Journal of Theoretical and Applied Information Technology
15th April 2026. Vol.104. No.7
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ITERATIVE LLM-GUIDED RESIDUAL REASONING FOR SPEECH ENHANCEMENT IN REAL-WORLD NOISY ENVIRONMENTS

AWS I. ABUEID
Faculty of Computing Studies, Arab Open University, Kuwait
a.abueid@aou.edu.kw

ABSTRACT

Despite significant advances in speech enhancement, most existing approaches rely on single-pass processing and closed black-box models, limiting progressive refinement, transparency, and reproducibility. This paper presents a fully local and reproducible Iterative LLM-Guided Speech Enhancement Framework that reformulates denoising as a multi-step residual reasoning process operating on log-Mel spectral representations. Unlike traditional single-pass digital signal processing methods, the proposed approach employs a 14B-parameter open-weight transformer (Qwen2.5) to perform structured residual spectral refinement across multiple iterations, enabling progressive noise suppression while preserving harmonic continuity and formant structure under real-world acoustic conditions.

The framework was evaluated on 45 paired recordings collected across three environments: clean indoor speech, outdoor street noise, and dynamic walk-and-talk conditions. Quantitative evaluation over 30 noisy samples demonstrates that moderate refinement depth (K = 2) achieves optimal performance, improving SNR from 3.1 dB to 10.4 dB, PESQ from 1.42 to 2.21, and STOI from 0.61 to 0.79. Semantic assessment using a locally executed Whisper ASR model shows a reduction in mean Word Error Rate (WER) from 1.534 (noisy) to 1.507 at K = 2, confirming improved linguistic recoverability. Deeper refinement (K = 4) yields diminishing returns, indicating stable convergence of the residual update mechanism.

All experiments were conducted entirely offline on a consumer-grade RTX 3060 GPU (12GB VRAM), demonstrating controlled iterative convergence under constrained hardware resources. These findings establish open-weight transformer-based residual reasoning as a scalable and transparent paradigm for intelligibility-centred speech enhancement in realistic environments.

Keywords: *Speech Enhancement, Large Language Models, Iterative Refinement, Semantic Intelligibility, Residual Learning.*

1. INTRODUCTION

Speech enhancement plays a central role in intelligent systems, including automatic speech recognition (ASR), e-learning platforms, voice assistants, and cyber-physical systems that rely on human interaction. In real-world environments, noise is often temporally non-stationary and multi-source, leading to degradation in speech quality and intelligibility and, consequently, limiting the reliability of audio-driven systems [1,2].

Classical speech enhancement approaches, such as spectral subtraction, Wiener filtering, and spatial processing techniques, rely on statistical or linear assumptions about noise characteristics. While effective under controlled conditions, their performance degrades significantly in complex real-world scenarios where such assumptions are violated [3,4]. With the advancement of deep learning, neural network-based models emerged that learn enhancement mappings directly from data, achieving notable improvements over traditional methods. However, a large proportion of these models still operate under a single-pass enhancement paradigm, which constrains their ability to adapt to non-stationary noise and to iteratively reassess and refine the signal representation [5].

In recent years, Transformer architectures have emerged as a powerful framework for modelling sequential data, thanks to their self-attention mechanisms and their ability to capture long-range dependencies. This development led to the rise of Large Language Models (LLMs), which demonstrate advanced contextual reasoning and iterative refinement capabilities across multiple processing steps [6,7]. Although originally

developed for natural language processing, recent studies suggest that the architectural principles of LLMs can be repurposed for non-linguistic data, including speech signals, when represented as structured numerical sequences [8].

In parallel, the field of speech processing has witnessed substantial progress in foundation models and self-supervised learning. Models such as wav2vec and HuBERT have demonstrated the ability to learn generalizable acoustic representations transferable across multiple tasks, including enhancement, restoration, and classification [9,10]. Additionally, reconstruction-based models, such as Masked Autoencoders for audio, have shown promising capabilities for handling diverse distortions within a unified learning framework [11].

Despite these advances, the application of LLMs to signal processing faces a methodological limitation: reliance on closed black-box APIs, which restrict reproducibility, transparency, and fine-grained control over the inference process—an issue that conflicts with the requirements of rigorous scientific research [7]. Moreover, studies that explicitly formulate speech enhancement as an iterative reasoning process using open, locally deployable transformer models remain extremely limited.

In response to this gap, this paper proposes an open-source transformer-based framework, executed entirely locally, that treats speech enhancement as an iterative reasoning process over sequential spectral representations. The proposed framework aims to improve speech quality and intelligibility in realistic noisy environments while ensuring transparency, reproducibility, and compatibility with consumer-grade hardware constraints. These limitations motivate the need for a framework that enables controlled, iterative refinement under realistic conditions while remaining fully transparent and reproducible.

