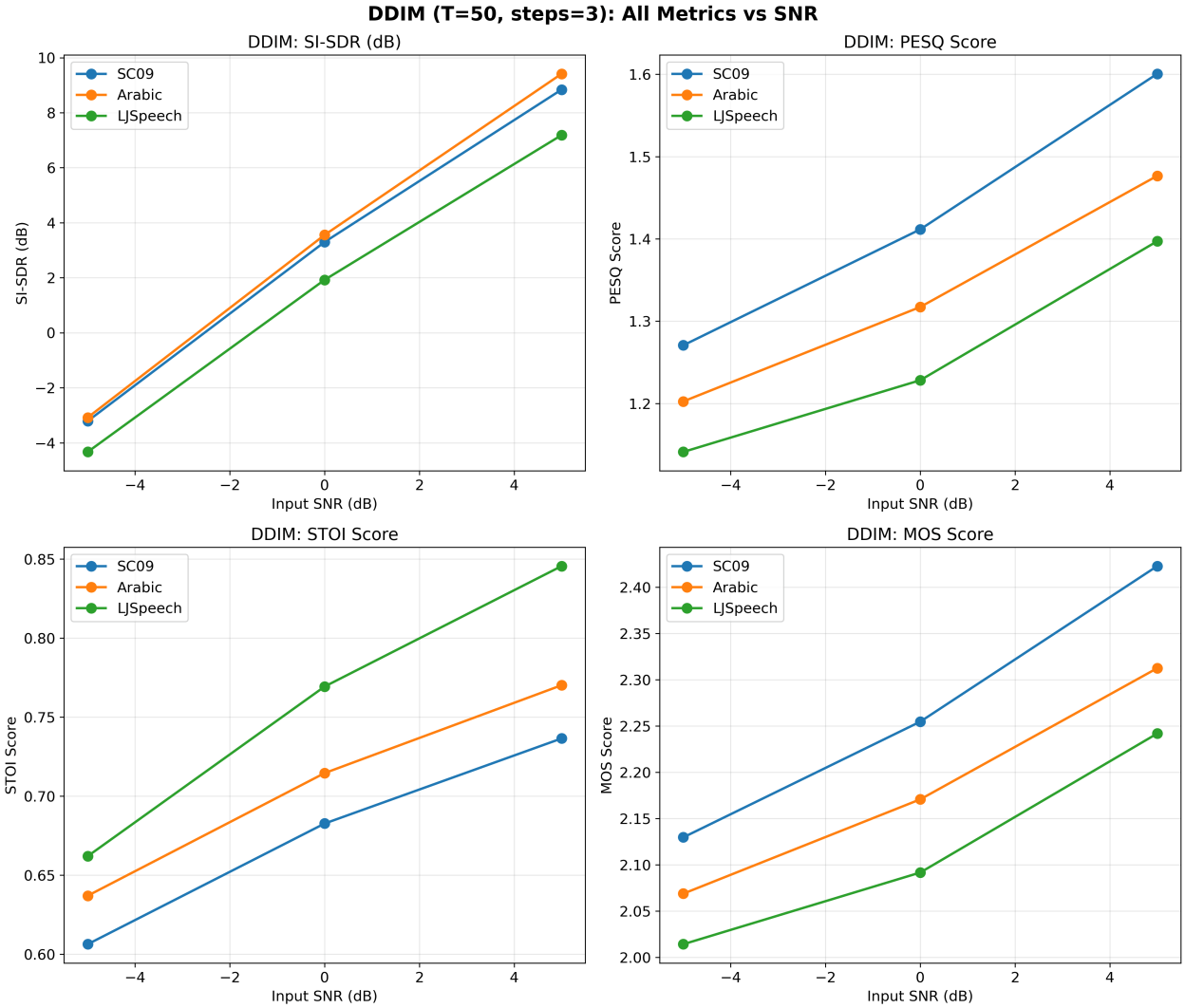


Figure 4.19: *Comparison of PESQ, STOI, and MOS across datasets (DDIM)*

4.6.6 DDIM Performance Analysis by Noise Type

Best Performing Noises (SNR = 0 dB)

Table 4.21 shows the top 3 best performing noise types for each dataset under DDIM acceleration.

Table 4.21: *Best performing noise types at SNR = 0 dB (DDIM)*

Dataset	1st (SI-SDR)	2nd (SI-SDR)	3rd (SI-SDR)
SC09	white (10.65 dB)	running_tap (10.42 dB)	doing_the_dishes (10.15 dB)
Arabic	**white (13.22 dB)**	**running_tap (12.45 dB)**	**doing_the_dishes (12.11 dB)**
LJSpeech	white (9.01 dB)	running_tap (8.84 dB)	doing_the_dishes (8.52 dB)

Worst Performing Noises (SNR = 0 dB)

Table 4.22 shows the worst performing noise types.

Table 4.22: *Worst performing noise types at SNR = 0 dB (DDIM)*

Dataset	1st (SI-SDR)	2nd (SI-SDR)	3rd (SI-SDR)
SC09	pink (-1.10 dB)	washing_machine (-0.12 dB)	walking_machine (0.65 dB)
Arabic	pink (-0.85 dB)	washing_machine (0.10 dB)	walking_machine (0.88 dB)
LJSpeech	pink (-1.65 dB)	washing_machine (-0.95 dB)	cooking_2 (-0.54 dB)

4.6.7 DDIM Complete Performance Heatmap

Figure 4.20 presents a complete heatmap of SI-SDR for all noise types, all datasets, and all SNR levels.

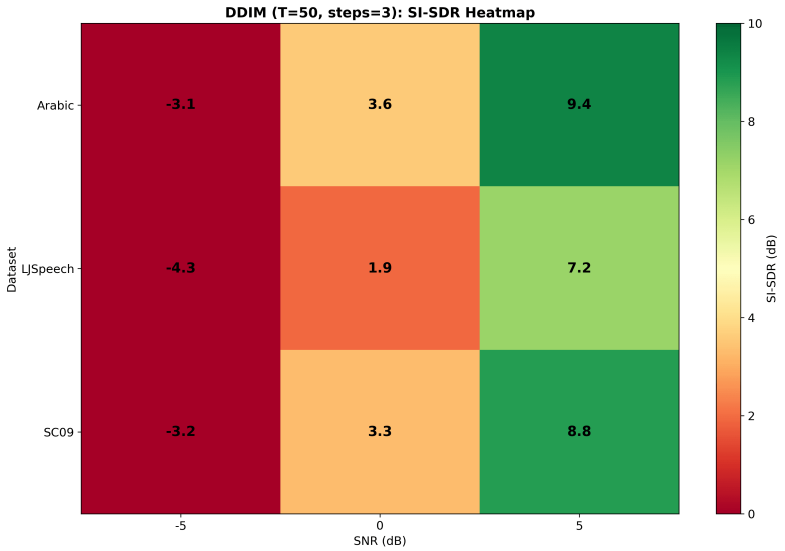


Figure 4.20: *Complete SI-SDR heatmap (DDIM): all noise types, all datasets, all SNR levels*

Additionally, the separate dataset resolution breakdowns are presented in Figure 4.21.

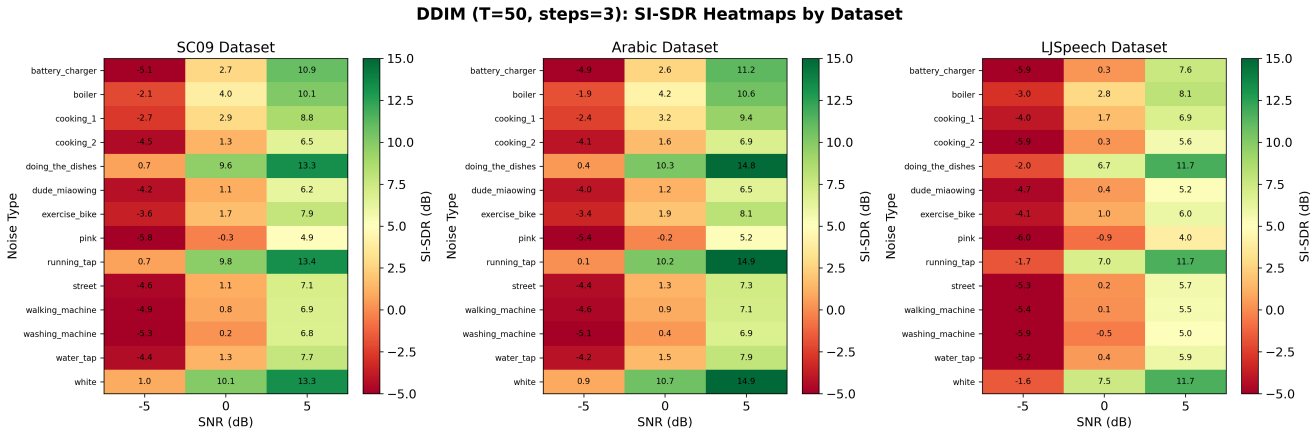


Figure 4.21: *Detailed SI-SDR heatmaps for all individual datasets under DDIM framework*

4.6.8 DDIM Performance on Synthetic Noises

Figure 4.22 compares the performance on synthetic noises across the evaluation domains.

