Low SNR Speech Perception with HuBERT: A Discussion on Visual-Audio Fusion and Domain-Specific Modeling

Jayasree Saha
UPES, University of Tomorrow
Dehradun, Uttarakhand, India

Vinay P. Namboodiri
University of Bath
Claverton Down, Bath, UK

C. V. Jawahar
IIIT Hyderabad
Hyderabad, Telengana, India

ABSTRACT

Audio-visual speech enhancement (AVSE) leverages visual cues, such as lip motion, to improve speech intelligibility in noisy conditions. While recent models have focused on using explicit visual input, our prior work introduced a pseudo-visual approach that synthesizes lip motion from noisy audio—addressing cases where visual input is unavailable. In this study, we revisit the role of visual cues in the context of recent advancements in speech representation models such as HuBERT. Specifically, we investigate: (i) the relative contribution of lip motion compared to robust pre-trained features like HuBERT across varying SNR levels, (ii) the necessity of training separate models for high and low SNR conditions, and (iii) the trade-offs between using a general model versus noise-specific models tailored to distinct real-world environments.

To this end, we incorporate HuBERT features, experiment with LoRA-based fine-tuning, and evaluate performance across five realistic noise types, including railway stations, road traffic, and festivals. Our findings offer insights into designing efficient and adaptable AVSE systems for practical deployment.

KEYWORDS

Speech enhancement, low SNR, Hubert, LoRA, PEFT

ACM Reference Format:
Jayasree Saha, Vinay P. Namboodiri, and C. V. Jawahar. 2025. Low SNR Speech Perception with HuBERT: A Discussion on Visual-Audio Fusion and Domain-Specific Modeling. In *Proceedings of 12th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'25).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Speech is a fundamental modality for human-to-human and human-machine communication. However, real-world acoustic environments often degrade speech signals through background noise, reverberation, and recording distortions, adversely impacting the performance of speech-driven systems [9]. Speech enhancement (SE) techniques are widely used as a front-end to improve intelligibility and quality in downstream tasks such as automatic speech recognition (ASR) [11, 21, 34], speaker recognition [26], hearing aids [13, 20], and cochlear implants [7]. Recent studies in neuroscience and speech perception have shown that humans rely heavily on visual cues—such as lip movements—to enhance auditory attention and comprehension [27, 28, 30]. Unlike acoustic signals, visual information remains unaffected by background noise, making it especially valuable in low signal-to-noise ratio (SNR) scenarios [31]. This observation has catalyzed interest in audio-visual speech enhancement (AVSE), where visual signals complement degraded audio to produce more intelligible speech [2, 8, 35].

Traditional AVSE methods commonly depend on frontal facial inputs for lip-reading, but these approaches are brittle under dynamic conditions involving head movement or occlusions. To address this, Sindhu et al. [14] proposed a pseudo-lip generation technique that synthesizes lip motion from noisy speech and a static face image, offering robustness in uncontrolled environments.

Despite these advancements, most existing AVSE approaches are evaluated under relatively high SNR settings, leaving the low SNR domain under-explored. Real-world conditions-such as train stations, urban roads, industrial environments, and crowded festivals—are rife with intense and fluctuating background noise, where conventional methods often fail. Furthermore, a single model trained for broad generalization may either overfit to high SNR scenarios or underperform due to over-regularization in extreme noise cases. This raises the question: can we build speech models that are robust, adaptive, and effective across a wide SNR spectrum?

At the same time, large pre-trained speech models like HuBERT have shown promise for downstream SE tasks. However, full fine-tuning of these models is computationally expensive and inefficient. Parameter-efficient fine-tuning (PEFT) strategies such as Low-Rank Adaptation (LoRA) [16] offer a lightweight alternative by freezing base parameters and introducing learnable low-rank matrices. This technique significantly reduces computational cost and retains the generalization ability of the base model. Recent extensions like Conv-LoRA [36] have also demonstrated improved performance in vision tasks by embedding convolutional priors. Although LoRA has been primarily explored in natural language and vision domains, its application to audio-visual speech processing remains limited. Biderman et al. [5] suggest that while LoRA may underperform compared to full fine-tuning, it acts as a strong regularizer and helps retain general performance across domains—an attribute especially useful in speech systems designed for deployment across diverse acoustic environments.

Contributions. In this study, we conduct a focused analysis of audio-visual speech enhancement strategies under variable SNR conditions. Our contributions are threefold:

(1) We extend the pseudo-lip approach proposed by Sindhu et al. [14] by integrating it with HuBERT speech representations. We analyze its performance across high and low SNR settings using the LRS3 dataset and noise profiles from the VGG-Sound dataset.

ICVGIP'25, December 2025, Mandi, India

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(2) We explore the effectiveness of LoRA-based fine-tuning for adapting pre-trained model to new noise conditions. Our investigation seeks to determine whether LoRA offers a viable, efficient alternative to full fine-tuning, especially in low SNR scenarios.
(3) Recognizing the limitations of single-model generalization, we experiment with five specialized models trained on noise from critical real-world settings: railway stations, Indian roads, office environments, factories, and Indian festivals. We highlight domain-specific performance and offer insights into the benefits of targeted adaptation.