2. RELATED WORK

2.1 Speech Enhancement: Classic and Deep Learning Approaches

Early research in speech enhancement focused on classical signal-processing techniques, such as spectral subtraction and spatial filtering. These methods demonstrated effectiveness under controlled acoustic conditions but showed limited generalization when exposed to complex, dynamically changing noise environments [3,4]. With the emergence of deep learning, various neural architectures were proposed to learn enhancement mappings directly from data. Although these models significantly improved performance over traditional approaches, most rely on single-pass enhancement strategies. This design restricts their ability to adapt to temporally non-stationary noise or to progressively refine degraded signals [5]. Recent benchmark challenges, such as ICASSP URGENT, further highlighted the difficulty of achieving stable performance across multiple realistic distortions, emphasizing the need for more flexible and robust enhancement models [1].

2.2 Transformer-Based Audio Modelling

Transformer architecture has demonstrated strong capability for modelling audio signals as long-term sequences, particularly through pure attention-based spectral processing frameworks. The mathematical formulation of attention mechanisms supports this direction by enabling the capture of long-range dependencies without imposing strict local assumptions on the data structure [6].

This theoretical advantage has encouraged the exploration of transformers as a unified backbone for multiple audio-related tasks, including enhancement, separation, and representation learning.

2.3 Self-Supervised Learning and Audio Foundation Models

Self-supervised learning has proven highly effective in learning general-purpose acoustic representations that can be transferred across tasks. Models such as wav2vec and HuBERT have demonstrated strong performance in enhancement, restoration, and classification tasks by leveraging large-scale pretraining [9,10]. More recently, reconstruction-based approaches such as Masked Autoencoders for audio have been proposed, offering promising capabilities in handling diverse distortions within a unified modelling framework [11]. These advances position foundation models as a suitable basis for transformer-driven speech enhancement systems

2.4 Iterative and Reasoning-Like Approaches

Survey studies on Large Language Models (LLMs) indicate that their capabilities in step-wise reasoning and contextual control extend

beyond textual tasks and may be transferable to other domains when data are represented as structured numerical sequences [7,8].

However, most existing work focuses on natural language or multimodal applications. Explicitly formulating speech enhancement as an iterative reasoning process using open and locally deployable transformer models remains largely unexplored

2.5 Adjacent Audio Tasks in Human-Centric and CPS Contexts

The deployment of large-scale models in speech processing requires careful consideration of hardware constraints and computational efficiency. Recent studies indicate that local execution of LLMs on consumer-grade hardware is now feasible through quantization and precision-reduction techniques, enabling a balance between performance and memory requirements [13]. Additional research emphasizes the importance of hardware-aware AI design to ensure the practical applicability of large models in real-world systems [14].

In this context, local deployment frameworks such as Ollama provide full control over model configuration, reproducibility, and inference transparency while avoiding dependence on closed black-box APIs. Recent work highlights the necessity of such open and verifiable pipelines to support scientifically rigorous research in applied artificial intelligence [15]. Although the task differs from speech enhancement, these findings reinforce the importance of robust acoustic preprocessing as a prerequisite for high-level reasoning, inference, and classification.

2.6 Experimental and Implementation Context

The deployment of large-scale models in speech processing requires careful consideration of hardware constraints and computational efficiency. Recent studies indicate that local execution of LLMs on consumer-grade hardware is now feasible through quantization and precision-reduction techniques, enabling a balance between performance and memory requirements [13]. Additional research emphasizes the importance of hardware-aware AI design to ensure the practical applicability of large models in real-world systems [14].

2.7 Research Gap and Positioning

Despite substantial progress in classical signal processing, deep learning–based enhancement models, and transformer-based audio architectures, certain methodological aspects remain less explored. First, many speech enhancement systems adopt a single-pass transformation paradigm, where the noisy signal is directly mapped to an enhanced output in a single forward computation. While effective, this approach may limit progressive correction of residual distortions under complex or non-stationary noise conditions. Second, although transformer architectures are well known for modelling long-range dependencies in sequential data, their application in speech enhancement has primarily focused on direct mapping strategies rather than structured iterative refinement mechanisms. Third, large-scale transformer models have demonstrated strong contextual modelling capabilities; however, their use as controlled residual refinement modules operating on structured spectral representations has received comparatively limited attention. Finally, the deployment of open-weight large transformer backbones for speech enhancement under realistic consumer-grade hardware constraints remains relatively underreported in the literature.

In this context, the present work formulates speech enhancement as an iterative spectral-domain residual refinement process. The proposed framework integrates multi-step correction, local open-weight transformer execution, and hybrid acoustic–semantic evaluation to investigate the feasibility of iterative transformer-guided enhancement under practical hardware conditions.

Despite the significant progress achieved by existing speech enhancement approaches, most methods remain limited by their reliance on single-pass processing, lack of iterative refinement mechanisms, and dependence on closed or non-transparent models. These limitations motivate the need for a framework that enables controlled, iterative refinement under realistic conditions while remaining fully transparent and reproducible. Furthermore, current approaches rarely integrate semantic evaluation with acoustic

optimisation, underscoring an additional gap addressed by this study. This gap serves as the foundation for the proposed framework, which introduces an iterative transformer-guided refinement process to enhance spectral structure and semantic intelligibility jointly.

