Noise-Robust Speech Transcription with Quantized Language Model Correction for Industrial Settings

Marco Murgia[1,2], Marco Fontana[3], Alberto Pes[2], Diego Reforgiato Recupero[1,2], Giuseppe Scarpi[1,2] and Leonardo Daniele Scintilla[3]

[1]*Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, Cagliari, Italy*

[2]*R2M Solution S.r.l., Via Fratelli Cuzio, 42, Pavia, Italy*

[3]*Fontana Group, Via Alcide de Gasperi 16, Calolziocorte, Italy*

Keywords: Language Models, Automatic Speech Recognition, Synthetic Dataset, Noisy Environment.

Abstract: In this paper, we propose a robust and computationally efficient pipeline for transcribing speech in noisy environments, such as workshops and industrial settings. The pipeline is designed to operate offline, making it suitable for resource-constrained scenarios. It begins with a noise filtering module that preprocesses audio recordings to suppress background noise and enhance speech clarity. The filtered audio is then passed to an Automatic Speech Recognition (ASR) model, which generates initial transcription outputs. Given the potential for transcription errors in challenging acoustic conditions, we incorporate a quantized Small Language Model (SLM) trained on an ontology of defects related to the industrial environment to post-process and correct these errors. The quantization of the SLM significantly reduces its computational footprint while maintaining correction accuracy, enabling the pipeline to function effectively on low-resource devices. Experimental evaluations demonstrate the effectiveness of the proposed approach in improving transcription quality in noisy conditions, highlighting its practicality for offline and resource-limited applications. In fact, preliminary validation on a synthetic dataset of 200 sentences in Italian and English showed a consistent F1 score of 87.04% for SNR as challenging as -5 dBW (Decibels Watt) in Italian sentences and 91.25% in English sentences, with the least computationally expensive version of Whisper (Whisper Tiny) and the SLM correction.

1 INTRODUCTION

Automatic Speech Recognition (ASR) systems often struggle in noisy industrial environments like workshops, where machinery sounds and reverberations lead to frequent transcription errors (Li et al., 2014; Virtanen et al., 2017; Mehrish et al., 2023). Furthermore, many state-of-the-art ASR systems rely on cloud processing, which is often unfeasible for scenarios with limited hardware resources or privacy-sensitive ones that require offline operation (Bodepudi et al., 2019). While noise suppression and robust models have improved (Zhang et al., 2018; Ephraim and Malah, 1984), a comprehensive solution for accurate, offline transcription remains a critical challenge for industrial applications in quality control and safety compliance (Huang et al., 2014). To address these challenges, we propose a robust and computationally efficient pipeline for transcribing speech commands in noisy settings. In collaboration with Fontana Group, an Italian leader in automotive manufacturing, we developed a toolchain to help operators recognize car body defects. The main challenges are the noisy workshop environment and the need for a solution that works locally on systems with limited hardware, given the potential lack of a reliable internet connection. Our pipeline integrates three components: (i) a noise filtering module, (ii) an ASR model for the transcription, and (iii) a quantized Small Language Model (SLM) trained on a domain-specific defect ontology to correct errors. This quantized SLM ensures a reduced computational footprint, enabling effective on-device performance. Preliminary validation on a synthetic test set demonstrates the effectiveness of our approach, achieving F1-scores of 87.04% for Italian and 91.25% for English sentences in scenarios with an SNR as low as -5 dBW.

