Rethinking the Role of Environmental Sounds as Situational Context for Speech Recognition

Debrina Veisha Rashika[1], *Bagus Tris Atmaja*[2], *Sakriani Sakti*[2], *Dessi Puji Lestari*[1]

[1] Institut Teknologi Bandung, Indonesia
[2] Nara Institute of Science and Technology, Japan

13522025@std.stei.itb.ac.id, bagus.tris@naist.jp, ssakti@is.naist.jp, dessipuji@staff.stei.itb.ac.id

Abstract

Environmental sounds are typically treated as noise in automatic speech recognition (ASR), although they also provide situational context that may improve recognition. We propose a hierarchical multitask framework that jointly performs Acoustic Scene Classification (ASC) and ASR, using learned environmental embeddings with confidence-gated Feature-wise Linear Modulation (FiLM) to incorporate environmental context into ASR. Experiments on noisy speech demonstrate that leveraging environmental context consistently improves recognition accuracy across diverse acoustic scenes and noise conditions. Ablation studies show that most of the improvement arises from joint multitask learning, while explicit FiLM conditioning provides only a modest additional benefit. We further conduct an exploratory LLM-based evaluation to examine whether predicted acoustic scene information benefits downstream spoken language understanding. Overall, the results suggest that environmental sounds should be viewed not merely as interference, but as contextual signals that benefit both speech recognition and downstream spoken language understanding.

Index Terms: Automatic Speech Recognition, Environmental Context, Acoustic Scene Classification, Multitask Learning, Spoken Language Understanding, Large Language Models.

1. Introduction

Automatic Speech Recognition (ASR) has become a fundamental component of many real-world applications, including virtual assistants, conversational agents, and automatic transcription systems [1]. Recent advances in self-supervised speech representation learning and transformer-based architectures, such as wav2vec 2.0 [2], HuBERT [3], and Whisper [4], have substantially improved recognition performance across diverse speech recognition tasks.

Despite these advances, ASR systems remain vulnerable to environmental noise. Conventional approaches therefore treat background sounds primarily as interference to be suppressed or removed before recognition [5]. While this often improves intelligibility, it may also discard information about the speaker's surroundings that could be useful for recognition [6]. Environmental sounds naturally encode situational context—for example, recordings from a café, metro station, or bus provide cues about where an utterance occurs—and prior work has shown that contextual information can improve recognition in location-dependent settings [7]. However, environmental sounds themselves are rarely modeled as contextual information for ASR.

Previous work has explored environmental representations for speech enhancement [8], contextual awareness in dialogue recognition [9], and multitask learning for robust ASR [10, 11]. Feature-wise Linear Modulation (FiLM) has also emerged as an effective mechanism for conditioning neural representations on auxiliary information [12]. However, prior studies typically combine FiLM with the auxiliary task that produces the conditioning signal, making it difficult to distinguish the benefit of explicit feature conditioning from that of joint multitask representation learning [13].

To address this question, we propose a hierarchical multitask framework that jointly performs Acoustic Scene Classification (ASC) and ASR while conditioning the ASR branch on learned environmental embeddings through confidence-gated FiLM. We further perform controlled ablation experiments to disentangle the contributions of multitask learning and FiLM-based conditioning.

Beyond ASR itself, we also investigate whether environmental context benefits downstream spoken language understanding. As ASR increasingly serves as the front-end for language models, transcription alone may provide insufficient situational information for accurate interpretation [14]. We therefore conduct an exploratory LLM-as-Judge evaluation [15], in which a Qwen2.5 instruction-tuned model [16] interprets ASR transcripts with and without predicted acoustic scene labels, and a second instance evaluates whether scene grounding improves interpretation quality.

The main contributions of this work are:

- We propose a hierarchical multitask framework that jointly performs Acoustic Scene Classification (ASC) and Automatic Speech Recognition (ASR), incorporating environmental context through confidence-gated FiLM conditioning.
- We provide a controlled ablation study showing the relative contributions of joint multitask learning and explicit FiLM-based conditioning for context-aware ASR.
- We conduct an exploratory LLM-based evaluation to investigate whether predicted acoustic scene information improves downstream interpretation of ASR transcripts.
- We present comprehensive experiments across diverse acoustic scenes and SNR conditions, together with embedding visualizations, to characterize when and how environmental context benefits ASR.

