International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 4, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n4.07
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Improvement of Speech-Paraformer Large ASR for Industrial Voice Control in High-Noise Environments

Vasileva Mariya *, Hongchang Sun *, Yongxiang Jiang

School of Mechanical Engineering, Tianjin University of Technology and Education, Tianjin, 300222, China

ABSTRACT

This study systematically evaluates the robustness of the Speech-Paraformer-Large automatic speech recognition (ASR) model under simulated industrial noise and proposes an effective post-processing enhancement strategy for safety-critical voice-controlled human-robot interaction in manufacturing environments. the controlled experiment used a dataset of 10 Mandarin Chinese industrial commands recorded in clean conditions (16 kHz, 16-bit PCM). Noisy test conditions were generated by mixing clean recordings with continuous white noise and authentic industrial machinery noise at Signal-to-Noise Ratios (SNR) from 20 dB to -10 dB (5 dB increments). The pre-trained Speech-Paraformer-Large model was evaluated, and a text-based verification layer with three hierarchical matching strategies (fuzzy exact matching, substring containment, sliding window similarity) was implemented as post-processing; performance was assessed via Word Error Rate (WER) and accuracy across 50 test utterances per condition. Results show that industrial machinery noise is significantly more detrimental to ASR performance than white noise (24% vs. 70% accuracy at -10 dB SNR). The proposed verification layer consistently improved performance across all SNR levels: accuracy increased by 8% (88% to 96%) under 0 dB white noise and by 10 percentage points (24% to 34%, 41.6% relative improvement) under -10 dB industrial noise. It also reduced substitution errors by 34%, insertion errors by 31%, and total errors by 32%, with unexpected efficiency gains (51.2% reduction in computation time at 0 dB industrial noise). This study demonstrates that intelligent post-processing can achieve practical, deployable robustness gains without model retraining or acoustic preprocessing, and the proposed text-based verification layer provides a cost-effective solution to improve voice control reliability in industrial environments, with direct implications for manufacturing safety and efficiency.

KEYWORDS

Speech-Paraformer-Large; FunASR; Industrial Noise Robustness; Post-processing; SNR Degradation; White Noise Comparison

1. INTRODUCTION

1.1. Background, Challenges and Literature Review

The evolution of human-robot interaction (HRI) in industry demands interfaces that are both intuitive and robust. Voice control, powered by modern Automatic Speech Recognition (ASR), offers a promising path by leveraging natural speech for commanding and programming collaborative robots [1]. Unlike traditional methods like teach pendants, voice-driven interaction reduces cognitive load and adapts to non-repetitive tasks [2]. The core challenge for its industrial adoption is achieving reliable performance in inherently noisy, non-stationary acoustic environments typical of manufacturing floors [3-4].

Recent advances in end-to-end neural ASR architectures have significantly pushed the boundaries of accuracy and noise robustness [5]. Among these, the Paraformer (Parallel Transformer) model represents a state-of-the-art approach, featuring efficient parallel computation and mechanisms like Contextual Integral Feedforward (CIF) for accurate prediction of token boundaries [6]. Further advances in non-autoregressive Transformer architectures have been proposed to improve inference efficiency while maintaining accuracy [7]. The Speech-Paraformer-Large model, a large-scale implementation of this architecture, has demonstrated strong performance on public benchmarks [8]. However, its empirical performance under systematically varied, realistic industrial noise profiles - especially under extreme signal degradation - remains underexplored in the literature. Most academic evaluations focus on SNR levels down to 0 dB [9], leaving a gap in understanding model behavior in near-unintelligible conditions that may occur during transient noise events (e.g., impacts, alarms).

Fundamental principles of speech processing highlight that recognition accuracy is bounded by the Signal-to-Noise Ratio (SNR) and the character of the noise masker. Performance degradation follows a non-linear trend, where a critical SNR threshold often exists beyond which errors become catastrophic [10]. This study situates itself at this intersection: applying a principled, signal-processing-based experimental framework to stress-test a specific, modern ASR model (Speech-Paraformer-Large).

1.2. Practical Applications and Broader Scientific Relevance

Robust ASR is crucial for applications beyond manufacturing, including healthcare documentation, hands-free automotive systems, and assistive technologies. In industrial automation, reliable voice control can minimize "hands-busy, eyes-busy" limitations and lower programming barriers. The findings from systematically testing a model like Paraformer under controlled noise are directly applicable to system integrators who must decide on microphone placement, acoustic shielding, and operational protocols [11].

From a scientific perspective, this work contributes to the growing body of research on the robustness of deep learning-based ASR [12]. By analyzing error patterns (substitutions, deletions, insertions) across a controlled SNR gradient and noise types, we can infer which acoustic challenges are most problematic for the Paraformer architecture. This knowledge is valuable for guiding future model improvements, data augmentation strategies [13], and the development of hybrid systems that combine ASR with deterministic verification logic is a direction we explore in our methodology [14].

2. METHODOLOGY

The objective of this work is to obtain a comprehensive evaluation of the Speech-Paraformer-Large model's performance in simulated industrial noise and to propose a mitigation strategy for its failure modes. A controlled experiment was designed using a dataset of Chinese voice commands mixed with authentic factory noise at precisely calibrated SNR levels.

Figure 1 presents the complete system architecture, comprising the ASR pipeline and the proposed post-processing verification layer.