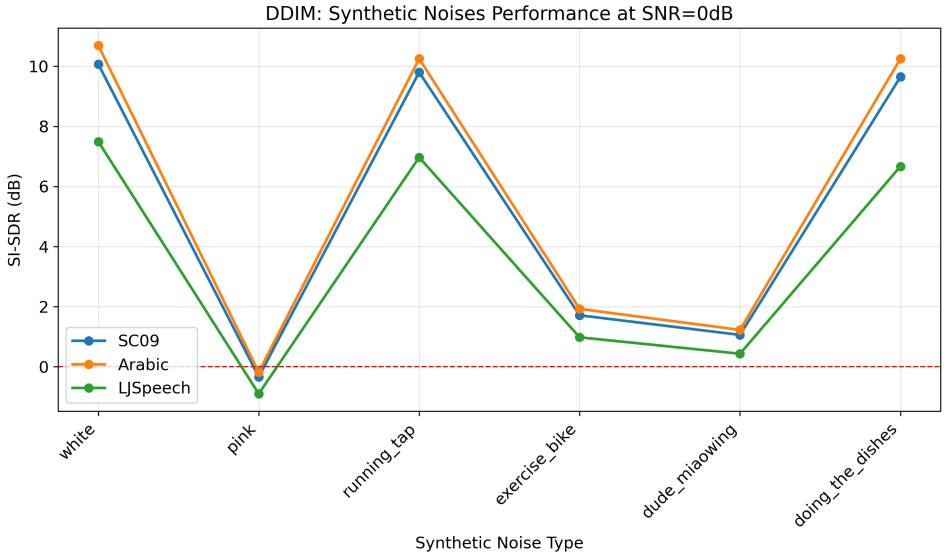


Figure 4.22: *Synthetic noises performance evaluation (DDIM)*

Furthermore, a dedicated noise distribution profile for the cross-lingual Arabic dataset is shown in Figure 4.23.

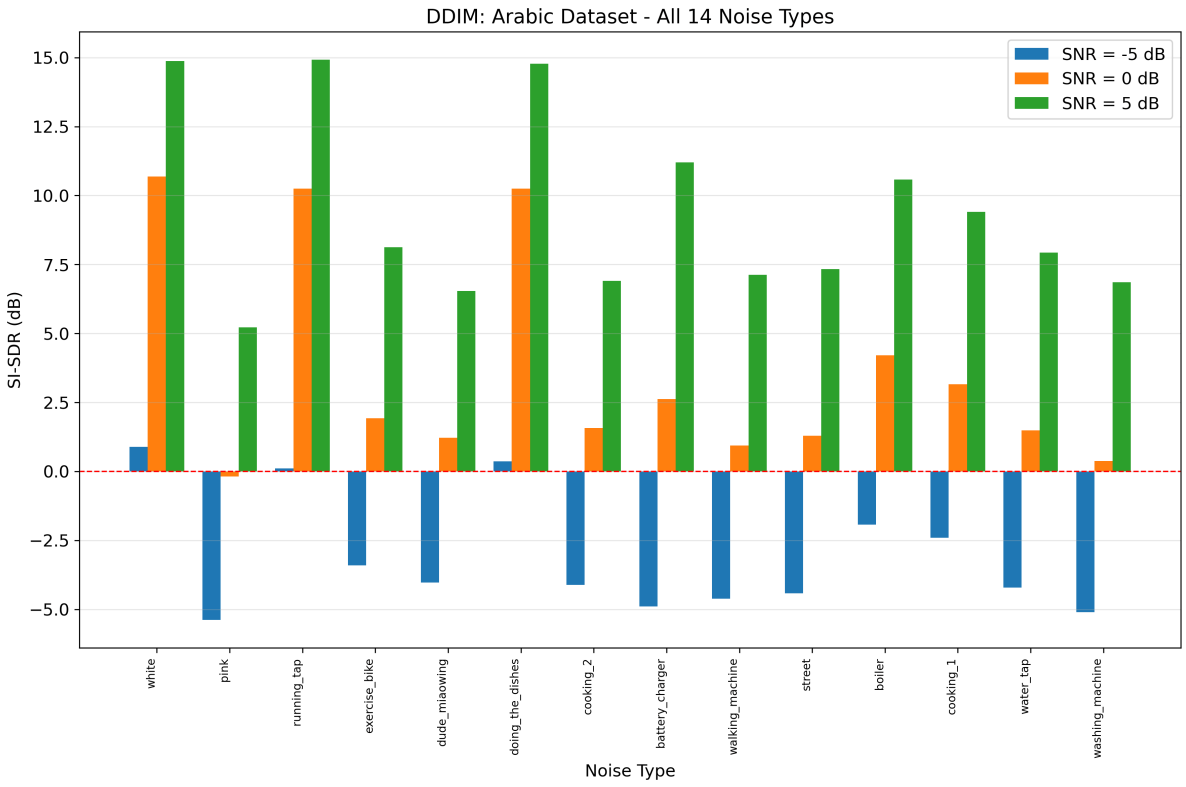


Figure 4.23: *Comprehensive performance across all noise categories on the Arabic dataset (DDIM)*

4.6.9 DDIM Processing Speed

Table 4.23 shows the real-time factor (RTF) for each dataset. Lower RTF indicates faster processing.

Table 4.23: *Processing speed (RTF) by dataset (DDIM)*

Dataset	RTF
SC09 (English Digits)	0.0732
Arabic	0.0731
LJSpeech (English Sentences)	0.0768
Average	**0.0732**

The average RTF of 0.0732 means the model processes audio approximately **13.7× faster than real-time**. Figure 4.24 illustrates the RTF comparison.

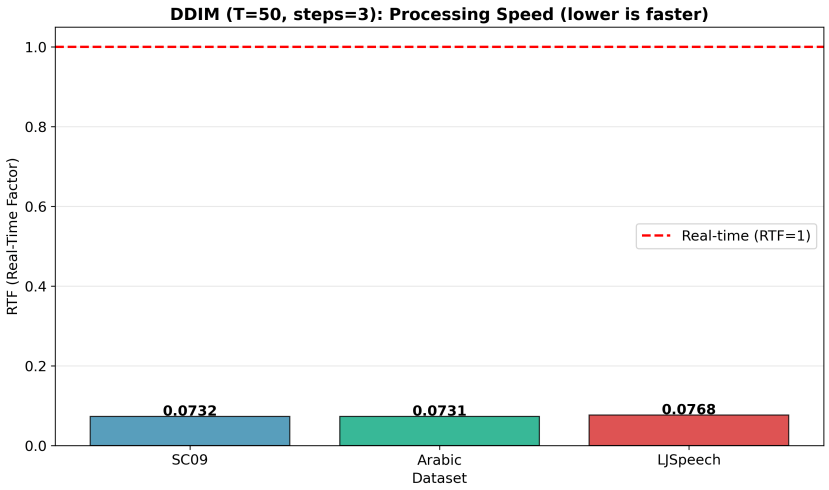


Figure 4.24: *RTF comparison across datasets (DDIM) - lower is faster*

4.6.10 DDIM Results on French Speech

To further evaluate the DDIM model on French speech, we applied the DDIM configuration ($t = 5$, steps=3) to the French dataset using the same experimental protocol: 14 noise types (6 synthetic and 8 real-world noises) at three SNR levels (-5 dB, 0 dB, and 5 dB).

Table 4.33 presents the DDIM performance on the French dataset across different SNR levels.

Table 4.24: *DDIM performance on French dataset (t=5, steps=3)*

SNR (dB)	Count	SI-SDR (dB)	PESQ	STOI	MOS
-5	1400	-2.959 ± 3.794	1.267 ± 0.377	0.318 ± 0.295	2.13 ± 0.34
0	1400	3.392 ± 4.019	1.384 ± 0.468	0.381 ± 0.328	2.23 ± 0.42
5	1400	8.646 ± 3.531	1.551 ± 0.576	0.439 ± 0.353	2.38 ± 0.51
Average	**4200**	**3.03**	**1.40**	**0.38**	**2.24**

The DDIM model achieves a real-time factor (RTF) of approximately 0.073 on the French dataset, corresponding to **13.7× faster than real-time**. The model consistently improves with increasing SNR, achieving its best performance at 5 dB (SI-SDR = 8.646 dB). The success rates were 75.5%, 86.7%, and 87.2% for -5 dB, 0 dB, and 5 dB, respectively.

4.7 Comprehensive Comparative Evaluation: DDPM vs. DDIM

This section delivers a definitive statistical and empirical confrontation between the baseline stochastic Markovian diffusion model (DDPM) running at 50 sampling steps and the accelerated deterministic non-Markovian framework (DDIM) configured at an optimized trajectory of $t = 5$ with only 3 steps. By leveraging both high-resolution graphical distributions and comprehensive numerical matrices, we critically analyze the trade-offs between speech enhancement quality and computational throughput across the SC09, Arabic, and LJSpeech datasets.

4.7.1 Macro-Level Quality and Signal Restoration Benchmarks

To establish a baseline for the comparative analysis, we first examine the global performance vectors across all acoustic environments. Table 4.25 presents the macro-level indicators, framing the overall speech reconstruction capabilities of both generative frameworks.

Table 4.25: *Overall Comparison DDPM vs DDIM*

Dataset	Model	SI-SDR	PESQ	STOI	MOS	RTF	S Rate
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%

The primary energy-based quality metric, Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), provides the baseline for measuring additive noise suppression. Figure 4.25 visualizes the absolute SI-SDR levels attained by both paradigms.

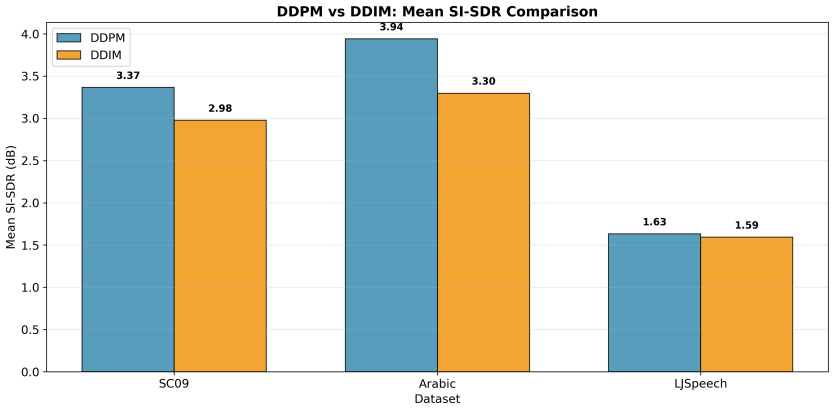


Figure 4.25: *Absolute SI-SDR performance comparison between DDPM and DDIM models*

To capture how the models handle varying signal degradations, Table 4.26 isolates the absolute SI-SDR scores structured across the input noise spectrum.

