WHERE DOES SPEECH ENHANCEMENT ADAPT? PROBING STUDY UNDER CONTROLLED DEGRADATION

Yair Amar, Amir Ivry, Israel Cohen

Technion Israel Institute of Technology
Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering

ABSTRACT

Speech enhancement (SE) models advance rapidly, yet it remains underexplored how degradation of input signals affects their internal representations. We introduce a probing process, aimed at modeling the behavior of internal representations in SE models under controlled degradations to input signals. We apply it to the MUSE SE model by extracting its layer activations under controlled Signal-to-Noise Ratio (SNR) and reverberation C50. We measure layer-wise representational similarity to clean input references using Centered Kernel Alignment (CKA) and regress it against the degradation level, yielding compact, robustness-adaptive profiles. Encoder layers maintain noise-invariant representations while decoder layers adapt strongly, with sensitivity increasing monotonically within blocks and skip-connection boundaries marking the sharpest transitions. The same structure emerges under reverberation and is reproduced independently by MP-SENet and Demucs, two structurally distinct architectures, suggesting that the tradeoff is induced by the enhancement objective rather than a particular model design. Together, these results characterize where SE models adapt to degradation. We then offer insight into how internal representations correlate with output-level performance metrics, e.g., PESQ. Code for reproducing the analysis is publicly available.[1]

***Index Terms*—** Speech Enhancement, Interpretability, Probing, Centered Kernel Alignment, Layer-wise Analysis

1. INTRODUCTION

Speech enhancement (SE) models are routinely evaluated using output-level metrics, yet the internal mechanisms by which they process degraded speech remain largely opaque. Understanding which layers preserve noise-invariant structure and which adapt to degradation conditions would complement output-level evaluation with a mechanistic account of model behavior. In text encoders and vision models, probing consistently reveals hierarchical specialization across depth [1, 2]. In self-supervised speech models, layer-wise analysis of speech foundation models, such as Wav2Vec 2.0 [3], has revealed a consistent acoustic-to-linguistic hierarchy connected to downstream performance [4]. Interpretability work in supervised SE has remained limited, focusing on residual connection analyses [5], linearized autoencoders [6], model dissection [7], and gradient-based attribution [8], none of which examine how internal activation representations of SE models adapt under controlled degradation.

We address this by probing the SE model MUSE [9] under controlled degradation conditions to the input signal, spanning additive noise with signal-to-noise-ratio (SNR) from -10 to 30 dB and reverberation with clarity index C50 [10] from -5 to 25 dB. We employ Centered Kernel Alignment (CKA) [11], which measures layer-wise representational similarity to clean references. A complementary measure uses the diffusion maps manifold learning technique [12, 13]. We map activation representations into a low-dimensional manifold, on which small Euclidean distances align with small diffusion distances of the activation representations.

Our findings reveal a systematic *robustness-adaptivity tradeoff* across depth: encoder layers maintain noise-invariant representations, whereas decoder layers adapt strongly, with adaptivity increasing monotonically toward the output. Skip-connection boundaries mark the sharpest increases in adaptivity. This tradeoff is characterized compactly via linear regression of CKA against degradation level. To assess whether the tradeoff is specific to additive noise, we repeat the analysis under reverberation; the same organizational pattern is recovered, albeit with compressed dynamic range. Examining the regression profiles of MP-SENet [14] and Demucs [15], two structurally distinct SE architectures, shows that the same tradeoff holds independently of the MUSE design. Diffusion distances independently corroborate these findings: Diffusion maps showed that representations of adjacent SNRs are closest in terms of diffusion distance. Together, these results characterize where and how SE models respond to degradation and link internal representations to output-level metrics such as PESQ.