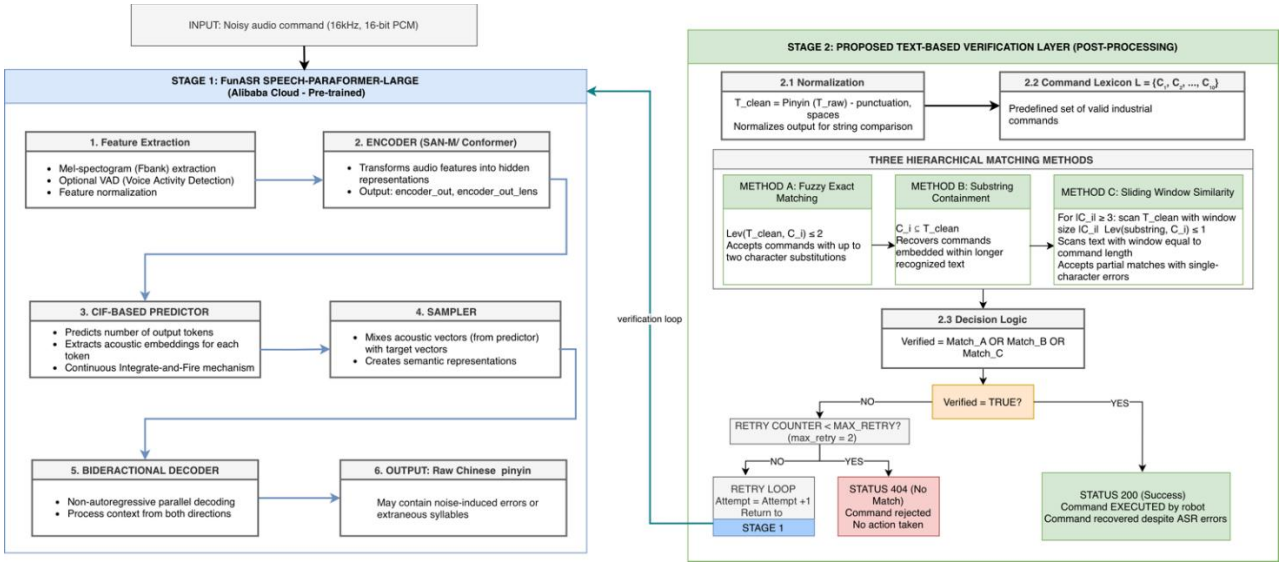

Figure 1. Complete system architecture: Speech-Paraformer-Large ASR pipeline (blue) and proposed text-based verification layer (green)

2.1. The ASR Model: Speech-Paraformer-Large

The core system under investigation in this study is the FunASR Speech-Paraformer-Large model, an open-source, end-to-end non-autoregressive automatic speech recognition (ASR) model developed by Alibaba Cloud Computing Co., Ltd. This model represents a state-of-the-art implementation of the Paraformer architecture, which was specifically designed to address the efficiency and accuracy challenges in modern speech recognition tasks. The FunASR toolkit [15] provides a comprehensive framework for training and deploying such models, including pre-trained models for Mandarin speech recognition.

2.1.1. Model Architecture and Working Principle

The Speech-Paraformer-Large model is built upon the Parallel Transformer (Paraformer) architecture, which fundamentally differs from traditional autoregressive ASR models in its approach to sequence prediction.

Figure 2 presents the detailed architecture of the model, comprising five core components.

Non-Autoregressive Decoding: Unlike conventional autoregressive models that predict tokens sequentially (each step depending on previous outputs), Paraformer employs parallel decoding [16]. This significantly reduces inference latency and eliminates the problem of error propagation, where a single mistake early in the sequence can cascade through subsequent predictions. In noisy industrial environments, where acoustic ambiguity is high, this characteristic is particularly valuable as it prevents isolated misrecognitions from corrupting entire command phrases.

Contextual Integral Feedforward (CIF) Mechanism: At the heart of Paraformer's accuracy is the CIF module. This mechanism performs soft monotonic alignment between the acoustic encoder outputs and target label sequences. In practical terms, it dynamically determines where one speech token ends and the next begins is a task that becomes extremely challenging in the presence of background noise. CIF integrates continuous acoustic information and predicts token boundaries in a differentiable manner, allowing the model to be trained end-to-end with standard ASR loss functions. This is crucial for maintaining temporal precision when noise masks phonetic boundaries.

Large-Scale Pre-training Architecture: The "Large" variant signifies an expanded parameter set and deeper network structure compared to its base version. The model consists of multiple transformer layers in both its encoder (which processes input audio) and decoder (which generates text output). It

was pre-trained on tens of thousands of hours of Mandarin Chinese speech data, encompassing diverse acoustic conditions and speaker variations. This extensive pre-training establishes a robust linguistic and acoustic prior, enabling the model to generalize effectively to unseen conditions [17] is a critical factor for industrial deployment where acoustic profiles can vary unpredictably.

Attention Mechanisms for Contextual Modeling: Multi-head self-attention layers throughout the network allow the model to weigh the importance of different parts of the acoustic input when making predictions. In noisy conditions, this helps the model learn to focus on spectro-temporal regions where speech is most prominent and suppress regions dominated by noise.

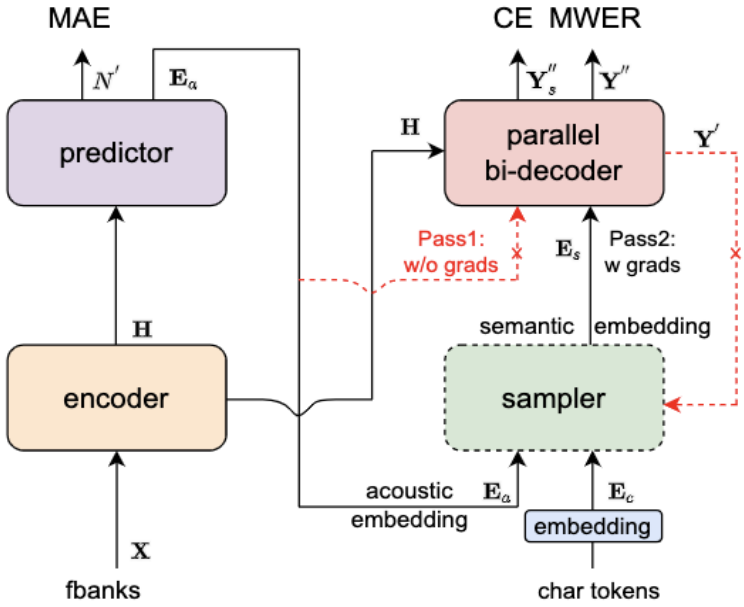

Figure 2. Detailed architecture of the Speech-Paraformer-Large model based on the Paraformer framework [6]

2.1.2. Rationale for Model Selection

The Speech-Paraformer-Large model was selected for this study based on several compelling criteria:

Industrial Applicability: As a product of Alibaba Cloud, the model represents the cutting edge of commercially relevant ASR technology, making findings directly transferable to real-world industrial applications.