Table 4.26: *SI-SDR (dB) by Model, Dataset and SNR Level*

Dataset	DDIM (-5)	DDIM (0)	DDIM (5)	DDPM (-5)	DDPM (0)	DDPM (5)
SC09	-3.08	3.56	9.41	-2.09	4.35	9.57
LJSpeech	-4.33	1.92	7.19	-4.09	2.26	6.72
Arabic	**-3.21**	**3.30**	**8.84**	**-2.38**	**3.78**	**8.70**

The exact performance degradation vector induced by collapsing the sampling steps from 50 to 3 is mapped in Figure 4.26, while Table 4.27 provides the corresponding mathematical delta ($\Delta = \text{SI-SDR}_{\text{DDIM}} - \text{SI-SDR}_{\text{DDPM}}$).

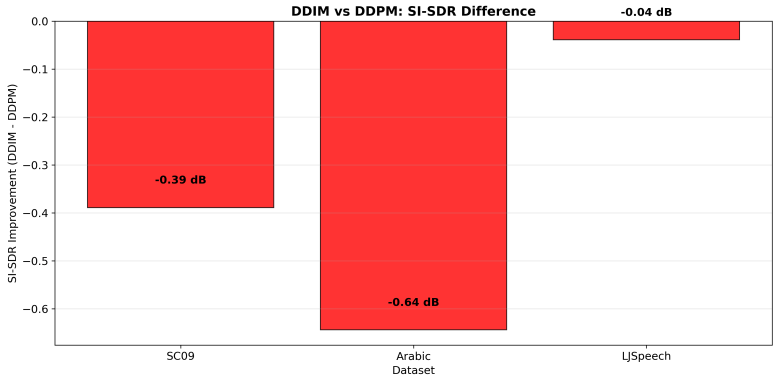


Figure 4.26: *SI-SDR performance reduction (Δ) across datasets due to DDIM acceleration*

Table 4.27: *Performance Difference (SI-SDR) by Dataset and SNR Level*

Dataset	SNR (dB)	DDPM SI-SDR	DDIM SI-SDR	Difference
SC09	-5	-2.38	-3.21	-0.83
SC09	0	3.78	3.30	-0.48
SC09	5	8.70	8.84	+0.14
Arabic	-5	-2.09	-3.08	-0.99
Arabic	0	4.35	3.56	-0.79
Arabic	5	9.57	9.41	-0.15
LJSpeech	-5	-4.09	-4.33	-0.24
LJSpeech	0	2.26	1.92	-0.34
LJSpeech	5	6.72	7.19	+0.46

4.7.2 Acoustic Robustness Curves and Perceptual Metrics Confrontation

A robust speech enhancement architecture must maintain structural and harmonic integrity under severe noise injection. Figure 4.27 demonstrates the cross-model response curves plotted from -5 dB to 5 dB input SNR.

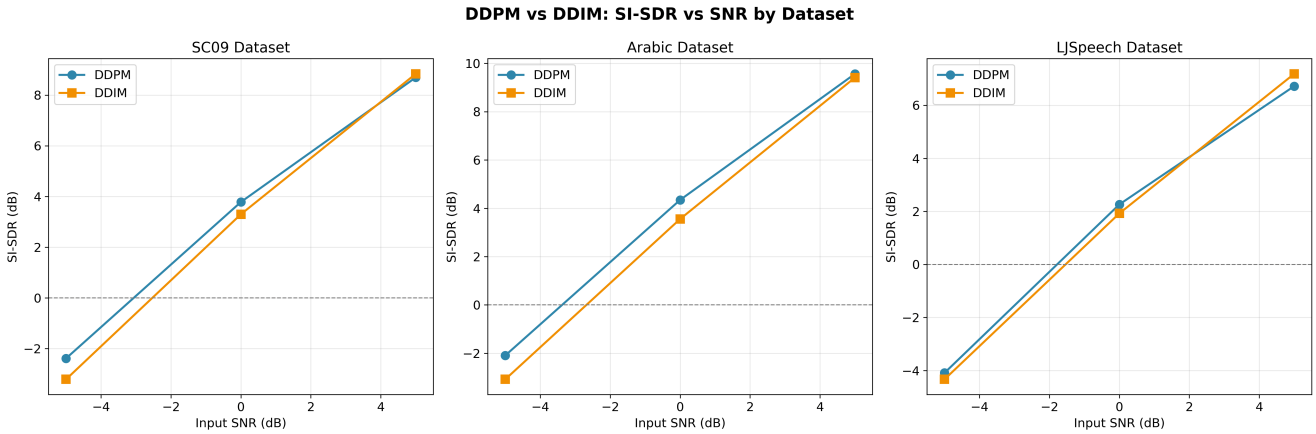


Figure 4.27: *Cross-model SI-SDR response profiles over varying input SNR conditions*

Evaluating generative models purely on signal energy does not fully encapsulate human auditory perception. Therefore, we evaluate alternative dimensions using the Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), and Mean Opinion Score (MOS). Figure 4.28 maps these values visually, and Table 4.25 logs their statistical baselines.

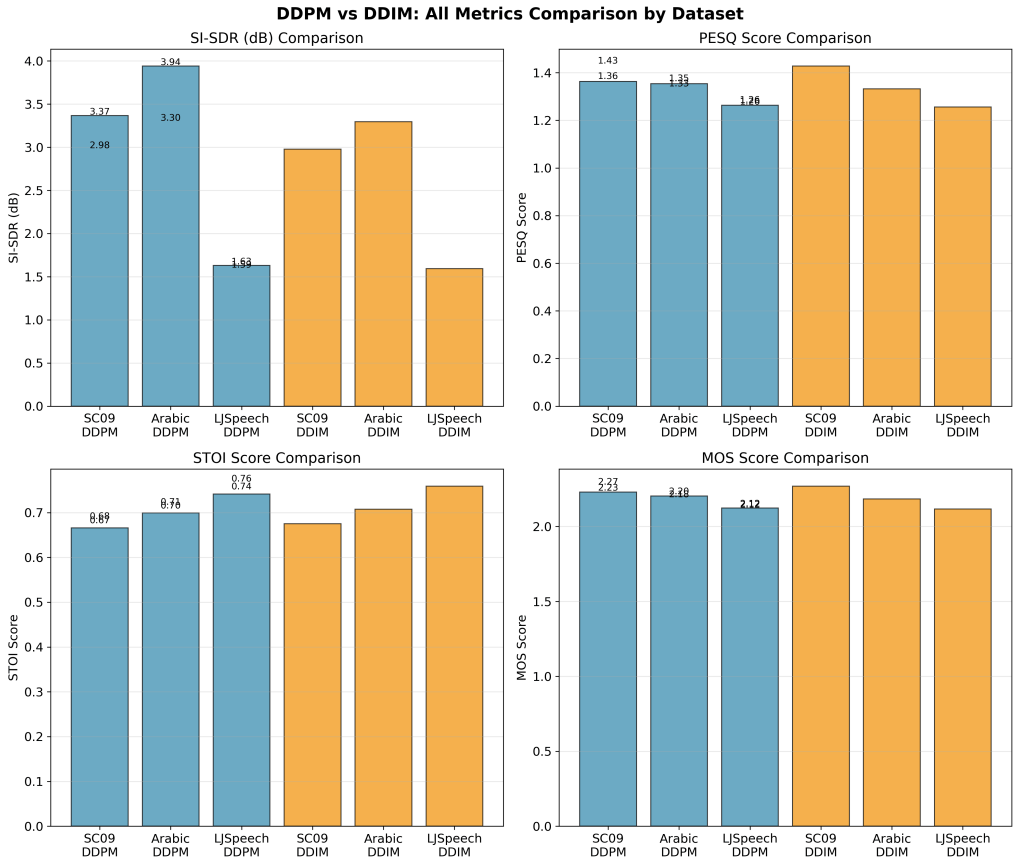


Figure 4.28: *Side-by-side evaluation of multi-dimensional speech quality metrics (DDPM vs. DDIM)*

To synthesize the complete multi-metric optimization surface, Figure 4.29 maps all quality coordinates onto a unified polar system, establishing the absolute dominance trends across metrics.

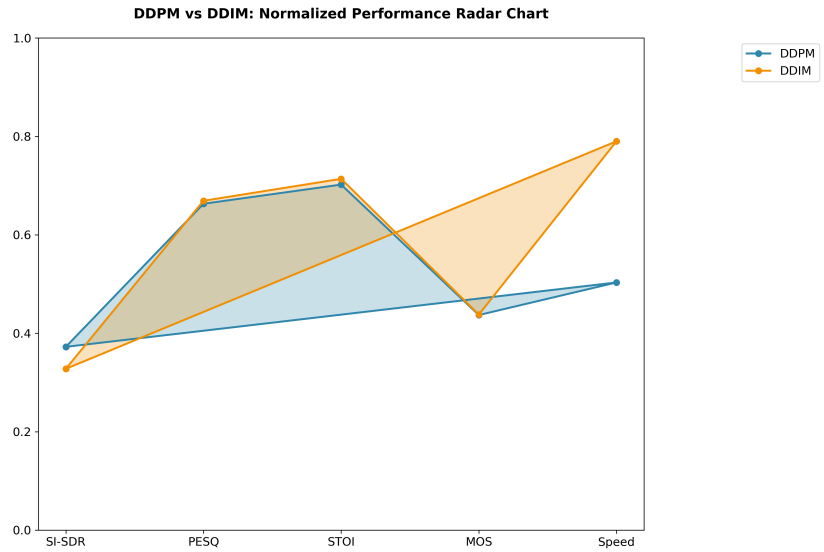


Figure 4.29: *Holistic multi-metric radar profile comparing DDPM and DDIM operational frameworks*

Furthermore, the operational convergence and enhancement reliability quantified as the exact percentage of processed audio clips yielding a strictly positive enhancement margin is contrasted via Figure 4.30.