2. PROBING FRAMEWORK

2.1. Probed Model and Activation Extraction

MUSE [9] is a transformer-convolutional (SE) model that follows a U-Net paradigm [16] and was trained on VoiceBank-DEMAND [17]. The architecture comprises a convolutional front end followed by hierarchical transformer blocks across four stages: encoder, latent, decoder, and refinement. Parallel-resolution-level blocks are connected via skip connections, as illustrated in Fig. 1. Each block consists of four transformer layers, yielding 24 probed layers across the magnitude pathway, which is the focus of this analysis. At each probed layer ℓ, the per-utterance activation tensor $\mathbf{A}^{(\ell)} \in \mathbb{R}^{C\times T\times F}$ (channels $\times$ time-frames $\times$ frequency-bins) is reduced to a representation matrix $\mathbf{H}^{(\ell)}$ by averaging over the time axis:

$$\mathbf{H}^{(\ell)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{A}_t^{(\ell)} \ \in \ \mathbb{R}^{C\times F}. \tag{1}$$

Activations are extracted at each of the 24 transformer layers, yielding a single $C \times F$ matrix per utterance per layer.

2.2. Degradation Model

We probe representations under two independent input degradation axes: additive noise and reverberation, each swept from perceptually

[1] https://github.com/YairAmar/seint

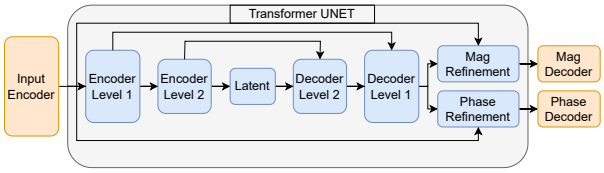

Fig. 1: Architecture of the MUSE model probed in this work. Each block consists of 4 transformer layers.

challenging to near-clean conditions.

For **additive noise**, the degraded signal is $y = x + s$, where x is the clean utterance and n is a noise signal scaled to a target SNR. SNR is swept in integer values from -10 to 30 dB. The corresponding clean utterances serve as the reference.

For **reverberation**, the degraded signal is given by the convolution operator $y = x * h$, where h is a room impulse response (RIR). Reverberation severity is characterized by C50, the logarithmic early-to-late energy ratio at the 50 ms boundary [10], which more directly quantifies the perceptual impact of late reverberation on the speech signal than RT60 [18]. To sweep C50, the RIR energy beyond 50 ms is rescaled to achieve a target C50. Thirteen values, linearly spaced, from -5 to 25 dB are evaluated. Reference activations are computed using the same RIR, with C50 set to 50 dB via late-tail scaling, thereby preserving early reflections, which have been shown to benefit speech perception [19].

2.3. Experimental Setup

Clean utterances are drawn from the VoiceBank test set (16 kHz) and paired with DEMAND noise recordings from the official VoiceBank-DEMAND evaluation setup [17]. Rather than relying on the pre-mixed test set, all mixtures were regenerated at the target integer SNRs to enable controlled degradation.

For the reverberation experiments, a subset of 88 RIRs from the AIR dataset [20] (16 kHz), spanning six room types, are used; for each utterance, five RIRs are selected at random and convolved at each target C50. Clean utterances are drawn from the VoiceBank-DEMAND test speakers. These RIRs are fully independent of the training set. Because the pretrained checkpoint was trained on non-reverberant speech, we fine-tuned MUSE on convolutive mixtures, initializing from pretrained weights with all parameters unfrozen, using AdamW (learning rate=10^{-4}, batch size 28, $\gamma = 0.99$) for 48 epochs on 11,572 utterances convolved on-the-fly with 758 RIRs from RIR-Mega [21], retaining the original loss. On held-out reverberant mixtures (RIR-Mega RIRs, VoiceBank-DEMAND test speakers), the fine-tuned model achieves PESQ=3.02 and STOI=0.944, compared with PESQ=2.17 and STOI=0.834 for the noise-only checkpoint, confirming that the model is suitable for probing in reverberant conditions. The objective of this procedure is to introduce in-domain reverberations into the model.

