A Front-End Adaptation Network for Improving Speech Recognition Performance in Packet Loss and Noisy Environments

Yehoshua Dissen, Shiry Yonash, Israel Cohen, *Fellow, IEEE*, and Joseph Keshet, *Senior Member, IEEE*

***Abstract*—Robust automatic speech recognition (ASR) in packet loss and noisy environments remains a significant challenge. Large pretrained transformer models have made notable strides in improving ASR performance across diverse domains. However, considerable room remains for improvement, even in moderate packet loss and noise conditions. Enhancing these models is particularly difficult because retraining is computationally prohibitive, and fine-tuning introduces the risk of domain shift, which can degrade performance in other languages or environments. We introduce a novel method that leverages a front-end adaptation network to improve word error rate (WER) performance in scenarios with packet loss and noise. Our approach addresses the constraints of working with large pretrained ASR models while avoiding retraining or fine-tuning. We connect an adaptation network to a frozen ASR model, where the network is trained to modify corrupted input spectra using both the loss function of the ASR model and an enhancement loss. This strategy allows the system to adapt to packet loss and noise without compromising the performance of the original ASR model or generalization across domains. The method focuses on improving WER rather than signal quality or intelligibility, targeting it for ASR applications. We conduct a comprehensive set of experiments on various types of noise. Our results demonstrate that the adaptation network significantly reduces WER in all conditions while preserving the foundational performance of the pretrained ASR model.**

***Index Terms*—Packet loss concealment, robust automatic speech recognition, speech enhancement.**

I. Introduction

RECENTLY, large transformer models have been utilized for achieving state-of-the-art results in automatic speech recognition (ASR) [1], [2], [3]. These models, trained on vast amounts of data across multiple domains, have demonstrated effectiveness in various domains and languages. In addition, due to the diversity and sheer volume of the training data, they have demonstrated robustness to various types of noise, including additive noise (e.g., random white noise, background noise), isotropic noise, reverberant speech, and even packet loss. In certain scenarios, such as recognition in white noise, results approach human-level performance [4]. However, its performance in challenging environments characterized by low signal-to-noise ratios (SNRs), high reverberation times (RT60s), or significant packet loss still leaves much room for improvement.

This work focuses on the prevalent scenario where a large pretrained transformer ASR model already exists, and the goal is to improve the model's performance under noisy and packet loss conditions. The paper primarily focuses on the packet loss scenario, but also evaluates other noises and layered noises. Enhancing the robustness of these models under adverse acoustic conditions is not a straightforward task. A primary challenge arises from the integrated nature of a pseudo-language model within the model, namely, the decoder part of the transformer. Fine-tuning these models not only adapts to the noise but also learns the domain of the fine-tuning data. Hence, it can easily overfit to the domain of data you fine-tune with. For example, if fine-tuned on noisy English data, the model might "forget" other languages [5]. Even within the same language, the model can improve on read speech while degrading performance on phone call speech.

Retraining foundational ASR models from scratch to handle adverse conditions such as noise, reverberation, or packet loss is also impractical due to the large size of the model and the vast amount of data required, which can demand excessive computational resources and time. An alternative approach to improving ASR performance under noisy conditions is to utilize a packet loss concealer or a speech enhancement model. Both packet loss concealment and speech enhancement have been extensively studied and shown to improve human speech intelligibility in various contexts [6], [7], [8], [9]. These models typically aim to fill gaps and remove noise from the signal by enhancing audio quality, which benefits human listeners. However, the distortions or artifacts introduced during enhancement can negatively affect ASR model performance [10], [11].

Most single-channel speech enhancement techniques focusing on ASR have shown limited performance improvements compared to ASR systems trained on multicondition data. Moreover, these require the ASR model to be retrained using the enhanced audio, which demands significant computational resources and data [12]. While there are ASR front-end Speech

Received 30 October 2024; revised 11 March 2025, 24 April 2025, and 18 May 2025; accepted 20 May 2025. Date of publication 29 May 2025; date of current version 9 June 2025. This work was supported in part by Technion Internal under Grant 2071452, in part by Israel Science Foundation under Grant 1449/23, and in part by Pazy Research Foundation. The associate editor coordinating the review of this article and approving it for publication was Dr. Jonas Borgstrom. *(Corresponding author: Shiry Yonash.)*

The authors are with the Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Haifa 3200003, Israel (e-mail: yehoshuad@campus.technion.ac.il; shiry.yonash@alumni.technion.ac.il; icohen@ee.technion.ac.il; jkeshet@technion.ac.il).

Digital Object Identifier 10.1109/TASLPRO.2025.3574840

Enhancement (SE) models [13], [14], [15], and jointly trained SE, ASR models in the time-frequency domain [16], [17], [18] or in the time domain [19]. Retraining the ASR backend becomes less practical with the introduction of large pretrained ASR models. This work focuses on the case where the ASR model already exists, and we only train a front-end network without touching the original ASR model.

We aim to develop a simple method for improving noise robustness for large transformer-based ASR models. We strive to achieve this without requiring in-domain data for fine-tuning, without compromising the model's generality across domains and languages, and without adding a significant number of trainable parameters to the existing model. Importantly, our objective is to improve ASR performance, prioritizing transcription accuracy over perceptual audio quality.

To this end, we focus on the ASR input features, specifically the spectral inputs of the model. Instead of altering the ASR model, we propose a lightweight adaptation network that modifies the input spectrum to recover missing frames and denoise the signal before it is processed by the ASR model. Drawing inspiration from architectures commonly seen in SE and packet loss concealment (PLC) models, we utilize a U-Net [20] architecture with skip connections. However, unlike traditional speech enhancement approaches, our focus is not on improving audio quality for human listeners, but rather on minimizing the word error rate (WER) for ASR. Therefore, rather than relying on perceptual loss functions, we use the frozen transformer ASR loss function to update the weights of the adaptation network.

This work extends our previous study [21], where we demonstrated the effectiveness of the model in handling packet loss scenarios. This paper expands our evaluation to include a broader range of noise conditions, including additive noise, reverberation, and packet loss, to create a more universal enhancement model. We include a study of how our method affects traditional intelligibility metrics, along with a webpage where users can subjectively evaluate the samples. We compare to LoRA fine-tuning. We also investigate the effects of reducing phoneme spans and joint training across various noise scenarios, and analyze how different noise types impact the WER. Using Whisper [1] as our pretrained ASR model, we demonstrate through our evaluations that our proposed framework enhances noise robustness across multiple domains without significantly degrading the original ASR performance.

This approach offers a practical and efficient solution to enhance the performance of large ASR models, such as Whisper, in noisy environments. By freezing the weights of the ASR model and only modifying the input spectrum, we avoid the common pitfalls of domain overfitting while maintaining the model's generalizability across languages and domains. Furthermore, our lightweight adaptation model requires minimal additional parameters relative to the ASR models, making it computationally efficient and suitable for large-scale deployment. In this work, we investigate the limitations of this method's effectiveness, examining the types of noise the model can handle most effectively and the extent to which the adaptation network can reconstruct packet loss or noise corruption without significantly impacting ASR performance. This comprehensive evaluation provides insights into the robustness of the approach across different acoustic conditions. The main contribution of this work is a method that enhances the robustness of foundational ASR models to various types of corruption without altering the underlying ASR model or compromising its results through a lightweight adapter model. Our implementation and trained models are available here[1] and a page with audio samples is available here[2]

The paper is organized as follows. Section II overviews the most relevant related work. Section III describes the formal problem setting. Section IV describes the evaluation setup, including the databases used and the training procedure, and Section V shows the results. We conclude the paper in Section VI.

II. Related Work

Single-channel noise-robust ASR is an active research area, with various approaches being developed to mitigate the impact of noise on automatic speech recognition performance. Many methods focused on augmenting training data by adding different types of noise, such as white noise, reverberation, or babble noise. This approach enables models to generalize more effectively in noisy environments, as demonstrated by models trained or fine-tuned on augmented data [22], [23], [24]. Another prominent strategy for improving robustness in noisy environments is using front-end speech enhancement models. These models focus on improving the audio input quality by learning a noise spectrum mask before passing it to the ASR system. Speech enhancement has been extensively studied in the context of improving perceptual speech quality for human listeners, often measured through metrics such as Perceptual Evaluation of Speech Quality (PESQ) [25] and Short-Time Objective Intelligibility (STOI) [12]. However, although these models improve human intelligibility, they introduce distortions or artifacts that can negatively impact the performance of the ASR model [10]. The challenge is that enhancing audio for humans does not always improve the transcription of ASR systems, resulting in limited improvements in word error rate (WER) despite audio enhancements.

Several recent works have addressed this issue by designing enhancement models targeting ASR performance rather than perceptual quality. For example, Subramanian et al. [26] introduced an end-to-end system that uses ASR objectives to guide the training of a speech enhancement model. By focusing on minimizing the WER rather than enhancing the perceptual quality of the audio, they achieved improvements in both ASR performance and traditional enhancement metrics. Iwamoto et al. [27] investigated the effectiveness of jointly training an SE front-end and an ASR backend, finding that this training can reduce ASR errors on SE artifacts. However, it increases the errors due to noise. Chang et al. [28] jointly trained ASR, SE, and self-supervised learning representation to achieve SOTA results on the CHiME-4 [29] benchmark. Yang et al. [19] proposed a joint training framework that integrates time-domain speech

[1] https://github.com/MLSpeech/WhisperDenoiser
[2] https://shuadissen.github.io/ASR_denoiser/

enhancement (SE) with an end-to-end ASR system using latent representations, eliminating the need for waveform reconstruction. To achieve this, they introduced a convolutional network to transform the SE encoder's latent representation into features compatible with ASR. To improve performance, they also modified Conv-TasNet [30] into an attention-based variant. However, none of these address the scenario where one has a foundational ASR model that cannot be changed.