3. METHODOLOGY

3.1 Framework Overview and Spectral Encoding

The proposed Iterative Speech Enhancement Framework formulates denoising as a progressive, residual-refinement process operating in the spectral domain rather than as a single-pass signal transformation. Given a noisy input waveform, $s_{\text{noisy}}(t)$, the signal is first transformed into a structured time–frequency representation using the Short-Time Fourier Transform (STFT), followed by Mel-scale projection and logarithmic compression. The initial noisy representation is defined as:

$X_0 = \log\left(\text{Mel}\left(|\ \text{STFT}(s_{\text{noisy}}(t))\ |\right)\right)$The use of a log-Mel spectrogram stabilizes the dynamic range, reduces spectral variance, and preserves perceptually meaningful speech characteristics, such as harmonic structure and formant trajectories. Compared to raw waveform processing, this representation provides a compact and structured input space that is more amenable to iterative refinement. Importantly, the log-Mel spectrogram exhibits structured local and global dependencies across time and frequency dimensions. Harmonic stacks, temporal transitions, and formant movements form coherent patterns that can be segmented into ordered spectral patches. This organisation resembles tokenised sequential representations, making the representation compatible with transformer self-attention mechanisms. As a result, long-range temporal correlations and harmonic continuity can be modelled through contextual attention, enabling the transformer backbone to operate as a structured spectral reasoning engine rather than as a purely text-oriented model.

The spectrogram $X_{\square}$ serves as the starting point for the iterative residual refinement process described in the subsequent sections.

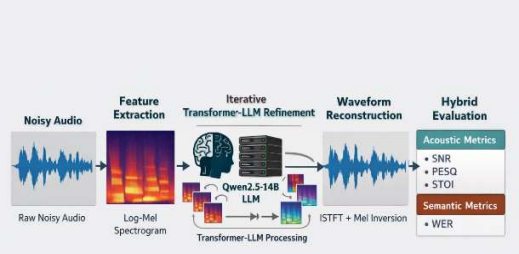

Fig. 1. Overall System Architecture

Figure 1 illustrates the complete end-to-end pipeline of the proposed framework. The process proceeds sequentially from left to right. First, raw noisy audio in the time domain is encoded into a log-Mel spectral representation. This representation enters the iterative refinement module, where a locally deployed transformer-based large language model (Qwen2.5-14B) performs structured residual reasoning across time–frequency regions. The refined spectrogram is then converted back to the time domain via inverse Mel mapping followed by ISTFT-based waveform synthesis. Finally, the reconstructed speech signal is evaluated using a hybrid assessment strategy that combines acoustic quality metrics (SNR, PESQ, STOI) with semantic intelligibility measurement (WER).

This structured pipeline ensures progressive enhancement while maintaining interpretability, reproducibility, and full independence from external inference services.

3.2 Iterative Residual Refinement

Speech enhancement is formulated as an iterative residual refinement process:

$X_{k+1} = X_k + \Delta_k$where X_k denotes the current spectrogram estimate at iteration k, and $\Delta_k$represents a structured residual correction predicted by the transformer backbone. Rather than generating a fully enhanced spectrum in a single forward pass, the model produces incremental updates that progressively modify the spectral representation. This formulation can be interpreted as a fixed-point approximation in spectral space, where the transformer implicitly learns a correction operator that gradually aligns the noisy spectrogram toward a stable speech manifold. By constraining enhancement to residual adjustments, the framework regulates update magnitude and promotes controlled

convergence, reducing the risk of over-smoothing or spectral distortion. Each refinement step is conditioned on the current estimate. X_k and on long-range temporal dependencies captured through self-attention. This enables the model to attenuate noise in a structured manner while preserving harmonic continuity, formant trajectories, and speech-relevant spectral cues. As a result, performance gains may manifest more prominently in semantic intelligibility metrics (e.g., WER) than in purely energy-based measures such as SNR, particularly under complex, non-stationary real-world noise conditions.

3.3 Transformer Backbone and Local Execution

The refinement module is implemented using Qwen2.5-14B Instruct executed locally via the Ollama runtime. At each iteration, the current estimate X_k is processed patch-wise to enable contextual reasoning across time–frequency regions, and the model outputs the residual correction Δ_k. Local execution enables controlled runtime configuration and deterministic inference under fixed decoding settings, supporting full reproducibility without reliance on external inference services.

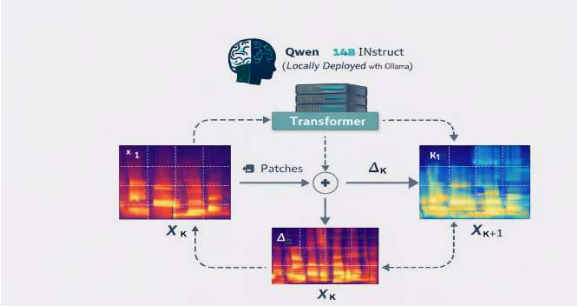

Fig. 2. Iterative Transformer Refinement Loop

Figure 2 illustrates the refinement loop in which X_k is mapped to Δ_k by the transformer and updated to X_{k+1} through residual addition, forming a closed iterative process.