2. Related Work

Noise-robust ASR. The dominant strategy for handling environmental noise in ASR is suppression, where speech enhancement or denoising front-ends are applied prior to recognition [5]. Although these approaches often improve intelligibility, recent studies have shown that denoising can also remove acoustic information that is beneficial for downstream recognition, particularly in domain-specific applications [6]. An alterna-

tive direction improves robustness through data augmentation, including masking-based methods such as SpecAugment [17] and interpolation-based approaches such as Mixup [18]. While these methods improve robustness to acoustic variability, they continue to treat environmental sounds as interference rather than as a potentially informative signal. In contrast, our work explicitly models environmental sounds as situational context that can complement speech recognition.

Context-conditioned ASR. Several studies have incorporated contextual information into ASR using external signals such as location metadata [7]. Related work has also explored environmental representations for dialogue recognition [9] and adaptive speech enhancement [8]. However, these approaches rely on external metadata or optimize representations for enhancement rather than learning contextual information jointly from the input audio itself. To the best of our knowledge, prior work has not explicitly distinguished the contribution of auxiliary multitask supervision from that of the conditioning mechanism used to inject contextual information into the ASR model. This distinction is important for understanding how environmental context should be integrated into robust speech recognition systems.

FiLM and multitask conditioning in speech. Feature-wise Linear Modulation (FiLM) [12] provides a general mechanism for conditioning neural representations using auxiliary embeddings. In parallel, multitask learning has become a widely adopted strategy for improving ASR robustness by jointly optimizing auxiliary objectives, including phonetic supervision [10] and self-supervised pretext tasks [11]. These auxiliary tasks primarily act as regularizers for the shared encoder rather than providing an explicit conditioning signal to the recognition branch. Acoustic scene classification has similarly been investigated as an auxiliary task within DCASE research [19], but not as a learned conditioning signal for ASR. Building on these ideas, our framework jointly learns environmental representations through acoustic scene classification and incorporates them into ASR via confidence-gated FiLM. More importantly, our controlled ablation study explicitly disentangles the contribution of joint multitask learning from that of FiLM-based conditioning, providing insight into how contextual information benefits ASR.

ASR as a front-end for language understanding. ASR is increasingly deployed as the front-end of spoken language understanding systems, where recognition outputs are consumed by downstream language models rather than read directly by humans. Consequently, recent work has investigated how transcription errors influence downstream reasoning and demonstrated that incorporating richer ASR hypotheses can improve robustness for tasks such as question answering and intent classification [14]. However, relatively little attention has been paid to whether situational context accompanying the speech signal can further improve downstream interpretation. To explore this question, we adopt the LLM-as-Judge paradigm [15] as an exploratory evaluation protocol. Specifically, we examine whether providing predicted acoustic scene information alongside ASR transcripts influences how a language model interprets speaker intent. This analysis is intended as a secondary investigation of the potential downstream value of environmental context rather than as the primary contribution of this work.



Figure 1: *Overview of the proposed context-aware ASR framework.*

3. Proposed Method

3.1. Overview

Figure 1 illustrates the overall architecture. Given a noisy speech signal, a pretrained wav2vec 2.0 model [2] is used as a shared feature extractor. The model consists of a convolutional feature encoder followed by a Transformer-based [21] contextual encoder, providing both local acoustic patterns and long-range contextual information.

The convolutional representations extracted by the wav2vec 2.0 feature encoder are processed by the Acoustic Scene Classification (ASC) branch, while the contextualized Transformer representations are used by the ASR branch. The ASC branch produces both scene-classification logits and a 128-dimensional environmental embedding that summarizes the acoustic environment. This embedding conditions the ASR branch through Feature-wise Linear Modulation (FiLM) [12], while the predicted scene label is also provided to the downstream LLM-based interpretation stage described in Section 3.5.

3.2. Acoustic Scene Classification

The ASC branch extracts contextual information from environmental sounds. Convolutional features from wav2vec 2.0 are processed by a multi-scale convolutional network capturing acoustic patterns at different temporal resolutions, and aggregated into an utterance-level representation via attentive statistical pooling [20], which computes weighted statistics across temporal frames so the model can focus on informative regions while suppressing irrelevant background variation. The resulting representation produces two outputs: scene-classification logits for ASC supervision, and a 128-dimensional environmen-

tal embedding used for conditioning, rather than the discrete scene label itself, so that richer, continuous environmental information is available to the ASR branch.