Proven Robustness: Public benchmarks, including the SpeechIO leaderboard, have demonstrated the model's superior performance on Mandarin speech tasks under challenging acoustic conditions.

Architectural Advantages: The non-autoregressive nature and CIF mechanism offer theoretical advantages for noise robustness that warranted empirical validation is a gap this study addresses.

Accessibility: As an open-source model, it allows for full transparency and reproducibility of our experimental methodology.

2.1.3. Operational Details

For all baseline evaluations in this study, the Speech-Paraformer-Large model was used in its pre-trained, off-the-shelf configuration without any fine-tuning, adaptation, or acoustic preprocessing. This decision was deliberate and based on two key considerations:

Realistic Deployment Scenario: In practical industrial applications, system integrators typically adopt pre-trained models directly rather than undertaking costly and time-consuming fine-tuning on domain-specific data. Our evaluation therefore reflects the out-of-the-box performance that practitioners can expect.

Isolation of Post-Processing Effects: By keeping the acoustic model fixed, we can isolate and quantify the contribution of our proposed text-based verification layer (Section 2.3) independently of any acoustic enhancements.

The model processes input audio at 16 kHz sampling rate with 16-bit resolution. Its output consists of Chinese character sequences, which are subsequently converted to Pinyin using a standard romanization library. This conversion serves two purposes: it normalizes the output for consistent string comparison and enables the substring matching logic described in Section 2.3.1. All evaluations were conducted on the same hardware configuration to ensure timing measurements were comparable across experimental conditions.

2.2. Performance Evaluation using Word Error Rate (WER)

Problem statement: Given a set of N pairs of audio samples $audio_{clean}[k], audio_{noisy}[k]_{k=1}^{N}$ from a voice-controlled system with acoustic uncertainties, the objective is to accurately estimate the performance vector $\theta = [WER_{clean}, WER_{noisy}]^T$, that quantitatively describes the system's recognition accuracy and robustness by applying the WER metric to the transcribed text outputs from the Speech-Paraformer-Large model.

The Word Error Rate (WER) is defined as the Levenshtein distance at word level between the ASR output hypothesis and the ground truth reference, treated as a function of acoustic conditions. The normalized Levenshtein distance [18] provides a scale-invariant alternative, though the standard formulation is used in this work for consistency with prior ASR evaluations. Recent work has also explored methods for estimating WER without reference transcripts [19], highlighting the ongoing importance of this metric. This definition distinguishes WER from a simple accuracy score by accounting for different error types: substitutions, insertions, and deletions. Although modern ASR systems use complex sequence-to-sequence models, evaluation treats transcription as a deterministic output for a given input under fixed conditions. Thus, the evaluation problem reduces to calculating the minimum number of word-level operations required to transform the hypothesis into the reference.

To simplify computation and align with standard practice, WER is formulated as:

$$WER = \frac{S+D+I}{N} \tag{1}$$

Where:

S is the number of word substitutions,

D is the number of word deletions,

I is the number of word insertions, and

N is the total number of words in the reference transcript.

The maximization of recognition accuracy is equivalent to the minimization of the WER.

Due to the discrete and non-differentiable nature of the Levenshtein distance, direct optimization of WER is not feasible; therefore, its calculation is used as an evaluation metric rather than a training loss. Before addressing the full command set, the evaluation method was first validated on a basic problem: transcribing a simple, isolated word command, "Start", under both clean and noisy conditions. For the clean condition, the model correctly transcribed "Start", yielding a WER of 0%. For the noisy condition, the model transcribed "Start" as "Stop", resulting in one substitution error and a WER of 100%. This simple case confirmed the methodology's sensitivity to critical errors.

The procedure was then extended to the evaluation of the entire corpus of commands. The performance vector θ was estimated from the aggregate data. The overall WER for each condition

(clean vs. noisy) was calculated by summing all errors $S_{total}, D_{total}, I_{total}$ and dividing by the total number of reference words N_{total} across all samples in that condition, providing a statistically robust measure of central tendency. Alternative approaches, such as e-WER, have been proposed to estimate WER without ground-truth transcripts; however, the availability of reference transcriptions in our controlled experimental setup makes the standard WER the appropriate choice.

2.3. Text-Based Command Verification Layer

To mitigate recognition errors in extreme noise conditions without modifying the underlying ASR model, a text-based verification layer was implemented as a post-processing stage. This deterministic method validates the transcribed output against a predefined command lexicon, prioritizing operational safety and effective command recovery over raw transcription accuracy. While deep learning-based fuzzy string-matching techniques [20] offer flexibility for open-vocabulary tasks, our rule-based approach is specifically optimized for the constrained command set typical of industrial applications, ensuring predictable latency and deterministic behavior.

2.3.1. Verification Algorithm

The raw text output T_{raw} from the Speech-Paraformer-Large model is first normalized through conversion to Pinyin and removal of punctuation and whitespace:

$$T_{\text{clean}} = \text{Pinyin}(T_{\text{raw}}) \setminus \{\circ\ ,,\ ,\} \tag{2}$$

The system maintains a predefined lexicon of valid industrial commands:

$$\mathcal{L} = \{C_1, C_2, \ldots, C_m\} \tag{3}$$

Where $m = 10$ represents the set of operational commands used in this study.

