Exponential Smoothing of Noisy Embeddings for Phoneme-Aware Speech Enhancement

Anonymous ACL submission

Abstract

Speech enhancement (SE) improves the robustness of downstream speech technologies under noisy conditions. Self-supervised models such as Wav2Vec 2.0 produce robust frame-level representations that capture phonetic information, but directly conditioning on noisy embeddings can propagate errors. In this work, we propose a temporal abstraction strategy that applies exponential smoothing to Wav2Vec 2.0 embeddings before conditioning a diffusion-based SE network via FiLM modulation. This approach reduces sensitivity to noise. We evaluate the method on a noisy dataset and demonstrate improvements in PESQ and STOI across multiple SNRs and model configurations. Ablations show that exponential smoothing outperforms both naive averaging and unconditioned diffusion models by a PESQ score of 0.35.

1 Introduction

Robust speech enhancement (SE) is a cornerstone front-end for many downstream speech technologies: improving signal quality before recognition or downstream processing often yields large gains in automatic speech recognition (ASR), keyword spotting, and other speech understanding tasks, especially in adverse acoustic conditions. Prior work (Nayem and Williamson, 2021) has shown that incorporating textual or linguistic information into SE can substantially improve intelligibility and downstream recognition performance by providing high-level constraints that guide denoising and spectral reconstruction.

Recent advances in large-scale weakly-supervised ASR, most notably OpenAI's Whisper family (Radford et al., 2023) demonstrate that models trained on massive, diverse, and partly noisy web-scale audio–transcript pairs learn representations and transcription robustness that generalize well across noise types, languages, and recording conditions. This robustness arises from scale and exposure to naturally noisy real-world data, which makes Whisper-style transcriptions an attractive source of supervision when clean transcripts are unavailable.

Motivated by these facts, several lines (Yang et al., 2023) of work have explored conditioning SE models on linguistic cues extracted from the speech signal, by fusing text with noisy waveform through an attention mechanism. Early text-informed SE methods (Wang et al., 2022) showed that supplying word-level transcripts can guide enhancement toward linguistically plausible spectra and improve intelligibility when the text is accurate. More recent work (Lu et al., 2023) shifts from lexical information to intermediate phonetic representations, such as phoneme posteriors and broad phonetic classes, which are easier to synchronize with acoustic features and remain reliable under higher noise levels. These studies collectively demonstrate both the promise and the limitations of linguistic conditioning: linguistic structure significantly helps enhancement, but real-world transcripts are often noisy or imperfectly aligned, which can propagate errors if used directly.

In this paper we leverage the robustness of a large-scale feature extractor like wav2vec 2.0 (Baevski et al., 2020) but intentionally avoid relying directly on noisy word-level transcripts at inference time. Instead, we extract noisy frame level phonetic information and use these noisy phonemes as conditioning for a neural speech enhancement model. Conditioning on phonemes (rather than raw text) reduces sensitivity to word-level errors, exploits phoneme-specific spectral priors as documented in prior work (Lu et al., 2020), and keeps the conditioning compact and aligned with typical frame-level SE architectures. They suggest that phonetic conditioning is effective even when phonetic inputs are imperfect, motivating our design choice.

We systematically investigate the use of large-

scale self-supervised speech representations for phoneme-aware speech enhancement. In particular, we propose a temporal abstraction strategy that applies exponential smoothing to Wav2Vec 2.0 derived embeddings before conditioning a diffusion-based SE network via FiLM (Perez et al., 2018). Unlike naive averaging, this smoothing preserves phonetic information across time, reducing sensitivity to noisy embeddings while minimizing additional computation. Through extensive experiments, we show that this simple yet effective design consistently improves perceptual quality (PESQ) and intelligibility (STOI) across diverse SNR conditions, highlighting the importance of temporal aggregation in embedding-based SE. This study frames temporal smoothing as a compute-efficient, phonetic-aware enhancement strategy, providing insights that generalize to other embedding-conditioned models.

2 Methodology

2.1 The Wav2Vec 2.0 Feature Extractor

Wav2Vec 2.0 (Baevski et al., 2020) is a self-supervised speech representation model that operates directly on raw audio sampled at 16 kHz and contains a convolutional feature encoder followed by a 12-layer (base, 95 M parameters) or 24-layer (large, 317 M parameters) transformer context network. The feature encoder reduces the temporal resolution by a factor of 320, producing latent representations at approximately 50 Hz. During pretraining, around 50% of the latent time steps are randomly masked, and the model is optimized with a contrastive loss that requires identifying the correct quantized target from a set of typically 100 negative samples, along with a codebook diversity loss. When fine-tuned with a CTC objective, Wav2Vec 2.0 achieves word error rates below 5% on LibriSpeech test-clean using only 10 minutes of labeled data and below 2% with full supervision, while its intermediate transformer layers yield strong phoneme-level discrimination for tasks such as forced alignment and pronunciation quality assessment.