3.3. Automatic Speech Recognition

Contextualized transformer representations from wav2vec 2.0 are projected through a convolutional layer and processed by a stack of Bidirectional GRUs [22], capturing long-range temporal dependencies in both directions. Transcription is produced via a CTC decoder [23], and a 4-gram KenLM language model [24] is incorporated through beam-search decoding at inference time to improve linguistic consistency, particularly under noisy conditions.

3.4. Context Integration via FiLM

The ASC embedding is integrated into the ASR branch via FiLM [12]. Given an environmental embedding c and an ASR feature representation x,

$$y = \gamma(c) \odot x + \beta(c), \tag{1}$$

where $\gamma(c)$ and $\beta(c)$ are learnable transformations of c and $\odot$ is element-wise multiplication.

To prevent unreliable scene predictions from degrading ASR performance, we extend FiLM with confidence gating (SceneGatedFiLM). Given the ASC branch's classification confidence $\hat{p} = \max(\mathrm{softmax}(\ell))$, where ℓ denotes the ASC logits, a binary gate is computed as

$$g = \mathbb{1}[\hat{p} > \tau], \tag{2}$$

and the final conditioned representation is

$$y = g \cdot (\gamma(c) \odot x + \beta(c)) + (1 - g) \cdot x, \tag{3}$$

with confidence threshold $\tau = 0.50$. Below this threshold, FiLM collapses to the identity ($y = x$), so low-confidence scene predictions do not modulate ASR representations. γ and β are identity-initialized ($\gamma_0 = 1$, $\beta_0 = 0$), so training begins from an unconditioned baseline and conditioning is learned as a refinement rather than an arbitrary transformation from initialization.

3.5. LLM-based Interpretation Evaluation

To probe whether environmental context is useful beyond the ASR objective itself, we couple the framework above with an exploratory downstream evaluation stage. Given a test utterance, we take the ASR transcript together with the scene label predicted by the ASC branch (Section III-B) and pass them to a Qwen2.5-0.5B-Instruct model [16], which is prompted to produce a short natural-language interpretation of what the speaker most likely meant or intended, under three conditions: (A) transcript only, (B) transcript with the predicted acoustic scene label, and (C) transcript with the predicted scene label plus a chain-of-thought instruction [25]. A separate instance of the same model is then used as an automatic judge, following the LLM-as-Judge protocol [15], to score each interpretation and to select the best of the three conditions per sample. Full scoring rubric and dataset details are given in Section 4.

4. Experimental Setup

A noisy speech dataset was constructed by mixing LibriSpeech utterances [26] with environmental recordings from the TAU Urban Acoustic Scenes dataset [27] at five SNR levels (-5, 0, 5, 10, and 15 dB), spanning conditions from severely degraded to relatively clean speech. The proposed framework was evaluated across 11 acoustic scene categories covering indoor, outdoor, and transportation environments. Speech utterances and environmental recordings were randomly paired to ensure that scene identity and linguistic content remained statistically independent, thereby preventing shortcut learning between lexical content and acoustic scene. Character Error Rate (CER) was used as the primary ASR evaluation metric, while scene classification accuracy was reported for the ASC branch.

Table 1: *Training Hyperparameters*

Parameter	Value
Optimizer	AdamW
ASR Learning Rate	3×10^{-4}
ASC Learning Rate	1×10^{-4}
Weight Decay	1×10^{-4}
Batch Size	16
Gradient Clipping	1.0
ASC Dropout (pre-classifier / pre-embedding)	0.45 / 0.35
MTL Loss Label Smoothing	0.15
FiLM Confidence Threshold (τ)	0.50
Environmental Embedding Dimension	128
ASC Freeze Period	First 5 epochs

Table 2: *Dataset Statistics*

Split	Utterances
Training	16,624
Validation	1,934
Test	1,921

To ensure a fair comparison, the baseline model is architecturally identical to the proposed framework except that the ASC branch, environmental embedding, and FiLM conditioning are removed. Both models share the same pretrained wav2vec 2.0 encoder, convolutional downsampling, BiGRU configuration, CTC decoder/vocabulary, KenLM decoding settings, AdamW optimizer [28], learning rates, weight decay, and gradient clipping (Table 1). The multitask model additionally applies label smoothing [29] to the ASC objective. This controlled design isolates the contributions of the auxiliary ASC objective and FiLM conditioning rather than incidental architectural or optimization differences.