Murgia, M., Fontana, M., Pes, A., Recupero, D. R., Scarpi, G. and Scintilla, L. D.
Noise-Robust Speech Transcription with Quantized Language Model Correction for Industrial Settings.
DOI: 10.5220/0013704200003982

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 2*, pages 479-486
ISBN: 978-989-758-770-2; ISSN: 2184-2809

2 RELATED WORK

Robust ASR in noisy industrial settings (Orel and Varol, 2023; Dua et al., 2023; Bandela et al., 2023) has been addressed through various techniques. Front-end approaches, ranging from traditional spectral subtraction (Moore et al., 2017) and Wiener filtering (Gomez and Kawahara, 2010) to modern deep learning denoisers (G. et al., 2024), aim to enhance speech intelligibility but can be computationally demanding for offline, resource-constrained contexts. On the ASR side, while large models like the Whisper family (Radford et al., 2022) offer impressive robustness, their computational cost remains a barrier for edge deployment, and lightweight variants like Whisper Tiny[1] often lack accuracy in challenging conditions, especially with domain-specific terminology. To bridge this gap, post-processing with language models (LLMs/SLMs) has emerged as a powerful error correction strategy (Ma et al., 2023; Yang et al., 2023; Jiang and Poellabauer, 2021), with domain-specific models showing particular promise for adapting to specialized vocabularies (Suh et al., 2024). Furthermore, optimizing SLMs via quantization techniques (Andreyev, 2025; Gholami et al., 2022) is critical for deployment on low-power devices. However, prior work has rarely focused on the combined application of these components: noise filtering, lightweight ASR, and quantized, domain-specific SLM correction within a offline pipeline (Rao et al., 2020; Kamahori et al., 2025). Our work bridges this gap by proposing and validating such a pipeline tailored for noise-robust, offline industrial applications.

3 BACKGROUND

3.1 The Targeted Task

Let S be the space of voice signals acquired during car body inspection, where each signal $s \in S$ is modeled as:

$$s = v + n \tag{1}$$

with v a spoken phrase containing a relevant term, and n background noise.

The goal is to find the correct transcription $\hat{t}$ of the term in v given an observed signal s via a function:

$$\hat{t} = f(s) \tag{2}$$

where $f : S \to T$ is a speech recognition and correction function, and T represents a finite set of domain-specific terms belonging to a *closed vocabulary*.

[1]https://huggingface.co/openai/whisper-tiny

Each phrase includes exactly one term from T, in either Italian or English:

- **English:** *positive bulge, negative bulge, waviness, deformation, reworking traces, crack, failure, scratches, glue residues, orange peel effect*
- **Italian:** *bollo positivo, bollo negativo, ondulazione, deformazione, tracce, buco, rottura, graffio, residuo di colla, effetto buccia d'arancia*

This is a single-label, multi-class classification task, where each input must be assigned to one class from the corresponding dictionary.

3.2 Material

Our pipeline leverages three key models, specifically optimized for an efficient workflow. For noise suppression, we employ DeepFilterNet[2] (Schröter et al., 2022), a speech enhancement framework. Initial transcription is performed by Whisper[3], using the optimized **Faster-Whisper**[4] implementation integrated with **WhisperX** (Bain et al., 2023) for fast inference and accurate alignment. For the final correction step, we use the lightweight SLM **Gemma 2 2B-it**[5]. To ensure its suitability for on-device deployment, we quantized the model to a 4-bit GGUF format[6] and ran it using **Llama.cpp**[7] with its Python bindings[8]. This setup significantly reduces the model's memory footprint, making it suitable for on-device inference.

4 THE PROPOSED PIPELINE

The pipeline operates in three steps. First, incoming audio, potentially containing background noise, is processed by a Denoiser Module based on DeepFilterNet to enhance speech clarity. Second, the filtered audio is passed to an ASR Module, which uses Whisper to generate an initial text transcription. Finally, this transcription is fed into a Corrector Module, which leverages a fine-tuned Gemma SLM to identify and extract the correct defect term from our predefined vocabulary, refining the final output.