2.2 Feature-wise Linear Modulation (FiLM).

Feature-wise Linear Modulation (FiLM) (Perez et al., 2018) conditions intermediate neural network representations on an external signal. Given an intermediate activation tensor $\mathbf{h} \in \mathbb{R}^{C \times T}$ and a conditioning variable $\mathbf{z}$, FiLM applies a channel-wise affine transformation

$$\mathrm{FiLM}(\mathbf{h}_c) = \gamma_c(\mathbf{z}) \odot \mathbf{h}_c + \beta_c(\mathbf{z}), \qquad (1)$$

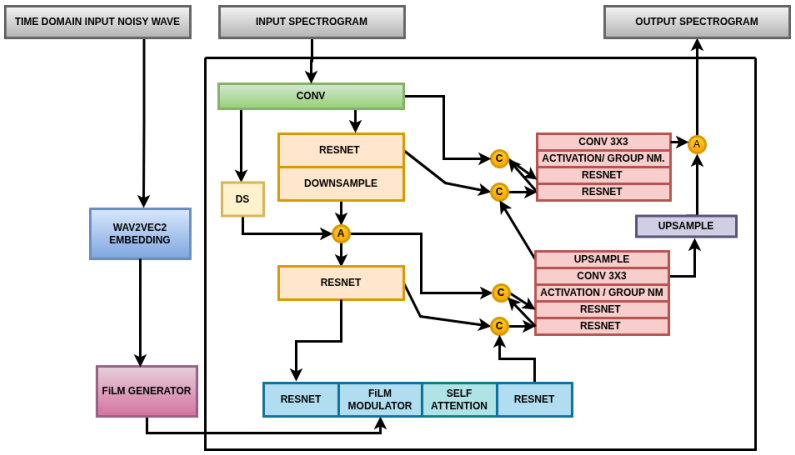
where $c \in \{1, \ldots, C\}$ indexes feature channels, $\odot$ denotes element-wise multiplication, and $\gamma_c(\cdot)$ and $\beta_c(\cdot)$ are learnable functions that generate scaling and bias parameters from the conditioning signal. This multiplicative-additive modulation allows the conditioning signal to selectively amplify, suppress, or shift feature activations at different layers, providing a parameter-efficient mechanism for injecting side information without altering the base network architecture.

2.3 The Proposed Model

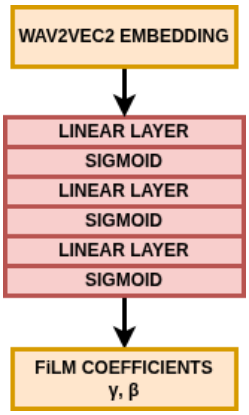
The backbone of the speech enhancement model that we propose in this work is inspired by the U Net style predictor-generator architecture employed by StoRM (Lemercier et al., 2023). Similar architectures have also been used by image segmentation models to separate the foreground from background. Our design is shown in Figure (a). The model can be broadly divided into three stages, the encoder, the bottleneck and the decoder. On the encoder side, the model cascades a Resnet layer followed by a downsampling stage to form an encoder block. At the end of this block, the block output is added to a downsampled version of the original noisy input spectrogram. The Resnet block consists of a linear layer, group norm/ activation, a dropout of 0.1 and a final 1 x 1 convolution layer to adjust the number of channels.

The bottleneck layer consists of two Resnet layers which are structured as described above, enclosing a standard attention layer on both sides. The bottleneck layer leaves the size of the input tensor unchanged, not contributing to the receptive field size but only using it's constituent attention layer to derive context from nearby acoustic features. The attention layer employed by this model is a single head attention which implements the projection operation as a linear layer.

On the decoder side, each block consists of five layers two Resnets, the swish activation function followed by Group Norm, a 3 x 3 convolution layer to reduce the number of channels in the final output to two and finally an upsampler to compensate for the downsampling on the encoder side. As with the encoder end, the final block does not have an upsampling stage. In the version of the model used in this work, we have four blocks on both the



(a) Shared speech enhancement backbone used in this work by both the predictive and generative stages.