The test set consists of 1,921 unique utterances, each evaluated once per SNR condition (-5, 0, 5, 10, 15 dB), giving $1{,}921 \times 5 = 9{,}605$ utterance-SNR evaluation pairs in total. In addition to ASR performance, we investigated whether the benefits of environmental context extend to downstream language understanding, using the LLM-based interpretation protocol described in Section 3.5.

4.1. LLM-based Interpretation Evaluation Protocol

Each of the four dimensions in Table 7 (transcription quality, interpretation quality, hallucination risk, overall score) was scored on a 1–5 scale, where higher indicates better quality and, for hallucination risk, higher indicates fewer unsupported details. A total of 200 test samples were randomly selected for evaluation using the three prompting conditions (A, B, C) and judge model introduced in Section 3.5.

Table 3: *ASR Performance At Different SNR Levels (Averaged Across All Acoustic Scenes)*

SNR (dB)	**ASC** Acc.	**Baseline** CER	**HC-MTL** CER	**ΔCER**
-5	56.27	48.12	**48.10**	-0.02
0	55.34	33.76	**32.33**	-1.43
5	55.23	21.75	**20.59**	-1.16
10	53.51	17.59	**16.73**	-0.86
15	50.91	16.35	**15.83**	-0.52
Avg.	54.25	27.51	**26.72**	-0.80

Table 4: *ASR Performance Across Acoustic Scenes (Averaged Across All SNR Levels)*

Scene	**ASC** Acc.	**Baseline** CER	**HC-MTL** CER	**ΔCER**
Beach	58.02	36.11	**34.86**	-1.25
Bus	46.36	20.60	**20.29**	-0.31
Cafe/Restaurant	51.19	33.06	**31.78**	-1.28
Car	60.82	18.70	**18.23**	-0.47
City Center	**89.94**	30.64	**29.04**	**-1.59**
Forest Path	77.86	30.18	**29.02**	-1.16
Grocery Store	39.10	27.64	**26.46**	-1.18
Home	47.36	**30.03**	30.18	+0.14
Metro Station	42.47	33.59	**32.49**	-1.10
Train	56.24	18.57	**18.52**	-0.05
Tram	27.47	23.92	**23.38**	-0.55
Avg.	54.26	27.55	**26.75**	-0.80

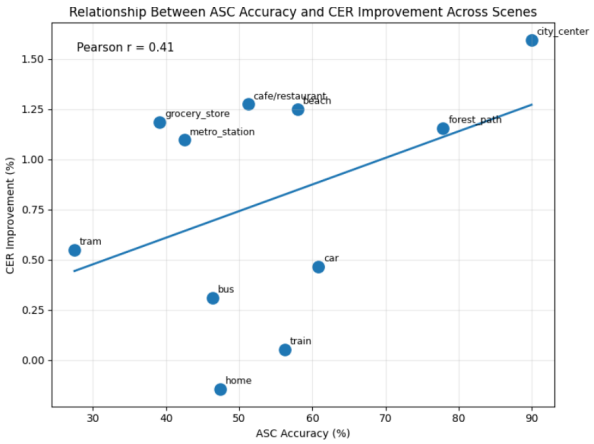

Figure 2: *ASC accuracy vs. CER improvement across acoustic scenes, computed directly from the per-utterance M1 (baseline) vs. M4 (proposed) ablation results in Section 5.3.*

5. Results and Discussion

5.1. Performance Across SNR Conditions

Table 3 summarizes ASR performance across SNR conditions. As expected, CER decreased as SNR increased for both models. HC-MTL consistently matched or outperformed the baseline across all SNR levels, although the magnitude of improvement varied substantially. The largest gain was observed at 0 dB (-1.43 percentage points), followed by progressively smaller improvements at 5 dB (-1.16), 10 dB (-0.86), and 15 dB (-0.52), while only a negligible improvement was obtained at -5 dB (-0.02). This pattern suggests that environmental context is most beneficial under moderate noise conditions, where both speech and environmental cues remain sufficiently informative. Under severe noise, speech degradation limits the usefulness of contextual information, whereas under high SNR the baseline already performs well, leaving less opportunity for improvement.