[2]https://github.com/Rikorose/DeepFilterNet
[3]https://github.com/openai/whisper
[4]https://github.com/SYSTRAN/faster-whisper
[5]https://huggingface.co/google/gemma-2-2b-it
[6]https://huggingface.co/docs/hub/en/gguf
[7]https://github.com/ggml-org/llama.cpp
[8]https://github.com/abetlen/llama-cpp-python

5 DATASET

5.1 Training Set

To fine-tune the Gemma SLM as a corrector, we constructed a synthetic dataset to map potentially erroneous transcriptions to our domain-specific vocabulary. The process began with the generation of a set of clean base sentences using the Claude-Sonnet 3.5 LLM[9]. For each of the 10 English and 10 Italian defect terms (see Section 3), we prompted the LLM to create 30 unique example sentences relevant to a car body inspection context (e.g., *"I see a deformation in the hood of the car"*), ensuring each sentence contained only one vocabulary term. This resulted in 600 unique base sentences (300 per language), which then served as the foundation for introducing programmatic perturbations.

Programmatic Perturbation for Realism

To ensure the SLM corrector is robust to real-world ASR errors, its training data have to reflect such imperfections. Therefore, we programmatically perturbed our 600 clean base sentences to simulate typical transcription errors. With a custom Python script, we introduced a set of perturbations, including character substitutions (e.g., deformation -> defomation), character deletions (e.g., scratches -> scrtches), and incorrect word splitting/merging. For each clean sentence, we generated 12 perturbed variants with a degree of distortion, from single minor alterations to multiple, disruptive errors. This process yielded a final training set of 7,200 noisy sentences (3,600 per language), exposing the SLM to a wide spectrum of potential input corruptions.

Final Dataset Structure

The complete training dataset comprises 7,800 samples (3,900 per language), consisting of the 600 clean base sentences and their 7,200 perturbed variants. Each sample is an input-output pair for fine-tuning the SLM: the input is a (clean or perturbed) sentence (e.g., *"Inspeckt the dor panel..."*), and the output is the corresponding canonical defect term (e.g., deformation). This structure enables the SLM to learn robust mappings from a wide range of corrupted inputs back to the correct labels, equipping it to handle real-world ASR transcription errors.

[9]https://www.anthropic.com/news/claude-3-5-sonnet

5.2 Test Set

The test set has been constructed by generating 20 new, distinct sentences for each term in the two vocabularies using the same Claude-Sonnet 3.5 LLM employed for the training set base sentences, ensuring no overlap between training and test samples. These sentences were then converted into speech using Bark, a text-to-speech model[10]. We also publicly share the link containing the text files used for the training set used for fine-tuning the SLM corrector and the synthetic voice files of the test set for the evaluation step we performed[11].

6 EXPERIMENTAL EVALUATION

6.1 Fine-Tuning

We carried out the fine-tuning of the SLM using the structured prompt shown in Figure 1 that instructs the model to extract the correct defect term from a noisy input sentence, given the closed vocabulary. During training, each input sentence (clean or perturbed) and its target term were embedded into this prompt. During inference, the ASR's output is placed into the same template, and the SLM generates the corresponding vocabulary term. Fine-tuning was carried out on an L4 GPU using the Unsloth framework[12]. We employed the Quantization-aware Low-Rank Adaptation (QLoRA)(Dettmers et al., 2023) PEFT technique(Mangrulkar et al., 2022). The model was trained for one epoch with a learning rate of 2e-4, a weight decay of 0.001, and the AdamW 8-bit optimizer.

6.2 Noise Injection for Realistic Evaluation

To evaluate the pipeline's robustness, we mixed the clean test audio with real-world environmental noise (e.g., car interior recordings). The Signal-to-Noise Ratio (SNR) was controlled at two levels, representing moderate (5 dB) and high (-5 dB) interference, allowing for an accurate evaluation of performance under varying acoustic conditions.

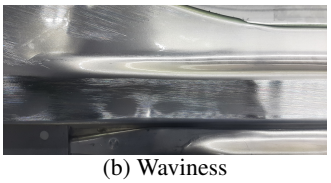
[10]https://huggingface.co/suno/bark

[11]https://github.com/Marcomurgia97/audios_NATTER_evaluation

[12]https://unsloth.ai/

Table 1: Examples of three defects.

