Incorporating Error Level Noise Embedding for Improving LLM-Assisted Robustness in Persian Speech Recognition

Anonymous ACL submission

Abstract

Automatic Speech Recognition (ASR) performance degrades severely under noise, a challenge that is particularly pronounced for low-resource languages such as Persian. Even state-of-the-art systems like Whisper exhibit substantial accuracy loss at low signal-to-noise ratios. We propose a noise-aware ASR error correction framework that combines multiple transcription hypotheses with explicit modeling of linguistic noise. From noisy Persian speech, we generate 5-best hypotheses using Whisper and introduce Error Level Noise (ELN), a representation that captures sentence- and token-level disagreement across hypotheses as a proxy for noise-induced uncertainty. ELN vectors are used to condition a fine-tuned LLaMA-2-7B model during post-hoc correction. Experiments show that ELN conditioning significantly reduces Word Error Rate (WER). On the Mixed Noise test set, our model lowers WER from 31.10% (Raw Whisper) to 24.84%, outperforming a text-only fine-tuned baseline (30.79%), while a zero-shot LLaMA-2 model fails to correct Persian ASR outputs. These results demonstrate the effectiveness of noise-aware multi-hypothesis correction for robust Persian ASR.

1 Introduction

ASR has become an integral part in the interaction of humans with computers in the modern world, and it finds applications in smart assistants, speech translation, and automatic captioning. While recent advances in end-to-end models and large language models (LLMs) have substantially improved ASR performance for high-resource languages, robustness under noisy conditions and for low-resource languages remains a major challenge. In particular, ASR systems for Persian still suffer significant degradation when exposed to environmental noise.

Previous studies on the robustness of automatic speech recognition (ASR) systems have primarily been centered around noise-aware training, multi-condition training, and data augmentation using synthetic noise (Ko et al., 2015; Watanabe et al., 2018; Li et al., 2016). While such methods have proven to be quite effective, they mainly work at the acoustic modeling level and usually involve the retraining or modification of the ASR system. Moreover, most existing methods are developed for high-resource languages and rely on single best hypotheses, ignoring uncertainty induced by noise. Post-processing approaches based on textual correction also generally lack explicit noise modeling and rarely leverage multi-hypothesis reasoning with large generative models.

In this work, we introduce a noise-robust, post-hoc generative error correction framework for Persian ASR that avoids retraining the underlying recognizer. We generate 5-best hypotheses from a Whisper-based Persian ASR system and propose *Error Level Noise* (ELN) embeddings, which capture sentence- and token-level disagreement across hypotheses as a proxy for noise-induced uncertainty. These embeddings act as inputs to a LLaMA-2-7B model (Touvron et al., 2023) that has been fine-tuned, allowing the noise-aware correction to be informed not only by the linguistic context but also by the variability of the hypothesis. Noisy Persian speech experiments derived from Common Voice (Ardila et al., 2019) and MUSAN (Snyder et al., 2015) show that the WER of the proposed approach is considerably decreased in comparison to both the ASR baseline and text-only correction models, thus it is a noise-robust ASR solution that is effective and scalable in low-resource scenarios.

2 Related Work

Automatic speech recognition (ASR) error correction has been approached from various angles, including rule-based methods, statistical language models, and more recently, neural and large lan-

guage model (LLM)-based methods, with continuous difficulties in low-resource and noisy environments.

Retrieval-augmented generation (RAG) has been used in ASR correction by relying on external examples. GEC-RAG (Robatian et al., 2025) helps Persian ASR by fetching closely matching error patterns for LLM-based correction. On the other hand, LA-RAG (Li et al., 2024a) and DARAG (Ghosh and et al., 2025) mix retrieval with generative modeling to increase resilience and generalization across different domains and acoustic scenarios.

Some of the works are dedicated entirely to the Persian ASR and spelling correction. PERCORE (Dashti et al., 2024) brings phonetic features to the table for tackling real-word and non-word spelling errors, whereas PSRB (Sedghiyeh et al., 2025) presents a benchmark dataset for assessing Persian ASR in varying linguistic and acoustic conditions.