In contrast, ASC accuracy decreased from 56.27% at -5 dB to 50.91% at 15 dB. This trend suggests that acoustic scene classification becomes more challenging as speech increasingly dominates the mixture, reducing the relative contribution of environmental sounds.

5.2. Performance Across Acoustic Scenes

Table 4 summarizes performance across acoustic scenes. HC-MTL reduced CER in 10 of the 11 evaluated scenes, with *home* being the only exception (+0.14 percentage points). This result indicates that environmental conditioning is broadly beneficial but not uniformly effective across acoustic environments. The limited improvement in the *home* category may reflect the relatively homogeneous acoustic characteristics of quiet indoor environments, making their environmental embeddings less distinctive than those of acoustically richer scenes such as *metro_station* or *tram*. Consistent with this observation, the ASC accuracy for *home* (47.36%) was among the lower-performing categories.

ASR improvement was not strongly associated with ASC classification accuracy across individual scenes. For example, *metro_station* achieved relatively low ASC accuracy (42.47%) yet reduced CER by 1.10 percentage points, whereas *train* achieved moderate ASC accuracy (56.24%) but showed almost no CER improvement (-0.05 pp). Figure 2 illustrates this relationship. Although the Pearson correlation was positive ($r = 0.41$), it was not statistically significant ($p = 0.21$, $n = 11$), suggesting that scene classification accuracy alone does not fully account for the observed ASR improvements.

These findings suggest that the learned environmental embedding captures information beyond the discrete scene label. Even when scene classification is imperfect, the embedding may retain acoustic characteristics, such as background spectral patterns and reverberation, that remain useful for ASR. Consequently, the effectiveness of the environmental representation may depend more on the quality of the learned embedding than on scene classification accuracy alone. This interpretation is consistent with the ablation results (Section 5.3), which indicate that joint multitask learning, rather than FiLM-based conditioning itself, is the primary contributor to the observed performance gains.

Transportation scenes such as *car* and *train* exhibited relatively low baseline CER despite only moderate ASC accuracy, whereas scenes such as *beach* and *cafe/restaurant* remained more challenging for ASR. This difference may reflect the extent to which environmental sounds overlap spectrally with speech, although a more detailed acoustic analysis is beyond the scope of this work.

To assess whether the observed improvement was broadly distributed across the evaluation set, we computed the CER difference for each utterance–SNR pair ($n = 9{,}605$). HC-MTL improved CER for 48.9% of the pairs, degraded CER for 41.8%, and produced no change for the remaining 9.2%. These results indicate that the overall improvement reflects a broad, albeit non-uniform, effect, with gains observed for nearly half of the evaluated samples rather than being driven by only a small sub-

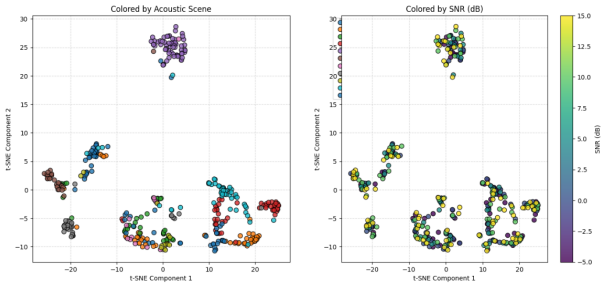

Figure 3: *t-SNE [30] visualization of the 128-dimensional ASC embeddings extracted from the test set: colored by ground-truth acoustic scene (left) and by SNR level (right).*

Table 5: *Ablation Study Results*

Model	**CER (%)**	**ASC Acc. (%)**
M1 (Baseline, no ASC)	27.51	–
M2 (ASC+MTL loss, no FiLM)	26.76	53.41
M3 (Frozen ASC, FiLM, no MTL grad.)	27.93	**56.60**
M4 (Proposed, fully joint)	**26.72**	54.25

set of utterances.

Fig. 3 (left) shows the ASC embeddings colored by ground-truth acoustic scene. The embeddings form partially separable clusters, with some scenes exhibiting clear grouping while others remain intermixed, consistent with the model's moderate ASC accuracy (53–56%) on the test set. Fig. 3 (right) shows the same embeddings colored by SNR level. In contrast to the scene labels, different SNR levels are distributed across the same embedding space rather than forming distinct clusters, suggesting that the learned representation is organized primarily by acoustic scene characteristics rather than noise level.