(a) Deformation



(b) Waviness

(c) Glue Residue

Prompt

Given the transcribed voice phrase: *'he deformaion is mre pronunced towads th pane edges'* extract and return only the correct term from the following vocabulary:

[positive bulge, negative bulge, waviness, deformation, reworking traces, crack, failure, scratches, glue residues, orange peel effect]

Note: There may be transcription errors in the original phrase. Identify the vocabulary term that best matches the intended meaning in the phrase, correcting any transcription errors.

Return only the correct term, without any additional explanation.

Example: given the phrase *'The bidwark shows a dfomaton here'*, the correct term to return would be 'deformation'. It may also happen that there are words that are correct from a lexical point of view but inappropriate for the sentence. For example, given the phrase *'I can see a primitive pulse'*, the correct term to return would be 'positive bulge'.

Answer: 'Deformation'.

Figure 1: An example of a prompt used for fine-tuning with the target answer *Deformation*.

6.3 Evaluation Metrics and System Variants

We evaluate our pipeline as a single-label multi-class classification task, where each defect term (e.g., deformation, waviness, see Table 1) is a distinct class. To assess the impact of the SLM corrector, we compare the performance of the pipeline with and without it for both English and Italian. System performance is measured using standard like Precision, Recall, and F1-Score.

7 DISCUSSIONS

In this section, we analyze the pipeline's performance by comparing the baseline ASR output against the results obtained with our SLM corrector. To assess the performance-efficiency trade-off for resource-constrained environments, we evaluate two ASR model sizes: the lightweight Whisper Tiny (39M parameters) and the more powerful Whisper Large (1.5B parameters). The analysis is conducted for both Italian (Tables 2, 3, 4) and English (Tables 5, 6, 7) across three noise conditions: no added noise, SNR 5 dBW, and SNR -5 dBW. Performance is measured using Precision, Recall, and F1-Score, based on a strict exact match criterion (case and singular/plural forms are ignored).

The analysis of Italian language performance (Tables 2-4) reveals two key findings. First, the SLM corrector is crucial for the effectiveness of the lightweight Whisper Tiny model. In noise-free conditions, the SLM boosts Tiny's F1-score from a poor 53.49 to an excellent 95.54, primarily by raising its low Recall (42.50) to 95.50. This corrected performance surpasses that of the uncorrected Whisper Large (F1 91.64), establishing the Tiny+SLM combination as a computationally efficient alternative. Second, as noise increases, the impact of SLM becomes even more pronounced. Under moderate noise (SNR 5 dBW), it mitigates Tiny's performance drop, raising its F1 from 44.16 to 90.53. In the most challenging scenario (SNR -5 dBW), where the Tiny model is ineffective (F1 34.45), the SLM corrector restores system usability, achieving a robust F1-score of 87.04. Whisper Large results show a similar trend, whose F1-score is maintained at near-perfect levels by the SLM (99.50 at no noise, 97.99 at -5 dBW). In all conditions, the SLM contribution allows a significant re-

Table 2: Italian Recognition Performance - No Noise.

Metric	Global	Defect Type									
		Bollo Negativo	Bollo Positivo	Buccia Di Arancia	Buco	Deformazione	Graffio	Ondulazione	Residuo Di Colla	Rottura	Traccia
Size: Tiny											
F1 SLM	95.54	95.00	95.00	97.44	90.91	97.56	97.56	95.00	94.74	94.74	97.44
F1 Whisper	53.49	0.00	26.09	62.07	85.71	94.74	40.00	57.14	18.18	62.07	88.89
Precision SLM	95.88	95.00	95.00	100.00	83.33	95.24	95.24	95.00	100.00	100.00	100.00
Precision Whisper	90.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	95.50	95.00	95.00	95.00	100.00	100.00	100.00	95.00	90.00	90.00	95.00
Recall Whisper	42.50	0.00	15.00	45.00	75.00	90.00	25.00	40.00	10.00	45.00	80.00
Size: Large											
F1 SLM	99.50	100.00	100.00	100.00	97.56	100.00	100.00	100.00	97.44	100.00	100.00
F1 Whisper	91.64	85.71	85.71	91.89	97.44	97.44	85.71	85.71	97.44	91.89	97.44
Precision SLM	99.52	100.00	100.00	100.00	95.24	100.00	100.00	100.00	100.00	100.00	100.00
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	99.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	95.00	100.00	100.00
Recall Whisper	85.00	75.00	75.00	85.00	95.00	95.00	75.00	75.00	95.00	85.00	95.00

Table 3: Italian Recognition Performance - SNR 5.