Packet loss concealment (PLC), the primary focus of this work, is another critical area within noise-robust ASR. This task involves recovering lost audio frames that are corrupted or dropped during transmission. Early methods for PLC, such as linear prediction or interpolation [31], [32], aimed to reconstruct the missing parts of the signal based on statistical models. With the rise of deep learning, more sophisticated approaches have emerged, utilizing neural networks to tackle this problem. Encoder-decoder frameworks have become a prevalent choice, with models like those introduced by Wang et al. [33] and Pascual et al. [34] using adversarial training to generate natural-sounding audio in the gaps created by packet loss. Westhausen and Meyer [35] developed a time-domain PLC model (tPLCnet) that uses recurrent networks to predict the next frame based on a short context buffer. Lin et al. [36] approached PLC as a generative regression problem, utilizing convolutional encoder-decoders with LSTM layers to predict future frames in the time domain. More recently, diffusion models have been used [37], [38] to generate more naturalistic sounds to fill in the gaps. However, despite these advancements, most PLC models still prioritize improving human intelligibility over optimizing ASR accuracy, which limits their effectiveness in scenarios where WER is the primary concern.

Recently, more focus has been put on ASR-based improvements to improve concealment and ASR robustness. Yang et al. [39] proposed an auxiliary loss function that encourages the latent representations of distorted and clean signals to align, improving packet loss concealment for ASR tasks by focusing on generating more accurate reconstructions of the input signal. Zhang et al. [40] introduced semantic awareness into PLC approaches, recognizing that linguistic information can guide more effective reconstruction of lost speech segments, facilitating more accurate reconstruction, particularly in scenarios characterized by extended burst packet losses. By incorporating semantic information, this approach provides the receiving system with contextual knowledge, enabling it to infer missing speech elements more accurately based on meaning rather than relying solely on acoustic reconstruction. Similarly, our method achieves a form of semantic awareness by directly optimizing for Word Error Rate (WER), which inherently incorporates linguistic and contextual information. Using WER as the guiding loss function, our adaptation network is trained to reconstruct missing or corrupted audio to maximize transcription accuracy.

III. Methods

This study proposes a technique that improves ASR robustness to packet loss and noisy scenarios while maintaining the pre-trained ASR architecture and weights. As stated earlier, one option would be to use a packet loss concealment or speech enhancement module to reconstruct the speech and subsequently apply ASR on the resulting signal. However, this solution is suboptimal, as these models can introduce artifacts detrimental to the performance of the ASR model. Here, we consider a different approach, replacing the PLC module with a front-end network that adapts the signal to improve ASR robustness under packet loss and unknown noise conditions, rather than enhancing the speech quality.

We start by presenting the notation and general setting. We denote the speech signal or its representation (e.g., Mel spectrum) by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t$ denotes a speech sample or mel frame and $\mathbf{X}$ is a sequence of T such elements. In our general setting we assume that signal $\mathbf{X}$ is corrupted by noise $\mathbf{N} = (\mathbf{n}_1, \ldots, \mathbf{n}_T)$, where each $\mathbf{n}_t$ is a vector of noise. Let us also define the packet-loss operator PL, which gets as input a speech signal and returns a corrupted speech signal. It also gets two parameters: the start frame K and the duration J for which these frames are lost. Namely,

$$\mathrm{PL}(\mathbf{X}; K, J) = \begin{cases} \mathbf{0}, & \text{for all t} \in [\mathrm{K}, \mathrm{K} + \mathrm{J} - 1] \\ \mathbf{x}_t, & \text{otherwise.} \end{cases} \tag{1}$$

Overall, in our generalized setting, the corrupted speech signal is $\tilde{\mathbf{X}} = \mathrm{PL}(\mathbf{X} + \mathbf{N}; K, J)$, where the parameters K and J, as well as the noise type, are considered in the empirical evaluation section. Note that we may apply the packet loss operators multiple times on a given input.

We assume a transcript $\mathbf{y} = (y_1, \ldots, y_U)$ is associated with the speech signal, where $\mathbf{y}$ represents a sequence of U words or sub-words (tokens). Note that T and U differ for each input and target sequence. Our objective is to propose a model that receives the corrupted speech $\tilde{\mathbf{X}}$ and outputs the target transcription $\mathbf{y}$ as if it had received the original (unobserved) signal $\mathbf{X}$.

Our model consists of two main components: a front-end *adaptation* network and a *frozen* ASR model. We denote the ASR model by $\mathbf{g}_\phi$ with parameter set ϕ. This function $\tilde{\mathbf{y}} = \mathbf{g}_\phi(\tilde{\mathbf{X}})$ takes as input a corrupted speech signal and predicts the word (token) sequence spoken. The ASR model was trained using a loss function denoted by $L_{\mathrm{ASR}}(\mathbf{g}_\phi(\mathbf{X}), \mathbf{y})$. Our objective is to enhance the performance of this pre-trained ASR on corrupted input signals, without applying any fine-tuning to the model itself.

We aim to design an adaptation network $\mathbf{f}_\theta$ with parameter set θ. This network receives the noisy speech $\tilde{\mathbf{X}}$ and outputs an adapted version $\hat{\mathbf{X}} = \mathbf{f}_\theta(\tilde{\mathbf{X}})$, which is used as input to the ASR model, yielding $\hat{\mathbf{y}} = \mathbf{g}_\phi(\hat{\mathbf{X}})$. Our goal is to have $\mathrm{WER}(\hat{\mathbf{y}}, \mathbf{y}) \leq \mathrm{WER}(\tilde{\mathbf{y}}, \mathbf{y})$. We note that the generated $\hat{\mathbf{X}}$ is tailored to improve the ASR's performance and may not necessarily enhance human intelligibility.

More specifically, we propose to train the adapter network by passing gradients from an ASR loss function through an ASR model (but keeping the weights of the ASR model frozen) and by adding a regularization function that ensures the corrected signal is close to the original uncorrupted signal.

Formally,

$$\min_{\theta} \lambda L_{\text{ASR}} \left(g_{\phi}(f_{\theta}(\tilde{\mathbf{X}})), \mathbf{y} \right) + (1-\lambda) L_{\text{reg}} \left(\mathbf{X}, f_{\theta}(\tilde{\mathbf{X}}) \right) \quad (2)$$

where L_{reg} is a regularization loss function. We emphasize that the minimization is over the adapter network's parameters θ, while the ASR model's parameters ϕ remain fixed. In the evaluation, we demonstrate the advantages of our model over fine-tuning the parameters ϕ of the ASR model.

In this work, we focus on the state-of-the-art supervised transformer-based ASR model Whisper [1]. In this case, the input $\mathbf{X}$ is the Mel spectrum, and the model is trained with the cross-entropy loss function $L_{\text{CE}}(\mathbf{g}_{\phi}(\tilde{\mathbf{X}}), \mathbf{y})$. For the adaptation network, we use a convolutional U-net architecture [20]. While our experiments focus on Whisper due to its strong performance, multilingual coverage, and public availability, the proposed framework is not restricted to Whisper or to mel-spectrogram inputs. For ASR systems such as wav2vec2.0 [41] or HuBERT [42] that operate on raw waveform inputs, our method can be adapted by replacing the mel-to-mel U-Net with a raw-to-raw front-end model architecture, such as Demucs [43] or Conv-TasNet [30]. Moreover, even though Whisper uses cross-entropy loss, models trained with connectionist temporal classification (CTC) losses, as is common with wav2vec 2.0 and HuBERT, are fully differentiable and can support joint training with our front-end adaptation strategy.

The adaptation network is trained with two loss functions. The first is the ASR model's principal loss function, which for Whisper is cross-entropy L_{CE}. This loss function guides the adapter network toward generating a spectrum that improves token classification accuracy. However, we found that training with ASR loss alone sometimes resulted in unstable convergence. While WER initially improves, it could suddenly degrade due to large gradients from the ASR model disrupting training or the model learning degenerate spectrogram transformations that momentarily reduce loss but ultimately harm ASR performance. To mitigate this, we introduced a secondary loss term: an L_1 loss component between the original signal $\mathbf{X}$ and the adapted signal $f_{\theta}(\tilde{\mathbf{X}})$. This regularization prevents the model from generating unrealistic spectrograms by ensuring that some of the loss penalizes extreme spectral deviations.

Empirically, we found that weighting the loss at approximately $1/50th$ of the ASR loss provided the best stability-performance tradeoff. This small weighting stabilized training without interfering with the primary optimization objective. So the final loss function takes the following form:

$$\min_{\theta} \lambda L_{\text{CE}} \left(g_{\phi}(f_{\theta}(\tilde{\mathbf{X}}), \mathbf{y}) \right) + (1-\lambda) L_1 \left(X, f_{\theta}(\tilde{\mathbf{X}}) \right). \quad (3)$$

The adaptation network is a fully convolutional U-net with skip connections and a residual block-based bottleneck. Downscaling is performed using max-pooling, and upscaling is achieved through nearest-neighbor resizing followed by a convolutional layer. The input to the ASR model is a mel-spectrogram; hence, the adapter network is designed to receive the mel-spectrogram of the noisy signal $\tilde{\mathbf{X}}$ and output an adapted mel-spectrogram $\hat{\mathbf{X}}$. This process is illustrated in Fig. 1.

