Rethinking Robust ASR Strategies: Can Textual In-Context Learning Improve Acoustic Robustness?

Benita Angela Titalim, Faisal Mehmood, and Sakriani Sakti
Nara Institute of Science and Technology, Japan
E-mail: benita.angela_titalim.ba9@naist.ac.jp, mehmood.faisal@naist.ac.jp, ssakti@is.naist.jp

***Abstract*—Automatic speech recognition (ASR) systems have achieved high accuracy under clean conditions, yet their performance deteriorates in real-world environments with acoustic variability. Most existing robustness methods focus on the acoustic side—using enhancement techniques, noise-aware training, or specialized architectures—but often struggle to generalize across diverse conditions, especially when background noise includes non-target human speech. In this work, we focus on ASR robustness against a range of irrelevant inputs that commonly occur in daily life—namely, stationary noise, background music with vocals, and human utterances not belonging to the target speaker. Crucially, when noise takes the form of intelligible speech, it introduces linguistic interference that purely acoustic methods may fail to suppress. This motivates our investigation of a novel, language-level approach: leveraging textual in-context learning with large language models (LLMs) to guide ASR output through prompting. Experiment results demonstrate that this strategy enhances robustness without requiring retraining or front-end modification, providing a flexible and scalable alternative to conventional acoustic-centric methods.**

I. Introduction

In recent years, automatic speech recognition (ASR) systems have achieved impressive accuracy under clean acoustic conditions, enabling their widespread adoption across various domains, including robotics, traffic control, and healthcare. However, ASR performance degrades significantly in real-world scenarios due to mismatches between training and test conditions, particularly when acoustic noise, reverberation, or speaker variability are present [1].

To address these challenges, a large body of research has focused on improving robustness at the acoustic level—using techniques such as front-end speech enhancement [2]–[4], noise-aware training [5]–[7], and robust acoustic modeling architectures [8], [9]. Many of these approaches have demonstrated effectiveness under specific types of noise, such as stationary background noise, reverberation, or music. However, one particularly challenging and underexplored type of interference is human speech that does not belong to the target speaker. In real-world environments—such as crowded spaces, smart homes, or meetings—non-target speech is common. Yet most ASR systems are designed to transcribe all speech present, rather than distinguishing and suppressing unintended speakers. In this work, we treat non-target human utterances—alongside stationary noise and singing music—as noise sources that should be filtered out to improve transcription of the primary speaker.

Despite the extensive progress in acoustic-side robustness, these methods typically require retraining or adaptation for each new environment, and often depend on supervised data or explicit speaker labels. This limits their flexibility in open, dynamic conditions. Moreover, robustness is rarely approached from the language-side, where higher-level contextual reasoning potentially could help distinguish signal from noise.

This motivates a new direction: *Can we use large language models (LLMs), through textual in-context learning, to improve ASR robustness—not at the acoustic level, but through language-level adaptation?* LLMs have shown strong generalization in various NLP tasks through in-context learning—the ability to perform new tasks by conditioning on a few examples at inference time, without additional training. Prompting strategies allow these models to flexibly adapt to new conditions or user intents, yet their application to speech recognition, especially for improving robustness, remains largely unexplored.

This work explores whether textual in-context learning can improve ASR robustness under noisy, real-world conditions. Specifically, we investigate: (1) whether prompting alone can handle background interference such as stationary noise, music with vocals, and competing speech; (2) how in-context prompting compares to fine-tuning-based adaptation for handling non-target speech; and (3) whether prompting strategies can be adapted to different noise types—especially competing speech—to improve ASR robustness further. This language-first approach requires no retraining or acoustic modification, offering a scalable path to noise-resilient speech recognition.

II. Related Work

A. Noise-Robust ASR

Improving the robustness of automatic speech recognition (ASR) in noisy, real-world environments remains a central challenge. Benchmark efforts such as CHiME Challenge [10], REVERB Challenge [11], and Aurora [12] have driven progress by providing standardized noisy datasets. Robust ASR research has evolved across multiple fronts:

- Front-end enhancement aims to clean the audio signal before recognition using beamforming or neural denoising. While effective under matched conditions, performance often drops with unseen noise types [13], [14].
- Multi-condition training exposes ASR models to diverse environments—noise, reverberation, speaker and channel variability—to improve generalization. This has become

979-8-3315-7206-8/25/$31.00

a standard technique when large, varied datasets are available [15], [16].