Overall, our work aims to provide a compact, empirical perspective on the challenges and opportunities in audio-visual speech enhancement, with a particular emphasis on underrepresented low SNR conditions and practical fine-tuning strategies.
The remainder of this paper is organized as follows: Section 2 reviews relevant literature in audio-visual speech enhancement and parameter-efficient fine-tuning. Section 3 outlines the methodological framework that forms the foundation for our key discussions. Section 4 presents the experimental setup, including details of the datasets used and curated for low SNR scenarios. Finally, we summarize our observations and insights in Section 5.

2 RELATED WORK

Audio-Visual Speech Enhancement. The integration of visual information into speech enhancement has gained substantial momentum in recent years, particularly with the rise of deep neural network (DNN)-based models. These audio-visual speech enhancement (AVSE) systems leverage complementary cues from both modalities to recover clean speech under challenging noise conditions, driving notable progress in the field. Yang et al. [35] leveraged audio-visual cues to guide the generation of neural speech codec representations, enabling the synthesis of clean and realistic speech from noisy inputs. Their work highlights the potential of using visual information to condition speech generation systems. The importance of speaker-specific cues in speech processing has led to the development of personalized AVSE models. Gogate et al. [12] proposed CochleaNet, a language-, noise-, and speaker-independent model designed to selectively enhance the target speaker's voice by suppressing background noise. Their architecture uses a cross-modal DNN that processes both noisy audio mixtures and lip images, producing time-frequency (T-F) masks based on audio-visual cues regardless of the SNR level. Human speakers instinctively adapt their speech patterns to increase intelligibility in noisy environments—a phenomenon known as the Lombard effect. Traditional SE systems, however, are typically trained on clean speech with synthetically added noise, which does not account for these natural adaptations. Michelsanti et al. [25] demonstrated that training deep learning models with Lombard speech can improve the intelligibility and quality of enhanced speech, especially in low SNR conditions. Their study also noted gender-based performance disparities, attributing them to differences in how males and females exhibit Lombard speech characteristics. Afouras et al. [2] redefined speech enhancement as a source separation problem—isolating a target speaker's voice from mixtures involving competing speakers and background noise. Their approach jointly conditions the enhancement on visual lip motion and voice embeddings. To address visual occlusions, they introduced artificial mouth-region masking during training, ensuring robustness when visual input is partially unavailable. Adeel et al. [1] showed that lip-reading-driven AVSE outperforms audio-only models in low SNR scenarios. However, they also observed that visual information offers diminishing returns at high SNRs, where clean audio is often sufficient. To address this, they proposed a context-aware AVSE framework featuring an adaptive switching module that dynamically selects among audio-only (A-only), visual-only (V-only), and combined audio-visual (AV) cues. This switching mechanism operates without explicit SNR estimation, learning to adapt based on contextual conditions. More recently, the advent of multimodal self-supervised learning has introduced powerful embeddings for AV tasks. AV-HuBERT, a self-supervised model trained on both audio and visual modalities, has demonstrated strong performance in tasks such as lip reading and automatic speech recognition. Chern et al. [8] extended AV-HuBERT for use in speech enhancement and separation by integrating it with a downstream SE module, thereby leveraging rich audio-visual representations to improve performance in both AVSE and AVSS tasks.
PEFT. The advent of foundation models has signaled a transformative paradigm shift, facilitating the tailored customization of versatile, general-purpose models for specialized downstream applications. Foundation models are expansive machine learning constructs trained on massive datasets, typically employing self-supervised learning techniques. Leveraging knowledge from these pre-trained foundation models can mitigate the data sparsity challenges in downstream tasks []. Nonetheless, finetuning the entire foundation model is resource-intensive and may result in overfitting and catastrophic forgetting [33]. In NLP, recent studies have introduced various parameter-efficient transfer learning techniques, including adapter tuning [29], prefix tuning [23], and Low-Rank Adaptation (LoRA) [16]. These methods achieve state-of-the-art performance by training only a small subset of model parameters, circumventing the challenges associated with full model finetuning. Unlike the extensive use of adapters in NLP and Computer Vision (CV), their application in the speech domain still needs to be explored. Existing research has predominantly focused on a limited set of tasks such as Automatic Speech Recognition (ASR) [18, 32] and Speech Translation [22]. Given the proven success of parameter-efficient methods in numerous downstream NLP applications, it is essential to systematically investigate their potential for enhancing performance in downstream speech processing tasks. In contrast to NLP and Computer Vision (CV), the application of adapters in the speech domain has thus far been limited to a few works pertaining to Automatic Speech Recognition (ASR) [18, 32] and Speech Translation [22]. Given the irrefutable success of parameter-efficient approaches in various downstream NLP tasks, it is imperative to evaluate their efficacy in solving downstream speech processing tasks. Since the approaches to solving NLP and speech processing tasks share many characteristics, we hypothesize that the parameter-efficient algorithms in NLP can yield similar gains in speech processing. Large PLMs in speech processing, such as Wav2Vec 2.0 [4], HuBERT [15], WaveLM [6] have evolved to encode effective representations of speech signals, reflecting in the state-of-the-art performances in downstream speech processing

tasks [3, 19]. Given the numerous similarities in solving NLP and speech processing tasks, we postulate that parameter-efficient algorithms effective in NLP can also drive significant improvements in speech processing. Advanced pre-trained language models (PLMs) such as Wav2Vec 2.0 [4], HuBERT [15], and WaveLM [6] have proven to encode robust speech signal representations. This capability has led to state-of-the-art results in various downstream speech processing