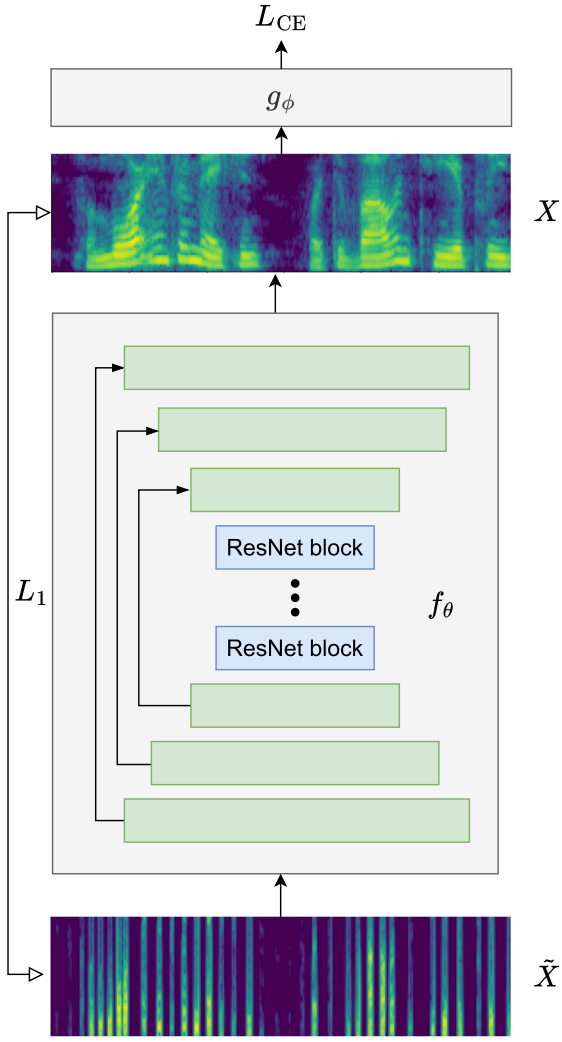


Fig. 1. Model Architecture where $\tilde{\mathbf{X}}$ is the corrupted spectrum, $\mathbf{X}$ is the clean spectrum, $\mathbf{g}_{\phi}$ is the frozen asr model and $\mathbf{f}_{\theta}$ is the trainable adapter model.

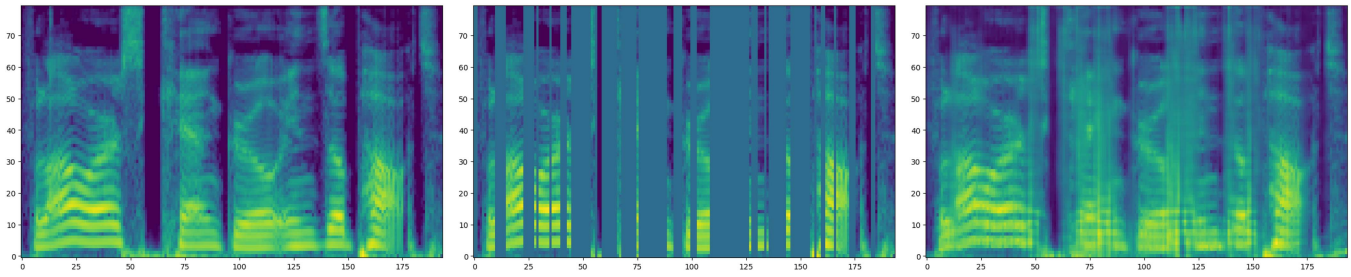
IV. Evaluation Settings

In this study, we aim to focus on the approach and ensure we do not confuse results with model improvements, complexity, or data-related artifacts, such as domain overfitting. For these reasons, we selected an existing U-Net architecture for the model and utilized Librispeech as our training data, while testing on a diverse range of domains. Essentially, we aim to explore the boundaries of this method by assessing its ability to reconstruct or denoise using ASR loss from a frozen model, without relying on in-domain data and without adding excessive parameters to the model 2.

A. Noise Types

In this study, we focus on these additive and interruption noises.

Packet Loss: In digital communications, packet loss results from the loss of bits of data during transmission, leading to incomplete auditory signals. This results in gaps or distortions



(a) Original spectrogram. (b) Corrupted spectrogram. (c) Recovered spectrogram.

Fig. 2. Example of original, corrupted, and recovered spectrograms of the utterance "flowers can grow in the pot" taken from the ALLSSTAR dataset. (a) Original spectrogram. (b) Corrupted spectrogram. (c) Recovered spectrogram.

in speech signals, significantly hampering the ASR system's ability to process spoken input accurately.

White Noise: White noise is uniformly distributed across all frequencies, presenting as a constant background sound.

Pub Noise: Characterized by the sound of multiple voices speaking simultaneously, pub noise is common in crowded environments. It introduces a complex mixture of speech-like sounds that can confuse ASR systems.

Reverberation: Reverberation is the persistence of sound after it is produced, caused by reflections from surfaces such as walls, ceilings, and floors. This prolongation of sound can blur speech signals, making it challenging for ASR systems to recognize words accurately.

B. Datasets

For training, we use the 960 hours of English LibriSpeech [44] data with various types of corruption applied to the data. We utilize multiple datasets from diverse domains to assess the method's robustness. A subset of ALLSSTAR [45], which is a collection of L1 Mandarin speakers speaking English [4], and Fleurs [46] for testing on multiple languages. Additionally, we evaluated the models on the blind set from the Interspeech 2022 Audio Deep PLC Challenge [47] and the validation set from the ICASSP 2024 Audio Deep PLC Challenge [48] for evaluating the PLC. We do not report improvements on the LibriSpeech test, as they are not significant since the improvements can be attributed to overfitting to the training domain.

1) Packet-Loss: For the packet-loss simulation, we randomly zero out frames based on two probabilities: a drop frequency (the percentage of frames that are zeroed out per utterance) and a probabilistic distribution that controls the length of consecutive frame losses. A single utterance may experience multiple spans of packet loss. During training, we utilize a drop frequency distribution, and each sample loaded is assigned a specific drop rate. For inference and reporting purposes, we duplicate the test set across multiple fixed drop frequencies. Due to the nature of the span length distribution, there may be slight variations (up to a tenth of a percent) from the fixed rate observed. When reporting the packet loss percentage, we refer to the total percentage of lost frames in the utterance. The Interspeech 2022 [47] and ICASSP 2024 [48] Audio Deep PLC Challenges have predetermined packet loss rates, which were created by collecting traces of packet losses from real Microsoft Teams calls, and we utilize these as is.

2) Additive Noise: In the additive noise scenarios, during training, each recording was either left alone or overlaid with random white noise or randomly sampled babble noise from YouTube videos. The noise was set at an SNR of either 2, 4, 6, or 8 dB. During inference we duplicate the evaluation sets and overlay them with random white noise or pub noise from the Audio Degradation Toolbox [49] used in [1] with SNR from -4 dB to 8 dB in steps of 2 dB (-4 dB, -2 dB, 0 dB, 2 dB, 4 dB, 6 dB, 8 dB) in addition to the original recordings without noise (quiet, Q). The set from [4] includes a prepared set of audio files with random white noise at various SNRs (from -4 to 8 dB), so we do not add anything here.

3) Reverberation: In the reverberation scenarios, each recording was convolved with real and simulated room impulse responses (RIRs) from the Multichannel Impulse Response Database [50] and the Room Impulse Response and Noise Database [22]. For inference, the evaluation sets were duplicated, and each set was convolved with RIRs of different reverberation times (RT60), ranging from mild (0.1 seconds) to severe (1.5 seconds). The original clean recordings, without reverberation, were also evaluated for comparison.

C. Frozen ASR Models

Whisper comes in five sizes with increasing parameters: tiny (39 M), base (74 M), small (244 M), medium (769 M), and large (1550 M). These are all transformer encoder-decoder models. Some also come in two varieties: English only or multilingual. For our experiments, we only use the multilingual models. For computational reasons, we ran most of the ablation experiments on the base model. However, we also report some on the Whisper large-V2 model, which is likely of more interest. In all evaluations, we used the same Whisper parameters for decoding: a beam size of 5, no-timestamps set to True, and manually set the language.

D. Model Specs

Our adapter model utilizes three downsampling and upsampling layers, each reducing the size by 50%, with six ResNet blocks serving as the bottleneck layers. Additionally, there are single-input and single-output convolutional layers that retain

the same dimensions. This model has 7.5 million trainable parameters, making it a negligible addition to Whisper.

V. Results

In this section, we comprehensively evaluate our model under various conditions and configurations. We assess its performance across distinct noise scenarios, including white noise, pub noise, reverberation, packet loss, and more complex combinations of layered noise. We compare the performance of various Whisper model sizes. Additionally, we benchmark our model against established speech enhancement systems and packet loss concealers, and evaluate its performance relative to fine-tuning and LoRA fine-tuning.

A detailed analysis of the implications of dropping different lengths of phoneme spans, measuring Word Error Rate (WER), and insertions, deletions, and substitutions is presented. We also examine how various loss functions affect the model's performance, and test whether adapting to the specific noise condition helps vs. a universally trained model. Furthermore, we demonstrate the versatility of our approach by testing it on multiple datasets, ensuring a thorough evaluation across diverse scenarios, and assessing its applicability to real-world conditions. For all our models, the only training data used was English LibriSpeech data. So all results on multilingual test sets are there to show that the model was able to retain its generality, and that our adapter model was able to isolate the noise, learn how to filter it, and not overfit to the training domain.

A. Stand Alone PLC

We start by evaluating the model as a stand-alone PLC. In this section, we present the evaluation of the proposed method and analyze the effect of different loss functions on WER in the packet loss scenario. We demonstrate the relative improvement of the proposed method over the unchanged baseline Whisper models, and recently published, open-source PLC models [35] and [51]. We then evaluate the model's robustness to different domains and compare it to fine-tuning Whisper. In all the experiments, the Whisper baselines refer to the original models with no PLC applied, to keep the dimensions the same, we use *zero-fill* for the dropped frames.

Figs. 3 and 4 present the performances of Whisper base and Whisper large-v2, respectively, on the original mel-spectrums in comparison with the spectrums generated by our adaptation networks and by PLC models FRN and tPLCnet. The graphs present WER for various packet loss rate (PLR) values on the ALLSSTAR dataset. We note that the vanilla Whisper large model is more robust to frame loss. It starts to seriously degrade only at PLRs larger than 20%, whereas the base model starts degrading immediately.

We present the effect of training the adaptation network with each loss function. Specifically, we compare the performance while training (i) solely using the CE loss function, L_{CE}, where the gradients flow from Whisper (noted as CE only); (ii) solely L_1 loss between the clean and lossy signals without referencing Whisper (noted as L1 only), which can be seen as similar to the TF-Unet in [52]; and (iii) a combined loss of L_{CE} and L_1 (CE + L1). We also compare packet loss training alone vs combined noisy and packet loss training indicated by the PL and all augmentation trains, respectively. We wanted to see if single-condition training would result in better results on that condition. We can see here that training with multiple augmentations resulted in better performance than sole packet loss training. The large-v2 model was trained solely on packet loss.