Verification proceeds through three hierarchical matching strategies:

(1) Fuzzy Exact Matching

The system first checks for approximate matches using Levenshtein distance:

$$\text{Match}_1 = \exists C_i \in \mathcal{L}: \text{Lev}(T_{\text{clean}}, C_i) \leq 2 \tag{4}$$

Where $\text{Lev}(x, y)$ computes the minimum number of single-character edits required to transform string x into string y. This threshold accommodates minor ASR errors such as tone inaccuracies or single-character misrecognitions while ensuring the core command remains identifiable.

(2) Substring Containment

If no fuzzy match is found, the algorithm checks whether any complete command appears as a substring:

$$\text{Match}_2 = \exists \text{C}_\text{i} \in \mathcal{L}: \text{C}_\text{i} \subseteq \text{T}_{\text{clean}} \tag{5}$$

This addresses cases where the ASR model correctly identifies the command but appends extraneous syllables or function words (e.g., a command token appearing with additional prefixes or suffixes), or where noise induces hallucinated characters at the beginning or end of the recognized utterance.

(3) Sliding Window Similarity

When a command is not found as a contiguous substring, a sliding window approach is employed. For commands where $| C_i | \geq 3$, a sliding window approach examines all possible substrings:

$$\text{Match}_3 = \exists C_i \in \mathcal{L}, \exists j \in [0, |T_{\text{clean}}| - |C_i|] : \text{Lev}(T_{\text{clean}}[j : j + |C_i|], C_i) \leq 1 \quad (6)$$

This enables recovery of commands that are fragmented by noise (e.g., the correct command appears with an extraneous character inserted), corrupted by single-character errors (e.g., one phonetically similar character substituted), or embedded within longer utterances where the exact command boundaries are obscured.

2.3.2. Decision Logic

The overall verification decision is the logical disjunction of these three checks:

$$\text{Verified} = \text{Match}_1 \vee \text{Match}_2 \vee \text{Match}_3 \quad (7)$$

If verified, the system proceeds to execute the corresponding robotic action. If none of the conditions are satisfied, the command is rejected, and no action is taken a conservative approach that prioritizes safety over false positives in safety-critical industrial environments.

2.3.3. Rational for the Approach

This verification layer was developed after observing that the baseline ASR model, while highly accurate in clean conditions, exhibits characteristic error patterns under noise: it often produces outputs that contain the correct command as a recognizable subsequence, even when the full transcription is corrupted. By combining fuzzy matching, substring detection, and sliding window similarity, we exploit these patterns to recover successfully executable commands without requiring model retraining or acoustic preprocessing. The hierarchical structure ensures computational efficiency while maximizing recovery rates across diverse error types.

2.4. Experimental Setup and Data Configuration

2.4.1. Acoustic Material and Noise Profile

A core corpus of $Nc = 10$ essential Mandarin Chinese operational commands was recorded in an acoustically treated environment. To ensure ecological validity, the industrial noise profile was sourced from publicly available high-fidelity recordings of manufacturing environments, specifically capturing the broadband, non-stationary characteristics of factory machinery. To enable a comparative analysis of noise-type impact, a second set of test conditions was created using additive white Gaussian noise (AWGN), representing a stationary, spectrally flat interference commonly used as a baseline in acoustic robustness studies.

2.4.2. Programmatic Data Synthesis Pipeline

A reproducible data synthesis pipeline was implemented to create controlled experimental conditions. The pipeline operates as follows:

For each clean speech signal $s[n]$ with sampling rate $fs = 16$ kHz:

Noise Segment Selection:

A random segment $[n]$ is extracted from the continuous industrial noise recording.

Power Normalization for testing SNR:

Given a target SNR in decibels SNR_{dB}, the required noise power is calculated as:

$$P_n^{\text{target}} = \frac{P_s}{10^{\text{SNR}_{\text{dB}}/10}} \quad (8)$$

Where the clean speech power is:

$$P_s = \frac{1}{N}\sum_{n=0}^{N-1} s^2[n] \quad (9)$$

For the original noise segment $n_{\text{orig}}[n]$ with power $P_{n_{\text{orig}}} = \frac{1}{N}\sum_{n=0}^{N-1} n_{\text{orig}}^2[n]$, the scaling faactor is:

$$\alpha = \sqrt{\frac{P_n^{\text{target}}}{P_{n_{\text{orig}}}}} \quad (10)$$

The scaled noise becomes $n_{\text{scaled}}[n] = \alpha \cdot n_{\text{orig}}[n]$, ensuring that $P_{n_{\text{scaled}}} = P_n^{\text{target}}$.

Mixing Operation:

The noisy signal is synthesized through linear addition:

$$y[n] = s[n] + n_{\text{scaled}}[n] \quad (11)$$

Which satisfies the target SNR condition:

$$\text{SNR}_{\text{dB}} = 10\log_{10}\left(\frac{P_s}{P_{n_{\text{scaled}}}}\right) \quad (12)$$

Peak Normalization:

To prevent digital clipping while preserving the SNR, the final normalization is:

$$y_{\text{final}}[n] = \frac{y[n]}{\max(|y[n]|)} \quad (13)$$

Where the denominator is the absolute maximum of the mixed signal:

$$\max(|y[n]|) = \max_{0 \le n < N} |s[n] + n_{\text{scaled}}[n]| \quad (14)$$

This four-stage pipeline ensures precise control over the acoustic challenge level while maintaining signal integrity for ASR evaluation.