3 METHOD

With the advent of foundational models, downstream tasks have achieved remarkable results. HuBERT is a prime example of such a foundational model. Leveraging a self-supervised approach, HuBERT develops representations resilient to noise and variations in speech, forming a robust foundation for fine-tuning on specific tasks with labeled data. Initially, HuBERT clusters unlabeled audio frames to create pseudo-labels using K-means. Subsequently, it employs a Transformer architecture to predict these pseudo-labels, refining its understanding of speech patterns and learning high-quality speech representations. By training on extensive amounts of raw audio data, HuBERT captures diverse speech characteristics, making it highly effective for various downstream tasks. Therefore, we utilize HuBERT features as a crucial cue from noisy speech. One motivation for this study is to understand the importance of visual cues in robust speech representations. To this end, we examined several combinations of feature inclusion prior to the decoder, as described in Section 4.2.1. The complete architecture is shown in Figure 1, where the fused audio-visual representation is passed to the speech decoder, which predicts an additive mask. When applied to the noisy speech, this mask enhances the speech signal.

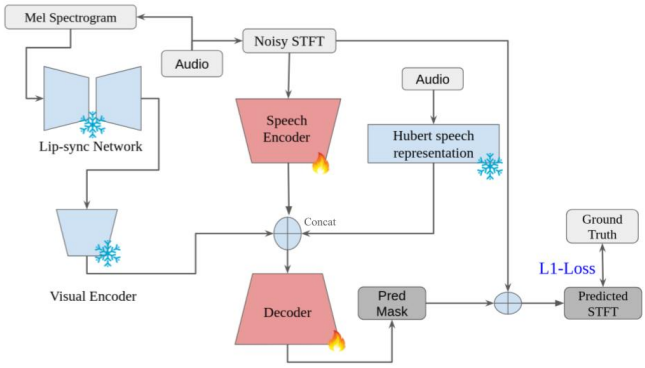


Figure 1: Schematic Diagram of Speech Enhancement Models used in our study.

3.1 LoRA-SE for low SNR condition

Low-Rank Adaptation (LoRA) [17] is a State-of-the-art Parameter-Efficient Fine-Tuning (PEFT) technique initially proposed for fine-tuning large language models. It works by freezing the weights of the pre-trained model and introducing trainable rank decomposition matrices into each layer of the transformer architecture. Figure 2 depicts the detailed structure of LoRA technique. This approach significantly reduces the number of trainable parameters needed for downstream tasks. Since its introduction, LoRA has been successfully applied to fine-tuning various foundation models for speech [24] and vision [36] related tasks. In a recent study, Biderman et al. [5] examined the performance of LoRA compared to full fine-tuning across two target domains: programming and mathematics. Their findings indicate that, in most scenarios, LoRA significantly underperforms compared to full fine-tuning. However, LoRA demonstrates a beneficial form of regularization by better preserving the base model's performance on tasks outside the target domain. We incorporated LoRA at every convolution layer in the Speech Encoder.It decomposes the incremental update of the each pretrained convolutional speech encoder's weights W into two matrices A and B. where $W \in R^{D\times K}$, $A \in R^{D\times r}$ $B \in R^{r\times K}$. Given a hidden representation $h = Wx$, the low-rank adapted representation is expressed as

$$\begin{aligned} h &= Wx + \Delta x \\ &= Wx + BAx \end{aligned} \tag{1}$$

The matrix A is initially drawn from a Gaussian distribution, while B is initialized to zero. This initialization ensures that Δ equals zero at the beginning of the training process.

The rationale behind this approach is to maintain the appropriate encoding for high SNR conditions while progressively enhancing the encoding for low SNR conditions. This allows the decoder to be fine-tuned to decode the newly learned embeddings for the speech enhancement task.

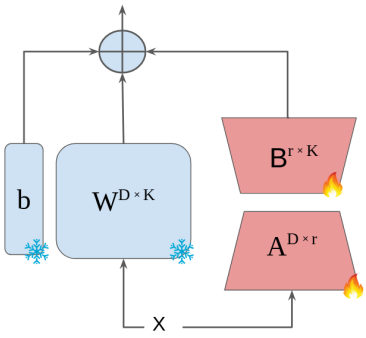


Figure 2: Detailed structure of the LoRA matrix. During fine-tuning, only the low-rank decomposed matrices A and B are trained, while the pretrained weights remain unchanged.

4 EXPERIMENTS

This section delineates the experimental setup and results. We employed two standardized evaluation metrics to gauge the SE performance: perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility measure (STOI). PESQ assesses the quality of processed speech, with scores ranging from -0.5 to 4.5. A higher PESQ score denotes superior speech quality. STOI, on the other hand, measures speech intelligibility, typically yielding scores between 0 and 1. A higher STOI value signifies enhanced speech intelligibility.

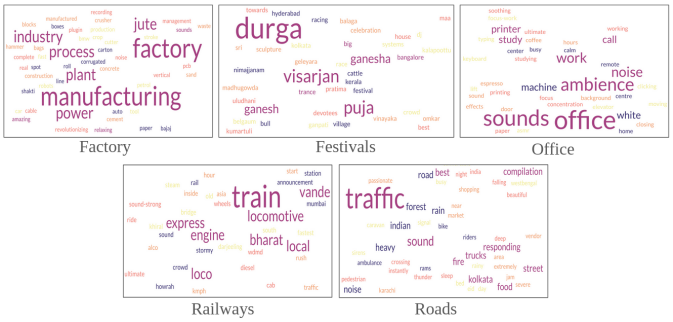


Figure 3: Word cloud depiction of noise video captions illustrating plausible noise types present in each of the five application scenarios.