All methods improve WER over the original Whisper model, with the cross-entropy (L_{CE}) loss yielding more significant improvements than the L_1 enhancement loss. However, the best performance is achieved when both losses are combined. The L_1 loss acts as a regularizer, preventing instability and

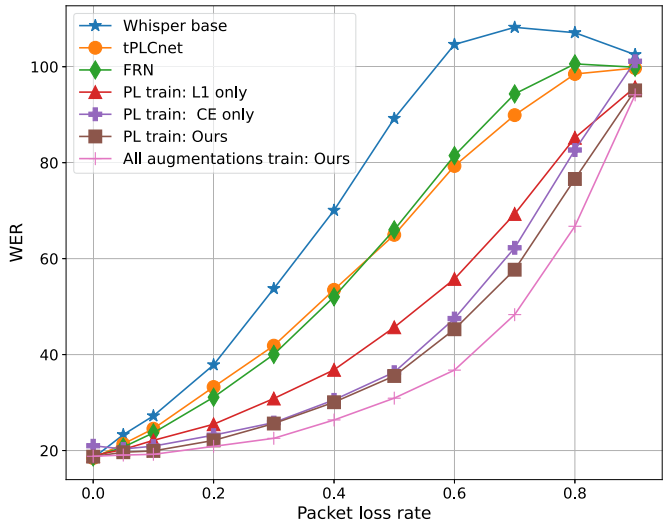


Fig. 3. WER of different models on ALLSSTAR with various PLRs. All the decoding is done with the Whisper base. FRN refers to Nguyen et al. [51], tPLCnet refers to Westhausen and Meyer [35]. PL train indicates packet loss only training. All augmentations refer to training with reverberations, white and pub noise, and packet loss.

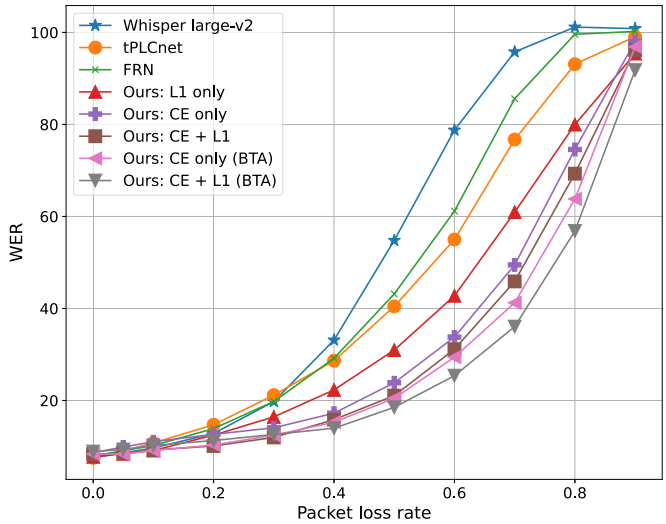


Fig. 4. WER of different models on ALLSSTAR with various PLRs. All the decoding is done with Whisper large-v2. FRN refers to [51], tPLCnet refers to [35]. BTA refers to adaptation networks trained on the Whisper base but connected to the Whisper large-v2.

TABLE I
COMPARISON OF WER FOR DIFFERENT LANGUAGES USING WHISPER BASE AND LARGE-V2

Whisper Size		Base				Large-V2			
Packet Loss Rate	Model	French	German	Russian	Spanish	French	German	Russian	Spanish
0%	Whisper	**24.7**	**17.2**	**20.3**	**10.3**	**7.2**	**4.6**	6.4	3.7
	Ours	25.9	17.8	21.6	10.6	7.6	4.7	6.4	3.7
5%	Whisper	29.1	20.8	24.6	12.7	**7.5**	**4.7**	**6.5**	**3.7**
	Ours	**27.4**	**18.9**	**22.8**	**11.1**	7.8	4.8	6.7	3.9
10%	Whisper	33.8	25.3	28.7	15.8	8.5	5.0	6.8	3.9
	Ours	**28.1**	**19.9**	**23.9**	**11.4**	**8.1**	5.0	6.8	3.9
20%	Whisper	48.6	39.0	39.1	24.7	10.3	6.0	7.9	4.2
	Ours	**31.0**	**23.1**	**26.7**	**12.6**	**8.8**	**5.7**	**7.5**	**4.1**
30%	Whisper	69.5	60.7	53.7	38.3	16.0	8.3	11.5	5.3
	Ours	**34.6**	**25.9**	**30.3**	**15.0**	**9.9**	**6.3**	**8.8**	**4.4**
40%	Whisper	104.9	102.2	75.6	57.2	27.5	13.4	18.2	7.5
	Ours	**39.5**	**31.2**	**35.8**	**17.7**	**12.3**	**7.6**	**9.9**	**5.2**
50%	Whisper	126.7	141.2	100.8	89.2	48.3	26.4	34.4	12.7
	Ours	**46.9**	**38.6**	**42.8**	**23.1**	**15.5**	**10.3**	**12.7**	**6.2**
60%	Whisper	124.6	140.4	121.1	127.9	71.7	51.8	61.5	26.6
	Ours	**57.6**	**48.3**	**54.1**	**30.9**	**20.6**	**15.5**	**19.1**	**9.4**

unrealistic spectrogram transformations, while the L_{CE} loss directly optimizes for ASR performance.

Empirically, we found that weighting the L_1 loss at approximately $1/50th$ of the ASR loss provided the best tradeoff. Assigning too much weight to L_1 degraded performance since L_{CE} is the primary driver of WER improvement. Conversely, using too little L_1 loss made it ineffective in stabilizing training. The selected weighting was small enough to avoid interfering with ASR optimization while preventing the model from entirely ignoring the L_1 loss. Finally, training with all noises denoted 'All augmentations train: Ours' improved the packet loss scenario over training with packet loss alone. In Fig. 4, we also presented the performance of the adaptation network trained on the gradients of the base model (the one that is depicted in Fig. 3) but connected and evaluated with Whisper large-v2. In Fig. 4, we denote these models as *Based Trained Adaptation (BTA)* and label them with *Ours: CE (BTA)* and *Ours: CE+L1 (BTA)*). Interestingly, training the model using Whisper base and connecting it to Whisper large-v2 gets better results than the models trained directly using Whisper large-v2. We assume this is because the gradients of the base model are easier to handle and, therefore, more effectively influence the adaptation networks. This suggests that better training parameters exist for the large model. We defer this issue for further research. This example underscores the broader principle that applying ASR metrics in PLC model training can significantly enhance ASR performance across various models.

Furthermore, the graph shows the WER of FRN [51] and tPLCnet [35]. tPLCnet is a time-domain many-to-one RNN model for PLC trained with a combined magnitude and complex mean absolute error loss in the time-frequency domain. We ran the large version of this model on the corrupted files and then decoded them with Whisper (base and large). FRN is an autoregressive RNN-based PLC model trained with a multi-resolution STFT loss. The model does not require additional inputs such as a loss mask, and conceals the signal in a blind fashion. The graph shows that these methods both improve the WER at a similar rate, with tPLCnet having the slight edge. However, the models trained using ASR metrics significantly improve the WER.

TABLE II
COMPARISON OF WER FOR DIFFERENT MODELS ON THE 2022 PACKET LOSS CHALLENGE BLIND SET

Model	WER% (base)	WER% (large-v2)
Whisper	24.0	15.4
tPLCnet [35]	20.4	16.2
FRN [51]	21.8	16.2
Ours	**18.1**	**14.2**

TABLE III
COMPARISON OF WER FOR DIFFERENT MODELS ON THE 2024 PACKET LOSS CHALLENGE VALIDATION SET

Model	WER% (base)	WER% (large-v2)
Whisper	36.1	23.1
tPLCnet [35]	30.0	25.3
FRN [51]	30.1	24.0
Ours	**29.7**	**20.5**

To ensure that our evaluation includes conditions reflecting modern wireless and VoIP networks, in Tables II and III, we compare the WER of the baseline Whisper models, tPLCnet [35] (large), FRN [51], and Ours on the blind set from the Interspeech 2022 PLC Challenge [47] and the validation set from the 2024 PLC Challenge [48]. The Interspeech 2022 and ICASSP 2024 Audio Deep PLC Challenges have predetermined packet loss rates derived from real packet traces captured in Microsoft Teams video conferencing sessions. As shown, tPLCnet and FRN reduce the WER compared to the baseline model, with tPLCnet having a slight edge. However, neither approach improves upon Whisper large, whereas Ours achieves the best performance in both scenarios.

To further illustrate the model's robustness across different domains and to demonstrate that this training method does not negatively impact the performance of the original Whisper models – unlike fine-tuning, which can degrade a model's performance in other domains or languages – we compare the WER of our model to that of the original Whisper models. This comparison is made using multiple languages randomly selected from the Fleurs dataset [46], as shown in Table I. Here, the pattern is similar to the results on the ALLSSTAR dataset, as shown in Figs. 3 and 4: the base model starts degrading immediately,

TABLE IV
COMPARISON OF WER FOR FINE-TUNING VS LORA OURS USING WHISPER BASE

Dataset	ALLSSTAR			Spanish		
PLR	0	0.2	0.4	0	0.2	0.4
Whisper	**18.4**	37.8	70.0	**10.3**	24.7	57.2
fine-tune	24.9	27.1	31.8	89.5	91.1	94.5
LoRA $\alpha = 16$	29.8	39.2	53.8	89.2	91.6	95.0
LoRA $\alpha = 4$	18.6	33.3	58.6	11.4	22.6	47.8
Ours	18.7	**20.8**	**26.3**	10.7	**12.6**	**17.7**

and the large only after 20% PLR, whereas Ours, other than a slight degradation in the zero PL scenario, improves results for all PLRs in all languages. It is important to note that if there is no packet loss, using the original Whisper model is preferable, as it has been trained on a vast amount of clean and diverse speech data. However, we include the zero PL results because our model is designed as a universal enhancement front-end that may still provide benefits in real-world scenarios where other degradations, such as reverberation or additive noise, are present. In such cases, even without packet loss, our approach could still improve ASR performance by mitigating these additional distortions.