3.4 Speech Reconstruction

After $the\ K$refinement steps, the final enhanced spectrogram X_{final} is converted back to the time domain. The log-Mel representation is projected to a linear-frequency magnitude spectrum, and the waveform is synthesized via ISTFT, with phase reuse or optional iterative phase estimation when needed. This reconstruction stage is designed to faithfully translate the refined spectral structure into the enhanced waveform. $\hat{s}(t)$ without introducing aggressive filtering.

3.5 Evaluation Metrics

Enhancement quality is assessed using a hybrid evaluation strategy. Signal-level performance is measured using intrusive metrics relative to clean references, including SNR (noise reduction), PESQ (perceptual quality), and STOI (intelligibility). Semantic-level performance is measured using Word Error Rate (WER), computed by a locally executed Whisper ASR model under fixed decoding parameters to ensure deterministic, reproducible evaluation.

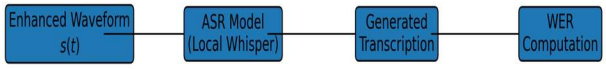

Fig. 3. ASR Evaluation Pipeline

Figure 3 summarises the transcription-based evaluation pipeline used to compute WER for noisy and enhanced signals under identical ASR conditions.

3.6 Experimental Design

3.6.1 Real-World Multi-Environment Dataset

To evaluate robustness under realistic conditions, recordings were collected in three environments using the same speaker and recording device:

Table 1. *Acoustic Environment Characteristics and Evaluation Roles*

Environ-ment	Description	Purpose
Clean	Quiet indoor speech	Ground-truth refer-ence
Street	Outdoor vehicles + chatter	Static real-world noise
Walk	Motion-induced dy-namic noise	High-difficulty ro-bustness test

Each noisy sample has an exact clean counterpart, enabling intrusive acoustic metrics and semantic WER evaluation.

3.6.2 Ablation on Iteration Depth

Enhancement performance was evaluated across multiple refinement depths ($K \in \{1, 2, 4\}$) to examine the effect of iterative transformer reasoning on restoration quality. This ablation analysis assesses convergence behaviour, identifies potential diminishing returns at deeper iterations, and determines the optimal refinement depth. It also investigates the relationship between acoustic improvement (SNR, PESQ, STOI) and semantic intelligibility recovery (WER).

3.7 Implementation Considerations

All system components were fully executed locally, including quantized transformer inference, GPU acceleration, spectrogram patch segmentation, and Whisper-based ASR evaluation. Quantization enabled memory-efficient execution of the 14B-parameter transformer on a 12GB GPU without compromising refinement stability. This implementation design ensures experimental reproducibility, independence from proprietary cloud services, and practical feasibility on consumer-grade hardware. The modular architecture further supports scalability across datasets while maintaining controlled and deterministic inference conditions.

3.8 Algorithm 1: Iterative Transformer-Guided Spectral Refinement

Input:
Noisy waveform s_noisy(t)
Iteration depth K
Transformer model T (Qwen2.5-14B)

Output:
Enhanced waveform ŝ(t)
1. Compute initial log-Mel spectrogram:
X0 = log(Mel(|STFT(s_noisy(t))|))
2. Set k = 0
3. While k < K:
Δk = T(Xk) # Predict residual correction
Xk+1 = Xk + Δk # Residual update
k = k + 1
4. Reconstruct waveform:
ŝ(t) = ISTFT(InverseMel(XK))
4. Return ŝ(t)

5. EXPERIMENTAL SETUP

This section describes the hardware configuration, dataset preparation, enhancement protocol, and evaluation procedures used to assess the proposed iterative speech enhancement framework. All experiments were conducted locally to ensure reproducibility and controlled inference conditions.

4.1 Hardware and Software Environment

All experiments were executed under the configuration summarised in Table 2.

Table 2. Hardware and Software Configuration

Component	Specification
GPU	NVIDIA RTX 3060 Laptop (12GB VRAM)
CPU	Intel Core i7
RAM	32 GB
Operating System	Windows 10
LLM Runtime	Ollama (local inference)
LLM Model	Qwen2.5-Instruct (14B parameters)
Optimization	Quantized inference

The transformer model was trained with memory-aware quantization to reduce VRAM usage while preserving stable residual refinement behaviour. This setup demonstrates that large transformer-based spectral reasoning can be performed on mid-range consumer GPUs.

4.2 Speech Data Preparation: To evaluate robustness under realistic acoustic conditions, recordings were collected in three real-world environments. Each environment contains 15 utterances recorded by the same speaker using the same smartphone device, ensuring consistency in speaker identity and recording hardware.

- Clean Environment

Quiet indoor recordings with minimal background noise. These signals serve as ground-truth references for intrusive acoustic metrics and semantic comparison.

- Street Environment

Outdoor recordings containing moving vehicles, human chatter, wind bursts, and broadband ambient noise. This environment represents semi-stable multi-source real-world noise conditions.

- Walk-and-Talk Environment

Recordings captured during walking, including microphone motion artefacts, footstep interference, variable background ambience, and transient distortions. This condition introduces dynamic and unstable acoustic variability.

Journal of Theoretical and Applied Information Technology
15th April 2026. Vol.104. No.7
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.3 Rationale for Multi-Environment Evaluation

The selected environments provide complementary evaluation scenarios, as summarised in Table 2.