- Fine-tuning adapts pre-trained models to specific domains or conditions using smaller targeted datasets. Systems like Whisper and MMS combine both multi-condition and domain-specific tuning, but adapting to new noise types still requires extra data and compute [17], [18].
- Language-level correction uses large language models (LLMs) to refine noisy ASR output post hoc, based solely on text. While promising, these methods rely on accurate initial transcriptions and operate separately from the ASR process [19]–[21].

In contrast, our work explores in-context learning within an LLM-based ASR model to improve robustness during inference. Rather than correcting errors afterward, we use prompts to steer the model toward the target speech—suppressing noise, music, or non-target voices—without requiring retraining or modification of the acoustic pipeline.

B. In-Context Learning LLM

Large language models (LLMs) are powerful generative models trained on massive text corpora, capable of performing a wide range of natural language tasks with minimal or no task-specific training. A key innovation behind their flexibility is in-context learning, the ability to condition on task instructions, and examples provided directly in the input prompt, without updating model parameters [22].

This capability enables users to guide the model behavior by prompt formatting rather than retraining, allowing for fast adaptation to new tasks. In-context learning encompasses several prompting paradigms, including zero-shot learning (only instructions), one-shot learning (a single example), and few-shot learning (a few examples), each offering distinct trade-offs between guidance and generalization [23]–[27]. The emerging practice of prompt engineering, crafting effective prompts to elicit desired behaviors, has further improved LLM performance on tasks such as summarization, question answering, translation, and reasoning [28], [29].

Although in-context learning has shown strong success in text-based domains, its potential in speech-related tasks remains underexplored. Most prior work involving LLMs and ASR uses the language model as a post-processing module to correct recognition errors after decoding. In contrast, our work investigates a novel use of in-context learning within an LLM-based ASR model itself. By injecting contextual cues through prompts at inference time, our aim is to dynamically guide the model's transcription behavior, suppressing irrelevant inputs such as non-target speech and background noise without requiring acoustic model modification or retraining.

III. Methodology

A. Model Description

To perform the ASR task, we utilize the SALMONN architecture [30], a multimodal audio-language model for general audio understanding. It integrates:

- Dual Auditory Encoders
  SALMONN utilizes two auditory encoders: OpenAI's Whisper-Large-v2 [31] for speech recognition and translation, extracting both speech content and background context, and the fine-tuned BEATs encoder [26], which captures high-level semantics from non-speech audio through self-supervised masked audio modeling.
- Window-level Q-Former
  Adapted from the Q-former structure [32], this module transforms variable-length audio features into textual tokens via trainable queries and cross-attention, applied in fixed-size windows. Each window is treated like an image patch and passed through the Q-former to generate textual tokens. This results in a variable number of output tokens that better preserve temporal resolution and ensure monotonic alignment with the original audio, which benefits tasks like speech recognition.
- LLM and low-rank adaptation (LoRA) Adaptation
  The Vicuna language model [33] (based on LLaMA) is used as the text encoder [34]. It is adapted via LoRA [35], a parameter-efficient method that updates only the query and value matrices in self-attention, while keeping the base model frozen.

SALMONN is trained through a three-stage process to improve generalization and reduce task-specific overfitting. In **Stage 1**, the Q-former and LoRA modules are pretrained on large-scale ASR and audio captioning datasets. **Stage 2** involves instruction tuning on diverse supervised audio-text tasks to teach the model to follow textual prompts, though this stage may introduce overfitting. To address this, **Stage 3** (activation tuning) reduces the LoRA scaling factor and applies self-supervised fine-tuning, leading to better generalization and more diverse outputs without degrading performance on learned tasks.