2.4. Analysis Tools

2.4.1. Centered Kernel Alignment

Representational similarity between degraded and clean activations is quantified using linear Centered Kernel Alignment (CKA). For each representation matrix $\mathbf{H}^{(\ell)}$ (Eq. 1), the linear kernel $\mathbf{K} = \mathbf{H}^{(\ell)}\mathbf{H}^{(\ell)\top}$ is centered; CKA is the cosine similarity between the two centered kernels in Frobenius norm, yielding a score in $[0, 1]$, invariant to orthogonal transformation and isotropic scaling [11]. CKA is computed per utterance on $\mathbf{H}^{(\ell)}$ (Eq. 1) for the degraded and clean conditions, then averaged across utterances and noise types per degradation level.

2.4.2. Linear Regression of CKA Profiles

To summarize representational behavior across degradation levels compactly, we fit a first-order linear model of CKA as a function of degradation level for each layer ℓ:

$$\widehat{\mathrm{CKA}}(\ell, s) = \alpha_\ell + \beta_\ell \cdot s \tag{2}$$

where s denotes degradation level in dB (SNR or C50), the slope β_ℓ quantifies how rapidly representations change with degradation level and serves as a measure of *adaptivity*. The intercept α_ℓ, corresponding to CKA at 0 dB, captures representational similarity under adverse conditions and serves as a measure of *robustness*. Linear fits achieve a coefficient of determination $R^2 > 0.95$ across all layers and conditions, justifying the use of $(\alpha_\ell, \beta_\ell)$ as a compact two-parameter profile for each layer. These profiles enable direct comparison across degradation types and architectures.

2.4.3. Diffusion Maps

To complement CKA similarity with a geometric perspective, we employ diffusion maps [12]. Given the set of centroid representations $\{\bar{\mathbf{H}}_s^{(\ell)}\}$ across degradation levels s, diffusion maps constructs a Markov chain over their pairwise affinities and embeds them into a low-dimensional space where Euclidean distances approximate diffusion distances - distances that reflect similarity between points in the context of the entire cloud's geometry, rather than pairwise proximity alone. For each layer ℓ and degradation level s, the centroid representation is computed by averaging over all N_s utterances at that level:

$$\bar{\mathbf{H}}_s^{(\ell)} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{H}_{s,i}^{(\ell)} \in \mathbb{R}^{C \times F}. \tag{3}$$

where $\mathbf{H}_{s,i}^{(\ell)}$ is the representation matrix (Eq. 1) of the i-th utterance under degradation level s. Given the set of centroids $\{\bar{\mathbf{H}}_s^{(\ell)}\}$ across degradation levels, diffusion maps constructs a Markov chain from a Gaussian affinity kernel with adaptive bandwidth and embeds them into a low-dimensional space where Euclidean distances approximate diffusion distances - distances that reflect similarity between points in the context of the entire cloud's geometry, rather than pairwise proximity alone.

3. RESULTS AND ANALYSIS

We present the analysis in four stages. First, we establish the core robustness-adaptivity tradeoff under additive noise using CKA and linear regression. We then test whether this tradeoff generalizes across degradation types and architectures. Finally, we validate these findings geometrically using diffusion maps.

3.1. The Robustness-Adaptivity Tradeoff Under Noise

The first encoder layers maintain high similarity to their reference activations across the full SNR range, as shown in Fig. 2, appearing as a near-uniform band. With increasing depth, a clear color gradient emerges along the SNR axis, indicating growing adaptivity

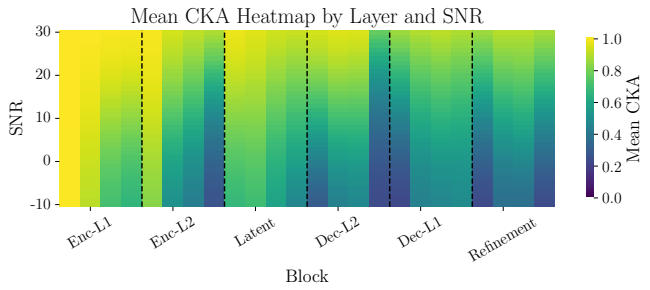

Fig. 2: CKA similarity between clean and noisy activations across layers, grouped by block. SNR increases along the vertical axis; layer depth increases along the horizontal axis.