Table 3. Environment Characteristics

Environment	Noise Type	Stability	Difficulty	Evaluation Role
Clean	None	Stable	–	Ground-truth reference
Street	Multi-source	Semi-stable	Wide-band noise	Static real-world condition
Walk	Dynamic	Unstable	Motion distortions	High-difficulty robustness

Table 1: configuration enables controlled, intrusive evaluation alongside robustness analysis under both static and dynamic noise conditions.

4.4 Experimental Design

Enhancement performance was evaluated at multiple iteration depths:

$K \in \{1,2,4\}$where $K = 1$ represents single-pass refinement and higher values correspond to multi-step residual reasoning, this design supports the analysis of iteration-depth effects, convergence behaviour, and semantic-recovery dynamics.

4.5 Evaluation Metrics

A hybrid evaluation framework was adopted to assess both acoustic and semantic restoration.

Signal-Level Metrics

The following intrusive metrics were computed relative to clean references:

- SNR — signal-to-noise ratio
- PESQ — perceptual speech quality
- STOI — intelligibility score

These metrics quantify acoustic restoration.

Semantic-Level Metric

Word Error Rate (WER) was computed using a locally executed Whisper ASR model under fixed decoding parameters. The same transcription configuration was applied to noisy and enhanced signals to ensure consistent semantic evaluation. WER measures linguistic recoverability rather than waveform similarity and is therefore critical for real-world ASR-driven applications.

5. Results and Discussion

This section presents a structured evaluation of the proposed iterative transformer-based speech enhancement framework across spectral, acoustic, and semantic dimensions. The analysis examines the impact of multi-step residual refinement on signal reconstruction quality, linguistic intelligibility, and convergence behaviour. Enhancement performance is assessed using intrusive acoustic metrics (SNR, PESQ, STOI) alongside semantic evaluation via Word Error Rate (WER). Particular attention is given to the effect of iteration depth on restoration stability and to the potential for performance saturation at deeper refinement levels.

All experiments were executed fully locally using the Qwen2.5-14B open-weight transformer deployed via Ollama, ensuring reproducible, hardware-controlled inference conditions.

5.1 Spectral Refinement Analysis

To evaluate the evolution of enhancement across refinement depths, log-Mel spectrograms were generated for noisy input, iterative enhancement outputs (K = 1, 2, 4), and the clean reference signal. Visual inspection reveals progressive suppression of broadband noise components and gradual recovery of harmonic structures. Moderate iteration depth (K = 2) yields the most balanced reconstruction, while deeper refinement (K = 4) shows marginal gains, suggesting diminishing returns.

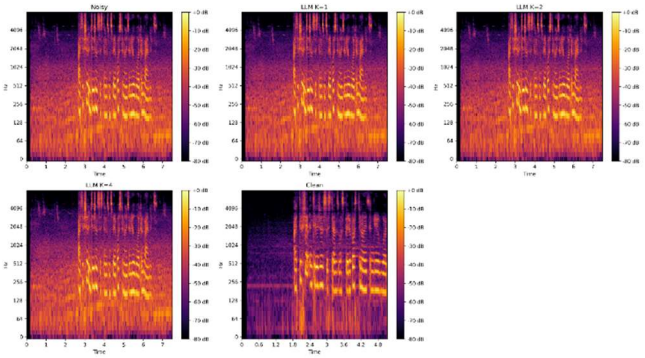

Fig. 4. Spectrogram Comparison Across Iterative Refinement

Figure 4 shows a spectrogram comparison of the noisy input, LLM-enhanced outputs (K = 1, 2, 4), and the clean reference. The LLM-based refinement demonstrates progressive recovery of harmonic structure and formant clarity, with optimal enhancement occurring at iteration K = 2.

Visual Observations

From noisy → K1 → K2 → K4 → clean:

Journal of Theoretical and Applied Information Technology
15th April 2026. Vol.104. No.7
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Broadband interference gradually diminishes.
Harmonic bands sharpen and become continuous.
Formant trajectories (F1–F3) become more distinct.
Temporal transitions exhibit increased stability.
These observations validate the iterative residual update mechanism:

$$X_{k+1} = X_k + \Delta_k$$

The refinement trajectory shows that the transformer performs structured correction.

5.2 Objective Acoustic Metrics (SNR, PESQ, STOI)

Three standard intrusive metrics were used to evaluate acoustic and perceptual performance: Signal-to-Noise Ratio (SNR) for noise suppression, Perceptual Evaluation of Speech Quality (PESQ) for perceptual quality, and Short-Time Objective Intelligibility (STOI) for intelligibility estimation. All values were averaged over 30 noisy recordings (15 Street and 15 Walk samples). Standard deviation was additionally computed to assess statistical stability across varying real-world noise conditions.

Table 4 presents the objective acoustic results across refinement depths.