1) Fine-Tuning: Next, to highlight the advantages of our method over both full fine-tuning and parameter-efficient fine-tuning (PEFT), we conducted experiments where Whisper was fine-tuned using the same noisy LibriSpeech-based training data that was used to train our adaptation network. The front-end adaptation model was not included in this setup. This allows for a direct comparison between our approach and a conventional fine-tuning baseline. The results aligned with our initial hypotheses. While fine-tuning led to substantial improvements on the LibriSpeech test set, it came with significant tradeoffs. As shown in Table IV, the fine-tuned model suffered a severe drop in performance on ALLSSTAR, an English dataset from a different domain, in the clean (no packet loss) condition. Additionally, it lost its ability to generalize across languages, as seen in Table IV on Spanish speech. However, under packet loss conditions, the fine-tuned model showed improvements, surpassing the baseline model when the packet loss rate became sufficiently high.

For PEFT, we employed Low-Rank Adaptation (LoRA) [53] where $rank = 16$. Unlike full fine-tuning, LoRA retains the original model parameters, meaning we get the original model's output if we scale the LoRA adaptors to zero. This should limit the model's degradation in domains different than the training. However, our experiments revealed that LoRA does not provide a viable middle ground. With a small alpha (low-rank adaptation weight), improvements under packet loss were only marginal. As we increased alpha, the degradation in the clean (no packet loss) condition grew much faster than the improvements in packet loss scenarios. This resulted in no optimal setting where LoRA effectively balances baseline performance and robustness to packet loss.

In contrast, our method consistently outperformed both fine-tuning and LoRA across all packet loss conditions while maintaining performance on clean speech and preserving generalization to new domains and languages. As seen in Table IV, our approach achieved the best WER across different packet loss rates, vastly outperforming the fine-tuning methods.

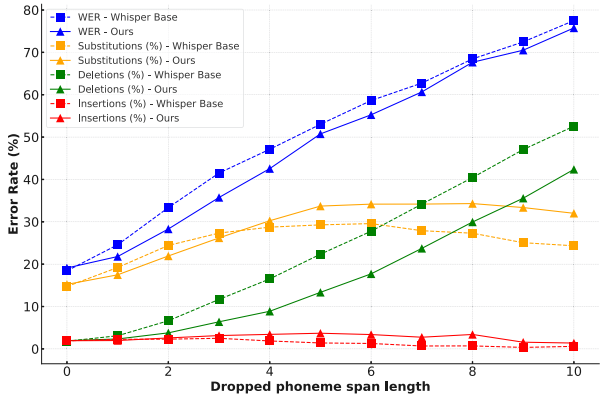


Fig. 5. WER and total substitutions, deletions, and insertions counts for dropped phoneme spans on the ALLSSTAR database using Whisper and our models.

2) Long Duration PL: To further investigate the in-painting capabilities of our model and test its performance in long-duration packet loss, we used the ALLSSTAR dataset. We processed it with the Montreal Forced Aligner [54]. Using the phoneme alignments generated by the forced aligner, we systematically dropped random sequences of consecutive phonemes from each utterance, varying the length from 1 to 10 phonemes. Each utterance was processed 10 times, once for each phoneme span length. We then evaluated the performance of the Whisper model both on the original mel spectrogram with dropped phonemes (i.e., missing phoneme spans are zeroed out) and on the adapted mel spectrogram produced by our front-end model, which attempts to reconstruct missing phoneme spans. For evaluation, we measured the overall WER and its components: deletions, insertions, and substitutions.

As shown in Fig. 5, our model consistently reduces the WER across all dropped span lengths compared to the baseline model. More interestingly, the breakdown of error types provides deeper insight into how our model handles different error patterns. Deletions: As expected, deletions increase sharply as the length of the dropped phoneme spans grows. However, with our model, the rate of increase is slower, indicating that it can incorporate some level of semantic understanding to infer and reconstruct at least part of the missing speech signal with the correct phonemes and words, thereby reducing outright deletions. Importantly, deletions contribute the most to WER: our model consistently reduces deletions by roughly an absolute 10% across different packet loss rates, demonstrating its effectiveness at recovering missing speech.

Insertions: The Whisper baseline shows a steady decline in insertions as span length increases. This is because, when phoneme spans are missing, Whisper produces fewer incorrect insertions, effectively leaving gaps in the transcription. In contrast, our model actively reconstructs missing content, sometimes inserting phonemes or words to compensate for the missing spans. This explains the slightly higher insertion rate: our model does not ignore gaps but attempts to predict what should have been there. However, insertions remain a very small component of the overall error. Even at its worst, the insertion rate of our model is only 3.5%, compared to Whisper's 2%, a negligible difference compared to the 10% absolute reduction in deletions.

TABLE V
EXAMPLES OF SUBSTITUTIONS IN RECONSTRUCTED SPEECH

Original Sentence	Modified Sentence
The mailman **brought a** letter	The mailman **put the** letter
The fire was **very** hot	The fire was **a bit** hot
The children **waved at** the train	The children **went to** the train
They car**ried some** shopping bags	They carr**y their** shopping bags
The dog is chasing the **cat**	The dog is chasing the **dog**

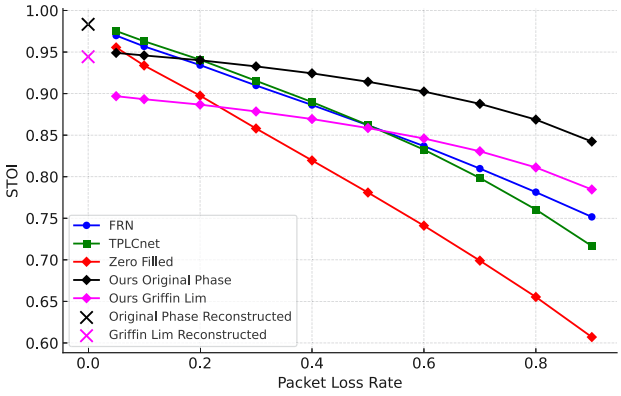


Fig. 6. STOI of different models on ALLSSTAR with various packet loss rates. FRN refers to Nguyen et al. [51], tPLCnet refers to Westhausen and Meyer [35].

Substitutions: Substitutions initially remain lower than the baseline for spans up to 4 phonemes, implying that our model effectively reconstructs short missing segments. However, the substitution rate increases beyond this point (approximately 2 words in length). This suggests that when the missing context is too large, for the same reason as the insertions, the model begins hallucinating incorrect phonemes or words, leading to more substitutions. Table V provides examples from the ALLSSTAR dataset, illustrating how the model occasionally substitutes words in reconstructed speech with words that are logical to insert based on the context. Despite this, substitutions are still secondary to deletions in contributing to WER. The increase in substitutions beyond 5-6 phoneme spans reflects the attempt of the model to recover from severe data loss, where the model can fill in the wrong words.

Overall, the analysis of error types aligns with expected behavior. As the model attempts to recover missing speech, deletions decrease since it successfully reconstructs lost phonemes. At the same time, the insertions and substitutions increase as a natural consequence of its objective. Unlike the Whisper baseline, which transcribes what remains in the input, our model actively reconstructs missing speech. This means it occasionally fills in gaps with words that seem contextually plausible but are incorrect, leading to a rise in insertions and substitutions.

3) Intelligibility Metrics: To further analyze the output of the model, we evaluate the adapted mel spectrograms using traditional intelligibility metrics. In Fig. 6, we compare the Short-Time Objective Intelligibility (STOI) [55] scores across different models: two open-source packet loss concealers (FRN [51] and tPLCnet [35]), the baseline Whisper model using zero-filled packet loss regions, and our adaptation network.

Unlike FRN and tPLCnet, which operate in the time domain, our model only produces a mel spectrogram, which is inherently a lossy representation. Because of this, we are at a disadvantage when evaluating STOI scores, as errors could stem either from (i) the lossy nature of the mel spectrogram reconstruction itself or (ii) the adaptation network modifying the spectrogram specifically to benefit ASR performance, potentially at the expense of intelligibility. However, since our goal is to improve Word Error Rate (WER) for ASR rather than human intelligibility, this tradeoff is not a concern.

To reconstruct a waveform from our adapted mel spectrograms, we estimate the power spectrum and apply two different phase reconstruction methods: (1) using the original clean audio's phase and (2) using the Griffin-Lim algorithm [56]. These reconstructions provide an approximate lower and upper bound for intelligibility metrics when compared to any trained vocoder. The degradation from this process can be seen in Fig. 6, noted as Original Phase Reconstructed and Griffin Lim Reconstructed, where we took the original clean signals and extracted mel spectrums and then reversed the process with either the original phase or Griffin Lim estimation and plotted their STOI scores compared with the original clean signals.

Fig. 6 shows that FRN and tPLCnet consistently outperform the zero-filled baseline across all packet loss rates. Our model surpasses the zero-filled audio in terms of STOI scores at packet loss rates above 10% for the original phase reconstruction and above 30% for the Griffin-Lim reconstruction. Notably, our model outperforms both FRN and tPLCnet in high packet loss scenarios, despite working within the constraints of a mel spectrogram representation. This suggests that while our approach is not optimized for intelligibility, it still offers benefits in packet loss conditions. However, our model achieves lower STOI scores in lower packet loss scenarios than FRN and tPLCnet. This aligns with the expectation that optimizing for ASR performance does not necessarily align with optimizing for perceptual intelligibility. Fig. 3 shows that our model achieves significantly better WER than the other PLC methods. Since our approach is explicitly geared toward improving ASR robustness rather than reconstructing high-quality waveforms, reducing STOI for lower packet loss rates is an expected and acceptable outcome.