5.3. Ablation Study

To disentangle the contributions of joint multitask learning and explicit FiLM-based conditioning, we compared four model configurations: M1 (baseline without the ASC branch), M2 (joint ASC–ASR multitask learning without FiLM), M3 (FiLM conditioning using a frozen pretrained ASC encoder), and M4 (the full proposed HC-MTL framework).

Adding the auxiliary ASC objective without FiLM conditioning (M2) recovers nearly the entire improvement over the baseline (CER 27.51 $\rightarrow$ 26.76), consistent with multitask learning acting as a regularizer for the shared wav2vec 2.0 encoder [13]. In contrast, applying FiLM conditioning using a frozen, separately pretrained ASC encoder (M3) degrades performance relative to the baseline (CER 27.51 $\rightarrow$ 27.93), indicating that environmental embeddings are ineffective when learned independently of the ASR objective. The full joint model (M4) achieves the best overall performance (CER 26.72), providing a modest additional improvement over M2.

Pairwise Wilcoxon signed-rank tests [31] on the per-utterance CER differences ($n = 9{,}605$) showed that each adjacent comparison was statistically significant: M1 vs. M2 ($p = 5.93\times10^{-18}$), M2 vs. M3 ($p = 1.97\times10^{-71}$), M3 vs. M4 ($p = 2.47\times10^{-78}$), and M1 vs. M4 ($p = 4.44\times10^{-21}$). Nevertheless, the absolute difference between M2 and M4 was only 0.04 CER points, indicating that statistical significance should not be interpreted as a practically meaningful improvement.

Taken together, these results indicate that the primary benefit of the proposed framework arises from joint multitask learning rather than FiLM-based conditioning alone. Simply conditioning the ASR branch with environmental embeddings (M3) is insufficient and even degrades performance when those embeddings are learned independently of the ASR objective. In contrast, jointly optimizing ASC and ASR (M2) recovers nearly the full performance gain, while FiLM provides only a modest additional improvement when trained end-to-end (M4). These findings suggest that environmental context is most effective when it shapes shared speech representations during training, rather than serving solely as an external conditioning signal during inference.

Table 6: *Example Interpretation Outputs Across Conditions*

Input	
Transcript	"What was that"
Label	home
Condition	**Generated Interpretation**
A	The speaker is most likely asking for clarification or confirmation about something previously mentioned.
B	The speaker is likely asking about something they heard in the home, as if they were unsure what just occurred there.
C	The speaker is likely asking about something they heard or observed in the home environment, as the acoustic scene suggests a domestic setting where such a query might naturally arise.

Table 7: *LLM-as-Judge Evaluation Results (n=200)*

Dimension	Mean	Std. Dev.
Transcription Quality	2.48	0.61
Interpretation Quality	3.38	0.68
Hallucination Risk	3.47	1.08
Overall Score	3.11	0.79

Table 8: *Best Interpretation Condition Selected by the Judge*

Condition	**Value**
A: Transcript Only	72 (36.0%)
B: Transcript + Scene	99 (49.5%)
C: Transcript + Scene + CoT	29 (14.5%)

5.4. Impact on Downstream Interpretation

Beyond ASR performance, we investigated whether environmental context also benefits downstream language understanding using the LLM-as-Judge protocol described in Section 3.5. Table 6 presents a representative example under the three prompting conditions. Without scene information (Condition A), the interpretation relies solely on the transcript and remains context-independent. Providing the predicted acoustic scene label (Condition B) grounds the interpretation in the surrounding environment, while adding chain-of-thought prompting (Condition C) produces a more elaborate interpretation but also introduces unsupported details.

Interpretation quality (3.38) was rated higher than transcription quality (2.48), suggesting that the semantic intent of an utterance was often recoverable despite transcription errors. At the same time, the relatively high hallucination score (3.47) indicates that the generated interpretations occasionally introduced unsupported details, highlighting a trade-off between contextual inference and factual reliability.

Condition B was selected as the best interpretation for

49.5% of the evaluated samples, outperforming both the transcript-only condition (A) and the chain-of-thought condition (C). This result suggests that providing the predicted acoustic scene label offers useful situational context for downstream interpretation without requiring additional reasoning. In contrast, Condition C achieved the lowest selection rate, indicating that chain-of-thought prompting did not provide additional benefit in this setting and may have introduced unsupported inferences, consistent with previous observations in ASR error correction [14]. Since both the generator and judge belong to the same model family and no independent human evaluation was conducted, these findings should be interpreted as preliminary evidence rather than definitive confirmation that scene grounding improves downstream language understanding.