4.1 Experimental Setup

In this section, the details of the dataset and the implementation steps of our work are introduced.

4.1.1 Dataset. We leverage the LRS3 dataset, which includes thousands of spoken sentences extracted from TED videos. For our training process, we utilize the "pre-train" subset, encompassing around 407 hours of video data and 119K utterances and 51k vocabulary. This dataset is notably challenging due to its extensive variety of speakers (5K), thereby facilitating the creation of a speaker-independent model. We selected 40K VGG-Sound noise which distribute over noises categorized under People, Animals, Instrument, Nature, Vehicle, Home, and Sports.
We also curate a set of datasets for experimenting with real-world noise environments. This process initiates by pinpointing real-world scenarios where low SNR speech enhancement is essential. Subsequently, we collect videos from the web using targeted keywords to encompass various situations within each noise category. These videos are then segmented into 10-second audio clips for training purposes. Figure 3 enumerates the keywords utilized for extracting audio from the web, while the corresponding visual data is depicted in Figure 4.

4.1.2 Training Details. We train our model on 1-second clips of clean speech randomly sampled from the target dataset. To simulate noisy speech during training, we randomly select noise samples from VGG-Noise and convolve them with clean speech at SNR levels of 0dB, 5dB, and 10 dB. VGG-Noise encompasses a wide range of noise models that reflect real-world acoustic environments. Further, we augment the noisy audio by adding segments of zero speech, reducing the ground truth speech, and doubling the noise levels. We adopt the pseudo lipsync model and its learned checkpoint from Sindhu et al. [14] to generate pseudo lips for noisy audio. The Adam optimizer is employed with a learning rate set to 10^{-3}. In addition to these audio-visual methods, we compare them with a well-known *Audio-Only* model [10] to highlight the significance of the visual stream in speech enhancement. For a fair comparison, all models are trained and evaluated on the same datasets.

4.2 Results

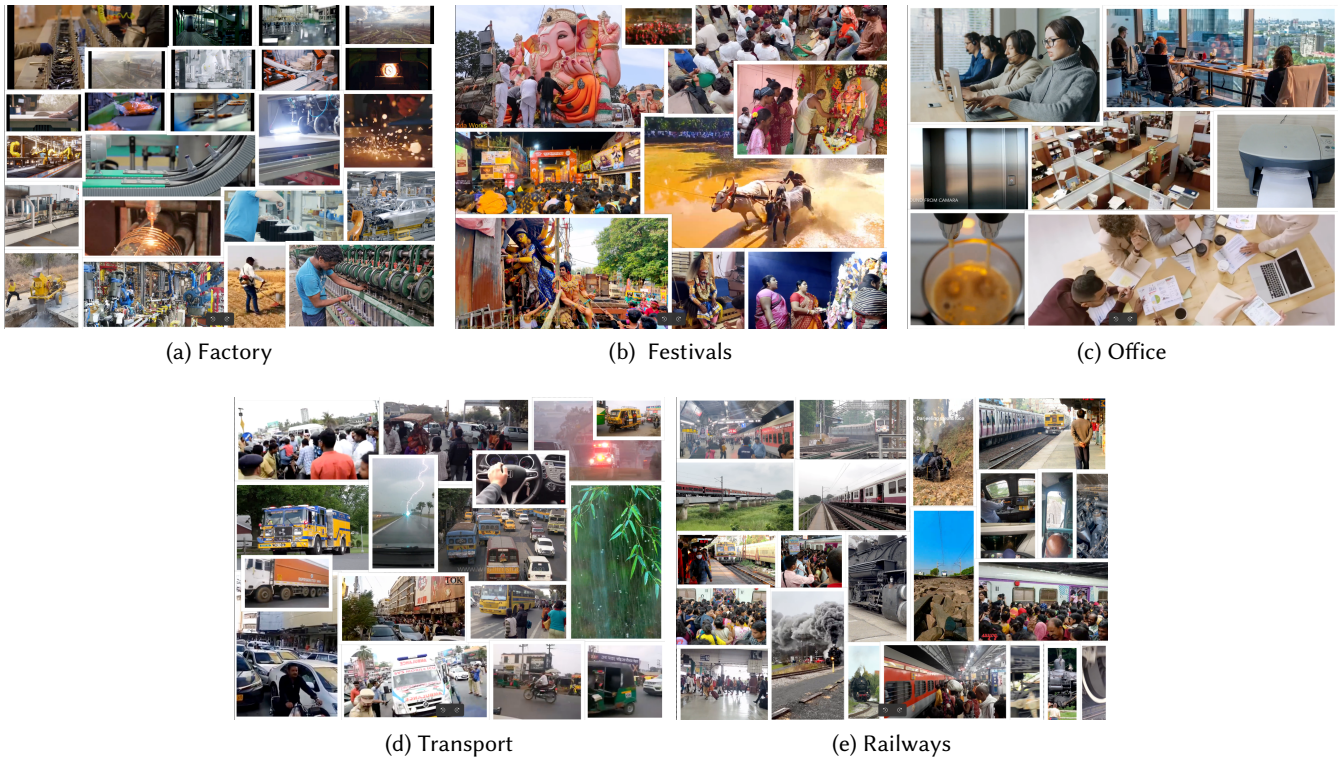
Our experiments pursue two main objectives: (1) to assess the importance of visual cues in comparison to robust speech representations from well-established pre-trained models under varying SNR conditions, and (2) to investigate the role of LoRA in fine-tuning speech enhancement models for low SNR environments, both across specific and multiple noise types.