In Fig. 7, we use Perceptual Evaluation of Speech Quality (PESQ) [25], a widely used objective metric that assesses speech quality by comparing degraded audio to a reference clean signal. An important observation is the degradation caused by the mel spectrogram reconstruction itself. The X markers in Fig. 7 indicate PESQ scores for the original clean waveforms and the two reconstructed versions. When the clean audio is converted to a mel spectrogram and then reconstructed using either the original phase or Griffin-Lim estimation, PESQ scores drop significantly, even though no actual packet loss corruption is present, and without applying our model. This demonstrates that the mel spectrogram inherently reduces perceptual quality, independent of our model. As shown in Fig. 7, tPLCnet achieves the highest PESQ scores across most packet loss rates, aligning with its strong STOI performance, while FRN performs similarly to the zero-filled baseline. Our model, however, yields consistently lower PESQ scores across all packet loss rates. This is expected,

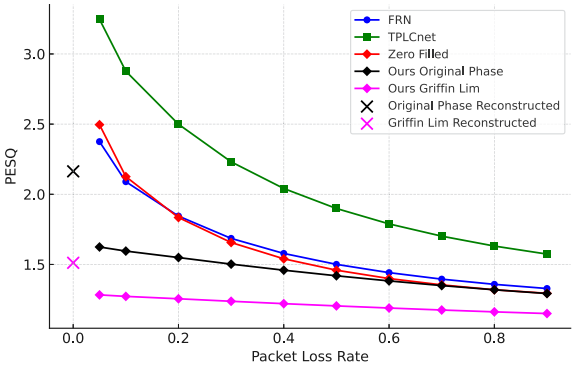


Fig. 7. PESQ of different models on ALLSSTAR with various packet loss rates. FRN refers to Nguyen et al. [51], tPLCnet refers to Westhausen and Meyer [35].

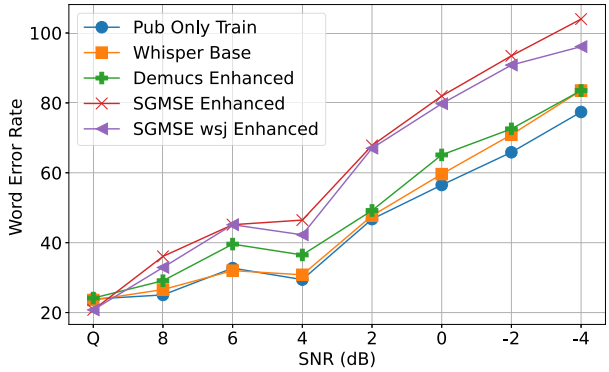


Fig. 8. WER of different models on ALLSSTAR with various pub noise SNRs. All the decoding is done with the Whisper base.

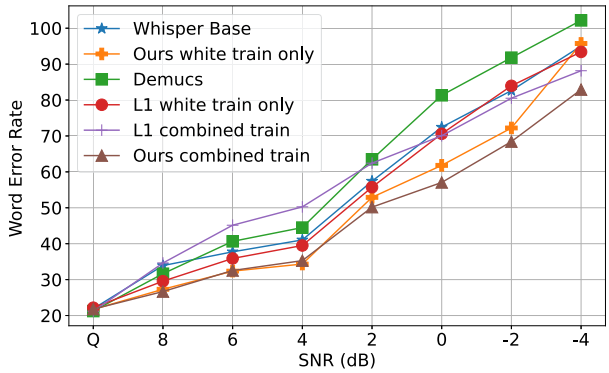


Fig. 9. WER of different models on ALLSSTAR with various white noise SNRs. All the decoding is done with the Whisper base.

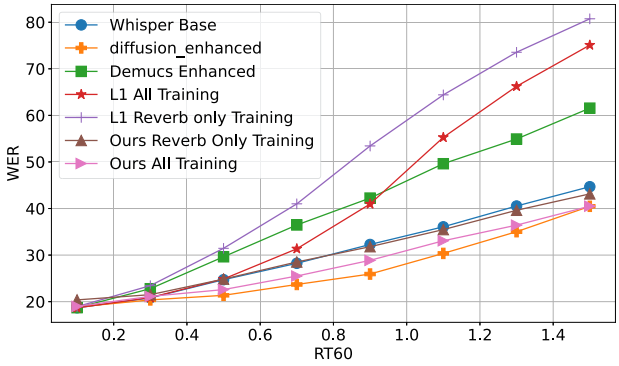


Fig. 10. WER of different models on ALLSSTAR with various RT60s. All the decoding is done with the Whisper base.

as the model is not designed to optimize perceptual speech quality but to enhance ASR robustness. Since PESQ penalizes unnatural modifications and artifacts, our approach – focused on improving ASR transcription rather than producing high-quality audio – naturally results in lower PESQ scores.

The conclusion drawn from these two metrics is that our model achieves higher STOI scores because it prioritizes intelligibility, which is beneficial for ASR performance. The model enhances phonetic and linguistic clarity, ensuring that speech remains recognizable even under packet loss conditions. Conversely, our model yields lower PESQ scores because it does not optimize for perceptual sound quality. Unlike traditional packet loss concealment models designed to produce natural-sounding speech, our approach focuses solely on improving ASR robustness. As a result, it does not attempt to preserve the fine details that contribute to perceptual quality, leading to lower PESQ scores. Ultimately, these results highlight the fundamental tradeoff in our approach: maximizing ASR performance at the expense of perceptual fidelity.

To complement these results, we provide audio samples of reconstructed speech at.[3] The page includes examples of our model, FRN, tPLCnet, and zero-filling under various packet loss conditions, reconstructed using both the original phase and Griffin-Lim methods. This allows for a qualitative assessment of the tradeoffs between intelligibility and ASR performance.

B. Additive Noise and Reverberation

To evaluate the performance of our adapter network in handling different types of noise, we first tested the model on isolated noise conditions: white noise, pub noise, and reverberation. These scenarios represent typical real-world environments, where a single type of distortion may dominate the audio signal.

As shown in Figs. 8, 9, and 10, we compare our model to the baseline Whisper model, to Demucs [43], a state-of-the-art real-time speech enhancer in the waveform domain and SGMSE [57] a diffusion model for dereverberation and noise supression. We also evaluated our model, which was trained only on L1 loss and explicitly trained on each noise type.

[3] https://shuadissen.github.io/ASR_denoiser/

Several key observations can be made in the white noise scenario, shown in Fig. 9. Our model consistently outperforms the baseline Whisper model, while applying Demucs speech enhancement degrades performance. Interestingly, training the model on various noise types, labeled *combined train*, rather than just white noise, labeled *white train only*, further improves performance in this setting, demonstrating the benefits of multi-noise training. Finally, training with ASR loss results in better performance than training with L1 loss alone, with the latter performing worse than the baseline Whisper model in higher SNRs.

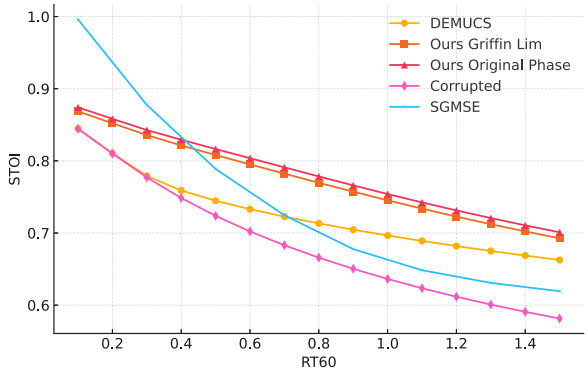


Fig. 11. STOI of different models on ALLSSTAR with various RT60s. DEMUCS refers to [43], SGMSE refers to [57].

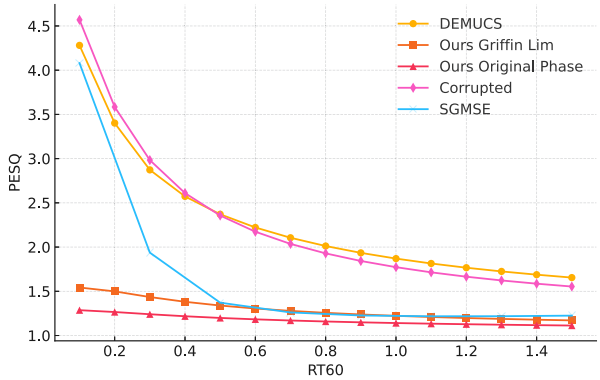


Fig. 12. PESQ of different models on ALLSSTAR with various RT60s.

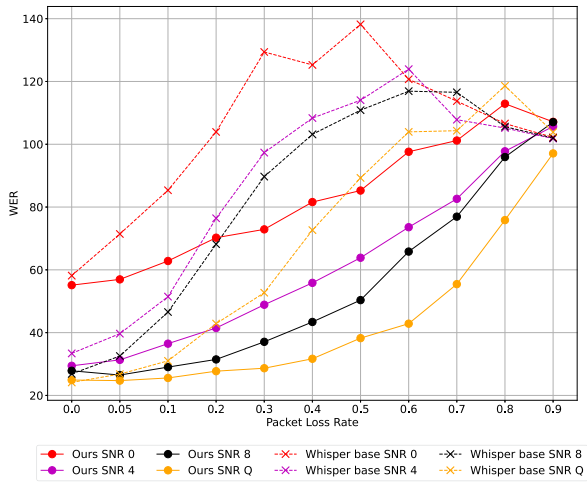


Fig. 13. WER of whisper base and our model on ALLSSTAR with various pub noise SNRs and PLRs. SNR Q refers to the quiet scenario with no added noise.

In the case of pub noise, shown in Fig. 8, characterized by overlapping speech and dynamic background sounds, our model indicates only minor improvement in lower SNRs and essentially no improvement in higher SNRs. This suggests that the Whisper ASR model is already quite robust to this type of noise. Additionally, the non-stationary nature of pub noise likely makes it difficult for our model to learn meaningful adaptations. We also compare with other enhancement models, Demucs, and two different SGMSE+ models, 1 trained on VoiceBank-DEMAND and the other trained on WSJ0-CHiME3. All 3 models degrade the WER results over the Whisper baseline.