(b) FiLM generator module

encoder and decoder ends. After the fourth upsampling block, the output of the Conv 3 x 3 layer in each block is added together, after appropriate upsampling to produce the final estimate of the clean waveform, to be used in the MSE loss function. While training, the model truncates a random portion of the training clean and noisy waveforms and takes a STFT of this portion to create spectrograms with a pre-specified number of time frames and frequency bins.

2.4 Augmentation with FiLM modulated features

In addition to the description in Section 2.3 our model also consists of an augmentation with a pretrained wav2vec 2.0 checkpoint and a FiLM generator module (Fig 1b). We proceed in two stages. In the first stage, we use the wav2vec 2.0 on a truncated chunk of noisy waveform to derive phonetic information of the relevant parts. In the second stage, the wav2vec 2.0 output feature is input to a FiLM generator module (Fig 1a) to derive the channel wise coefficients γ_c^t and β_c^t for each 20 ms time slice t, as described in Section 2.2. Since we need only one $\tilde{\gamma}_c$ and $\tilde{\beta}_c$ per channel we perform exponential averaging along time as follows:

$$\tilde{\gamma}_c^t = \alpha\,\gamma_c^t + (1-\alpha)\,\tilde{\gamma}_c^{t-1}, \tag{2}$$

$$\tilde{\beta}_c^t = \alpha\,\beta_c^t + (1-\alpha)\,\tilde{\beta}_c^{t-1}, \qquad 0 < \alpha \leq 1 \tag{3}$$

with initialization $\tilde{\gamma}_c^0 = \gamma_c^0$ and $\tilde{\beta}_c^0 = \beta_c^0$. The final values of $\tilde{\gamma}_c$ and $\tilde{\beta}_c$ are given by $\tilde{\gamma}_c^T$ and $\tilde{\gamma}_c^T$, respectively, with T being the total number of time slices in the wav2vec 2.0 output. Subsequently, $\tilde{\gamma}_c$ and $\tilde{\beta}_c$ are used to modulate the input h to the attention layer in the U-Net bottleneck as $\tilde{\gamma}_c \odot h + \tilde{\beta}_c$ (Fig 1a).

The key idea behind modulating with wav2vec 2.0 features along the channel dimension is to treat phonetic information as an additional view of the acoustic signal, parallel to the spectrogram. In a standard spectrogram-based U-Net, the network receives a 4-dimensional tensor (B, C, F, T) consisting of a batch dimension, a channel dimension, and two spatial dimensions corresponding to frequency bins and time frames. By modulating the bottleneck feature channels with phonetic information, we give the model an augmented input representation where each time frame is annotated not only with its acoustic energy distribution (from STFT) but also with a phonetic likelihood distribution.

An important property of this approach is that it remains strictly local in time, just like standard masking-based enhancement. The modulation scheme does not change the architecture in a disruptive way. They simply provide a richer representation of each time frame, grounded in linguistic structure, which allows the network to operate with a higher-level prior about which frequency patterns are speech-like versus noise-like. This is especially useful in challenging noise conditions or speech distortions, where the raw spectrogram alone does not reliably indicate what should be preserved.

The method also leverages the fact that phonetic states change more slowly than raw spectral coefficients. Wav2vec 2.0 representation carry information about phoneme identity, which persists across tens of milliseconds, while short-term spectral energy fluctuates rapidly.

A further practical benefit is that this fusion does not require architectural changes to the U-Net's en-

SNR	Ours (128 ch.)		StoRM (128 ch.)		Ours (32 ch.)		StoRM (32 ch.)	
	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ
-3 dB	**0.8757**	**1.7874**	0.7258	1.2611	0.7821	1.3743	0.8090	1.3210
0 dB	**0.9108**	**1.9972**	0.7870	1.3178	0.8390	1.5557	0.8626	1.4696
3 dB	**0.9360**	**2.2059**	0.8464	1.4218	0.8820	1.7714	0.9017	1.6457
6 dB	**0.9553**	**2.3958**	0.8885	1.4984	0.9146	1.9854	0.9306	1.8348
12 dB	**0.9778**	**2.7690**	0.9506	1.8613	0.9571	2.4080	0.9660	2.2392

Table 1: PESQ and STOI scores for different configurations of the proposed and baseline models.

coder–decoder or skip connections. As long as the phoneme information is temporally aligned with the STFT frames (and broadcast across frequency), they plug directly into the model. This makes the approach lightweight, architecture-neutral, and easy to scale. Different linguistic or articulatory side features (e.g., phoneme embeddings, bottleneck features, articulatory features) can be swapped in without redesigning the backbone.