Generative Error Correction (GER) (Ma et al., 2023) sees ASR post-processing as a sequence-to-sequence rewriting task and has spawned developments like Denoising GER (Liu et al., 2025) and RobustGER (Hu et al., 2024b), which increase noise robustness through language-space modeling and knowledge distillation respectively. ClozeGER (Hu et al., 2024a) also takes acoustic information into account by changing the correction problem into a cloze task with speech logits. HyPoradise (Chen et al., 2023) showcases the capability of LLMs for generative speech recognition, and multilingual LLMs have been adjusted as well for direct 1-best ASR correction without N-best hypotheses, exhibiting cross-lingual transfer advantages for low-resource languages (Li et al., 2024b).

Differently from previous research, our method utilizes disagreement among several ASR hypotheses to represent noise-induced uncertainty explicitly and uses an LLM to generate structured text-level noise representations without the need for acoustic inputs or changes to the ASR system.

3 Methodology

We propose **GER + Text Denoising (Ours)**, a noise-aware ASR error correction framework that conditions a large language model (LLM) on structured text-level noise representations. The method targets low-resource Persian ASR and operates as a post-hoc correction module without modifying or retraining the underlying recognizer. As shown in Figure 1 (Panel d), the framework uses the top-5 ASR hypotheses (*5-best list*) as input.

Our approach is inspired by RobustGER (Hu et al., 2024b) but is designed to be computationally lightweight. Unlike RobustGER, which relies on knowledge distillation and optional audio noise modeling, we focus exclusively on *text noise*. We introduce *Error Level Noise* (ELN) vectors that capture sentence- and token-level disagreement across hypotheses and use them as conditioning signals for the LLM. This enables noise-aware correction directly in the language space, without audio features or mutual information estimation.

3.1 System Overview

Figure 1 compares four architectures used for ASR noise Robustness. Panels (a) through (c) are adapted from(Hu et al., 2024b): (a) GER (Chen et al., 2023; Yang et al., 2023), a text-only generative correction model; (b) GER with audio denoising (Liu et al., 2025); (c) RobustGER (Hu et al., 2024b), which models language noise via knowledge distillation; and (d) our proposed method, which injects text-level noise embeddings into the LLM to improve robustness across noise conditions.

3.2 Hypothesis Generation

Given noisy speech, we generate an N-best list using the Whisper large ASR model (Radford and et al., 2023), employing a fine-tuned Persian variant[1]. Beam search decoding produces up to five unique hypotheses per utterance; samples with fewer hypotheses are padded to $n = 5$.

All hypotheses and references undergo Persian text normalization, including Unicode standardization, digit conversion, spacing correction, punctuation removal, and whitespace cleanup. This preprocessing ensures consistent evaluation using Word Error Rate (WER) and supports reliable computation of ELN vectors.

3.3 ELN Vector Extraction

For an utterance, let $\mathcal{H} = \{H_1, \ldots, H_n\}$ denote the $n = 5$ ASR hypotheses. ELN consists of sentence-level and token-level components, which together quantify linguistic disagreement across hypotheses and serve as text-level noise representations.