Lastly, in the reverberant setting shown in Fig. 10, our model offers a slight advantage over the baseline. SGMSE+ diffusion model trained on WSJ0-REVERB improves the results the most. However, the sharp deterioration in performance when using L1 loss or Demucs enhancement is more striking, both of which fall below the baseline results. This indicates that reverberation poses a particular challenge for approaches that rely on waveform enhancement or L1-based training, inserting enhancement artifacts that are difficult for the ASR to overcome.

1) Intelligibility Metrics: To further understand the results, we analyze traditional ineligibility metrics. Fig. 11 presents STOI scores across different reverberation times (RT60) for various models. This allows us to assess how well each method preserves speech intelligibility in increasingly reverberant environments. As RT60 increases, STOI scores steadily decline across all models, as expected. All models outperform the corrupted speech. DEMUCS achieves higher STOI scores for higher RT60 values. SGMSE, a diffusion-based model for speech enhancement, performs best in low RT60s but falls below DEMUCS in higher RT60s. Our model (Ours Original Phase and Ours Griffin Lim) both outperform DEMUCS for all RT60 and SGMSE for higher RT60s, indicating our model performs well at preserving intelligibility.

Fig. 12 presents PESQ scores across the same RT60 conditions. Unlike STOI, PESQ evaluates overall perceptual quality rather than just intelligibility. The results show that DEMUCS performs similarly to the corrupted version, and SGMSE degrades sharply, matching ours at around 0.5 RT60. Our model exhibits significantly lower PESQ scores, similar to our observations in the packet loss experiments. This is expected, as our approach is not designed to produce high-quality, natural-sounding speech but to optimize ASR performance. These results indicate that low PESQ scores don't affect the WER performance, as both ours and SGMSE improve WER with very low PESQ.

These results reinforce the same tradeoff in packet loss conditions: our model prioritizes intelligibility, leading to high STOI scores, but does not optimize for perceptual quality, resulting in lower PESQ scores.

C. Layered Noise

In this section, we examined how the model performs when exposed to both packet loss and random white noise, pub noise, or reverberant data simultaneously. The layered noise scenarios, shown in Figs. 13, 14, and 15.

Fig. 14 depicts the White Noise and Packet Loss scenario: Our model consistently outperforms the baseline Whisper model for white noise combined with packet loss. Our model makes better use of the available signal at lower packet loss rates, with WERs remaining lower even as noise increases. This is evident from the smaller SNR gaps between conditions (e.g., SNR 2 and SNR 4) in our model compared to the baseline. The baseline model shows a greater dependence on cleaner signals to recover, resulting in a more considerable WER increase when noise is introduced. In contrast, our model experiences a smaller

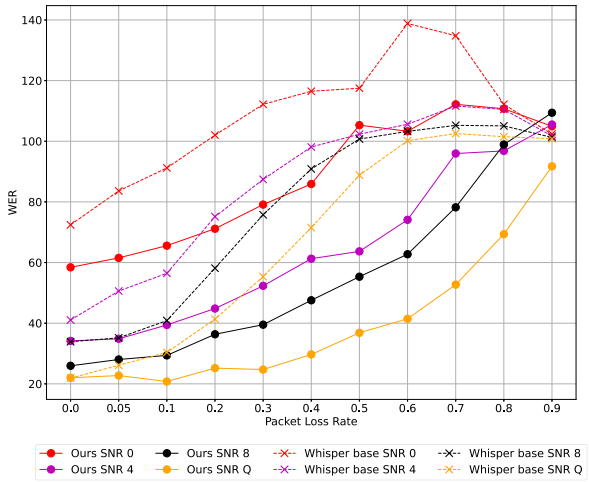


Fig. 14. WER of different models on ALLSSTAR with various white noise SNRs. All the decoding is done with Whisper base. SNR Q refers to the quiet scenario with no added noise.

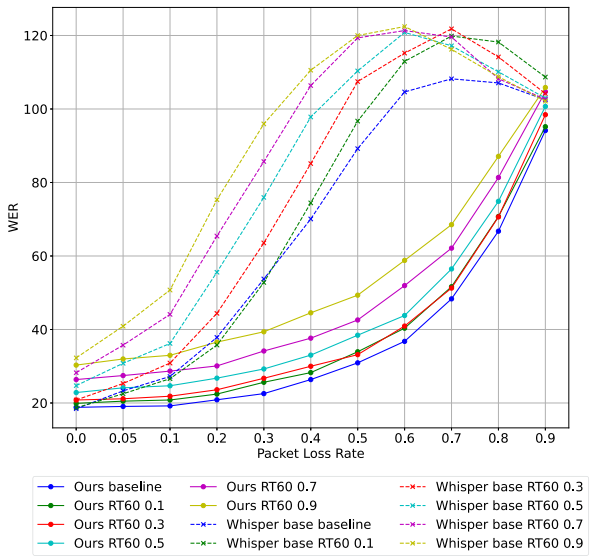


Fig. 15. WER of whisper base and our model on ALLSSTAR with various RT60 levels and PLRs.

rise in WER as noise intensifies, particularly in moderate SNR conditions. It highlights its ability to adapt to environments where moderate noise coexists with significant packet loss. However, as the packet loss rate exceeds 50%, WERs for both models begin to converge, and the differences between SNR levels become less meaningful, with both models approaching a WER of 100%, and the signal degradation becomes too severe for either model to recover effectively, regardless of the noise level.

Fig. 13 depicts the Pub Noise and Packet Loss scenario: In the zero packet loss scenario, both models exhibit slightly more resilience to pub noise than white noise. However, in the packet loss scenarios up to 30%, the baseline model experiences a sharp degradation even with a small amount of pub noise, compared to the no noise scenario. After this initial severe degradation, increasing the noise (i.e., lowering the SNR) does not degrade the baseline model much further, and its performance stabilizes after the initial drop.

In contrast, our model shows a much smaller degradation when pub noise is introduced compared to the no-noise scenario, and its performance gradually worsens as the SNR decreases, finally breaking at 0 SNR. This indicates that our model can recover from packet loss in moderate noise, a common real-world scenario.

Fig. 15 depicts the Reverberation and Packet Loss scenario: Reverberation combined with packet loss follows a similar trend, with our model outperforming the baseline consistently across RT60 values. However, the gaps between RT60 levels are smaller than those for white or pub noise. The gaps are also pretty consistent, adding a few absolute WER percentages for each RT60 across packet loss rates with no "breaking point." At lower packet loss rates (below 20%), the gap between low RT60s (under 0.5) is very small for the models.

We see general trends across all noises when layered with packet loss. Effectiveness in Moderate Conditions: Across all noise types, the most significant improvement from our model over the baseline occurs in moderate noise and packet loss conditions. For example, when packet loss is below 20%, and noise is moderate (e.g., SNR 4-8), our model maintains a much lower WER than the baseline, which breaks at this point. This highlights the practical utility of our model in real-world scenarios, where packet loss and noise often coexist but are not extreme.

Limited Gains in Severe Conditions: While the packet loss concealment alone is adequate even in very high loss rates, in the layered scenarios, as the packet loss rate rises beyond 40%, the absolute WER gaps between models shrink, and the performance of our model converges toward really high WERs and even with keeping a 50% difference in WER between our model and the baseline the improvements are less useful at such a high WER.

Noise-Specific Behavior: The nature of the noise also plays a crucial role. White noise allows for more substantial improvements in WER, likely due to its stationary nature, whereas pub noise presents more challenges due to its unpredictability and overlapping speech. Reverberation affects both models less, indicating a more baseline robustness to this type of noise.

Finally, to better reflect realistic deployment scenarios, we conducted an additional evaluation combining all three distortion types simultaneously. Specifically, we applied random RIRs between 0.1 and 1.5 RT60s, overlaid additive noise at random SNRs between 8 and -4, and introduced 20% packet loss. In this highly challenging condition, our model achieved a WER of 42% on the ALLSSTAR dataset, significantly outperforming the Whisper baseline, which yielded 66% . This result further supports the robustness and practical utility of our proposed method in real-world environments.

VI. Conclusion

This study introduced a novel approach for enhancing the robustness of large ASR models in both packet loss and noisy scenarios. The proposed method integrates a smaller adaptation model specifically designed to modify the input features of the

ASR system. The model is trained with the ASR model loss function while keeping the ASR model frozen. While our implementation utilizes Whisper and its cross-entropy loss, the approach is not inherently tied to Whisper or cross-entropy. The adapter network could be trained using gradients from any differentiable ASR loss, including CTC, transducer losses, or other token-level objectives. The only requirement is that the ASR model provides a gradient signal with respect to its input features, which is a property shared by most end-to-end ASR systems.

Our results demonstrate that this integration improves ASR robustness against packet loss and noise, particularly when packet loss is combined with noise. More moderate improvements are observed in other noise conditions when the adaptation network is trained using the gradients of a larger ASR model. We analyzed the model outputs and found that the model maintained or improved STOI scores while having very low PESQ. This indicates that the model preserves ineligibility that correlates with WER and does not prioritize sound quality perception.

The promising outcomes of this approach open up several avenues for future research in ASR development. Future studies could explore the applicability of this method in improving the robustness of ASR models against additional noise types, such as clipping and echo suppression. Another interesting direction would be to explore the role of the pseudo-language model in reconstruction by systematically dropping entire words and analyzing the perplexity of the reconstructed sentences versus the original sentences.

REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.

[2] K. Wei, P. Guo, and N. Jiang, "Improving transformer-based conversational ASR by inter-sentential attention mechanism," in *Proc. INTERSPEECH 2022*, 2022, pp. 3804–3808.

[3] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH 2020*, 2020, pp. 5036–5040.

[4] S. -E. Kim, B. R. Chernyak, O. Seleznova, J. Keshet, M. Goldrick, and A. R. Bradlow, "Automatic recognition of second language speech-in-noise," *JASA Exp. Lett.*, vol. 4, no. 2, 2024, Art. no. 025204.