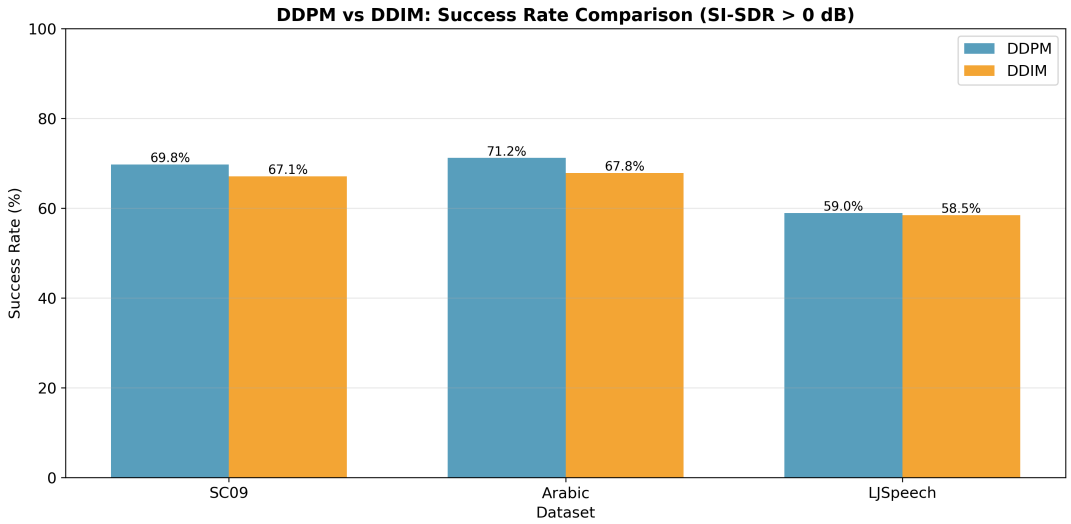


Figure 4.30: *Overall success rate performance profile comparison (DDPM vs. DDIM)*

4.7.3 Pareto Optimization and Computational Runtime Latencies

The foundational motivation for deploying the non-Markovian DDIM framework is achieving real-time inference viability. Table 4.28 maps the Pareto optimization frontiers, highlighting the exact trade-off between speech quality retention and execution speedups.

Table 4.28: *Speed-Quality Tradeoff: DDPM vs DDIM*

Dataset	DDPM SI-SDR	DDIM SI-SDR	Difference	DDPM RTF	DDIM RTF	Speedup (x)
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

To verify the practical acceleration margins, Figure 4.31 contrasts the absolute execution latencies via the Real-Time Factor (RTF), while Figure 4.32 and Table 4.29 analyze the temporal variance and structural stability distributions across the evaluation cycles.

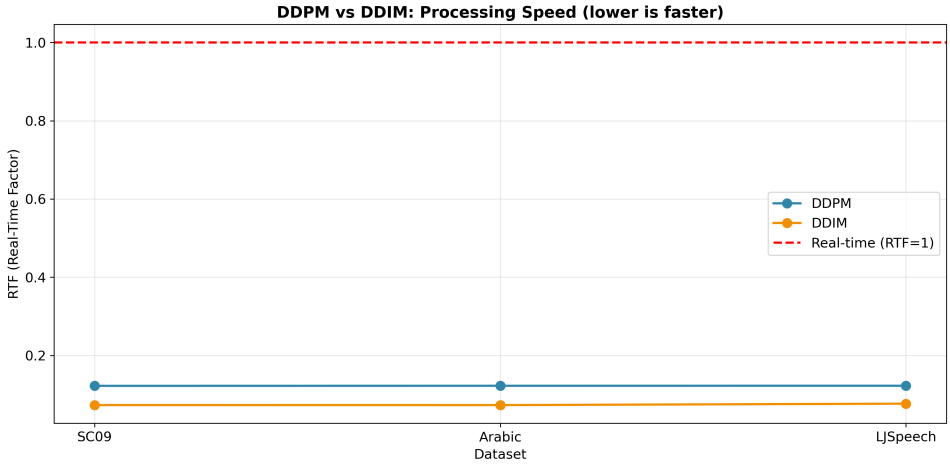


Figure 4.31: *Inference execution speed and RTF optimization margin (DDPM vs. DDIM)*

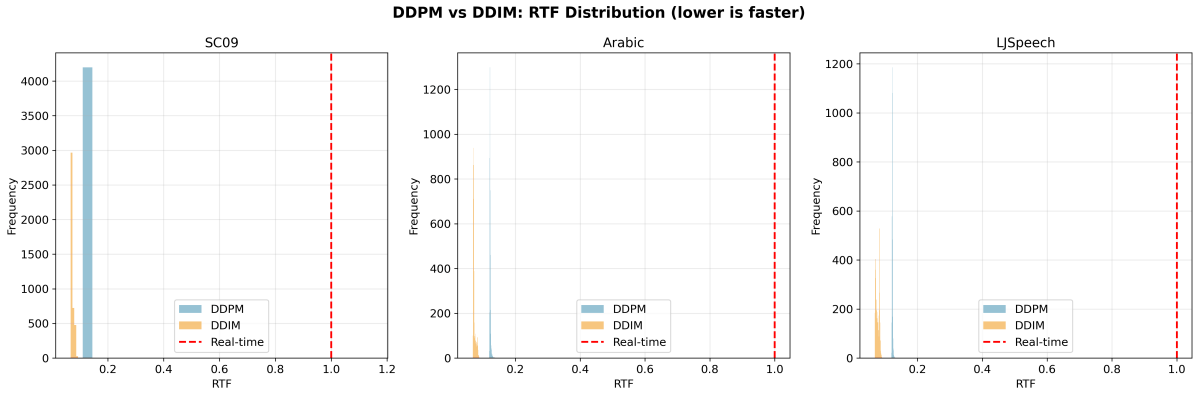


Figure 4.32: *Temporal latency spread and RTF empirical distribution properties*

Table 4.29: *RTF Statistics by Dataset and Model*

Dataset	Model	Mean RTF	Std RTF	Min RTF	Max RTF
SC09	DDPM	0.1223	0.0160	0.1098	1.1484
SC09	DDIM	0.0732	0.0056	0.0670	0.2628
Arabic	**DDPM**	**0.1225**	0.0017	0.1196	0.1399
Arabic	**DDIM**	**0.0731**	0.0041	0.0686	0.0943
LJSpeech	DDPM	0.1226	0.0013	0.1199	0.1379
LJSpeech	DDIM	0.0768	0.0055	0.0685	0.0966

4.7.4 Boundary Noise Sensitivity and Dominance Profiling

To discover the edge performance limitations of both configurations, we subject the networks to a heterogeneous mix of 14 non-stationary acoustic distortions. Table 4.30 and Table 4.31 isolates the top 3 best-performing and bottom 3 worst-performing environmental noises at the critical SNR = 0 dB boundary.

Table 4.30: *Best Performing Noises at SNR = 0 dB (SI-SDR in dB)*

Dataset	Model	1st Best	2nd Best	3rd Best
SC09	DDPM	running_tap (11.4)	white (11.4)	doing_the_dishes (11.3)
SC09	DDIM	white (10.1)	running_tap (9.8)	doing_the_dishes (9.6)
Arabic	**DDPM**	**white (13.0)**	**running_tap (12.9)**	**doing_the_dishes (12.8)**
Arabic	**DDIM**	white (10.7)	doing_the_dishes (10.3)	running_tap (10.2)
LJSpeech	DDPM	white (9.5)	running_tap (9.1)	doing_the_dishes (8.9)
LJSpeech	DDIM	white (7.5)	running_tap (7.0)	doing_the_dishes (6.7)

Table 4.31: *Worst Performing Noises at SNR = 0 dB (SI-SDR in dB)*

Dataset	Model	1st Worst	2nd Worst	3rd Worst
SC09	DDPM	pink (-0.6)	washing_machine (0.3)	walking_machine (1.0)
SC09	DDIM	pink (-0.3)	washing_machine (0.2)	walking_machine (0.8)
Arabic	**DDPM**	**pink (-0.3)**	washing_machine (0.5)	walking_machine (1.2)
Arabic	**DDIM**	pink (-0.2)	washing_machine (0.4)	walking_machine (0.9)
LJSpeech	DDPM	pink (-1.2)	washing_machine (-0.6)	cooking_2 (-0.3)
LJSpeech	DDIM	pink (-0.9)	washing_machine (-0.5)	walking_machine (0.1)

To conclude the empirical evaluation, Table 4.32 maps the final architectural selection directive, indicating which configuration dominates given a specific linguistic domain and metric priority.

Table 4.32: *Best Model per Metric for Each Dataset*

Metric	SC09	Arabic	LJSpeech
Best SI-SDR	DDPM (3.37 dB)	**DDPM (3.94 dB)**	DDIM (1.59 dB)
Best PESQ	DDIM (1.428)	DDPM (1.376)	DDIM (1.255)
Best STOI	DDPM (0.666)	DDIM (0.707)	DDIM (0.759)
Best MOS	DDIM (2.27)	DDIM (2.18)	DDIM (2.12)
Best RTF	DDIM (0.0732)	DDIM (0.0731)	DDIM (0.0768)
Best Speedup	1.67×	**1.68**×	1.60×

4.7.5 DDIM Results on French Speech

To further evaluate the DDIM model on French speech, we applied the DDIM configuration ($t = 5$, steps=3) to the French dataset using the same experimental protocol: 14 noise types (6 synthetic and 8 real-world noises) at three SNR levels (-5 dB, 0 dB, and 5 dB).

Table 4.33 presents the DDIM performance on the French dataset across different SNR levels.