[1]https://huggingface.co/vhdm/whisper-large-fa-v1

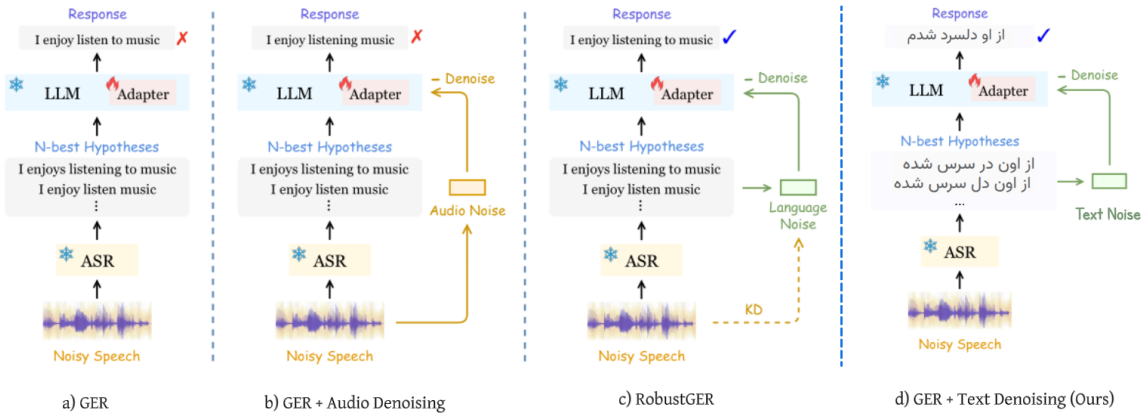

Figure 1: Overview of a) GER (Chen et al., 2023) b) GER + Audio Denoising (Liu et al., 2025) c) RobustGER (Hu et al., 2024b) and d) GER + Text Denoising (Ours).

Sentence-level ELN Each hypothesis is embedded using a sentence encoder (e.g., Sentence-BERT (Reimers and Gurevych, 2019)):

$$\mathbf{e}_i = \text{Embed}_{\text{sent}}(H_i) \in R^d.$$

The sentence-level ELN vector is computed as:

$$\mathbf{v}_{\text{sent}} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (\mathbf{e}_i - \mathbf{e}_j)^2$$

capturing semantic variance among hypotheses.

Token-level ELN Each hypothesis is tokenized: $H_i = [t_{i,1}, t_{i,2}, \ldots, t_{i,L_i}]$, and padded to the maximum length $L_{\max} = \max_i L_i$. Tokens are embedded into vectors of dimension d':

$$\mathbf{t}_{i,k} = \text{Embed}_{\text{tok}}(t_{i,k}) \in R^{d'}.$$

Token-level ELN is defined as:

$$\mathbf{v}_{\text{tok}} = \frac{1}{L_{\max}} \sum_{k=1}^{L_{\max}} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (\mathbf{t}_{i,k} - \mathbf{t}_{j,k})^2$$

Final ELN Representation The final ELN vector concatenates both components:

$$\mathbf{v}_{\text{ELN}} = \mathbf{v}_{\text{sent}} \| \mathbf{v}_{\text{tok}},$$

and is injected into the LLM via an adapter-based conditioning mechanism (Zhang et al., 2023).

ELN Magnitude We define ELN magnitude as the L_2 norm:

$$\|\mathbf{v}_{\text{ELN}}\|_2,$$

which provides a scalar measure of noise-induced disagreement, with larger values indicating greater transcription uncertainty.

4 Experimental Setup

We evaluate the proposed method using a unified prompt that instructs the large language model (LLM) to perform ASR error correction based on the top-5 hypotheses generated by the recognizer and considering the noise represented by ELN vector, while preserving meaning and avoiding synonym substitution.

4.1 Datasets

Common Voice 16.1 (Persian) (Ardila et al., 2019) was the source of clean speech in the creation of noisy Persian speech, whereas the noise corpus was from MUSAN (Snyder et al., 2015). The Persian part of Common Voice has nearly 90 hours of transcribed speech (40k utterances) recorded at 16 kHz with various speakers and accents. When it comes to noise, we have taken the ambient noise part of MUSAN which is a collection of around 6 hours of environmental sounds as well as artificial noises.

Noise Augmentation Noisy utterances were generated by adding MUSAN noise and low-amplitude Gaussian noise to clean speech: MUSAN noise was added at SNRs uniformly sampled between 0–15 dB, and Gaussian noise amplitudes were drawn from $[0.001, 0.015]$, yielding a wide range of realistic acoustic conditions.

4.2 Evaluation Metrics and Noise Conditions

Performance is measured using Word Error Rate (WER).