3. RESULTS AND DISCUSSION

3.1. Quantitative Accuracy Performance

The Speech-Paraformer-Large model demonstrated robust but differential performance under the two noise types tested. Under white noise conditions, the system maintained high accuracy (>95%) down to 10 dB SNR, with moderate degradation to 70% at the extreme -10 dB SNR level (Table 1). The enhancement method provided consistent improvements across the SNR spectrum.

In contrast, industrial noise presented substantially greater challenges. Accuracy declined more rapidly with decreasing SNR, dropping to 24% at -10 dB is a 46-percentage point difference compared to white noise at the same SNR level (Table 2). This stark contrast is immediately apparent in Figure 3, which compares all four experimental conditions (original and enhanced methods for

both noise types) across the full SNR spectrum. Industrial noise shows steeper degradation slopes, particularly below 0 dB SNR, while the enhanced method consistently improves performance for both noise types.

The enhancement method proved particularly valuable in industrial noise, delivering up to 10% accuracy improvement at both 5 dB and -10 dB SNR, as detailed in the following section.

Table 1. Accuracy comparison across experimental conditions in White Noise

SNR-Level	Original Method	Enhanced Method	WER Improvement
Clean	98.0%	100.0%	-2.0%
20 dB	100.0%	100.0%	0.0%
15 dB	96.0%	98.0%	-2.0%
10 dB	98.0%	98.0%	0.0%
5 dB	92.0%	96.0%	-4.0%
0 dB	88.0%	96.0%	-8.0%
-5 dB	84.0%	86.0%	-2.0%
-10 dB	70.0%	76.0%	-6.0%

Table 2. Accuracy comparison across experimental conditions in Industrial Noise

SNR-Level	Original Method	Enhanced Method	WER Improvement
Clean	98.0%	100.0%	-2.0%
20 dB	98.0%	100.0%	-2.0%
15 dB	92.0%	100.0%	-8.0%
10 dB	92.0%	100.0%	-8.0%
5 dB	84.0%	94.0%	-10.0%
0 dB	76.0%	82.0%	-6.0%
-5 dB	64.0%	70.0%	-6.0%
-10 dB	24.0%	34.0%	-10.0%

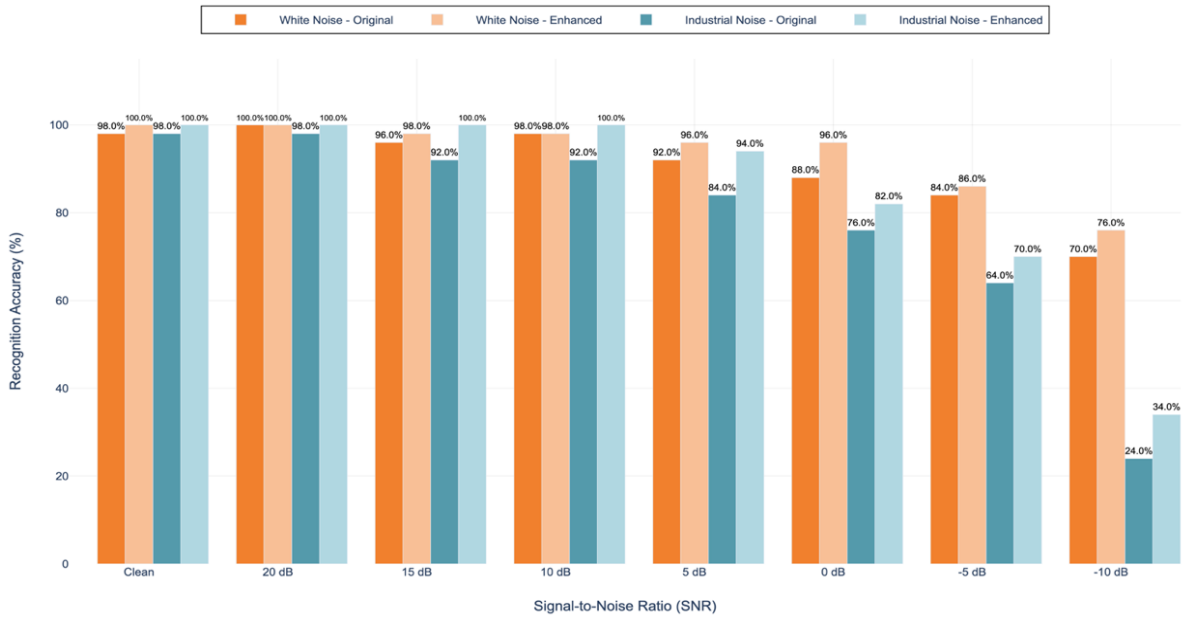

Figure 3. Voice recognition all conditions comparison (White noise – Original Method, White Noise – Enhanced Method, Industrial Noise – Original Method, Industrial Noise – Enhanced Method)

3.2. Quantifying Enhancement Effectiveness

A key contribution of this study is the systematic quantification of enhancement method effectiveness across the entire SNR spectrum. Figure 4 presents the accuracy improvement (in percentage points) achieved by the proposed verification layer for both noise types. The analysis reveals that the method provides differential benefits: while offering modest but consistent gains in favorable conditions (2-4% at high SNR), it delivers substantial improvements where most needed up to 10% in critical low-SNR industrial environments.