B. Dataset Construction

In this paper, we construct a customized dataset based on the Clarity Enhancement Challenge (CEC3) corpus[1]. The original dataset features real acoustic scenes recorded in a controlled room using hearing aid shells with three microphones (front, mid, back), worn by a listener. Each scene includes a target sentence and two or three interferers—competing (non-target) speech, music, or domestic appliance noise (referred to as stationary noise)—played through a 13-loudspeaker array surrounding the listener. The corpus contains 6,000 training, 2,500 development, and 1,500 evaluation scenes, with SNRs ranging from –12 dB to +6 dB. Recordings were made at 48 kHz and synchronized with head-tracking data.

For our experiments, we used only the training and development sets. Since the original evaluation set lacks transcriptions, the development set was split into 60% for validation and 40% for testing. To facilitate a detailed analysis of how different noise types affect performance, we constructed new noisy signals by mixing clean speech with randomly selected noise (stationary, music, or competing speech) at SNRs of –5, 0, 10, and 20 dB. All audio was resampled to 16 kHz to match typical ASR input requirements. For fine-tuning, each noisy

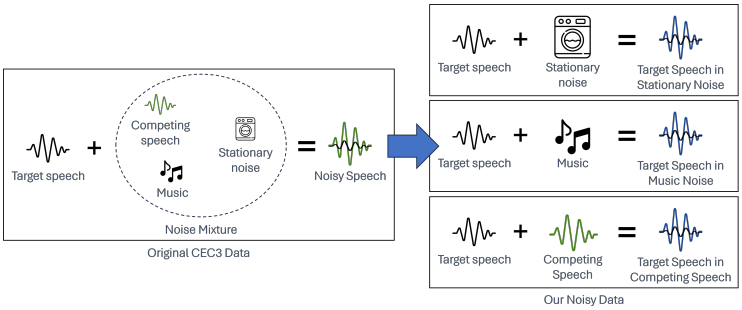

Fig. 1. Dataset construction. Left: Original CEC3 dataset scenes, which include a target sentence mixed with two or three real-world interferers, such as speech, music, or stationary noise. Right: Synthetic noisy data used in the experiment. Clean speech was mixed with a single noise type—stationary noise, music, or competing speech— selected from the CEC3 dataset.

signal was paired with its transcription and used in training.

C. Analytical Focus

In this study, the experiments are structured based on research questions: (1) whether prompting alone can handle background interference such as stationary noise, music with vocals, and competing speech; (2) how in-context prompting compares to fine-tuning-based adaptation for handling non-target speech; and (3) whether prompting strategies can be adapted to different noise types—especially competing speech—to further improve ASR robustness. These analyses aim to provide a comprehensive understanding of the effectiveness, generalization, and limitations of in-context prompting for robust ASR.

D. Experiment Setup

We conducted in-context inference using a multi-modal model that combine speech and language understanding. The language backbone was Vicuna-13B, and speech features were extracted using frozen Whisper-Large-v2 and BEATs encoders. A window-level Q-Former processed speech with one query token per 0.333-second window. The model was adapted using LoRA fine-tuning and conditioned via multi-prompt in-context learning, with user-assistant style prompts loaded from JSON files. Inference used beam search (`num_beams=4`) with temperature-based sampling (`temperature=1.0`, `top_p=0.9`) and a maximum of 200 new tokens.

In contrast, the fine-tuning approach updated parameters of the Q-Former, speech-to-language projection layer, and applied LoRA adaptation (rank 8, alpha 32, dropout 0.1) to Vicuna-13B, while keeping Whisper and BEATs encoders frozen. Fine-tuning optimized the model over up to 30 epochs with a starting learning rate of 3e-5, using batch size 4 and multi-prompt templates for conditioning. Unlike in-context learning, which relies solely on prompt conditioning at inference without updating model weights, fine-tuning directly adapts model parameters to improve task-specific performance.

IV. Experimental Results

A. Evaluating In-Context Learning with Input/Output-Aware Prompting under Diverse Acoustic Interference

In this experiment, we investigate how prompt engineering influences the ability of a large multimodal language model (SALMONN) to transcribe speech under severe acoustic interference using in-context (zero-shot) learning. We evaluate four prompts of increasing specificity and structure, ranging from minimal guidance to more constrained instructions, each varying in whether they focus on the input only, the output only, or both. These prompts, detailed in Table I, emphasize different transcription goals, including sentence-level fluency and word-level precision.