[5] G. Sun, X. Zheng, C. Zhang, and P. C. Woodland, "Can contextual biasing remain effective with Whisper and GPT-2?," in *Proc. INTERSPEECH 2023*, 2023, pp. 1289–1293.

[6] H. Phan et al., "Improving GANs for speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1700–1704, 2020.

[7] M. M. Mohamed, M. A. Nessiem, and B. W. Schuller, "On deep speech packet loss concealment: A mini-survey," 2020, *arXiv:2005.07794*.

[8] M. Fujimoto and H. Kawai, "One-pass single-channel noisy speech recognition using a combination of noisy and enhanced features," in *Proc. INTERSPEECH*, 2019, pp. 486–490.

[9] S. J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, 2018, vol. 2018, pp. 1571–1575.

[10] K. Iwamoto et al., "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR," in *Proc. Interspeech*, 2022, pp. 5418–5422, doi: 10.21437/Interspeech.2022-318.

[11] T. Menne, R. Schlüter, and H. Ney, "Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR," in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6660–6664.

[12] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *Proc. ICASSP 2020-2020 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7009–7013.

[13] F. Weninger et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Latent Variable Anal. Signal Separation, 12th Int. Conf., LVA/ICA 2015, Liberec, Czech Republic, Aug. 25-28, 2015, Proc. 12*, Springer, 2015, pp. 91–99.

[14] Z. -Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[15] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 826–835, Apr. 2014.

[16] T. Gao, J. Du, L. -R. Dai, and C. -H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4375–4379.

[17] Z. -Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 796–806, Apr. 2016.

[18] L. Li, Y. Kang, Y. Shi, L. Kürzinger, T. Watzel, and G. Rigoll, "Adversarial joint training with self-attention mechanism for robust end-to-end speech recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, pp. 1–16, 2021.

[19] D. -H. Yang and J. -H. Chang, "Attention-based latent features for jointly trained end-to-end automatic speech recognition with modified speech enhancement," *J. King Saud Univ.- Comput. Inf. Sci.*, vol. 35, no. 3, pp. 202–210, 2023.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.–MICCAI 2015, 18th Int. Conf., Munich, Germany, Oct. 5–9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[21] Y. Dissen, S. Yonash, I. Cohen, and J. Keshet, "Enhanced ASR robustness to packet loss with a front-end adaptation network," in *Proc. INTERSPEECH 2024*, 2024, pp. 5008–5012.

[22] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5220–5224.

[23] J. Balam, J. Huang, V. Lavrukhin, S. Deng, S. Majumdar, and B. Ginsburg, "Improving noise robustness of an end-to-end neural model for automatic speech recognition," 2020, *arXiv:2010.12715*.

[24] A. Narayanan et al., "Toward domain-invariant speech recognition via large scale training," in *Proc. 2018 IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 441–447.

[25] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. 2001 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.

[26] A. S. Subramanian et al., "Speech enhancement using end-to-end speech recognition objectives," in *Proc. 2019 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 234–238.

[27] K. Iwamoto et al., "How does end-to-end speech recognition training impact speech enhancement artifacts?," in *Proc. ICASSP 2024-2024 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 11031–11035.

[28] X. Chang, T. Maekaku, Y. Fujita, and S. Watanabe, "End-to-End integration of speech recognition, speech enhancement, and self-supervised learning representation," in *Proc. Interspeech*, 2022, pp. 3819–3823, doi: 10.21437/Interspeech.2022-10839.

[29] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.

[30] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

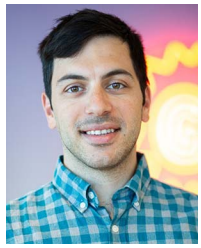
[31] E. Gunduzhan and K. Momtahan, "Linear prediction based packet loss concealment algorithm for PCM coded speech," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 778–785, Nov. 2001.

[32] J. -H. Chen, "Packet loss concealment based on extrapolation of speech waveform," in *Proc. 2009 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 4129–4132.

[33] J. Wang, Y. Guan, C. Zheng, R. Peng, and X. Li, "A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission," *J. Acoust. Soc. America*, vol. 150, no. 4, pp. 2577–2588, 2021.

[34] S. Pascual, J. Serrà, and J. Pons, "Adversarial auto-encoding for packet loss concealment," in *Proc. 2021 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 71–75.

[35] N. L. Westhausen and B. T. Meyer, "tPLCnet: Real-time deep packet loss concealment in the time domain using a short temporal context," in *Proc. INTERSPEECH 2022*, 2022, pp. 2903–2907.
[36] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A time-domain convolutional recurrent network for packet loss concealment," in *Proc. ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7148–7152.
[37] D. -H. Yang and J. -H. Chang, "Diff-PLC: A diffusion-based approach for effective packet loss concealment," in *Proc. 2024 IEEE Spoken Lang. Technol. Workshop*, 2024, pp. 357–363.
[38] D. -H. Yang and J. -H. Chang, "Flow-PLC: Towards efficient packet loss concealment with flow matching," *IEEE Signal Process. Lett.*, vol. 32, pp. 1430–1434, 2025.
[39] D. -H. Yang and J. -H. Chang, "Towards robust packet loss concealment system with ASR-guided representations," in *Proc. 2023 IEEE Autom. Speech Recognit. Understanding Workshop*, 2023, pp. 1–8.
[40] J. Zhang, Z. Zhao, Y. Liu, J. Liu, Z. He, and K. Niu, "TD-PLC: A semantic-aware speech encoding for improved packet loss concealment," in *Proc. INTERSPEECH 2024*, 2024, pp. 1745–1749.
[41] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12449–12460.
[42] 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 Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
[43] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. INTERSPEECH 2020*, 2020, pp. 3291–3295.
[44] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.
[45] A. Bradlow, L. Ackerman, L. Burchfield, L. Hesterberg, J. Luque, and K. Mok, "ALLSSTAR: Archive of l1 and l2 scripted and spontaneous transcripts and recordings," in *Proc. Int. Congr. Phonetic Sci.*, 2010, pp. 356–359.
[46] A. Conneau et al., "FLEURS: Few-shot learning evaluation of universal representations of speech," in *Proc. 2022 IEEE Spoken Lang. Technol. Workshop*, 2023, pp. 798–805.
[47] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "INTERSPEECH 2022 audio deep packet loss concealment challenge," in *Proc. INTERSPEECH 2022*, 2022, pp. 580–584.
[48] L. Diener, S. Branets, A. Saabas, and R. Cutler, "The ICASSP 2024 audio deep packet loss concealment grand challenge," *IEEE Open J. Signal Process.*, vol. 6, pp. 231–237, 2025.
[49] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proc. 14th ISMIR Conf.*, 2013.
[50] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. 14th Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 313–317.
[51] V. -A. Nguyen, A. H. T. Nguyen, and A. W. H. Khong, "Improving performance of real-time full-band blind packet-loss concealment with predictive network," in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[52] A. A. Nair and K. Koishida, "Cascaded time + time-frequency Unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps," in *Proc. ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7153–7157.
[53] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9
[54] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. INTERSPEECH*, 2017, vol. 2017, pp. 498–502.
[55] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[56] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.
[57] J. Richter, S. Welker, J. -M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2351–2364, 2023.



Yehoshua Dissen received the M.Sc. degree in computer science from Bar-Ilan University, Ramat Gan, Israel, in 2016. He is currently working toward the Ph.D. degree with the Faculty of Electrical and Computer Engineering, Technion—Israel Institute of Technology, Haifa, Israel. He was a Speech Researcher in several industry positions. His research interests include speech enhancement, automatic speech recognition, and speaker diarization.

Shiry Yonash received the B.Sc. degree from the Technion—Israel Institute of Technology, Haifa, Israel, and the M.Sc. degree in computer science from the University of Haifa, Haifa. She is currently a Speech Researcher. Her research interests include automatic speech recognition, speaker verification, speaker diarization, and voice-controlled systems.

Israel Cohen (Fellow, IEEE) received the B.Sc. (*Summa Cum Laude*), M.Sc., and the Ph.D. degrees in electrical engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 1990, 1993, and 1998, respectively. From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with Computer Science Department, Yale University, New Haven, CT, USA. In 2001, he joined the Electrical Engineering Department, Technion– Israel Institute of Technology. He is currently Louis and Samuel Seidan Professor of electrical and computer engineering with the Technion – Israel Institute of Technology. He is the coeditor of the Multichannel Speech Processing Section of the *Springer Handbook of Speech Processing* (Springer, 2008), and the coauthor of *Fundamentals of Signal Enhancement and Array Signal Processing* (Wiley-IEEE Press, 2018). His research interests include array processing, statistical signal processing, deep learning, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, and system identification and adaptive filtering. Dr. Cohen was awarded an Honorary Doctorate from Karunya Institute of Technology and Sciences, Coimbatore, India in 2023, Norman Seiden Prize for Academic Excellence in 2017, SPS Signal Processing Letters Best Paper Award in 2014, Alexander Goldberg Prize for Excellence in Research in 2010, and the Muriel and David Jacknow Award for Excellence in Teaching in 2009. He was an Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, a member of IEEE Audio and Acoustic Signal Processing Technical Committee and IEEE Speech and Language Processing Technical Committee, and a Distinguished Lecturer of the IEEE Signal Processing Society.

Joseph Keshet (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Tel Aviv University, Tel Aviv, Israel, in 1994 and 2002, respectively, and the Ph.D. degree in computer science from the School of Computer Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel, in 2008. From 2008 to 2009, he was a Postdoctoral Researcher with EPFL- Swiss Federal Technology Institute of Lausanne, Lausanne, Switzerland and IDIAP Research Institute, Martigny, Switzerland. He was a Research Assistant Professor with TTIC from 2009 to 2012. From 2013 and 2022, he was an Associate Professor with the Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel. Since 2022, he has been an Associate Professor with the Faculty of Electrical and Computer Engineering, Technion – Israel Institute of Technology, Haifa, Israel. His research interests include speech recognition, speech synthesis, and speech processing.