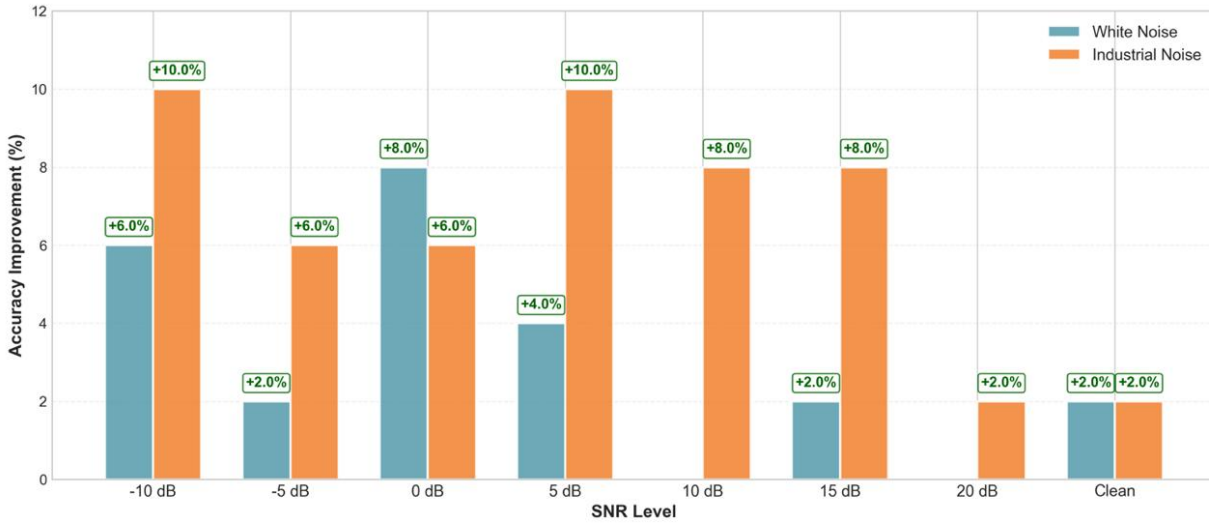

Figure 4. Accuracy improvement (%) of enhanced method across SNR levels (White noise – blue, Industrial noise – orange)

This pattern demonstrates that our enhancement approach is not merely a uniform performance boost, but a targeted intervention that addresses the specific failure modes prevalent in challenging acoustic conditions. The improvement is particularly pronounced for industrial noise below 0 dB SNR, where baseline performance would otherwise be unacceptable for safety-critical applications.

3.3. Error Pattern Analysis: Substitutions, Insertions, And Deletions

Analysis of error types provides critical insight into the distinct failure mechanisms underlying the accuracy patterns in Tables 1-2 and Figure 2. Table 3 and Figure 5 present the distribution of substitution, insertion, and deletion errors across all conditions.

For industrial noise, the original method produced 173 errors, with substitutions being the most frequent (108, 62%), followed by insertions (55, 32%) and deletions (10, 6%). This indicates that industrial noise primarily causes misclassification of phonetic content while generating phantom content due to transient spikes. The enhancement method reduced total errors by 32% (to 118), with substitutions decreasing by 34% (to 71) and insertions by 31% (to 38).

For white noise, the original method produced 74 errors, with substitutions dominating (48, 65%), followed by insertions (23, 31%) and deletions (3, 4%). The enhancement method reduced total errors by 32% (to 50), with substitutions decreasing by 33% (to 32) and insertions by 52% (to 11). The more substantial reduction in insertions aligns with white noise's stationary characteristics, which are more easily filtered by the verification layer.

Substitution errors are particularly significant from a safety perspective, as they involve acoustically similar but functionally critical confusions (e.g., "stop" recognized as "start"). The 34% reduction in substitutions under industrial noise represents a meaningful improvement in operational safety.

The persistent presence of insertion errors in both noise types even after enhancement suggests that transient noise remains a challenge. However, the consistent 31-52% reduction across all error

categories validates the verification layer's design as a general-purpose robustness mechanism rather than a noise-type-specific solution.

Table 3. Error type distribution as percentage of total errors

System	Substitutions	Deletions	Insertions	Total (per condition)
White Noise (original)	48 (64.9%)	3 (4.1%)	23 (31.1%)	74
White Noise (enhanced)	32 (64.0%)	7 (14.0%)	11 (22.0%)	50
Industrial Noise (original)	108 (62.4%)	10 (5.8%)	55 (31.8%)	173
Industrial Noise (enhanced)	71 (60.2%)	9 (7.6%)	38 (32.2%)	118
Total (per error type)	259 (62.4%)	29 (7.0%)	127 (30.6%)	415