4.2.1 The Significance of Visual Cues vs. Pre-trained HuBERT Features. To investigate the relative contributions of visual cues and pre-trained speech representations, we consider a model composed of three distinct components: (i) a frozen lip generation module that produces visual features, (ii) a trainable speech encoder, and (iii) a frozen HuBERT model. This architecture leverages visual inputs alongside robust HuBERT-based speech embeddings, with the speech encoder adapting to the target environment using both modalities. The full system architecture is illustrated in Figure 1.

We construct two variants of this model: *Model A*, trained on high SNR noisy audio (0, 5, and 10 dB), and *Model C*, trained on low SNR audio (ranging from 0 to −10 dB), to examine whether exposure to low-SNR conditions during training is essential for effective speech enhancement. To further assess the role of visual cues, we remove the lip generation module and evaluate performance using the same training setups, resulting in *Model B* and *Model D*, corresponding to training on high and low SNR data, respectively. The performance of these four models—A, B, C, and D—is evaluated across multiple SNR levels using three standard metrics: PESQ (perceptual quality), STOI (intelligibility), and Word Error Rate (WER). Results are summarized in Tables 1 and 2. When comparing Model A and Model B, we observe that the inclusion of visual cues leads to marginal improvements in PESQ, reflecting slightly enhanced speech quality. However, STOI and WER remain virtually unchanged, suggesting that visual information contributes little to intelligibility when robust pre-trained features such as those from HuBERT are already present. A comparison between Model C and Model D, both lacking visual input, highlights the effect of training data. Model D, trained on low SNR audio, achieves slightly better STOI and WER scores than Model C, particularly under challenging noise conditions. This indicates that training on low SNR data provides some benefit in preserving intelligibility, though the gains are modest.

Overall, all HuBERT-based models (A–D) significantly outperform the audio-only baseline across all SNR levels, underscoring the strength of HuBERT representations for noisy speech enhancement. Nevertheless, the performance differences among the HuBERT-based models themselves are relatively minor. These findings suggest that while visual cues and low-SNR-specific training offer incremental improvements, the dominant factor influencing model performance is the use of a robust, pre-trained speech representation. This highlights HuBERT's resilience to noise and points to diminishing returns from additional input modalities or environment-specific training in already well-initialized systems.

4.2.2 Evaluating Model Robustness Across VGG-Sound Noise Categories. To gain deeper insights into model behavior under varied



(a) Factory (b) Festivals (c) Office

(d) Transport (e) Railways

Figure 4: Illustration of five application scenarios, each represented by several internet-sourced videos. Each scenario (a)–(e) highlights a common type of noise typical to that application.

acoustic conditions, we conduct a fine-grained evaluation of Model-A and Model-B across eight distinct noise categories sourced from the VGG-Sound dataset. These categories represent diverse real-world noise environments, such as Nature, Vehicle, Animals, and Industry, each of which embodies different spectral-temporal characteristics and practical SNR variability. As shown in Table 3, both models exhibit gradual performance degradation as the SNR decreases from 10 dB to -10 dB across all categories. However, we observe consistent trends in relative performance: Model-A, which incorporates visual features, consistently outperforms Model-B in both PESQ and STOI metrics, albeit with modest margins. Furthermore, the results reveal that the noise category significantly influences enhancement quality. For example, Animals and People categories yield higher intelligibility and perceptual scores compared to more acoustically complex categories like Sports and Vehicle. This aligns with our earlier observation that the nature of background noise plays a critical role in determining model robustness. These findings underscore that while visual features offer incremental gains, the inherent difficulty posed by certain noise types—due to either overlapping frequency content or impulsive characteristics—can still challenge the model's generalization capabilities, particularly under low SNR conditions.

4.2.3 Analysis of models on Speech separation. Lip motion serves as a critical visual cue in speech separation, particularly in challenging multi-speaker scenarios. To investigate this, we constructed two test conditions using the LRS-3 test set: one involving a two-speaker mixture and the other a three-speaker mixture, without any added environmental noise. These settings simulate overlapping speech scenarios commonly encountered in real-world conversations. To evaluate the effectiveness of visual inputs in disambiguating speech sources, we employed a two-stage strategy. First, the mixed speech was fed into a pseudo lip-sync generator, simulating lip movement from the audio. Separately, the clean target speech was used to generate a target-driven pseudo lip representation. Both of these lip motion cues were subsequently provided to the speech enhancement module to assess their impact on separation performance. As seen in Table 4, the results demonstrate a consistent trend: using pseudo lips generated from clean target speech slightly improves performance over those generated from the mixed speech, particularly under low-SNR conditions (e.g., -5 dB and -10 dB). This suggests that the pseudo lip generator struggles when the input audio includes multiple overlapping voices, leading to inaccurate or ambiguous visual guidance. Furthermore, performance degradation becomes more pronounced as we move from two-speaker to three-speaker mixtures, underscoring the increased difficulty of the task. Despite these observations, the overall contribution of visual cues—whether from clean or mixed speech—remains relatively limited in the presence of strong audio representations such as those from HuBERT, suggesting that robust pre-trained audio

Table 1: Quantitative comparison of SE models under high- and low-SNR conditions. Visual cues offer slight PESQ gains at high SNR, while low-SNR training provides modest STOI improvements in noisy settings. Models A/C: full model trained on high/low SNR speech. Models B/D: variant without the synthetic lip generation module, trained under the same conditions.