Table II presents the transcription performance of prompts p1–p4 under varying signal-to-noise ratios (SNRs): 20dB, 10dB, 0dB, and –5dB, across four conditions: clean speech, stationary noise, background music with vocals, and competing non-target speech. Transcription accuracy is evaluated using Character Error Rate (CER) and Word Error Rate (WER).

Under background stationary noise, prompts that incorporate more structure and contextual guidance lead to significantly better performance. Among the single-aspect prompts, considering only the input (**p1**) consistently outperforms the output-only prompt (**p2**), yielding a 28.0% relative improvement in average CER and 21.7% in average WER across all SNR levels. This suggests that grounding the transcription in the acoustic signal helps reduces irrelevant or hallucinated content. The most robust performance, however, is achieved by **p4**, which integrates both input and output considerations, emphasizing full-sentence transcriptions. Compared to **p1**, **p4** achieves a 42.2% relative reduction in CER and 27.6% in WER, on average across SNRs. Furthermore, **p4** outperforms the word-level prompt **p3**, with a 53.3% relative reduction in CER and 36.2% in WER, highlighting the importance of well-formed sentence structures even in noisy conditions.

With background music containing vocals (i.e., singing), the interference presents a unique challenge. Although vocals are speech-like in content, their musical rhythm, pitch, and harmonic variation introduce acoustic patterns that are distinct from those of natural speech. In this context, the prompt focusing on the human speaking voice (**p1**) outperforms the output-only prompt (**p2**), yielding a 56.8% relative improvement in average CER and 43.6% in WER, averaged across all levels of SNR. This suggests that grounding the model in the modality of the intended source is particularly beneficial when background content contains linguistic elements but lacks a natural conversational structure. As in the previous condition, **p4** remains the overall strongest performer, achieving a relative

[1]https://claritychallenge.org/docs/cec3/task_2/cec3_task2_data

TABLE I
PROMPTING STRATEGIES BASED ON INPUT-OUTPUT CONTEXT AND THEIR FOCUS

Prompt	Instruction Text	Focus Description
p1	*"Give me the transcription for the human speaking voice."*	Considers only the input source as the human speaking voice.
p2	*"Give me the transcription in a complete English sentence."*	Considers only the output text, requiring well-formed English sentences.
p3	*"Give me word-level English transcription for the human speaking voice."*	Considers both input speech and output text, with emphasis on word-level sentence output.
p4	*"Give the transcription of the human speaking voice in a complete English sentence."*	Considers both input speech and output text, with emphasis on requiring well-formed English sentences.

TABLE II
AVERAGE CER/WER IN VARIOUS NOISE CONDITIONS WITH PROMPTING STRATEGIES BASED ON INPUT AND OUTPUT CONTEXT

Stationary Noise				
Prompt	**p1**	**p2**	**p3**	**p4**
SNR 20dB	1.324 / 2.223	3.750 / 5.078	1.263 / 2.126	0.384 / 0.899
SNR 10dB	1.592 / 2.600	3.780 / 5.200	1.856 / 3.086	0.663 / 1.349
SNR 0dB	3.835 / 5.892	6.505 / 8.784	6.133 / 8.359	2.707 / 4.362
SNR -5dB	13.828 / 18.054	15.262 / 20.253	16.701 / 21.310	9.293 / 13.656
Background Music with Vocals				
Prompt	**p1**	**p2**	**p3**	**p4**
SNR 20dB	0.640 / 1.482	3.149 / 4.167	1.339 / 2.308	0.339 / 0.741
SNR 10dB	0.789 / 1.604	3.046 / 4.143	1.414 / 2.478	0.467 / 0.972
SNR 0dB	3.958 / 6.050	10.659 / 11.712	4.453 / 5.783	1.773 / 2.782
SNR -5dB	8.516 / 11.250	15.313 / 19.645	17.010 / 20.945	6.118 / 8.808
Competing Non-target Speech				
Prompt	**p1**	**p2**	**p3**	**p4**
SNR 20dB	21.074 / 21.103	3.948 / 4.775	8.655 / 8.930	4.629 / 5.151
SNR 10dB	80.418 / 76.637	26.313 / 26.631	44.147 / 43.616	40.867 / 38.926
SNR 0dB	181.279 / 182.116	153.329 / 159.191	168.189 / 171.899	167.067 / 169.190
SNR -5dB	194.327 / 202.163	178.045 / 188.689	184.502 / 194.533	186.918 / 194.436