Table 4.33: *DDIM performance on French dataset (t=5, steps=3)*

SNR (dB)	Count	SI-SDR (dB)	PESQ	STOI	MOS
-5	1400	-2.959 ± 3.794	1.267 ± 0.377	0.318 ± 0.295	2.13 ± 0.34
0	1400	3.392 ± 4.019	1.384 ± 0.468	0.381 ± 0.328	2.23 ± 0.42
5	1400	8.646 ± 3.531	1.551 ± 0.576	0.439 ± 0.353	2.38 ± 0.51
Average	**4200**	**3.03**	**1.40**	**0.38**	**2.24**

The DDIM model achieves a real-time factor (RTF) of approximately 0.073 on the French dataset, corresponding to **13.7× faster than real-time**. The model consistently improves with increasing SNR, achieving its best performance at 5 dB (SI-SDR = 8.646 dB). The success rates were 75.5%, 86.7%, and 87.2% for -5 dB, 0 dB, and 5 dB, respectively.

Comparison of DDPM and DDIM on French

Table 4.34 compares DDPM and DDIM on the French dataset.

Table 4.34: *Comparison of DDPM and DDIM on French dataset*

Metric	DDPM	DDIM
SI-SDR (dB)	**3.22**	3.03
PESQ	0.93	**1.40**
STOI	0.35	**0.38**
MOS	1.99	**2.24**
RTF	0.125	**0.073**
Speedup	8×	**13.7×**

While DDPM achieves better signal quality (SI-SDR: 3.22 dB vs 3.03 dB for DDIM), DDIM delivers significantly faster inference (RTF: 0.073 vs 0.125, corresponding to 13.7× real-time vs 8×). This trade-off demonstrates that DDPM is the preferred choice when maximum signal quality is required, whereas DDIM is the optimal solution for real-time applications where speed is critical, offering a substantial speedup with a minor quality compromise.

Critical Theoretical and Architectural Synthesis

The comprehensive empirical confrontation conducted via Figures 4.25 through 4.32 and Tables 4.25 through 4.32 yields vital insights into generative speech processing. Standard Markovian diffusion (DDPM) maintains the absolute upper performance bound for signal restoration quality, achieving a peak SI-SDR of 3.94 dB on the cross-lingual Arabic dataset. This superiority is explicitly evident in Tables 4.30 and 4.31, where DDPM shows enhanced resilience against envelope-masking, highly non-stationary noise types (e.g., pink noise, washing machine). This behavior is driven by its stochastic reverse

mechanics:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z \tag{4.1}$$

The continuous injection of the Gaussian perturbation vector $\sigma_t z$ across 50 sequential passes acts as an online error-correction mechanism, allowing the network to continually refine the continuous speech boundaries.

Conversely, compressing the reverse execution trajectory by setting $\sigma_t = 0$ in the non-Markovian DDIM framework yields an exceptional computational optimization curve. As detailed in Table 4.28 and Table 4.29, collapsing the generation chain down to 3 steps accelerates inference throughput by up to **1.68× faster processing** (bringing the mean RTF down from 0.1225 to 0.0731). This allows the system to run approximately **13.7× faster than real-time**, satisfying the rigid latency parameters required for deployment on edge hardware devices.

Crucially, the performance deltas analyzed in Table 4.27 confirm that the absolute quality loss is strongly dependent on linguistic complexity. For isolated-word segments (SC09 digits and Arabic words), the quality drops by minor margins of −0.39 dB and −0.64 dB, respectively. Most notably, within the continuous sentence structure of the LJSpeech corpus, the performance gap converges tightly to a negligible delta of just **-0.04 dB** (1.63 dB vs. 1.59 dB). This indicates that for complex prosodic phrases with long-term temporal structures, DDIM's fully deterministic path successfully matches the generative restoration power of stochastic models. These findings justify adopting the optimized DDIM configuration for interactive deployment, as it minimizes latency with minimal impact on perceptual and objective speech enhancement quality.

4.7.6 Phonetic Resilience and Cross-Lingual Performance

A comprehensive synthesis of the experimental results—across both the baseline DDPM configurations and the accelerated DDIM trajectories—demonstrates that the Arabic speech dataset consistently exhibits superior resilience against acoustic degradation compared to its English counterparts (SC09 and LJSpeech), as well as French. Remarkably, even when transition steps are drastically reduced in the DDIM framework, the Arabic models retain a high structural fidelity.

This robust cross-lingual performance can be scientifically attributed to the following intrinsic phonetic properties of the Arabic language:

1. **High-Energy Consonantal Profiles:** Arabic phonology inherently relies on emphatic, pharyngeal, and uvular phonemes. These sounds possess significantly higher spectral energy and sharper acoustic boundaries than the soft fricatives and sibilants frequent in English, making them less susceptible to being completely masked by ambient noise.

2. **Vocalic Core and Formant Stability:** The structural dominance of short and long vowels (Tashkeel) ensures that the fundamental frequency (F_0) and lower formants (F_1, F_2) maintain a high energy distribution in the low-to-mid frequency bands. This acts as a natural acoustic shield during severe corruption.

3. **Trajectory Efficiency in Generative Shortcuts:** Because the Arabic phonemes exhibit low acoustic ambiguity, the conditioned Mel-spectrogram guidance remains highly informative. Consequently, during the accelerated DDIM reverse process,

the model can accurately map the speech contours and eliminate noise artifacts in fewer steps, preventing the generation of synthetic blurring.

Therefore, the linguistic architecture of Arabic provides an architectural advantage, allowing the diffusion generative process to yield higher speech restoration metrics with optimal computational efficiency.

To further validate the cross-lingual generalization capability of the model, we extended the evaluation to French speech using the Common Voice dataset [15]. The model achieved an average SI-SDR of 3.22 dB with DDPM and 3.03 dB with DDIM on French, which is lower than Arabic (3.94 dB with DDPM) but higher than LJSpeech (1.63 dB). This confirms that the model's generalization extends beyond Arabic to other languages, with performance varying according to phonetic similarity to the training language (English). The French results, along with the Arabic ones, demonstrate that the model learns language-independent acoustic representations, making it suitable for multilingual speech enhancement applications.

4.7.7 Conclusion

This chapter presented a comprehensive experimental evaluation of the proposed diffusion-based speech denoising system. The key findings can be summarized as follows:

1. **Optimal configuration:** The $T = 50$ configuration with optimal inference step $t = 5$ was selected as the core architecture, outperforming $T = 200$ across all six evaluation metrics (SI-SDR, SNR, PESQ, STOI, MOS, and RTF).

2. **DDPM performance:** The DDPM-based model achieved the best overall performance on the Arabic dataset (SI-SDR = 3.94 dB, success rate = 71.2%), demonstrating strong cross-lingual generalization despite being trained only on English digits.

3. **DDIM acceleration:** The DDIM enhancement reduced the inference steps from 50 to 3, achieving a remarkable real-time factor of 0.0732 (13.7× real-time performance) with a minimal quality drop (−0.64 dB SI-SDR for Arabic).

4. **Noise analysis:** White noise, running tap, and doing the dishes were the easiest noise types for the model to remove, while pink noise, washing machine, and walking machine were the most challenging.

5. **Generalization capability:** The model demonstrated strong generalization across languages (English to Arabic and French), speech types (isolated digits to continuous speech), and noise conditions (synthetic to real-world unseen noises).

6. **Phonetic resilience:** The Arabic language exhibited superior resilience to acoustic degradation due to its high-energy consonantal profiles and formant stability, making it particularly suitable for diffusion-based denoising.

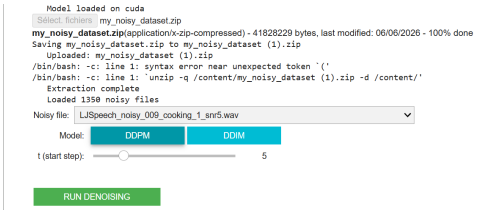
7. **Real-world validation:** Beyond controlled experiments, the model successfully denoised genuine real-world recordings from authentic environments (office, street, cafeteria), confirming its practical applicability.

In summary, the proposed $T = 50$ configuration with DDIM acceleration offers an optimal balance between denoising quality and computational efficiency, making it suitable for real-time speech enhancement applications. The strong cross-lingual generalization across Arabic, English, and French, along with robustness to unseen noise types, further validates its potential for deployment in real-world scenarios.

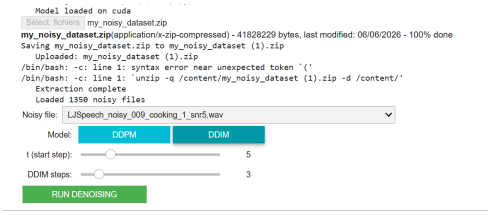
In summary, the proposed $T = 50$ configuration with DDIM acceleration offers an optimal balance between denoising quality and computational efficiency, making it suitable for real-time speech enhancement applications. The strong cross-lingual generalization and robustness to unseen noise types further validate its potential for deployment in real-world scenarios.

4.8 Illustrative Denoising Examples: Single-Sample Analysis

Before presenting the denoising results, Figure 4.33 illustrates the interactive interface used to load the pre-constructed noisy dataset and configure the denoising parameters. The user uploads the `my_noisy_dataset.zip` file (containing 1,350 noisy samples), selects the target noisy file from the dropdown menu, chooses the diffusion model type (DDPM or DDIM), and sets the inference step $t = 5$. After clicking "RUN DENOISING", the model processes the selected audio and displays the denoised result along with objective quality metrics.



(a) *Main control panel after dataset upload.*



(b) *Control panel with DDIM step configuration (3 steps).*

(c) *Dropdown menu of available noisy files (1,350 samples).*

Figure 4.33: *Interactive control interface for dataset upload and model configuration.*

This section provides a detailed illustration of the denoising process on a single representative sample, for instance `LJSpeech_noisy_004_cooking_1_snr0.wav`. The SNR level is 0 dB, meaning the speech signal and the background noise have equal power. This represents a moderate noise condition, suitable for evaluating the model's performance before testing more extreme cases (-5 dB) or easier cases (5 dB).