SNR	Model-A		Model-B		Model-C		Model-D		Audio only	
(dB)	PESQ ↑	STOI ↑	PESQ ↑	STOI ↑	PESQ ↑	STOI ↑	PESQ ↑	STOI ↑	PESQ ↑	STOI ↑
10	2.714	0.932	2.640	0.932	2.603	0.909	2.589	0.927	2.259	0.940
5	2.459	0.917	2.383	0.917	2.391	0.896	2.363	0.913	1.938	0.907
0	2.169	0.896	2.095	0.895	2.166	0.879	2.122	0.895	1.654	0.853
-5	1.878	0.868	1.810	0.865	1.928	0.856	1.885	0.870	1.409	0.760
-10	1.624	0.833	1.569	0.829	1.698	0.828	1.659	0.840	1.243	0.598

Table 2: Word Error Rate (WER) analysis for Models A–D. A/C correspond to the full audio-visual model trained with high/low SNR speech; B/D are trained identically but omit the synthetic lip generation stage.

SNR level (dB)	Model-A	Model-B	Model-C	Model-D
10	0.140	0.134	0.162	0.137
5	0.162	0.153	0.161	0.155
0	0.177	0.176	0.162	0.153
-5	0.192	0.195	0.172	0.169
-10	0.213	0.206	0.189	0.181

features may already capture much of the information needed for separation in these settings.

4.2.4 Analyzing LoRA Placement and Rank in Fine-Tuning for Robust Audio Denoising. In our earlier experiments, training models under low SNR conditions led to improvements in challenging noise scenarios but caused a noticeable degradation in performance for high SNR cases. This revealed a trade-off between robustness and generalization. To address this, we propose incorporating Low-Rank Adaptation (LoRA) modules at every convolutional layer in the speech encoder, aiming to enhance performance under low SNR conditions while preserving the model's capabilities at high SNR levels.

Our approach builds upon Model A, which was initially trained on high SNR data. By fine-tuning this model using LoRA with low SNR audio, we aim to retain its performance in clean conditions while improving robustness in noisier environments. This method is appealing because LoRA introduces only a small number of additional parameters, enabling effective adaptation without full retraining.

To evaluate this strategy, we conducted a comprehensive ablation study by experimenting with various configurations of LoRA integration. The designs include: (1) LoRA-Enc, where LoRA is applied only to the encoder; (2) LoRA-Enc-Dec Frozen, where LoRA is added to the encoder while freezing the decoder during fine-tuning; (3) LoRA-Dec, which applies LoRA only to the decoder; (4) LoRA-Enc-Dec, where LoRA is included in both encoder and decoder; and (5) Full Fine-tuning, where Model A is fine-tuned without LoRA. All configurations used LoRA with rank 8 and $\alpha = 16$.

The results in Table 5 show that the LoRA-Enc configuration consistently achieves the best PESQ and STOI scores across all SNR levels, particularly under more challenging conditions. This validates our hypothesis that adapting only the encoder via LoRA is more effective than full model fine-tuning or modifying the decoder. Furthermore, we also evaluated the impact of different LoRA rank values, as shown in Table 6, and found that a lower rank (rank=4) provides a better balance of performance and efficiency for our task.

4.2.5 Fine-tuning LoRA on real-world noise models. To assess the practical value of fine-tuning under realistic acoustic scenarios, we evaluated LoRA-based fine-tuned models against their pre-trained VGG counterparts across five real-world noise environments: Office, Factory, Railways, Indian Roads, and Indian Festival. As shown in Table 7, both PESQ and STOI scores were recorded across multiple SNR levels ranging from 10 dB to −10 dB.

While the fine-tuned models generally achieve slightly better scores than their pre-trained versions, the improvements remain consistently marginal. For example, under the most challenging noise condition of −10 dB SNR in the Railway environment, the fine-tuned model achieves a PESQ of 1.454, compared to 1.318 for the pre-trained one—an improvement of just 0.136. Similarly, the STOI score improves from 0.793 to 0.817, a minor gain of only 0.024. Such narrow margins are typical across the dataset, regardless of the noise type or SNR level.

These results suggest that the pre-trained VGG model already generalizes well to diverse real-world noises, and the added benefit of LoRA fine-tuning is minimal in this context. Therefore, fine-tuning may not be necessary unless the target environment exhibits highly unique or domain-specific acoustic characteristics that are underrepresented in the original training data.

5 CONCLUSION

This study investigates the role of visual cues and robust pre-trained speech representations, particularly HuBERT, in enhancing speech under low SNR conditions. Our experiments show that while visual features and low-SNR-specific training offer modest improvements, the dominant contributor to performance is the use of robust, pre-trained audio models like HuBERT. We also find that visual input significantly aids in complex speech separation scenarios involving multiple speakers.

To improve low-SNR robustness without sacrificing generalization, we employ LoRA modules in the encoder, achieving a favorable trade-off in performance and parameter efficiency. However, fine-tuning on real-world noise conditions offers negligible benefit over

Table 3: Performance breakdown of Models A–D on selected VGGSound noise categories. Models with visual cues (A/C) are compared against those without (B/D), each trained on high- and low-SNR speech respectively.