3 Dataset

To train the speech enhancement models, 19,992 audio waveforms were taken from the train split of Librispeech *train-clean-100* (Panayotov et al., 2015). These files were divided equally among 27 noises, which ranged from everyday noises such as baby cries to ambient noises from military settings such as helicopters. For each noise type, an equal number of waveforms were randomly chosen to be mixed with it at 5 different SNRs: -3dB, 0dB, 3dB, 6dB and 12dB. These SNRs were selected to model the wide range of scenarios that a general-purpose speech enhancement system might encounter. Owing to the relatively large inference time for diffusion models such as StoRM, we kept our validation set to a minimum. The validation set consisted of 10 waveforms from the Librispeech validation set. The set was divided into 5 groups of 2 each for one SNR. The noises used for each set were from the office and crowd domains, as background chatter is considered to be one of the most challenging in the domain of speech enhancement due to the irregular structure of these noises. For the test set, we used 8 noises from the training set and 2 unseen noises. Each noise was combined with 20 utterances from the Librispeech test set at the same five SNRs. Thus, the total number of utterances in the test set was 1000, totaling around 4 hours in duration. Model checkpoints, noisy and enhanced test set can be found here.

4 Experiments

In this section, we evaluate the effectiveness of phoneme-based conditioning by comparing the proposed model against a baseline enhancement system that does not incorporate phonetic cues. In particular, we analyze both (i) the perceptual quality of the reconstructed speech (PESQ scores) and (ii) the intelligibility by measuring how well the enhanced signal preserves short-time temporal envelope (STOI). For the smoothing scheme in Section 2.4, we experimented with various values of the hyperparameter α and found that $\alpha = 0.1$ performs best. We experiment with different widths of the proposed model, defining width as the number of output channels of the green convolution layer in Figure 1a. The PESQ and STOI scores obtained for different channel counts and SNR are summarized in Table 1.

Further, we experiment with a different averaging scheme (naive averaging) and present the enhancement results in Table 2 in Appendix A. We also train and test on the VoiceBank+DEMAND dataset (Valentini-Botinhao et al., 2016). The details and results of this experiment can be found in Appendix B.

5 Conclusions

In this work, we proposed a diffusion-based speech enhancement framework augmented with phonetic representations derived from wav2vec 2.0. We investigated the impact of different channel configurations as well as multiple strategies for fusing phonetic information within the enhancement pipeline. Experiments conducted on both a standard and a non-standard dataset demonstrate the effectiveness of the proposed approach across diverse acoustic conditions. We evaluate performance using objective metrics, including PESQ for perceptual quality and STOI for intelligibility, and observe consistent improvements over baseline configurations.

Limitations

While our proposed speech enhancement system demonstrates improved robustness and perceptual quality, it has several limitations. First, the use of a diffusion-based architecture results in high computational cost and longer inference times compared to conventional feedforward models, which may limit real-time deployment. Second, conditioning on Wav2Vec 2.0 features increases the overall parameter count, even though these features are frozen during training. Reducing the number of channels can mitigate the parameter and memory footprint; however, this also leads to a measurable degradation in the quality of the enhanced waveform. Third, our experiments are conducted on a non-standard dataset, which may limit comparability with other methods. We plan to extend this study to publicly available speech enhancement datasets such as DNS Challenge, and LibriSpeech-derived noisy corpora to better evaluate generalization and reproducibility. Finally, as with most supervised speech enhancement methods, performance may vary across unseen noise types, speakers, and recording conditions.

Ethics Statement

This work studies speech enhancement methods aimed at improving speech quality and intelligibility in noisy conditions. The proposed approach does not involve the collection of new human subject data. Speech enhancement systems may present ethical risks if performance varies across speakers, accents, languages, or recording conditions. To mitigate potential bias, we train our model on diverse speakers and noise conditions where possible, and we report aggregate performance rather than claims about individual speakers. Nonetheless, residual biases may remain, particularly for underrepresented accents or speaking styles, and should be considered when deploying such systems in real-world applications. The proposed method is not intended for surveillance, speaker identification, or forensic use. While enhanced speech could potentially be misused to improve the intelligibility of recordings without a speaker's consent, this risk is inherent to speech enhancement technologies broadly and not specific to our approach. We encourage responsible use consistent with privacy laws and ethical guidelines.