We report results under four conditions: **Clean** (5000 samples), **Mixed Noise** (random MUSAN noise at 0–15 dB, 5000 samples), **SNR = 5 dB**

Method	Clean	Mixed Noise	SNR = 5 dB	SNR = 10 dB
Raw Whisper	24.80	31.10	42.70	38.30
Base Model (Zero-shot)	62.43	64.58	70.63	67.75
Fine-tuned (No ELN)	24.06	30.79	39.76	31.59
Fine-tuned + ELN (Ours)	**24.39**	**24.84**	**32.34**	**28.02**

Table 1: WER (%) comparison of Raw Whisper, Base LLaMA2 (zero-shot), LLaMA2 fine-tuned without ELN, and the ELN-conditioned model across four acoustic conditions.

(random MUSAN noise at 5 dB,1000 samples), and **SNR = 10 dB** (random MUSAN noise at 10 dB,1000 samples). All noisy settings combine MUSAN ambient noise with low-level Gaussian noise.

4.3 Compared Models

We compare three configurations:

1. Zero-shot: LLaMA-2-7B without fine-tuning.
2. Fine-tuned (Text-only): LLaMA-2-7B fine-tuned with LoRA (Hu et al., 2022) on 5-best hypotheses.
3. Fine-tuned + ELN (Ours): LLaMA-2-7B fine-tuned with LoRA incorporating ELN vectors. ELN vectors were mapped through a small MLP to match the LLM embedding dimension and prepended as prefix embeddings.

4.4 Training Details

All models were fine-tuned using 4-bit quantization, a learning rate of 2×10^{-4} for 3 epochs, gradient accumulation and checkpointing, and a cosine learning-rate scheduler with weight decay.

5 Results and Discussion

5.1 Quantitative Results

Table 1 reports the Word Error Rate (WER, %) of all evaluated models under four noise conditions: Clean, Mixed Noise (covering 0–15 dB SNR plus Gaussian noise), SNR = 5 dB, and SNR = 10 dB. For reference, the Raw Whisper baseline follows Radford et al. (Radford and et al., 2023), while the Base Model (Zero-shot) corresponds to LLaMA-2 (Touvron et al., 2023). All WER values are expressed in percentage.

5.2 Comparison with RobustGER on VB-DEMAND

We further evaluate noise robustness and cross-lingual generalization on the VoiceBank-DEMAND (VB-DEMAND) benchmark (Veaux et al., 2017; Thiemann et al., 2013) using the 16 kHz version. Whisper `small.en` is used to generate N-best hypotheses, and the ELN-conditioned LLaMA2-7B model is trained on the official training split and evaluated on the test set.

Method	Baseline WER	Improved WER
RobustGER	13.00	10.70
Ours (Fine-tuned + ELN)	7.93	**3.96**

Table 2: Comparison of WER (%) between the proposed ELN-conditioned model and RobustGER (Hu et al., 2024b) on the **VB-DEMAND dataset**.

As shown in Table 2, our method reduces WER from 7.93% to 3.96%, substantially outperforming RobustGER (Hu et al., 2024b) (10.70%). This demonstrates that ELN-based text denoising generalizes effectively beyond Persian and provides strong robustness under real-world noise.

For clarity, in Table 2, Baseline WER refers to the Word Error Rate of the Whisper ASR system prior to any correction, while Improved WER indicates the WER after applying the respective error correction model.

5.3 Key Findings

Our results show that: (i) task-specific fine-tuning is essential for Persian ASR correction; (ii) ELN conditioning significantly reduces WER, particularly under moderate and severe noise; and (iii) ELN magnitude strongly correlates with recognition difficulty, validating its role as a noise-aware signal for generative ASR error correction.

6 Conclusion

We explored using large language models (LLMs) for post-hoc correction of ASR outputs, focusing on low-resource Persian. We proposed a computationally efficient framework that conditions a fine-tuned LLaMA-2-7B on an Error Level Noise (ELN) vector derived from n-best ASR hypotheses.

By explicitly modeling linguistic noise via ELN, our approach improves transcription robustness without modifying the original ASR. Conditioning LLMs with structured noise thus offers a practical, scalable way to enhance ASR reliability in low-resource, noisy settings.