Noise Model	SNR (dB)	Model-A PESQ	Model-A STOI	Model-B PESQ	Model-B STOI	Noise Model	Model-A PESQ	Model-A STOI	Model-B PESQ	Model-B STOI
	10	2.732	0.931	2.657	0.931		2.761	0.930	2.699	0.930
	5	2.483	0.916	2.403	0.916		2.513	0.916	2.449	0.915
Nature	0	2.209	0.896	2.137	0.895	People	2.246	0.896	2.171	0.894
	-5	1.931	0.870	1.866	0.867		1.970	0.870	1.900	0.867
	-10	1.675	0.836	1.610	0.832		1.710	0.838	1.649	0.835
	10	2.612	0.928	2.538	0.928		2.708	0.930	2.639	0.929
	5	2.343	0.911	2.266	0.910		2.450	0.914	2.377	0.913
Vehicle	0	2.048	0.887	1.974	0.885	Home	2.163	0.893	2.087	0.891
	-5	1.745	0.854	1.682	0.851		1.871	0.864	1.806	0.862
	-10	1.493	0.815	1.447	0.811		1.617	0.829	1.567	0.825
	10	2.769	0.932	2.700	0.932		2.617	0.928	2.546	0.928
	5	2.521	0.918	2.445	0.917		2.344	0.912	2.269	0.910
Animals	0	2.254	0.899	2.177	0.897	Sports	2.046	0.888	1.972	0.886
	-5	1.979	0.873	1.907	0.870		1.732	0.855	1.670	0.852
	-10	1.724	0.840	1.667	0.836		1.478	0.815	1.420	0.811
	10	2.671	0.926	2.595	0.925		2.695	0.936	2.622	0.936
	5	2.406	0.909	2.321	0.907		2.429	0.922	2.351	0.921
Industry	0	2.120	0.886	2.040	0.883	Instrument	2.137	0.902	2.052	0.900
	-5	1.824	0.856	1.759	0.853		1.819	0.872	1.742	0.869
	-10	1.579	0.820	1.520	0.816		1.556	0.8362	1.500	0.832

Table 4: Performance of the model for speech separation

	2-voice mix with mixed speech's pseudo lip		2-voice mix with target speech's pseudo lip	
SNR (dB)	PESQ	STOI	PESQ	STOI
10	2.555	0.929	2.565	0.931
5	2.244	0.909	2.238	0.912
0	1.862	0.878	1.879	0.883
-5	1.526	0.839	1.561	0.847
-10	1.348	0.799	1.377	0.810

	3-voice mix with mixed speech's pseudo lip		3-voice mix with target speech's pseudo lip	
SNR (dB)	PESQ	STOI	PESQ	STOI
10	2.420	0.919	2.435	0.921
5	2.089	0.894	2.097	0.898
0	1.684	0.857	1.716	0.862
-5	1.380	0.807	1.408	0.816
-10	1.214	0.761	1.236	0.773

pretrained models, suggesting limited need for domain-specific adaptation unless the noise characteristics are highly atypical.

Overall, our findings highlight the strength of pre-trained speech models and the selective value of visual cues and fine-tuning under specific conditions. Future work should explore dynamic modality fusion and adaptive training strategies to further bridge the gap in extreme noise scenarios.

REFERENCES

[1] Ahsan Adeel, Mandar Gogate, and Amir Hussain. 2020. Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. *Information Fusion* 59 (2020), 163–170.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2019. My lips are concealed: Audio-visual speech enhancement through obstructions. In *Proc. Interspeech 2019*.

[3] Alexei Baevski and Abdelrahman Mohamed. 2020. Effectiveness of self-supervised pre-training for ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7694–7698.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020).

[5] Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. 2024. LoRA Learns Less and Forgets Less. arXiv:2405.09673 [cs.LG]

[6] Anyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[7] F. Chen, Y. Hu, and M. Yuan. 2015. Evaluation of noise reduction methods for sentence recognition by mandarin-speaking cochlear implant listeners. *Ear Hear.* 36, 1 (2015), 61–71.

[8] I-Chun Chern, Kuo-Hsuan Hung, Yi-Ting Chen, Tassadaq Hussain, Mandar Gogate, Amir Hussain, Yu Tsao, and Jen-Cheng Hou. 2023. Audio-Visual Speech Enhancement and Separation by Utilizing Multi-Modal Self-Supervised Embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[9] Gillian M Davis. 2018. *Noise reduction in speech applications.* CRC press.

[10] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real Time Speech Enhancement in the Waveform Domain. In *Interspeech*.

[11] A. El-Solh, A. Cuhadar, and R. A. Goubran. 2007. Evaluation of speech enhancement techniques for speaker identification in noisy environments. In *Proc. 9th IEEE Int. Symp. Multimedia Workshops*. 235–239.

[12] Mandar Gogate, Kia Dashtipour, Ahsan Adeel, and Amir Hussain. 2020. CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement. *Information Fusion* 63 (2020), 273–285.

[13] E. W. Healy, M. Delfarah, Eric M. Johnson, and DeLiang Wang. 2019. A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation. *J. Acoustical Soc. Amer.* 145, 3 (2019), 1378–1388.