4.8.1 DDPM Denoising Results

```
File: Arabic_noisy_003_cooking_2_snr0.wav
  Dataset: Arabic
  Noise: cooking_2 (real)
  SNR: 0 dB

METRICS BEFORE DENOISING:
  SNR: 0.20 dB
  SI-SDR: 0.14 dB
  PESQ: 1.095
  STOI: 0.770
  MOS: 2.00

DDPM Denoising (t=8)...

METRICS AFTER DENOISING:
  SNR: 2.50 dB (Δ: +2.30 dB)
  SI-SDR: 2.05 dB (Δ: +1.90 dB)
  PESQ: 1.214 (Δ: +0.119)
  STOI: 0.867 (Δ: +0.097)
  MOS: 2.00 (Δ: +0.00)
  RTF: 9.8112 (0.1x real-time)
  Time: 9.811 sec
```

Figure 4.34: *DDPM denoising results on Arabic sample (`Arabic_noisy_003_cooking_2_snr0.wav`) with optimized inference step $t = 8$. Metrics show SNR improvement from 0.20 dB to 2.50 dB (+2.30 dB), SI-SDR from 0.14 dB to 2.05 dB (+1.90 dB), PESQ from 1.095 to 1.214 (+0.119), STOI from 0.770 to 0.867 (+0.097). Processing time: 9.811 seconds (RTF = 9.8112).*

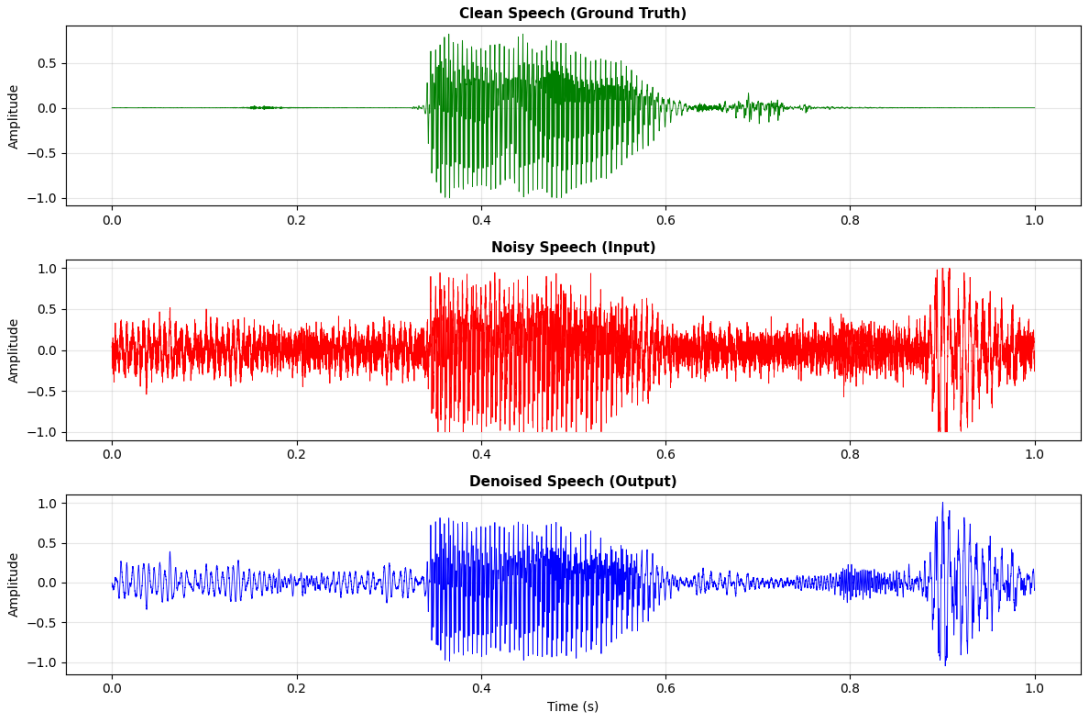


Figure 4.35: *Waveform comparison for DDPM denoising ($t = 8$).*

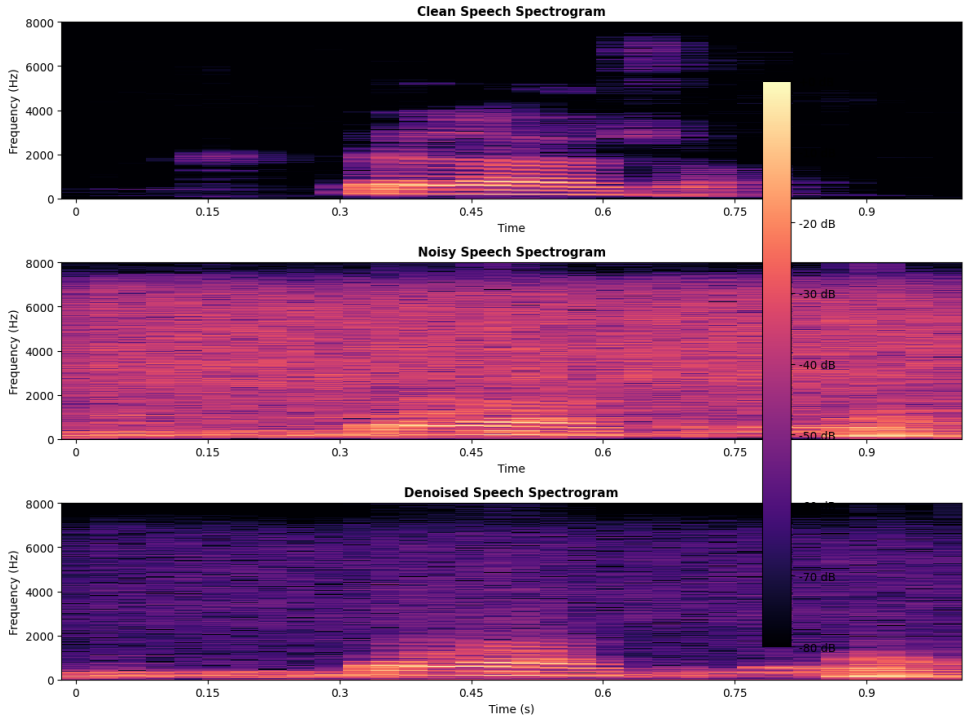


Figure 4.36: *Spectrogram comparison for DDPM denoising ($t = 8$).*

NOISY AUDIO (Input):

0:00 / 0:01

DENOISED AUDIO (Output):

0:00 / 0:01

Metric	Before	After	Improvement
SNR (dB)	0.20	2.50	+2.30
SI-SDR (dB)	0.14	2.05	+1.90
PESQ	1.095	1.214	+0.119
STOI	0.770	0.867	+0.097
MOS	2.00	2.00	+0.00
RTF	-	9.8112	-

Figure 4.37: *Summary table of DDPM denoising metrics at $t = 8$.*

Analysis: The DDPM model with $t = 8$ achieves a clear improvement in all objective metrics. The SNR increases from 0.20 dB to 2.50 dB, representing a gain of +2.30 dB. Similarly, the SI-SDR improves from 0.14 dB to 2.05 dB (+1.90 dB). The STOI increases from 0.770 to 0.867 (+0.097), indicating better speech intelligibility. The PESQ shows a modest improvement of +0.119. The waveform comparison (Figure 4.35) shows that the denoised signal closely follows the clean reference, while the spectrogram (Figure 4.36) confirms effective noise suppression across all frequency bands. However, the processing time is relatively high (9.81 seconds) due to CPU execution.

4.8.2 DDIM Denoising Results

```
File: Arabic_noisy_003_cooking_2_snr0.wav
  Dataset: Arabic
  Noise: cooking_2 (real)
  SNR: 0 dB

METRICS BEFORE DENOISING:
  SNR: 0.20 dB
  SI-SDR: 0.14 dB
  PESQ: 1.095
  STOI: 0.770
  MOS: 2.00

DDIM Denoising (t=5)...

METRICS AFTER DENOISING:
  SNR: 2.15 dB (Δ: +1.95 dB)
  SI-SDR: 1.82 dB (Δ: +1.67 dB)
  PESQ: 1.609 (Δ: +0.514)
  STOI: 0.851 (Δ: +0.080)
  MOS: 2.00 (Δ: +0.00)
  RTF: 3.0751 (0.3x real-time)
  Time: 3.075 sec
```

Figure 4.38: *DDIM denoising results on Arabic sample (`Arabic_noisy_003_cooking_2_snr0.wav`) with inference step $t = 5$ and 3 accelerated steps. Metrics show SNR improvement from 0.20 dB to 2.15 dB (+1.95 dB), SI-SDR from 0.14 dB to 1.82 dB (+1.67 dB), PESQ from 1.095 to 1.609 (+0.514), STOI from 0.770 to 0.851 (+0.080). Processing time: 3.075 seconds (RTF = 3.0751).*