TABLE III
AVERAGE CER/WER IN COMPETING NON-TARGET SPEECH CONDITION WITH FINE-TUNING

Prompt	**p1**	**p2**	**p3**	**p4**
SNR 20dB	7.879 / 8.018	6.060 / 6.427	4.357 / 4.799	4.629 / 5.151
SNR 10dB	49.285 / 46.106	41.359 / 39.473	34.576 / 33.641	40.867 / 38.926
SNR 0dB	169.714 / 170.684	165.254 / 167.610	161.037 / 164.208	167.067 / 169.190
SNR -5dB	185.978 / 192.893	185.029 / 192.723	183.344 / 192.346	186.918 / 194.436

reduction of 59.5% in CER and 40.6% in WER compared to **p3**, and a 60.6% CER reduction relative to **p1**, again based on average performance across SNRs.

In the case of competing human speech, where both the target and background signals are intelligible spoken language, prompts that rely on identifying speech alone (**p1**, **p3**, **p4**) become less effective. The average CER and WER for these prompts range from 99.9% to 119.8%, indicating substantial degradation. Interestingly, the best performance in this setting is achieved by **p2**, which emphasizes producing a complete English sentence without explicitly referencing the input modality. Compared to **p4**, **p2** yields a 10.0% relative reduction in CER and 6.9% in WER, averaged across all SNR levels. This suggests that, under high linguistic interference, encouraging coherent output may help suppress irrelevant or fragmented non-target speech. The results also highlight a counterintuitive finding: In highly ambiguous acoustic environments, output-focused prompting can offer a robustness advantage.

B. Addressing Competing Non-Target Speech

While in-context prompting shows promising results across various noise conditions, transcription performance degrades significantly under competing non-target speech—a particularly challenging form of interference due to its linguistic similarity to the target signal. In this section, we focus specifically on this difficult case and explore whether further gains can be achieved through two strategies: (1) fine-tuning the acoustic model on in-domain data, and (2) extending prompting beyond input/output awareness to incorporate noise-type information.

1) Fine-Tuning for Competing Speech: In this experiment, we fine-tuned the model using a fixed prompt format and evaluated its performance across different prompts (**p1**–**p4**) during testing. This setup allowed us to train a single model and assess its robustness to prompt variation during inference, without retraining for each prompt. Results in Table III show that fine-tuning outperforms in-context learning at higher SNRs (20dB and 10dB), where speech is less corrupted. However, at lower SNRs (0dB and –5dB), both approaches suffer significant degradation, with fine-tuning yielding minimal or even negative gains. This decline is attributed not just to increased noise levels but also to the nature of the interference: the background consists of competing speech, which is linguistically structured and intelligible. In such case, the model struggles to isolate the target speaker, indicating that fine-tuning alone does not resolve this linguistic ambiguity.

2) Noise-Aware Prompting: To address the limitations of input/output-aware prompting under competing speech conditions, we introduce three additional prompts—**p5**, **p6**, and **p7**—that incorporate explicit knowledge about the structure of

TABLE IV
NOISE-AWARE PROMPTING STRATEGIES AND THEIR FOCUS

Prompt	Instruction Text	Focus Description
p5	*"Consider the first person speaking as noise. Transcribe the other person's voice."*	Introduces noise-type awareness by treating the first speaker as interfering speech and guiding the model to focus on the target speaker.
p6	*"Two people are speaking. Focus on the second speaker only. Do not transcribe the voice of the first speaker."*	Uses noise-aware guidance by identifying the first speaker as non-target speech and prompting selective transcription of the second speaker.
p7	*"Given audio with two speakers, transcribe only the second speaker's utterance. The transcription should be a complete English sentence. Disregard the first speaker entirely."*	Combines noise-type awareness and output structure by specifying both the speaker to transcribe and the requirement for a complete sentence.