Metric	Global	Defect Type									
		Bollo Negativo	Bollo Positivo	Buccia Di Arancia	Buco	Deformazione	Graffio	Ondulazione	Residuo Di Colla	Rottura	Traccia
Size: Tiny											
F1 SLM	90.53	92.31	90.91	92.68	79.17	97.56	91.89	82.35	85.71	97.44	95.24
F1 Whisper	44.16	0.00	0.00	62.07	78.79	88.89	40.00	46.15	0.00	40.00	85.71
Precision SLM	92.26	94.74	83.33	90.48	67.86	95.24	100.00	100.00	100.00	100.00	90.91
Precision Whisper	70.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	0.00	100.00	100.00
Recall SLM	90.50	90.00	100.00	95.00	95.00	100.00	85.00	70.00	75.00	95.00	100.00
Recall Whisper	34.50	0.00	0.00	45.00	65.00	80.00	25.00	30.00	0.00	25.00	75.00
Size: Large											
F1 SLM	98.24	97.56	100.00	100.00	92.68	100.00	100.00	100.00	97.44	94.74	100.00
F1 Whisper	89.88	85.71	82.35	88.89	94.74	97.44	85.71	85.71	97.44	88.89	91.89
Precision SLM	98.57	95.24	100.00	100.00	90.48	100.00	100.00	100.00	100.00	100.00	100.00
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	98.00	100.00	100.00	100.00	95.00	100.00	100.00	100.00	95.00	90.00	100.00
Recall Whisper	82.00	75.00	70.00	80.00	90.00	95.00	75.00	75.00	95.00	80.00	85.00

Table 4: Italian Recognition Performance - SNR -5.

Metric	Global	Defect Type									
		Bollo Negativo	Bollo Positivo	Buccia Di Arancia	Buco	Deformazione	Graffio	Ondulazione	Residuo Di Colla	Rottura	Traccia
Size: Tiny											
F1 SLM	87.04	92.68	87.18	82.05	70.83	100.00	87.80	88.89	85.71	92.31	82.93
F1 Whisper	34.45	0.00	0.00	26.09	51.85	88.89	40.00	33.33	0.00	33.33	70.97
Precision SLM	88.63	90.48	89.47	84.21	60.71	100.00	85.71	100.00	100.00	94.74	80.95
Precision Whisper	70.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	0.00	100.00	100.00
Recall SLM	86.50	95.00	85.00	80.00	85.00	100.00	90.00	80.00	75.00	90.00	85.00
Recall Whisper	25.00	0.00	0.00	15.00	35.00	80.00	25.00	20.00	0.00	20.00	55.00
Size: Large											
F1 SLM	97.99	100.00	97.56	100.00	95.00	100.00	97.56	100.00	100.00	94.74	95.00
F1 Whisper	77.88	62.07	57.14	62.07	66.67	97.44	85.71	75.00	91.89	88.89	91.89
Precision SLM	98.05	100.00	95.24	100.00	95.00	100.00	95.24	100.00	100.00	100.00	95.00
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	98.00	100.00	100.00	100.00	95.00	100.00	100.00	100.00	100.00	90.00	95.00
Recall Whisper	66.00	45.00	40.00	45.00	50.00	95.00	75.00	60.00	85.00	80.00	85.00

covery of the Recall, preserving system functionality even in severe noise.

The results for English (Tables 5-7) mirror the trends observed for Italian, confirming the SLM corrector's critical role. The baseline Whisper Tiny starts with a higher F1-score than its Italian counterpart (72.06 in no-noise conditions), but is still hampered by low Recall (61.50). The SLM corrector el-

Table 5: English Recognition Performance - No Noise.