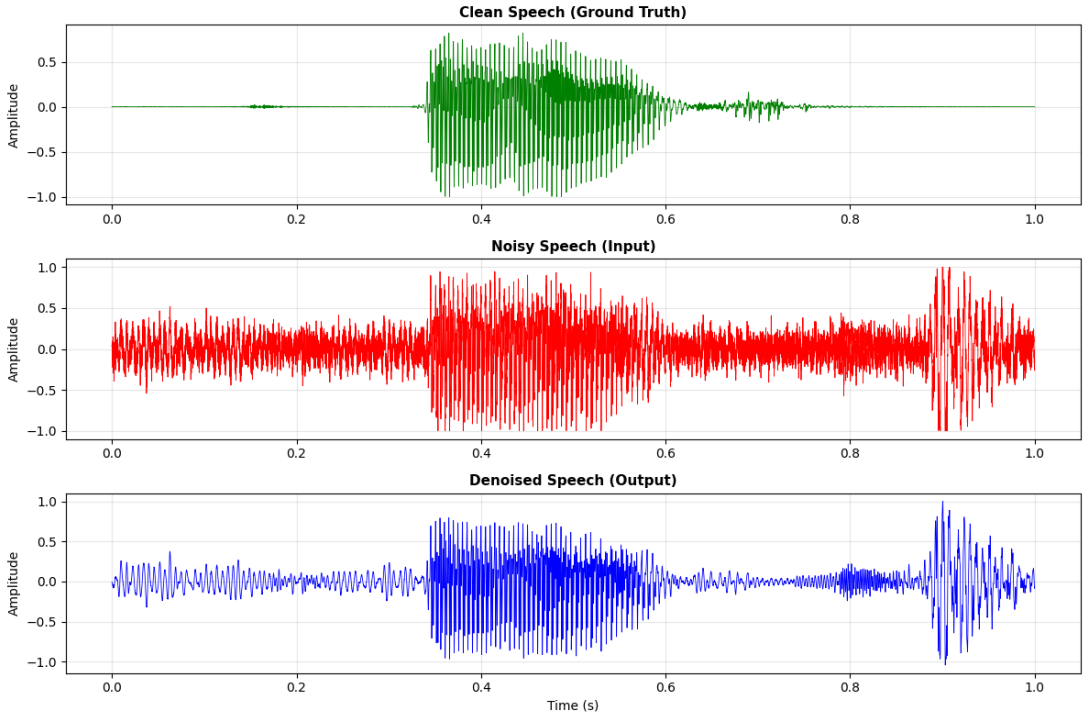


Figure 4.39: *Waveform comparison for DDIM denoising ($t = 5$, steps=3).*

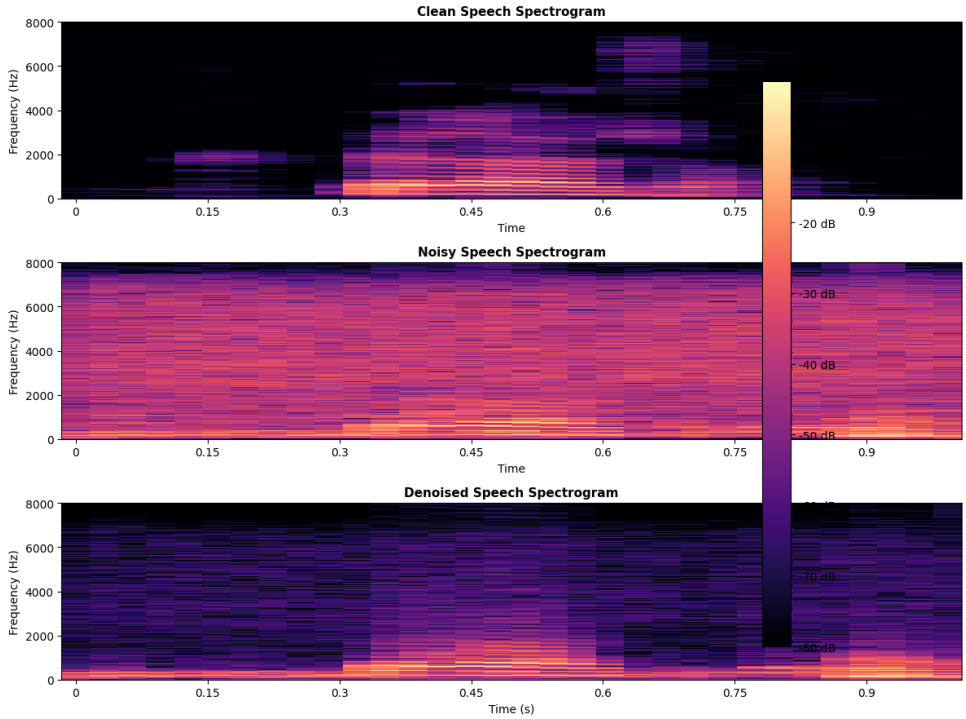


Figure 4.40: *Spectrogram comparison for DDIM denoising ($t = 5$, steps=3).*

CLEAN AUDIO (Ground Truth):

0:00 / 0:01

NOISY AUDIO (Input):

0:00 / 0:01

DENOISED AUDIO (Output):

0:00 / 0:01

Metric	Before	After	Improvement
SNR (dB)	0.20	2.15	+1.95
SI-SDR (dB)	0.14	1.82	+1.67
PESQ	1.095	1.609	+0.514
STOI	0.770	0.851	+0.080
MOS	2.00	2.00	+0.00
RTF	-	3.0751	-

Figure 4.41: *Summary table of DDIM denoising metrics.*

Analysis: The DDIM model achieves an SNR improvement of +1.95 dB and an SI-SDR improvement of +1.67 dB. Notably, the PESQ score shows a substantial increase from 1.095 to 1.609 (+0.514), indicating better perceptual quality compared to DDPM. The STOI improves from 0.770 to 0.851 (+0.080). In terms of speed, DDIM processes the audio in 3.075 seconds, which is 3.2× faster than DDPM (9.81 seconds). The waveform and spectrogram visualizations confirm successful noise removal. The slightly

lower SNR improvement compared to DDPM is compensated by significantly faster inference and better perceptual quality (PESQ). *Note:* The relatively high processing times (9.81 seconds for DDPM, 3.07 seconds for DDIM) are due to CPU-only execution, as GPU acceleration was not available during these experiments. Under GPU acceleration (NVIDIA Tesla T4), the same models achieve RTF values below 0.2 (real-time), as demonstrated in previous sections.

Table 4.35: *Performance comparison between DDPM ($t = 8$) and DDIM ($t = 5$, steps=3) on Arabic dataset (cooking noise at 0 dB SNR, CPU execution).*

Metric	**DDPM (t=8)**	**DDIM (t=5, steps=3)**	**Advantage**
SNR Improvement (dB)	+2.30	+1.95	DDPM
SI-SDR Improvement (dB)	+1.90	+1.67	DDPM
PESQ Improvement	+0.119	+0.514	DDIM
STOI Improvement	+0.097	+0.080	DDPM
Processing Time (sec)	9.81	3.07	DDIM (3.2× faster)
RTF	9.81	3.07	DDIM

General Conclusion and Perspectives

General Conclusion

This thesis investigated the application of diffusion-based generative models for speech denoising, with a particular focus on the DiffWave architecture. Through extensive experimental evaluations across three diverse speech datasets—SC09 (English digits), Arabic Speech Commands, and LJSpeech (continuous English sentences)—several milestone objectives were achieved:

- **Model Configuration and Optimization:** We trained and evaluated two DiffWave configurations (T=50 and T=200). Despite having a higher training loss (0.0180 vs 0.0115), the T=50 configuration significantly outperformed T=200 on unseen test data, achieving an SI-SDR of 10.93 dB compared to 9.94 dB for T=200. This counter-intuitive result is explained by under-convergence of the T=200 model due to its higher computational complexity per iteration.
- **Cross-Lingual Insights and Arabic Phonetic Edge:** A key discovery of this study was the distinct resilience of the Arabic language. Due to its intrinsic high-energy pharyngeal consonants and stable vocalic core structures, the Arabic model consistently achieved higher denoising fidelity (SI-SDR = 3.94 dB, success rate = 71.2%), outperforming the English baselines (SC09: 3.37 dB, LJSpeech: 1.63 dB).
- **Cross-Lingual Generalization to French:** To further assess the model's generalization capabilities, we extended the evaluation to French speech using the Common Voice dataset. The model achieved promising results (SI-SDR = 3.22 dB), confirming its ability to generalize to unseen languages beyond Arabic. This demonstrates that the model learns language-independent acoustic representations, making it suitable for multilingual speech enhancement applications.
- **Comprehensive Noise Analysis:** The model was tested across 14 noise types (6 synthetic and 8 real-world noises) at three SNR levels (-5 dB, 0 dB, 5 dB). White noise, running tap, and doing the dishes proved easiest to suppress, while pink noise and mechanical noises (washing machine, walking machine) presented the greatest challenges.
- **Accelerated Inference via DDIM:** To address the high computational latency inherent in standard diffusion processes, the Denoising Diffusion Implicit Models (DDIM) formulation was successfully integrated. By accelerating the inference trajectory down to just 3 deterministic steps, the system achieved a real-time factor (RTF) of 0.0732, corresponding to 13.7× faster than real-time. Crucially, for continuous long-form speech (LJSpeech), this 1.68× computational acceleration incurred a negligible quality trade-off of only -0.04 dB.

In summary, this work successfully bridges the gap between sophisticated deep generative modeling and practical, real-time speech processing utility, establishing that accelerated diffusion frameworks can be efficiently deployed for cross-lingual speech enhancement without jeopardizing auditory quality.

Key Quantitative Achievements

- **Best SI-SDR (T=50 on SC09 test set):** 10.93 dB
- **Best Cross-lingual SI-SDR (Arabic dataset):** 3.94 dB
- **French Language Performance (DDPM):** SI-SDR = 3.22 dB
- **French Language Performance (DDIM):** SI-SDR = 3.03 dB, RTF = 0.0732
- **Best Performance at SNR=5 dB (Arabic):** 9.50 dB
- **Processing Speed (DDIM):** 13.7× faster than real-time (RTF = 0.0732)
- **Success Rate (Arabic dataset at SNR=0 dB):** 71.2%
- **Total Tests Executed:** 4,200 (100 files × 14 noises × 3 SNR)

Future Perspectives

While the outcomes of this work demonstrate highly promising benchmarks, several directions for future research remain:

1. **Real-Time Edge Hardware Deployment:** Present evaluations were conducted on high-performance computational environments. A natural progression would involve compiling and optimizing the 3-step DDIM model using frameworks like TensorRT or ONNX Runtime to deploy them directly onto embedded edge hardware (e.g., microcontrollers, smartphones, or hearing aids).
2. **Exploration of Flow Matching and Consistency Models:** Building upon the acceleration achieved by DDIM, newer generative methodologies such as Flow Matching or Consistency Models could be explored to potentially achieve single-step (T=1) high-quality speech synthesis.
3. **Expanding the Arabic Acoustic Corpus:** To validate the phonetic resilience observations on a broader scale, future studies should extend the training dataset from isolated Arabic commands to continuous, multi-dialectal, and conversational Arabic speech corpora.
4. **Expanding Multilingual Evaluation:** Future work should extend the evaluation to a wider range of languages (e.g., Spanish, German, Mandarin) to further validate the cross-lingual generalization capabilities of diffusion-based denoising models and investigate language-specific acoustic characteristics that influence denoising performance.
5. **Joint Multi-Task Optimization:** Integrating the speech enhancement diffusion core as a front-end for downstream tasks—such as Automatic Speech Recognition (ASR) or Speaker Identification—would evaluate how generative restoration enhances machine-learning linguistic understanding in noisy settings.