TABLE V
AVERAGE CER/WER IN COMPETING NON-TARGET SPEECH CONDITION WITH NOISE AWARE PROMPTING STRATEGIES

Prompt	Input/Output Aware Prompting		Noise Aware Prompting		Noise+Output Aware Prompting
	p1	**p2**	**p5**	**p6**	**p7**
SNR 20dB	21.074 / 21.103	3.948 / 4.775	7.344 / 7.909	3.481 / 4.204	4.812 / 6.682
SNR 10dB	80.418 / 76.637	26.313 / 26.631	23.734 / 24.456	18.005 / 19.074	18.078 / 20.180
SNR 0dB	181.279 / 182.116	153.329 / 159.191	122.109 / 126.534	120.491 / 126.886	112.696 / 120.690
SNR -5dB	194.327 / 202.163	178.045 / 188.689	150.941 / 159.981	152.312 / 162.775	144.782 / 155.911

the audio. In these scenarios, we assume prior knowledge that the input contains two speakers: the first acts as background interference, and the second—whose speech occurs later—is the intended target. As listed in Table IV, prompts **p5** and **p6** direct the model to treat the first speaker as noise and focus on the second speaker's voice. Prompt **p7** further enhances this strategy by requiring the output to be a complete English sentence, thus combining noise-aware and output-aware guidance.

As shown in Table V, noise-aware prompting (**p5** and **p6**) leads to substantial improvements under severe interference, especially in low-SNR settings. Although gains in input/output-sensitive prompting (e.g. **p2**) are modest at 20 dB, they grow with increasing noise. At -5 dB, the most challenging condition, prompt **p6** reduces CER and WER by approximately 14.5% and 13.7%, respectively, compared to **p2**. Furthermore, prompt **p7**, which combines noise-type awareness with structured output, achieves the best overall performance at low SNR, with CER and WER reduced by 18.7% and 17.4% compared to **p2**. These results suggest that incorporating the nature of the interference and constraints on the transcription output is a promising strategy for handling highly degraded, speech-like noise.

V. CONCLUSIONS

This work explored improving ASR robustness through in-context prompting with large language models (LLMs), targeting realistic noise conditions including stationary noise, music with vocals, and competing human speech. We found that input–output-aware prompts consistently outperformed those considering only input or output, particularly under moderate noise. Fine-tuning improved performance at higher SNRs but was less effective under low-SNR competing speech, where the noise closely resembles the target voice. To address this, we introduced noise-aware prompts (**p5**, **p6**) that explicitly distinguish between target and non-target speakers, yielding substantial gains under severe interference. Building on this, prompt **p7** combines noise awareness with structured output requirements—guiding the model to produce coherent, sentence-level transcriptions from the target speaker only. These results show that incorporating knowledge about the noise structure into prompt design can meaningfully enhance ASR robustness in conditions where traditional methods fall short.

Furthermore, while this study has focused on a CEC3-based dataset and primarily evaluated within the SALMONN framework, future work will extend our approach to a broader range of datasets and benchmarks. This will enhance our ability to evaluate the generalizability and robustness of the proposed method.

ACKNOWLEDGMENT

Part of this work was supported by JST SPRING Grant Number JPMJSP2140 and JSPS KAKENHI Grant Numbers JP21H05054, JP23K21681, and JP25H01139.

REFERENCES

[1] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Chapter 1 - introduction," pp. 1–7, 2016.

[2] Y. Yang, A. Pandey, and D. Wang, *Towards decoupling frontend enhancement and backend recognition in monaural robust ASR*, 2024. arXiv: 2403.06387.

[3] T. O'Malley, A. Narayanan, Q. Wang, A. Park, J. Walker, and N. Howard, "A conformer-based ASR frontend for joint acoustic echo cancellation, speech enhancement and speech separation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 304–311.

[4] T. Yoshioka and M. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech Language*, vol. 31, no. 1, pp. 65–86, 2015, ISSN: 0885-2308.