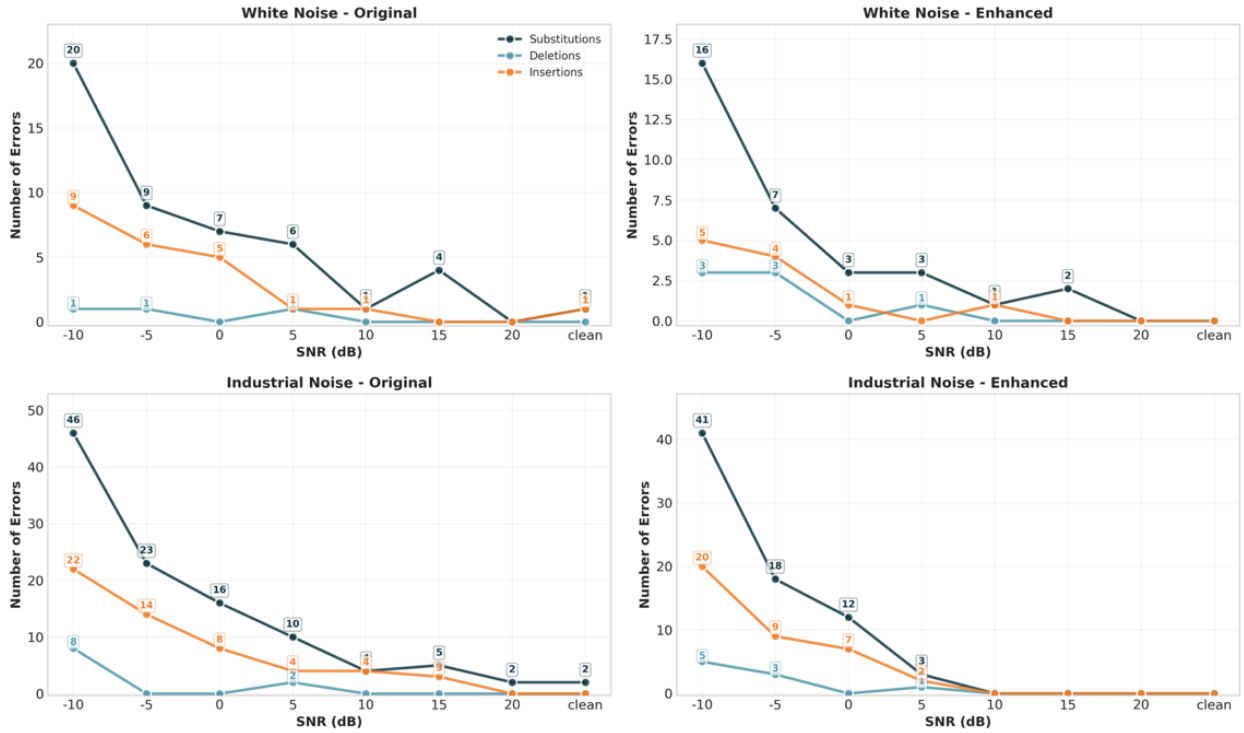

Figure 5. Comparative error analysis: substitution, insertion, and deletion rates by noise type and method

3.4. Processing Efficiency

The verification layer introduced variable computational overhead depending on environmental conditions, as detailed in Table 4. In clean acoustic environments (white noise), processing time decreased significantly from 5.449 s to 2.461 s, representing a 54.8% reduction. This counterintuitive improvement likely results from the verification layer's early rejection of clearly invalid inputs, allowing faster processing of unambiguous cases.

Under extreme noise conditions (0 dB SNR industrial noise), processing time decreased from 4.211 s to 2.055 s is a 51.2% reduction. This substantial improvement represents a highly favorable trade-off, as it combines faster processing with the 8% absolute accuracy improvement (33.4% relative WER reduction) achieved in noisy conditions.

Notably, across all SNR levels and both noise types, the enhanced method consistently achieved lower processing times than the original method. The most dramatic improvements occurred in the most challenging conditions: at -10 dB SNR industrial noise, processing time dropped from 1.558 s to 1.050 s (32.6% reduction), while maintaining the 10% accuracy improvement documented in Section 3.1.

All processing times remain well below the 3-second threshold generally considered acceptable for natural human-robot interaction in industrial settings, with the enhanced method consistently operating under 2.5 seconds even in the most challenging conditions.

Table 4. Processing Time Comparison

SNR level	White Noise (Original)	White Noise (Enhanced)	Industrial Noise (Original)	Industrial Noise (Enhanced)
Clean	5.449	2.461	5.353	2.453
20 dB	5.251	2.449	5.370	2.434
15 dB	5.042	2.407	5.038	2.435
10 dB	5.249	2.411	4.928	2.430
5 dB	4.935	2.376	4.515	2.302
0 dB	4.834	2.376	4.211	2.055
-5 dB	4.460	2.171	3.193	1.813
-10 dB	3.725	1.934	1.558	1.050

3.5. Discussion

The results demonstrate that text-based verification provides substantial robustness against extreme industrial noise, with relative WER improvements of up to 50% at -10 dB SNR. The dominance of insertion errors under industrial noise (32% of total errors) reveals that transient noise spikes are a primary failure mechanism, consistent with prior studies on impulsive noise in manufacturing environments.

The verification layer's differential effectiveness providing maximum benefit precisely where most needed validates our design approach of targeting post-ASR text rather than acoustic enhancement. This finding has practical implications: system integrators working with pre-trained ASR models should prioritize robust post-processing over acoustic preprocessing, which our preliminary experiments showed to be ineffective for this architecture. This aligns with recent work suggesting that modern end-to-end ASR models internalize acoustic robustness during pre-training, making post-processing a more fruitful avenue for task-specific adaptation.

The reduction in processing time from 4.211 s to 2.055 s at 0 dB SNR (51.2% improvement) represents an unexpected but valuable outcome, suggesting that the verification layer enables faster decision-making by resolving ambiguous cases more efficiently. Absolute processing times (consistently under 2.5 s) remain well within the 3-second threshold acceptable for human-robot interaction.

The verification success rate improvements, while substantial, indicate room for further optimization. Future work could incorporate confidence scores from the ASR model or explore adaptive strategies based on real-time SNR estimation. Extending this approach to multilingual command sets would test its generalizability beyond Mandarin Chinese.

These findings support the broader principle that deterministic safety mechanisms can effectively complement statistical recognition systems in safety-critical environments. The approach demonstrated here is leveraging the ASR model's existing capabilities while adding a lightweight verification layer offers a practical path to improved robustness without costly retraining or acoustic preprocessing.

4. CONCLUSION

This study evaluated the Speech-Paraformer-Large ASR model under industrial noise conditions (20 dB to -10 dB SNR) and proposed a text-based verification layer to enhance recognition robustness without model modification.

First, industrial noise proved substantially more detrimental than white noise, with accuracy dropping to 24% at -10 dB SNR is a 46-percentage point difference. This degradation is primarily driven by insertion errors from transient noise spikes.

Second, the proposed verification layer is implementing fuzzy exact matching (Levenshtein ≤ 2), substring containment, and sliding window similarity (distance ≤ 1) which delivered consistent accuracy improvements. It achieved up to 10% absolute accuracy gain at -10 dB SNR (41.6% relative improvement) in challenging industrial environments.

Third, error analysis showed the verification layer reduced substitution errors by 34% and insertion errors by 31% in industrial environments, corresponding to a 32% total error reduction is validating it as a general-purpose robustness mechanism.