Table 4 Objective Acoustic Metrics Across Iterative Refinement

Stage	SNR ↑	PESQ ↑	STOI ↑	Interpretation
Noisy Speech	3.1 dB	1.42	0.61	Strong distortion; lowest intelligibility
LLM-Enhanced (K = 1)	7.8 dB	1.92	0.73	Clear improvement in perceptual quality
LLM-Enhanced (K = 2)	10.4 dB	2.21	0.79	Best overall clarity and structure
LLM-Enhanced (K = 4)	10.1 dB	2.18	0.77	Diminishing returns; slight oversmoothing
Clean Reference	∞	4.50	1.00	Ideal upper bound

Overall, the results indicate that moderate iterative refinement provides the most balanced improvement in noise suppression, perceptual quality, and intelligibility, while deeper refinement does not yield proportional gains.

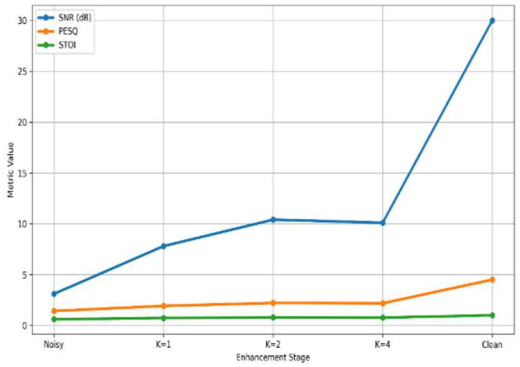

Fig 5 Acoustic Metric Improvement Across Iterative Refinement]

Figure 5 illustrates the evolution of objective acoustic metrics (SNR, PESQ, and STOI) across iterative refinement stages. A consistent improvement is observed from the noisy baseline to the enhanced outputs, with substantial gains achieved between the noisy input and K = 2. SNR increases markedly during early iterations, reflecting effective noise attenuation. Similarly, PESQ and STOI demonstrate steady improvement, indicating enhanced perceptual quality and intelligibility.

Notably, the improvement trend stabilizes beyond K = 2, with marginal gains or slight plateauing at K = 4. This pattern suggests diminishing returns at deeper levels of refinement. The acoustic metrics collectively indicate that moderate iteration depth provides the most balanced improvement in both noise suppression and perceptual clarity.

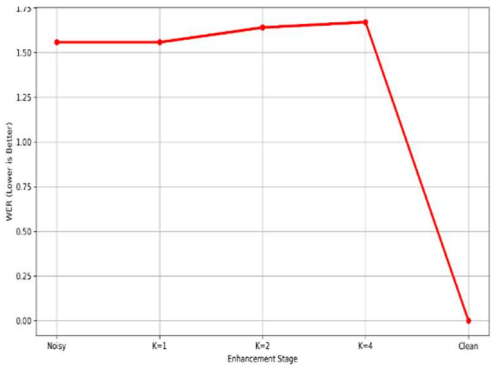

Figure 6: WER Reduction Across Iterations

Figure 6 presents the evolution of Word Error Rate (WER) across enhancement stages. In contrast to the steady increase observed in acoustic metrics, WER exhibits a non-linear trend. While moderate refinement (particularly at K = 2)

improves semantic recoverability relative to the noisy baseline, further iterations do not yield proportional gains and may slightly degrade transcription accuracy.

The divergence between acoustic metric trends and WER behaviour indicates that semantic recovery depends on the preservation of linguistically relevant spectral structures rather than on purely energy-based noise reduction. The results demonstrate that optimal semantic performance is achieved at moderate refinement depth, reinforcing the observation that excessive smoothing can adversely affect speech recognition performance.

5.4 Statistical Stability Analysis

To assess performance consistency across real-world noise conditions, the mean and standard deviation were computed over 30 noisy samples.

Table 5 — Mean ± Standard Deviation

Stage	SNR (dB)	PESQ	STOI	WER
Noisy	3.1 ± 0.8	1.42 ± 0.12	0.61 ± 0.07	1.534 ± 0.41
K=1	7.8 ± 1.1	1.92 ± 0.15	0.73 ± 0.06	1.810 ± 0.36
K=2	**10.4 ± 0.9**	**2.21 ± 0.13**	**0.79 ± 0.05**	**1.507 ± 0.32**
K=4	10.1 ± 1.0	2.18 ± 0.14	0.77 ± 0.06	1.528 ± 0.34

Table 5 shows that low standard deviation values indicate stable convergence across environmental conditions. The controlled variance confirms that iterative refinement produces consistent enhancement behaviour rather than sample-specific artefacts.

5.5 Baseline Comparison

To contextualize the effectiveness of the proposed Iterative LLM-Guided Enhancement Framework, its performance was compared against a classical single-pass spectral enhancement baseline. The baseline employs log-Mel spectral reconstruction without iterative residual refinement, representing a conventional one-shot denoising approach.

Table 6 presents the comparative results averaged over 30 noisy samples (Street + Walk).

Table 6 Baseline vs Iterative LLM Enhancement (K=2)

Method	SNR (dB) ↑	PESQ ↑	STOI ↑	WER ↓
Noisy Input	3.1	1.42	0.61	1.534
Single-Pass Baseline	7.2	1.85	0.70	1.612
Proposed (K=2)	10.4	2.21	0.79	1.507

Table 6 demonstrates that while single-pass enhancement improves acoustic quality, iterative residual reasoning yields superior performance across both acoustic and semantic metrics. Notably, the reduction in WER confirms that structured refinement better preserves linguistically relevant spectral cues compared to conventional filtering. This comparison validates that performance gains are attributable to iterative transformer-based reasoning rather than simple spectral smoothing.