[5] H.-S. Lee, P.-Y. Chen, Y.-F. Cheng, Y. Tsao, and H.-M. Wang, *Speech-enhanced and noise-aware networks for robust speech recognition*, 2022. arXiv: 2203.13696.

[6] D. Raj, J. Villalba, D. Povey, and S. Khudanpur, *Frustratingly easy noise-aware training of acoustic models*, Nov. 2020.

[7] P. Karanasou, Y. Wang, M. J. F. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Interspeech 2014*, 2014, pp. 2180–2184.

[8] M. Fujimoto and H. Kawai, "Noise robust acoustic modeling for single-channel speech recognition based on a stream-wise transformer architecture," Aug. 2021, pp. 281–285.

[9] K. J. Han, J. Pan, V. K. N. Tadala, T. Ma, and D. Povey, *Multistream cnn for robust acoustic modeling*, 2021. arXiv: 2005.10470.

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[11] K. Kinoshita, M. Delcroix, S. Gannot, *et al.*, "A summary of the reverb challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, 2016.

[12] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, vol. 4, 29–32.

[13] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, *Far-field automatic speech recognition*, 2020. arXiv: 2009.09395.

[14] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel nmf-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 960–971, May 2019, ISSN: 2329-9304.

[15] F. Weninger, H. Erdogan, S. Watanabe, *et al.*, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust ASR," vol. 9237, Aug. 2015.

[16] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 116–120.

[17] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource ASR," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, Jun. 2024.

[18] V. Pratap, A. Tjandra, B. Shi, *et al.*, *Scaling speech technology to 1,000+ languages*, 2023. arXiv: 2305.13516.

[19] Y. Hu, C. Chen, C.-H. H. Yang, *et al.*, *Large language models are efficient learners of noise-robust speech recognition*, 2024.

[20] Y. Bai, J. Chen, J. Chen, *et al.*, *Seed-ASR: Understanding diverse speech and contexts with llm-based speech recognition*, 2024.

[21] S. Radhakrishnan, C.-H. H. Yang, S. A. Khan, *et al.*, *Whispering llama: A cross-modal generative error correction framework for speech recognition*, 2023.

[22] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language models are few-shot learners*, 2020.

[23] T. Li, X. Ma, A. Zhuang, Y. Gu, Y. Su, and W. Chen, "Few-shot in-context learning on knowledge base question answering," Association for Computational Linguistics, 2023, pp. 6966–6980.

[24] R. L. IV, I. Balazevic, E. Wallace, F. Petroni, S. Singh, and S. Riedel, "Cutting down on prompts and parameters: Simple few-shot learning with language models," Association for Computational Linguistics, 2022, pp. 2824–2835.

[25] J. Shin, Y. Ahn, S. Won, and S. J. Choi, "Learning to adapt large language models to one-shot in-context intent classification on unseen domains," Association for Computational Linguistics, 2024, pp. 182–197.

[26] W.-L. Chen, C.-K. Wu, Y.-N. Chen, and H.-H. Chen, *Self-icl: Zero-shot in-context learning with self-generated demonstrations*, 2023.

[27] J. He, L. Wang, Y. Hu, *et al.*, *Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction*, 2023.

[28] D. Li, Z. Liu, X. Hu, Z. Sun, B. Hu, and M. Zhang, *In-context learning state vector with inner and momentum optimization*, 2024.

[29] X. L. Do, Y. Zhao, H. Brown, *et al.*, *Prompt optimization via adversarial in-context learning*, 2024.

[30] C. Tang, W. Yu, G. Sun, *et al.*, *Salmonn: Towards generic hearing abilities for large language models*, 2024.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," JMLR.org, 2023, pp. 1–27.

[32] J. Li, D. Li, S. Savarese, and S. Hoi, *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*, 2023.

[33] W.-L. Chiang, Z. Li, Z. Lin, *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality," vol. 2, p. 6, 2023.

[34] H. Touvron, T. Lavril, G. Izacard, *et al.*, *Llama: Open and efficient foundation language models*, 2023.

[35] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," OpenReview.net, 2022.