References

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

Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. 2023. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2724–2737.

Y.-J. Lu, C.-F. Liao, X. Lu, J.-w. Hung, and Y. Tsao. 2020. Incorporating broad phonetic information for speech enhancement. In *Proc. Interspeech*, pages 2417–2421.

Yen-Ju Lu, Chia-Yu Chang, Cheng Yu, Ching-Feng Liu, Jeih-weih Hung, Shinji Watanabe, and Yu Tsao. 2023. Improving speech enhancement performance by leveraging contextual broad phonetic class information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2738–2750.

Khandokar Md. Nayem and Donald S. Williamson. 2021. Towards an asr approach using acoustic and language models for speech enhancement. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7123–7127.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi. 2016. Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In *Proc. Interspeech*.

Wei Wang, Wangyou Zhang, Shaoxiong Lin, and Yanmin Qian. 2022. Text-informed knowledge distillation for robust speech enhancement and recognition. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 334–338.

SNR	Ours (128 ch.)		StoRM (128 ch.)		Ours (32 ch.)		StoRM (32 ch.)	
	Naive	Exp.	Naive	Exp.	Naive	Exp.	Naive	Exp.
-3 dB	1.4351	**1.7874**	1.2611	-	1.2062	1.3743	1.3210	-
0 dB	1.6087	**1.9972**	1.3178	-	1.3229	1.5557	1.4696	-
3 dB	1.8268	**2.2059**	1.4218	-	1.4938	1.7714	1.6457	-
6 dB	2.0365	**2.3958**	1.4984	-	1.6824	1.9854	1.8348	-
12 dB	2.4065	**2.7690**	1.8613	-	2.0755	2.4080	2.2392	-

Table 2: PESQ scores for exponential smoothing and naive averaging for different configurations of the proposed and baseline models. Since baseline StoRM has no averaging we report it's score under the "Naive" column.

	Ours (128 ch.)	StoRM (128 ch.)	Ours (32 ch.)	StoRM (32 ch.)
PESQ	2.06	**2.41**	2.29	**2.49**
STOI	**0.84**	0.84	**0.82**	0.82

Table 3: PESQ and STOI scores for VoiceBank+DEMAND noisy test set.

Yufeng Yang et al. 2023. Time-domain speech enhancement for robust automatic speech recognition. In *Proceedings of Interspeech*, pages 4913–4917.

A Exponential Smoothing v/s Naive Averaging

In this section, we replace the running average scheme described in Section 2.4 with a naive averaging scheme. More specifically, we have

$$\tilde{\gamma}_c = \frac{1}{T}\sum_{t=1}^{T}\gamma_c^t, \qquad \tilde{\beta}_c = \frac{1}{T}\sum_{t=1}^{T}\beta_c^t \tag{4}$$

A distinct disadvantage of this scheme is that it cannot be pipelined unlike the exponential averaging discussed earlier. Hence it precludes real time usage of the enhancement model since we require the knowledge of γ_c^t, β_c^t for all t to calculate $\tilde{\gamma}_c$ and $\tilde{\beta}_c$ for modulation. Nevertheless, we enhance using a trained checkpoint from naive averaging and compare PESQ scores with exponential smoothing in Table 2. It is evident that exponential smoothing provides superior enhancement performance across all SNRs and channel configurations.

B Enhancement Performance on VoiceBank+DEMAND dataset

In the interest of reproducibility and comparison with other well known models we train the proposed model on VoiceBank+DEMAND 28 speaker training dataset (Valentini-Botinhao et al., 2016) and enhance the noisy test set. The PESQ and STOI scores are summarized in Table 3. We down sample the training and test splits of the dataset to 16kHz from 48kHz as the pretrained wav2vec 2.0 checkpoint we are using in our model cannot process 48kHz waveforms. We trained and inferred using both 128 and 32 channel variants of our model and compare it's enhancement capabilities with the baseline model.

Although the baseline model beats us on PESQ, we are equally good on STOI. This is in contrast with the performance trends observed with our custom dataset which we attribute to the different noise types and SNR levels used in preparing this dataset. Nevertheless, we observe an interesting phenomenon from this experiment. It is evident from Table 3 that a larger model does not guarantee superior enhancement performance. The 32 channel variant of the proposed and baseline models beat the 128 channel variant, respectively, on PESQ scores.