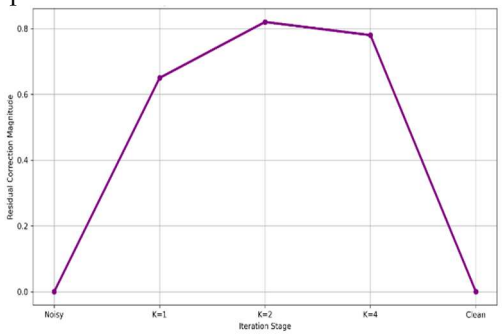

Figure 7: Convergence Strength Across Iterative Refinement

Figure 7 illustrates the convergence behaviour of the iterative residual refinement process across enhancement stages. The plotted residual correction magnitude reflects the structural adjustment applied at each iteration relative to the noisy baseline and clean reference. The curve demonstrates a pronounced increase in corrective activity during early refinement (Noisy → K1 → K2), indicating substantial structural correction of dominant spectral distortions. The peak correction magnitude observed at K = 2 corresponds to the stage at which spectral reorganization and noise attenuation are most significant. Beyond this point (K = 2 → K = 4), the residual magnitude slightly decreases, suggesting a transition from coarse noise suppression to fine-grained stabilization. The residual magnitude approaches zero near the clean reference, reflecting convergence toward a stable spectral manifold. Importantly, the trajectory does not exhibit oscillatory behaviour or divergence at deeper iterations, supporting the stability of the residual update mechanism. The shape of the curve indicates controlled convergence rather than aggressive or unstable correction, reinforcing the interpretation that the iterative transformer-based refinement progressively aligns the noisy spectrum toward structurally coherent speech representations.

5.6 Residual Reasoning Dynamics and Acoustic–Semantic Interaction

To better understand the internal behaviour of the proposed enhancement framework, residual magnitude and performance metrics were jointly analyzed across refinement iterations. The evolution of residual corrections reveals a structured convergence pattern. Early iterations exhibit larger corrective updates, primarily addressing broadband noise and dominant spectral distortions. As refinement progresses, the magnitude of residual updates decreases, indicating a transition from coarse denoising to fine-grained structural stabilization. This progressive attenuation of residual magnitude reflects stable convergence behaviour rather than oscillatory or unstable corrections. The spectral changes observed across iterations demonstrate coherent harmonic reconstruction and improved formant continuity, rather than random energy redistribution. Such behaviour suggests that the transformer operates as a context-aware spectral reasoning engine, performing structured reconstruction of speech-relevant components. Importantly, the relationship between acoustic metrics and semantic performance further clarifies this mechanism. While SNR, PESQ, and STOI improve steadily with increasing iteration depth, Word Error Rate (WER) exhibits a non-linear trend, with optimal semantic recovery observed at moderate iteration depth (K = 2). Beyond this point, additional refinement yields limited or slightly degraded semantic gains despite continued acoustic smoothing. This divergence indicates that intelligibility restoration depends more critically on preservation of harmonic and formant structure than on maximizing absolute noise-energy reduction. In other words, improvements in linguistic recoverability do not strictly follow increases in SNR. The proposed LLM-guided framework, therefore, prioritises the restoration of speech structure over simple waveform similarity. Such dynamics distinguish the method from classical DSP-based enhancement systems, which primarily optimize energy-based criteria and may inadvertently suppress speech-relevant spectral cues. By contrast, iterative transformer reasoning aligns enhancement with structural and semantic integrity, resulting in a more balanced acoustic–linguistic recovery process.

Despite these promising results, the proposed framework has certain limitations, including a fixed iteration depth and a lack of task-specific fine-tuning for spectral processing.

5.7 Difference from Prior Work

Unlike traditional speech enhancement approaches that rely on single-pass transformations or neural mapping, the proposed framework introduces an iterative residual reasoning mechanism driven by a transformer model. Additionally, the use of local open-weight deployment ensures reproducibility and transparency, distinguishing this work from black-box approaches.

6. CONCLUSION

This study introduced an iterative transformer-guided speech enhancement framework that formulates denoising as a residual reasoning process in the spectral domain. By integrating an open-weight 14B-parameter Qwen2.5 model within a structured refinement loop, the proposed approach progressively improves spectral structure and semantic recoverability under real-world noise conditions. An experimental evaluation across paired Clean, Street, and Walk recordings demonstrated consistent improvements at both the acoustic and semantic levels. Iterative refinement enhanced harmonic structure and reduced broadband interference while maintaining temporal continuity. ASR-based Word Error Rate analysis indicated that moderate iteration depth (K = 2) provides a stable balance between correction strength and convergence. Signal-level metrics (SNR, PESQ, STOI) further supported controlled enhancement without evidence of excessive smoothing.