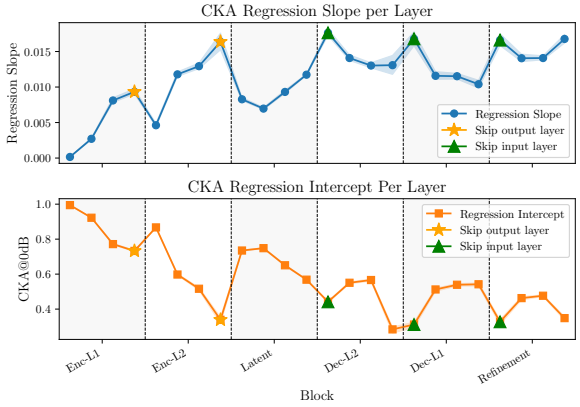

Fig. 3: Linear regression slopes (top) and intercepts (bottom) of CKA as a function of SNR for each probed layer. Skip-connection output and input layers are marked. All fits achieve $R^2 > 0.95$. Shaded regions indicate 95% bootstrap confidence intervals (1000 resamples).

to noise conditions. The latent block occupies an intermediate position, whereas the decoder and refinement layers exhibit the most pronounced gradients. This broad pattern suggests that depth governs the extent to which representations depend on degradation conditions.

Linearization of the CKA-SNR relationship via Eq. 2 for each layer reveals finer patterns within each block (Fig. 3). First, adaptivity increases monotonically with depth within each encoder block but resets at block boundaries. In the decoder and refinement blocks, the within-block trend reverses: the first layer of each block exhibits the highest adaptivity, which then decreases with depth. This aligns with the intuition from the classical Wiener gain $G = \xi/(1+\xi)$, which departs from unity as the *a priori* SNR ξ decreases [22] - optimal suppression is inherently condition-dependent. Second, the intercept mirrors this pattern inversely. Third, layers at decoder skip-connection boundaries produce local slope maxima, exhibiting the highest adaptivity values in the network. This positions skip-connection junctions as sites where condition-dependent processing concentrates. All fits achieve $R^2 > 0.95$, confirming that the linearization captures the dominant trend.

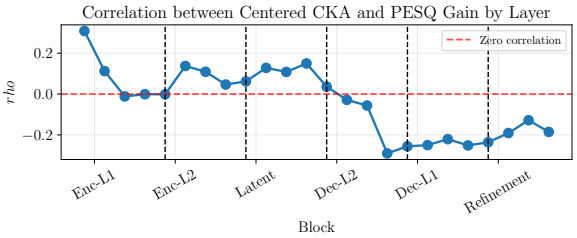

Fig. 4: Layer-wise correlation between CKA and PESQ gain after removing SNR-dependent mean effects via within-group centering.

3.1.1. Connection to Perceptual Quality

Fig. 4 examines whether these internal representational trends carry information on perceptual quality. For each utterance we compute the PESQ improvement $\Delta\text{PESQ} = \text{PESQ}(\hat{y}, y) - \text{PESQ}(x, y)$, where $\hat{y}$ is the enhanced output, x the degraded input, and y the clean reference. Since both CKA and ΔPESQ are strongly driven by SNR, we control for this factor before correlation analysis. We remove SNR-dependent mean effects via within-group centering for each layer ℓ and SNR level s, corresponding to standard residualization of a confounder [23]:

$$\text{CKA}^c_{\ell_i} = \text{CKA}_{\ell_i} - \mathbb{E}[\text{CKA} \mid \ell_i, s_i] \quad (4)$$

$$\Delta\text{PESQ}^c = \Delta\text{PESQ} - \mathbb{E}[\Delta\text{PESQ} \mid s_i] \quad (5)$$

The conditional expectations are estimated by sample means within each group, and per-layer Pearson correlations are computed on the centered quantities. This approach removes any mean-level effect of SNR without imposing a parametric form (e.g., linear dependence). CKA is centered per (ℓ, s), while ΔPESQ per s only. Encoder and latent layers show near-zero positive correlations. Decoder and refinement layers, however, exhibit increasingly negative values with depth in the first decoder block, which then saturate through the other decoder block and the refinement layer. This likely reflects the decoder's functional role, exploiting input-specific structure rather than preserving similarity to the clean reference, which is associated with better enhancement, even as it distances internal states from the clean reference. This links the adaptivity observed in deeper layers directly to perceptual improvement, grounding the robustness-adaptivity tradeoff in output-level performance. Taken together, these results show that MUSE's encoder and decoder serve complementary representational roles: the encoder maintains relatively degradation-invariant structures while the decoder adapts to input conditions. The decoder's stronger adaptation is associated with greater perceptual improvement, as measured by PESQ.