Metric	Global	Defect Type									
		Negative Bulge	Positive Bulge	Orange Peel Effect	Crack	Deformation	Scratch	Waviness	Glue Residue	Failure	Reworking Trace
Size: Tiny											
F1 SLM	96.01	95.24	90.48	92.68	97.44	100.00	100.00	94.74	94.74	94.74	100.00
F1 Whisper	72.06	57.14	40.00	62.07	91.89	100.00	100.00	40.00	51.85	91.89	85.71
Precision SLM	96.78	90.91	86.36	90.48	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	95.50	100.00	95.00	95.00	95.00	100.00	100.00	90.00	90.00	90.00	100.00
Recall Whisper	61.50	40.00	25.00	45.00	85.00	100.00	100.00	25.00	35.00	85.00	75.00
Size: Large											
F1 SLM	99.00	100.00	100.00	100.00	95.00	100.00	100.00	100.00	97.44	100.00	97.56
F1 Whisper	88.54	62.07	78.79	94.74	94.74	97.44	100.00	91.89	70.97	100.00	94.74
Precision SLM	99.02	100.00	100.00	100.00	95.00	100.00	100.00	100.00	100.00	100.00	95.24
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	99.00	100.00	100.00	100.00	95.00	100.00	100.00	100.00	95.00	100.00	100.00
Recall Whisper	81.50	45.00	65.00	90.00	90.00	95.00	100.00	85.00	55.00	100.00	90.00

Table 6: English Recognition Performance - SNR 5.

Metric	Global	Defect Type									
		Negative Bulge	Positive Bulge	Orange Peel Effect	Crack	Deformation	Scratch	Waviness	Glue Residue	Failure	Reworking Trace
Size: Tiny											
F1 SLM	94.22	95.24	92.68	95.24	88.37	97.44	95.00	97.44	88.89	91.89	100.00
F1 Whisper	68.56	46.15	40.00	57.14	82.35	100.00	97.44	40.00	51.85	91.89	78.79
Precision SLM	94.99	90.91	90.48	90.91	82.61	100.00	95.00	100.00	100.00	100.00	100.00
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	94.00	100.00	95.00	100.00	95.00	95.00	95.00	95.00	80.00	85.00	100.00
Recall Whisper	57.00	30.00	25.00	40.00	70.00	100.00	95.00	25.00	35.00	85.00	65.00
Size: Large											
F1 SLM	98.00	100.00	97.56	97.44	92.68	100.00	100.00	100.00	94.74	100.00	97.56
F1 Whisper	84.91	51.85	66.67	88.89	94.74	97.44	94.74	91.89	70.97	100.00	91.89
Precision SLM	98.10	100.00	95.24	100.00	90.48	100.00	100.00	100.00	100.00	100.00	95.24
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	98.00	100.00	100.00	95.00	95.00	100.00	100.00	100.00	90.00	100.00	100.00
Recall Whisper	76.50	35.00	50.00	80.00	90.00	95.00	90.00	85.00	55.00	100.00	85.00

Table 7: English Recognition Performance - SNR -5.

Metric	Global	Defect Type									
		Negative Bulge	Positive Bulge	Orange Peel Effect	Crack	Deformation	Scratch	Waviness	Glue Residue	Failure	Reworking Trace
Size: Tiny											
F1 SLM	91.25	97.56	88.89	95.24	85.71	100.00	94.74	97.44	85.71	78.79	88.37
F1 Whisper	52.83	40.00	18.18	51.85	62.07	97.44	82.35	18.18	33.33	78.79	46.15
Precision SLM	93.06	95.24	80.00	90.91	81.82	100.00	100.00	100.00	100.00	100.00	82.61
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	91.00	100.00	100.00	100.00	90.00	100.00	90.00	95.00	75.00	65.00	95.00
Recall Whisper	40.50	25.00	10.00	35.00	45.00	95.00	70.00	10.00	20.00	65.00	30.00
Size: Large											
F1 SLM	95.98	95.00	93.02	100.00	92.68	100.00	97.44	97.44	91.89	92.31	100.00
F1 Whisper	73.42	40.00	40.00	66.67	88.89	97.44	91.89	82.35	46.15	91.89	88.89
Precision SLM	96.72	95.00	86.96	100.00	90.48	100.00	100.00	100.00	100.00	94.74	100.00
Precision Whisper	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Recall SLM	95.50	95.00	100.00	100.00	95.00	100.00	95.00	95.00	85.00	90.00	100.00
Recall Whisper	62.50	25.00	25.00	50.00	80.00	95.00	85.00	70.00	30.00	85.00	80.00