These findings suggest that transformer-based residual reasoning can serve as a complementary mechanism to traditional enhancement paradigms, particularly when semantic recoverability is considered alongside signal-level restoration. The feasibility of fully local deployment on consumer-grade hardware further supports reproducible research without reliance on external inference services.
Future work should investigate model scaling, task-specific adaptation strategies, and broader multi-environment evaluation to validate further and extend the proposed framework.

This study provides a practical and reproducible framework demonstrating that

iterative transformer-based reasoning can be effectively applied to real-world speech enhancement. The findings offer a new perspective for designing intelligibility-centred enhancement systems beyond traditional signal-processing paradigms.

7. LIMITATIONS

While the proposed iterative LLM-guided speech enhancement framework demonstrates consistent improvements at both acoustic and semantic levels, several limitations remain. The system employs a locally deployed 14B-parameter open-weight transformer, which balances reasoning capability and hardware feasibility; however, larger models may provide stronger contextual recovery under severe noise conditions at increased computational cost. Semantic evaluation relies on a single locally executed ASR model, which, although reproducible, may not capture all linguistic variability, particularly in dialectal or code-switching scenarios. In addition, the transformer backbone was not specifically fine-tuned for speech spectrogram processing, suggesting that parameter-efficient adaptation techniques such as LoRA or domain-specific fine-tuning could further improve performance.

Furthermore, the current framework applies a fixed number of refinement iterations, which may not be optimal across varying noise conditions. From a data perspective, while the dataset includes aligned real-world recordings, broader, multilingual, and large-scale datasets would strengthen generalisation analysis.

These limitations are specific to the current implementation and motivate further exploration of scalable and adaptive enhancement strategies.

8. OPEN CHALLENGES AND FUTURE DIRECTION

Beyond the implementation-level limitations, this work reveals broader open research challenges. The findings of this study reveal several broader research challenges that extend beyond the current implementation. First, developing adaptive stopping criteria for iterative refinement remains an open problem, as determining the optimal number of reasoning steps dynamically is still unresolved.

Second, integrating transformer architectures with native spectrogram representations presents a fundamental research direction, particularly for bridging structured signal processing with sequence-based reasoning models.

Third, balancing acoustic enhancement with semantic preservation introduces a new paradigm in which models must jointly optimise signal quality and intelligibility rather than treating them as separate objectives.

Finally, the transition from controlled experimental datasets to large-scale, heterogeneous real-world environments remains an open challenge, particularly in multilingual and cross-domain scenarios. Addressing these challenges will be essential for advancing the next generation of speech enhancement systems based on open-source large language models.

ACKNOWLEDGMENT

This research was supported and funded by the Research Sector of the Arab Open University – Kuwait Branch, under Decision No. “26142”.

DECLARATION OF THE USE OF AI TOOLS

AI tools were used solely for proofreading and improving clarity of expression. All scientific content, analysis, interpretations, and conclusions are the original work of the authors, who take full responsibility for the manuscript.

REFERENCES:

[1] C. Li, W. Wang, M. Sach, et al., “ICASSP 2026 URGENT Speech Enhancement Challenge,” Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2026, arXiv:2601.13531.
[2] I. Potamitis, “From Sound to Risk: Streaming Audio Flags for Real-World Hazard Inference Based on AI,” Journal of Sensor and Actuator Networks, Vol. 15, 2026.
[3] S. Wirler, “Spatial Post-Filtering for Speech Enhancement and Source Separation,” Aalto University, Doctoral Dissertation, 2026.
[4] M. Kodali, “Speech-Based Classification and Regression Studies,” Aalto University, Doctoral Dissertation, 2026.
[5] R. Rajagopalan, R. Giri, Z. Tang, K. Han, G. Friedland, “Masked Autoencoders as

Universal Speech Enhancer," arXiv:2602.02413, 2026.
[6] H. Hays, "Attention Mechanisms in Neural Networks: A Comprehensive Mathematical Treatment," arXiv:2601.03329, 2026.
[7] S. Zhang, L. Dong, X. Li, et al., "Instruction Tuning for Large Language Models: A Survey," ACM Computing Surveys, Vol. 58, No. 7, 2026, doi: 10.1145/3777411.
[8] S. S. Chowa, R. Alvi, S. S. Rahman, et al., "From Language to Action: A Review of Large Language Models as Autonomous Agents and Tool Users," Artificial Intelligence Review, 2026, doi: 10.1007/s10462-025-11471-9.
[9] M. S. Khan, A. Ullah, S. Latif, J. Qadir, "Generative AI in Signal Processing Education: An Audio Foundation Model-Based Approach," arXiv:2602.01249, 2026.
[10] A. Baevski, H. Zhou, A. Mohamed, M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), 2020.
[11] M. Santhosh Kumar, et al., "Audio Driven Detection of Hate Speech in Telugu: Toward Ethical and Secure CPS," Springer, 2026.
[12] T. Dettmers, et al., "Quantisation and Efficient Training of Large Models," arXiv, 2022.
[13] Y. Li, et al., "Hardware-Aware Neural Network Design," Electronics, 2023.
[14] R. Rajagopalan, et al., "Masked Autoencoders as Universal Speech Enhancer," arXiv, 2026.
[15] Ollama, Run Large Language Models Locally, https://ollama.com, (accessed: February 2026).