[14] Sindhu B. Hegde, K.R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2021. Visual Speech Enhancement Without a Real Visual Stream. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer*

Table 5: Ablation study of LoRA finetuning on model design

Models	SNR=10dB		SNR=5dB		SNR=0dB		SNR=-5dB		SNR=-10dB	
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
LoRA-Enc	2.761	0.934	2.515	0.920	2.255	0.901	1.988	0.877	1.734	0.846
LoRA-Enc-Dec-Frozen	2.625	0.932	2.384	0.917	2.110	0.896	1.833	0.868	1.587	0.834
LoRA-Decoder	2.698	0.931	2.443	0.916	2.172	0.895	1.891	0.867	1.638	0.834
LoRA Enc-Dec	2.700	0.932	2.443	0.917	2.168	0.896	1.886	0.869	1.635	0.835
Full Finetune	2.675	0.943	2.339	0.915	2.224	0.892	1.829	0.892	1.618	0.834

Table 6: Performance of the model with respect to several rank

Rank	2		4		8	
SNR	PESQ	STOI	PESQ	STOI	PESQ	STOI
10	2.709	0.933	2.716	0.932	2.698	0.932
5	2.449	0.919	2.716	0.932	2.698	0.931
0	2.165	0.898	2.180	0.896	2.172	0.895
-5	1.871	0.869	1.897	0.869	1.891	0.867
-10	1.613	0.835	1.644	0.836	1.638	0.834

Table 7: Comparison of Finetuned Models with Their pretrained VGG Counterparts. This figure presents the performance results of the finetuned models in comparison to their corresponding VGG models.

SNR level (dB)	Model type	Office		Factory		Railways		Indian Roads		Indian Festival	
		PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
10	Fine-tuned	2.487	0.924	2.479	0.923	2.547	0.929	2.770	0.939	2.628	0.931
	Pre-trained	2.473	0.924	2.470	0.922	2.516	0.928	2.703	0.934	2.584	0.930
5	Fine-tuned	2.209	0.905	2.219	0.905	2.284	0.913	2.536	0.927	2.367	0.916
	Pre-trained	2.182	0.903	2.169	0.901	2.239	0.909	2.460	0.921	2.318	0.912
0	Fine-tuned	1.919	0.879	1.944	0.880	2.016	0.891	2.277	0.911	2.079	0.895
	Pre-trained	1.863	0.872	1.843	0.871	1.928	0.882	2.187	0.902	2.010	0.888
-5	Fine-tuned	1.620	0.844	1.656	0.847	1.723	0.859	2.007	0.888	1.796	0.866
	Pre-trained	1.526	0.830	1.511	0.830	1.583	0.841	1.903	0.876	1.706	0.854
-10	Fine-tuned	1.353	0.799	1.390	0.803	1.454	0.817	1.745	0.857	1.538	0.830
	Pre-trained	1.266	0.781	1.259	0.781	1.318	0.793	1.620	0.840	1.444	0.812

Vision (WACV). 1926–1935.

[15] Wei-Ning Hsu et al. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[18] Anjuli Kannan et al. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model. In *Proc. Interspeech 2019*. 2130–2134.

[19] Cheng-I Lai et al. 2021. Semi-supervised spoken language understanding via self-supervised speech and language model pretraining. In *ICASSP*. IEEE.

[20] H. Levit. 2001. Noise reduction in hearing aids: An overview. *J. Rehabil. Res. Develop.* 38, 1 (2001), 111–121.

[21] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong. 2015. *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Academic Press, New York, NY, USA.

[22] Xian Li et al. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *ACL*.

[23] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[24] Wei Liu, Ying Qin, Zhiyuan Peng, and Tan Lee. 2024. Sparsely Shared LoRA on Whisper for Child Speech Recognition. In *ICAASP*.

[25] Daniel Michelsanti, Zheng-Hua Tan, Sigurdur Sigurdsson, and Jesper Jensen. 2019. Deep-learning-based audio-visual speech enhancement in presence of Lombard effect. *Speech Communication* 115 (2019), 38–50.

[26] J. Ortega-Garcia and J. Gonzalez-Rodriguez. 1996. Overview of speech enhancement techniques for automatic speaker recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 2. 929–932 vol.2.

[27] M. L. Patterson and J. F Werker. 2003. Two-month-old infants match phonetic information in lips and voice. *Developmental Science* 6 (2003), 191–196.

[28] Summerfield Q. 1979. Use of visual information for phonetic perception. *Phonetica* 34 (1979), 314–331.

[29] Sylvestre-Alvise Rebuffi et al. 2017. Learning multiple visual domains with residual adapters. *Advances in Neural Information Processing Systems* 30 (2017).

[30] W. H. Sumby and I. Pollack. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26 (1954), 212–215.

[31] Wenxin Tai. 2022. A Repetitive Spectrum Learning Framework for Monaural Speech Enhancement in Extremely Low SNR Environments (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (2022), 13063–13064.

[32] Bethan Thomas, Samuel Kessler, and Salah Karout. 2022. Efficient adapter transfer of self-supervised speech models for automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7102–7106.

[33] Steven Vander Eeckt et al. 2022. Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition. *arXiv e-prints* (2022), arXiv–2203.

[34] E. Vincent, T. Virtanen, and S. Gannot. 2018. *Audio Source Separation and Speech Enhancement.* Wiley, Hoboken, NJ, USA.

[35] Karren Yang, Dejan Marković, Steven Krenn, Vasu Agrawal, and Alexander Richard. 2022. Audio-Visual Speech Codecs: Rethinking Audio-Visual Speech Enhancement by Re-Synthesis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 8217–8227.

[36] Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. 2024. Convolution Meets LoRA: Parameter Efficient Finetuning for Segment Anything Model. In *International Conference on Learning Representations.*