evates its performance to a competitive 96.01, once again making the Tiny+SLM setup a powerful, low-resource alternative to the uncorrected Whisper Large (F1 88.54). As noise levels increase, the SLM's corrective power ensures system robustness. Under moderate noise (SNR 5 dBW), it lifts Tiny's F1-score from 68.56 to 94.22. In the high-noise scenario (SNR -5 dBW), where the uncorrected performance of Tiny

degrades significantly (F1 52.83), the SLM provides a substantial boost to 91.25, effectively mitigating the noise-induced errors. For Whisper Large, the SLM consistently maintains near-perfect accuracy, raising its F1 from 73.42 to 95.98 in the -5 dBW condition. As with Italian, the primary mechanism for this improvement is a significant recovery in Recall (e.g., for Tiny at -5 dBW, from 40.50 to 91.00), confirming that the Tiny+SLM combination is a reliable and efficient solution even under adverse conditions.

Our analysis consistently reveals three key trends across both languages and all noise levels. First, as expected, ASR performance degrades with increasing noise. Second, the SLM corrector significantly enhances performance in all conditions, especially noisy ones. Third, and most importantly, the SLM narrows the performance gap between Whisper Tiny and Whisper Large, establishing the Tiny+SLM combination as a computationally efficient alternative. For instance, at SNR -5 dBW, the corrected Tiny model processed samples in just 0.24 seconds, compared to 0.79 seconds for the corrected Large variant on our test system (Intel i7, RTX 3070). This significant reduction in computational implies also lower energy consumption, a critical factor for battery-powered devices on the factory floor. While our work establishes a strong proof-of-concept, future work will address further real-world deployment challenges, including on-device integration and user interface design.

Our choice to use an SLM for post-correction, rather than fine-tuning the ASR model itself, is a pragmatic one. Fine-tuning an ASR model requires extensive and costly domain-specific audio data collection. Our approach, which leverages a pre-trained ASR and focuses adaptation on an SLM trained with easily generated synthetic text, is far more data-efficient. The strong performance gains, especially with Whisper Tiny, validate this strategy. This modularity also allows for rapid adaptation to new domains by simply retraining the SLM corrector with a new text dataset, a much simpler task than acquiring new audio recordings. An interesting phenomenon was observed: for some terms, the F1-score was higher under severe noise (SNR -5 dBW) than moderate noise (SNR 5 dBW). This counterintuitive result likely stems from the nature of the ASR errors. At -5 dBW, the highly corrupted output may paradoxically produce error patterns that more closely match the synthetic perturbations in the SLM's training data (see Section 5), enabling more effective correction. Conversely, errors at 5 dBW, though milder, might be of a type less represented in our training set, thus limiting the SLM's corrective ability. This highlights the critical role of the training data's composition in the SLM's effectiveness at mitigating specific types of noise-induced errors.