Limitations

Although the results were encouraging, a number of the study's limitations should be considered. First, the experiments used the Common Voice (Persian)

dataset which was most probably augmented with synthetic MUSAN noise. This might not be a perfect representative of the diversity and the complexity of the real Persian speech, especially when it comes to conversational, dialectal, and domain-specific contexts. Second, the evaluation mainly depended on the Word Error Rate (WER) metric. Of course, WER is a valuable quantitative tool, but the next research should involve human-based assessments or semantic similarity metrics to more deeply evaluate changes in fluency, coherence, and meaning preservation. Lastly, the current method only models noise through textual proxies with the help of Error Level Noise (ELN) vectors, without any direct acoustic evidence. It will also be very useful to add explicit acoustic features to the model that will enable a complete understanding of the influence of noise and thus better robustness of the language space noise model could be achieved.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng Siong Chng. 2023. Hyporadise: An open baseline for generative speech recognition with large language models. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems (Datasets and Benchmarks Track)*.

Seyed Mohammad Sadegh Dashti, Amid Khatibi Bardsiri, and Mehdi Jafari Shahbazzadeh. 2024. Percore: A deep learning-based framework for persian spelling correction with phonetic analysis. *International Journal of Computational Linguistics and Applications*, 17(1):1–20.

Suman Ghosh and et al. 2025. Darag: Improving generative error correction for asr with synthetic data and retrieval-augmented generation. *Findings of the Association for Computational Linguistics*, 3:125–134.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yuchen Hu, Chen Chen, Chengwei Qin, Qiushi Zhu, Eng Siong Chng, and Ruizhe Li. 2024a. Listen again and choose the right answer: A new paradigm for automatic speech recognition with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and Eng Siong Chng. 2024b. Large language models are efficient learners of noise-robust speech recognition. In *International Conference on Learning Representations (ICLR) 2024*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proceedings of Interspeech*, pages 3586–3589.

Qiang Li, Yan He, and Li Deng. 2016. *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Academic Press.

Shaojun Li, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Zongyao Li, Xianghui He, Min Zhang, and Hao Yang. 2024a. La-rag: Enhancing llm-based asr accuracy with retrieval-augmented generation. *arXiv preprint arXiv:2409.08597*.

Sheng Li, Chen Chen, Chin Yuen Kwok, Chenhui Chu, Eng Siong Chng, and Hisashi Kawai. 2024b. Investigating asr error correction with large language model and multilingual 1-best hypotheses. In *Proc. Interspeech*, Kos, Greece.

Yanyan Liu, Minqiang Xu, Yihao Chen, Liang He, Lei Fang, Sian Fang, and Lin Liu. 2025. Denoising ger: A noise-robust generative error correction with llm for speech recognition. *Preprint*, arXiv:2509.04392.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark J. F. Gales, and Kate M. Knill. 2023. Can generative large language models perform asr error correction? *CoRR*, abs/2307.04172.

Alec Radford and et al. 2023. Robust speech recognition via whisper. In *arXiv preprint arXiv:2212.04356*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Amin Robatian, Mohammad Hajipour, Mohammad Reza Peyghan, Fatemeh Rajabi, Sajjad Amini, Shahrokh Ghaemmaghami, and Iman Gholampour. 2025. Gec-rag: Improving generative error correction via retrieval-augmented generation for automatic speech recognition systems. *arXiv preprint arXiv:2501.10734*.

Nima Sedghiyeh, Sara Sadeghi, Reza Khodadadi, Farzin Kashani, Omid Aghdaei, Somayeh Rahimi, and Mohammad Sadegh Safari. 2025. Psrb: A comprehensive benchmark for evaluating persian asr systems. *arXiv preprint arXiv:2505.21230*.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *CoRR*, abs/1510.08484.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. The DEMAND dataset of realistic noise recordings for speech processing. In *Proceedings of the 21st International Conference on Acoustics (IWAENC)*, pages 1–4. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. Version 0.92.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Rui Zhang, Songlin Yang, Chenguang Zhu, and Yue Zhang. 2023. LLM-Adapters: Parameter-efficient fine-tuning of large language models with adapter layers. *arXiv preprint arXiv:2304.01933*.