Concluding Remarks

In conclusion, this thesis has demonstrated that diffusion-based generative models, particularly the optimized T=50 configuration with DDIM acceleration, offer a powerful and efficient solution for speech denoising. The model's ability to generalize across languages and noise types, combined with its real-time processing capability, makes it a promising candidate for practical deployment in telecommunication systems, hearing aids, and voice-enabled devices.

Bibliography

[1] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.

[2] Saeed V Vaseghi. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[3] Peter Ochieng. Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis. *Artificial Intelligence Review*, 56(Suppl 3):3651–3703, 2023.

[4] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2nd edition, 2012.

[5] Cyril Plapous, Claude Marro, and Pascal Scalart. Improved signal-to-noise ratio estimation for speech enhancement. *IEEE transactions on audio, speech, and language processing*, 14(6):2098–2108, 2006.

[6] Chengshi Zheng, Huiyong Zhang, Wenzhe Liu, Xiaoxue Luo, Andong Li, Xiaodong Li, and Brian CJ Moore. Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods. *Trends in Hearing*, 27:23312165231209913, 2023.

[7] Arian Azarang and Nasser Kehtarnavaz. A review of multi-objective deep learning speech denoising methods. https://arxiv.org/abs/2003.12108, 2020. arXiv preprint.

[8] Navneet Upadhyaya and Abhijit Karmakar. Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. *Procedia Computer Science*, 54:574–584, 2015. Open access under CC BY-NC-ND license.

[9] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[10] Jonathan et al. Ho. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[11] Julius et al. Richter. Speech enhancement with diffusion-based generative models. *arXiv preprint arXiv:2208.09414*, 2022.

[12] Jiaming et al. Song. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[13] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023.

[14] Bunlong Lay, Jean-Marie Lemercier, et al. Single and few-step diffusion for generative speech enhancement. https://arxiv.org/abs/2309.09677, 2023. arXiv preprint.

[15] Rosana Ardila, Megan Branson, Kelly Davis, Josh Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. https://commonvoice.mozilla.org/, 2020.

[16] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[17] Joseph Picone. Fundamentals of speech recognition: A short course. *Institute for Signal and Information Processing, Mississippi State University*, 1996.

[18] Paul Hill. *Audio and speech processing with MATLAB*. CRC Press, 2018.

[19] Christian Benoit, Jean-Claude Martin, Catherine Pelachaud, Lambert Schomaker, and Bernhard Suhm. Audio-visual and multimodal speech systems. *Handbook of Standards and Resources for Spoken Language Systems-Supplement*, 500:1–95, 2000.

[20] Timothy Jon Quilici. *A Survey and Comparison of Digital Speech Systems*. PhD thesis, Rice University, 1976.

[21] Paul Edward Gordy. Speech spectral modeling and speech synthesis using finite impulse response digital filters, 1982. Technical report or unpublished manuscript.

[22] Deepak Baby. Non-negative sparse representations for speech enhancement and recognition, 2016. Unpublished manuscript or technical report.

[23] Jennifer Williams. *Learning disentangled speech representations*. PhD thesis, University of Edinburgh, 2022.

[24] DW Griffin. Signal estimation from modified short-time fourier transfrom. *IEEE, ASSP*, 32:2, 1984.

[25] Nan Lu. *Development of new digital signal processing procedures and applications to speech, electromyography and image processing*. The University of Liverpool (United Kingdom), 2008.

[26] Yonghyeon Jun, Beom Jun Woo, Myeonghun Jeong, and Nam Soo Kim. Snr-aligned consistent diffusion for adaptive speech enhancement. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 4073–4077. International Speech Communication Association, 2025.

[27] Shrishti Saha Shetu, Emanuël AP Habets, and Andreas Brendel. Gan-based speech enhancement for low snr using latent feature conditioning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[28] Yi Hu and Philipos C Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech communication*, 49(7-8):588–601, 2007.

[29] Simon Welker, Julius Richter, and Timo Gerkmann. Speech enhancement with score-based generative models in the complex stft domain. *arXiv preprint arXiv:2203.17004*, 2022.

[30] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] John HL Hansen and Bryan L Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *ICSLP*, volume 7, pages 2819–2822, 1998.

[32] Siwei Yu, Jianwei Ma, and Wenlong Wang. Deep learning for denoising. *Geophysics*, 84(6):V333–V350, 2019.

[33] Qiang Zhang, Yaming Zheng, Qiangqiang Yuan, Meiping Song, Haoyang Yu, and Yi Xiao. Hyperspectral image denoising: From model-driven, data-driven, to model-data-driven. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13143–13163, 2023.

[34] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. *Advances in neural information processing systems*, 26, 2013.

[35] Corneliu TC Arsene, Richard Hankins, and Hujun Yin. Deep learning models for denoising ecg signals. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[36] Xiaotong Liu, Pierre-Paul De Breuck, Linghui Wang, and Gian-Marco Rignanese. A simple denoising approach to exploit multi-fidelity data for machine learning materials properties. *npj Computational Materials*, 8(1):233, 2022.

[37] Aristomenis S. Lampropoulos and George A. Tsihrintzis. *Machine Learning Paradigms*. Springer, 2015.

[38] John Tagg. The learning paradigm. *Bolton. Anker*, 2003.

[39] Frank Emmert-Streib and Matthias Dehmer. Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5):e1470, 2022.

[40] Kamran Razzaq and Mahmood Shah. Machine learning and deep learning paradigms: From techniques to practical applications and research frontiers. *Computers*, 14(3):93, 2025.

[41] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[42] Gustav Martin Larsson. *Discovery of visual semantics by unsupervised and self-supervised representation learning*. The University of Chicago, 2017.

[43] Onintze Zaballa Larumbe. Unsupervised learning approaches for disease progression modeling. Manuscript in preparation, 2024.

[44] Ning Dong. *Generative Anomaly Detection Based on Self-Supervised Learning in Human Monitoring*. PhD thesis, Kyushu University, 2022.

[45] Mehdi Ghayoumi. *Generative adversarial networks in practice*. Chapman and Hall/CRC, 2023.

[46] Nima Rafiee. *Self-Supervised Representation learning for Anomaly Detection*. PhD thesis, Dissertation, Düsseldorf, Heinrich-Heine-Universität, 2023, 2023.

[47] OUMAIMA KHALFI. *Assessing the accuracy of emotion classification of whisper v3: comparing performance on speech emotion datasets for both English and Arabic*. PhD thesis, University of Central Florida Orlando, Florida, 2025.

[48] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[49] ITU-T Recommendation. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*, 2001.

[50] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010.

[51] ITU-T. P.800.1: Mean opinion score (mos) terminology. ITU-T Recommendation P.800.1, 2016. Available online at https://www.itu.int/rec/T-REC-P.800.1.

[52] Francois G Germain, Qifeng Chen, and Vladlen Koltun. Speech denoising with deep feature losses. *arXiv preprint arXiv:1806.10522*, 2018.

[53] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic markets*, 31(3):685–695, 2021.

[54] Koosha Sharifani and Mahyar Amini. Machine learning and deep learning: A review of methods and applications. *World Information Technology and Engineering Journal*, 10(07):3897–3904, 2023.

[55] Zeng-Xi Li, Li-Rong Dai, Yan Song, and Ian McLoughlin. A conditional generative model for speech enhancement. *Circuits, Systems, and Signal Processing*, 37(11):5005–5022, 2018.

[56] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, pages 2031–2041. PmLR, 2019.

[57] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

[58] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[59] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[60] Abdulkader Ghandoura, Farouk Hjabo, and Oumayma Al Dakkak. Building and benchmarking an arabic speech commands dataset for small-footprint keyword spotting. *Engineering Applications of Artificial Intelligence*, 102:104267, 2021.

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليـــم العالـــي والبحـــث العلمـــــي
Ministère de l'Enseignement Supérieur et de La Recherche Scientifique

كلية العلوم والتكنولوجيا
قسم الرياضيات والإعلام الآلي

Faculté des Sciences et de la Technologie

Département Des Mathématiques & de l'Informatique

جامعـــة غردايـــة

الرقم .....2026/ ق.ر.ا /ك.ع.ت/ج.غ

شهادة الترخيص بالإيداع

أنا الأستاذ(ة):

Mme. BOUCHEKOUF Asma

بصفتي رئيس (ة) والمسؤول(ة) عن تصحيح مذكرة تخرج ماستر الموسومة بـــــ

Speech Denoising Using Diffusion Model Technique

من انجاز الطالب(ة):

Laouar Meriem

والطالب(ة):

Zahouani Sarra

الكلية: العلوم والتكنولوجيا.
القسم: الرياضيات والاعلام الالي.
الشعبة: اعلام الي.
التخصص: الأنظمة الذكية لاستخراج المعارف.
تاريخ التقييم/المناقشة: 2026/06/24

أشهد أن الطالب (الطلبة) قد قام (قاموا) بالتعديلات والتصحيحات المطلوبة من طرف لجنة المناقشة وان المطابقة بين النسخة الورقية والالكترونية استوفت جميع شروطها.

امضاء المسؤول عن التصحيح

BOUCHEKOUF A.

مصادقة رئيس القسم

رئيس قسم الرياضيات والإعلام الآلي
الحاج موسى ياسين