8 CONCLUSIONS

We presented a three-stage pipeline for robust speech transcription in noisy industrial settings, combining noise filtering (DeepFilterNet), ASR (Whisper), and a post-correction module. The core of our contribution is a lightweight fine-tuned SLM (Gemma 2 2B-it) that refines ASR outputs based on a domain-specific defect ontology. Our results show that this SLM corrector substantially improves F1-scores across all tested conditions, mitigating noise-induced transcription errors. It enables the lightweight Whisper Tiny to achieve performance comparable to the much larger Whisper Large variant, confirming the pipeline's suitability for offline deployment on resource-limited hardware. The evaluation was conducted on a synthetic dataset, which, while allowing for rigorous testing, does not capture the full complexity of real-world scenarios. Key limitations include a lack of speaker variability (e.g., accents, speaking rates) and environmental noise diversity (e.g., impulsive sounds, overlapping speech). The strong performance observed therefore establishes a promising proof-of-concept. Our planned mitigation strategy is to conduct further real-data validation in collaboration with our industrial partner. This important next step, involving the collection and testing of on-field data, will be essential for bridging the gap between our results and a practical deployment.

REFERENCES

Andreyev, A. (2025). Quantization for openai's whisper models: A comparative analysis.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, pages 4489–4493.

Bandela, S. R., Sharma Sadhu, S., Rathore, V. S., and Jagini, S. K. (2023). Development of noise robust automatic speech recognition system. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Bodepudi, A., Reddy, M., Gutlapalli, S., and Mandapuram, M. (2019). Voice recognition systems in the cloud networks: Has it reached its full potential? *Asian Journal of Applied Science and Engineering*, 8:51–60.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in*

Neural Information Processing Systems, volume 36, pages 10088–10115. Curran Associates, Inc.

Dua, M., Akanksha, and Dua, S. (2023). Noise robust automatic speech recognition: review and analysis. *Int. J. Speech Technol.*, 26(2):475–519.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121.

G., T. Y., B.G., N., and Jayanna, H. (2024). Development of noise robust real time automatic speech recognition system for kannada language/dialects. *Engineering Applications of Artificial Intelligence*, 135:108693.

Gholami, A., Kim, S., Zhen, D., Yao, Z., Mahoney, M., and Keutzer, K. (2022). *A Survey of Quantization Methods for Efficient Neural Network Inference*, pages 291–326.

Gomez, R. and Kawahara, T. (2010). Optimizing spectral subtraction and wiener filtering for robust speech recognition in reverberant and noisy conditions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4566–4569.

Huang, X., Baker, J., and Reddy, R. (2014). A historical perspective of speech recognition. *Commun. ACM*, 57(1):94–103.

Jiang, Y. and Poellabauer, C. (2021). A sequence-to-sequence based error correction model for medical automatic speech recognition. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3029–3035.

Kamahori, K., Kasai, J., Kojima, N., and Kasikci, B. (2025). Liteasr: Efficient automatic speech recognition with low-rank approximation.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777.

Ma, R., Qian, M., Manakul, P., Gales, M., and Knill, K. (2023). Can generative large language models perform asr error correction?

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. (2022). Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., and Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869.

Moore, A., Peso Parada, P., and Naylor, P. (2017). Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures. *Computer Speech & Language*, 46:574–584.

Orel, D. and Varol, H. A. (2023). Noise-robust automatic speech recognition for industrial and urban environments. In *IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

Rao, M., Raju, A., Dheram, P., Bui, B., and Rastrow, A. (2020). Speech to semantics: Improve asr and nlu jointly via all-neural interfaces. In *Interspeech 2020*, page 876–880. ISCA.

Schröter, H., Escalante-B., A. N., Rosenkranz, T., and Maier, A. (2022). DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering. In *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Suh, J., Na, I., and Jung, W. (2024). Improving domain-specific asr with llm-generated contextual descriptions. In *Proc. Interspeech 2024*, pages 1255–1259.

Virtanen, T., Plumbley, M. D., and Ellis, D. (2017). *Computational Analysis of Sound Scenes and Events*. Springer Publishing Company, Incorporated, 1st edition.

Yang, C.-H. H., Gu, Y., Liu, Y. C., Ghosh, S., Bulyko, I., and Stolcke, A. (2023). Generative speech recognition error correction with large language models and task-activating prompting.

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E.-D., Jin, W., and Schuller, B. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Trans. Intell. Syst. Technol.*, 9(5).