6. Conclusion

This paper investigated the use of environmental context for automatic speech recognition through a hierarchical conditional multitask framework that jointly performs Acoustic Scene Classification (ASC) and ASR with confidence-gated FiLM conditioning. Experimental results demonstrated consistent improvements over a strong baseline across diverse acoustic scenes and noise conditions. A controlled ablation study further showed that the primary benefit arises from joint multitask learning rather than FiLM-based conditioning, indicating that environmental context is most effectively exploited by shaping shared speech representations during training.

Beyond speech recognition, an exploratory LLM-based evaluation suggested that environmental context may also benefit downstream spoken language understanding, motivating future investigation through human evaluation and larger downstream benchmarks.

Overall, our findings suggest that environmental sounds should be viewed not merely as noise to suppress, but as contextual signals that can benefit both speech recognition and downstream spoken language understanding. We hope this work encourages future research on context-aware, multilingual speech-language systems that more effectively integrate environmental information.

7. Acknowledgment

This research is partially funded by Hibah P2MI ITB (Project ID: STEI.PPMI-1-5-2026).

8. AI-Generated Content Disclosure

All study design, experiments, analyses, and scientific conclusions were originally performed by the authors. Claude (developed by Anthropic) was subsequently used only for language refinement and editorial assistance. The authors reviewed and verified all AI-assisted edits and take full responsibility for the final content of the manuscript.

9. References

[1] N. Jain and S. Rastogi, "Speech recognition systems–A comprehensive study of concepts and mechanism," *Acta Informatica Malaysia*, vol. 3, no. 2, pp. 25–30, 2019.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 12449–12460, 2020.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2303.00733, 2023.

[5] M. Medani, N. Saleem, F. Fkih, M. A. Alohali, H. Elmannai, and S. Bourouis, "End-to-end feature fusion for jointly optimized speech enhancement and automatic speech recognition," *Scientific Reports*, vol. 15, Art. no. 22892, 2025.

[6] S. Chondhekar, V. Murukuri, R. Vasani, S. Goyal, R. Badami, A. Rana, S. SN, K. Pandia, S. Katiyar, N. Jagadeesh, and S. Gulati, "When de-noising hurts: A systematic study of speech enhancement effects on modern medical ASR systems," arXiv preprint arXiv:2512.17562, 2025.

[7] Siri Speech Recognition Team, "Finding local destinations with Siri's regionally specific language models for speech recognition," Apple Research and Development, 2018.

[8] N. Krishnamurthy and J. H. L. Hansen, "Environment dependent noise tracking for speech enhancement," *International Journal of Speech Technology*, vol. 16, pp. 303–312, 2013.

[9] W. Lee, S. Kim, and G. G. Lee, "Enhancing dialogue speech recognition with robust contextual awareness via noise representation learning," arXiv preprint arXiv:2408.06043, 2024.

[10] S. Toshniwal, H. Tang, L. Lu, and K. Livescu, "Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition," in *Proc. Interspeech*, 2017, pp. 3532–3536.

[11] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6989–6993.

[12] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2018.

[13] S. Ruder, "An overview of multi-task learning in deep neural networks," arXiv preprint arXiv:1706.05098, 2017.

[14] K. Everson, Y. Gu, H. Yang, P. Gurunath Shivakumar, G.-T. Lin, J. Kolehmainen, I. Bulyko, A. Gandhe, S. Ghosh, W. Hamza, H.-y. Lee, A. Rastrow, and A. Stolcke, "Towards ASR robust spoken language understanding through in-context learning with word confusion networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[15] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

[16] Qwen Team, "Qwen2.5 Technical Report," arXiv preprint arXiv:2412.15115, 2024.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[19] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic Scene Classification in DCASE 2020 Challenge: Generalization Across Devices and Low Complexity Solutions," in *Proceedings of the DCASE Workshop*, 2020.

[20] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech*, 2018.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[22] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proc. EMNLP*, 2014.

[23] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006.

[24] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011.

[25] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[27] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, 2018, pp. 9–13.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[31] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.