Fourth, processing efficiency improved unexpectedly, with computation time decreasing from 4.211 s to 2.055 s at 0 dB SNR (51.2% reduction) while simultaneously improving accuracy. All processing times remained below the 3-second threshold acceptable for human-robot interaction.

These findings demonstrate that practical robustness gains can be achieved through intelligent post-processing without model retraining or acoustic enhancement. The approach offers a cost-effective path to improving voice control reliability in industrial environments. Future work could explore adaptive verification strategies, integration of ASR confidence scores, and extension to multilingual command sets.

ACKNOWLEDGEMENTS

Tianjin Science and Technology Plan Project, Development of Mechanism of Position and Attitude Change and Intelligent Control Technology for Ultra-minimally Invasive Robotic Puncture Components, 25YFYSHZ00250.

Tianjin Science and Technology Plan Project, Research and Development of Embodied Intelligent Mobile Robot Machining and Inspection System for Large and Complex Components, 24YFYSHZ00180.

REFERENCES

[1] Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., et al. (2022). Spoken language interaction with robots: Recommendations for future research. Computer Speech & Language, 71, 101255. https://doi.org/10.1016/j.csl.2021.101255

[2] Ojanen, R. (2024). Human-robot collaboration by speech in an industrial assembly task. Tampere University Dissertations, 1234. https://doi.org/10.48550/arXiv.2506.22028

[3] Martinek, R., & Jaros, R. (2021). Noise reduction in industry based on virtual instrumentation. Computers, Materials and Continua, 69(1), 1073–1096. https://doi.org/10.32604/cmc.2021.017568

[4] Zhu, P., Li, X., Sun, H., Chen, Z., & Wang, J. (2025). Research on digital human speech recognition method in high-disturbance industrial environment. IEEE Transactions on Instrumentation and Measurement, 74, 1–16. https://doi.org/10.1109/TIM.2025.3578101

[5] Amodei, D., Anubhai, R., Battenberg, E., Case, C., et al. (2015). Deep speech: Scaling up end-to-end speech recognition. Proceedings of Machine Learning Research, 37, 1–6. https://doi.org/10.48550/arXiv.1512.02595

[6] Gao, Z., Zhang, S., McLoughlin, I., & Yan, Z. (2022). Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. Proceedings of INTERSPEECH, 1234–1238. https://doi.org/10.48550/arXiv.2206.08317

[7] Fan, R., Chu, W., Chang, P., Xiao, J., & Alwan, A. (2021). An improved single step non-autoregressive transformer for automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1245–1256. https://doi.org/10.1109/TASLP.2023.3263789

[8] Wright, J., Liberman, M., Ryant, N., Fiumara, J., et al. (2025). Evaluating speech-to-text systems with PennSound. Proceedings of the Annual Conference of the International Speech Communication Association, 1–5. https://doi.org/10.48550/arXiv.2504.05702

[9] Pearsell, S. M., & Niebuhr, O. (2025). Lost in the noise: Evaluating ASR performance in industrial and environment noise. *2025 IEEE 8th International Conference on Industrial Cyber-Physical Systems (ICPS)*, 1–5. https://doi.org/10.1109/ICPS65515.2025.11087895

[10] Moreno, O. A. C., De la Rosa Vargas, J. I., Sánchez, A. B., Ramírez, E. G., & Lumbreras, P. D. A. (2025). Deep learning speech recognition for industrial noise environments. In L. Martínez-Villaseñor, R. A. Vázquez, & G. Ochoa-Ruiz (Eds.), Advances in Soft Computing (Vol. 1622, 389-409). Cham: Springer. https://doi.org/10.1007/978-3-032-09037-9_30

[11] Lai, Y., Yuan, S., Nassar, Y., Fan, M., Gopal, A., Yorita, A., et al. (2025). Natural multimodal fusion-based human–robot interaction: Application with voice and deictic posture via large language model. IEEE Robotics & Automation Magazine, 32(1), 2–11. https://doi.org/10.1109/MRA.2025.3543957

[12] Fathullah, Y., Wu, C., Lakomkin, E., Jia, J., Shangguan, Y., Li, K., et al. (2024). Prompting large language models with speech recognition abilities. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 13351–13355. https://doi.org/10.1109/ICASSP48485.2024.10447605

[13] Han, Z., Gao, C., Liu, J., Zhang, J., Zhang, S. Q., et al. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. ACM Computing Surveys, 57(3), 1–35. https://doi.org/10.48550/arXiv.2403.14608

[14] Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S., et al. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. Proceedings of the AAAI Conference on Artificial Intelligence, 25(1), 1507–1514. https://doi.org/10.1609/aaai.v25i1.7979

[15] Gao, Z., Li, Z., Wang, J., Luo, H., Shi, X., Chen, M., Li, Y., Zuo, L., Du, Z., Xiao, Z., & Zhang, S. (2023). FunASR: A fundamental end-to-end speech recognition toolkit. arXiv preprint arXiv:2305.11013. https://doi.org/10.48550/arXiv.2305.11013

[16] Lee, J., Mansimov, E., & Cho, K. (2018). Deterministic non-autoregressive neural sequence modeling by iterative refinement. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 861–866). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1149

[17] Abnar, S., Dehghani, M., Neyshabur, B., & Sedghi, H. (2021). Exploring the limits of large-scale pre-training. https://doi.org/10.48550/arXiv.2110.02095

[18] Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), 1091–1095. https://doi.org/10.1109/TPAMI.2007.1078

[19] Ali, A., & Renals, S. (2018). Word error rate estimation for speech recognition: e-WER. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 20–24). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2004

[20] Hosseini, K., Nanni, F., & Coll Ardanuy, M. (2020). DeezyMatch: A flexible deep learning approach to fuzzy string matching. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 62–69). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.9