3.2. Generalization Across Degradation Types

To determine whether the observed tradeoff in 3.1 is specific to additive noise or a more general architectural property, we repeated the regression analysis under controlled reverberation, sweeping C50 from -5 to 25 dB (Fig. 5). As in the noise setting, early encoder layers remain close to their reference representations across the full C50 range. Deeper layers, particularly in the decoder and refinement blocks, exhibit increasing adaptivity to degradation. The same depth-dependent progression of slopes (adaptivity) emerges, with intercepts following the inverse pattern. Skip-connection boundaries again produce local slope maxima. The dynamic range compresses,

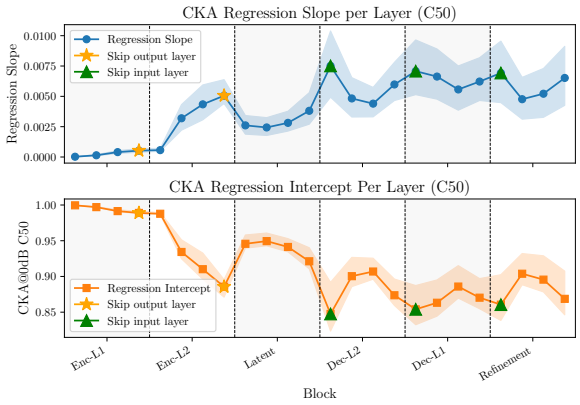

Fig. 5: Linear regression slopes (top) and intercepts (bottom) of CKA as a function of C50 for each probed layer. Shaded regions indicate 95% bootstrap confidence intervals (1000 resamples).

with intercepts remaining above 0.85 across all layers compared to values approaching 0.2 in the deepest decoder layers under noise. Taken together, both inter-block and intra-block trends closely mirror those observed under additive noise. This supports the interpretation that the tradeoff reflects how the architecture partitions enhancement functionally, rather than an exploitation of noise-specific cues.

3.3. Generalization Across Architectures

To determine whether the tradeoff reflects a property of the enhancement objective or a particular architectural choice, we applied the same probing pipeline to two additional architectures: MP-SENet [14], and Demucs [15], both pretrained for speech enhancement on VoiceBank-DEMAND, and DNS Challenge corpus [2] accordingly. CKA was computed on per-block $[T, C]$ activations (time frames $\times$ channels) for Demucs, analogously to the $[C, F]$ matrices used for MUSE and MP-SENet.

Figure 6 plots robustness against adaptivity for all three models, per layer. MUSE and MP-SENet both show strong, highly significant negative correlations. Demucs exhibits a significant but weaker trend. Spearman correlations confirm that the monotonic relationship is robust across all three models. That all three independently trained architectures - spanning transformer, convolutional-recurrent, and hybrid designs - share this negative relationship suggests the tradeoff is a property of the speech enhancement objective, not an artifact of any particular model topology.

3.4. Geometric Corroboration

CKA regression quantifies how each layer's representations differ from clean references as a function of degradation level. Figure 7 extends the analysis from scalar similarity to manifold geometry by measuring pairwise diffusion distances between centroid representations across the SNR grid for each probed block. Representations order consistently by SNR along the diffusion trajectory (Spearman $\rho > 0.95$ across all blocks), indicating that the observed CKA differences reflect structured geometric variation rather than unstructured drift.

[2]The DNS Challenge corpus includes VCTK speakers overlapping with VoiceBank-DEMAND; this does not affect the probing analysis, which examines representational organization rather than enhancement performance.

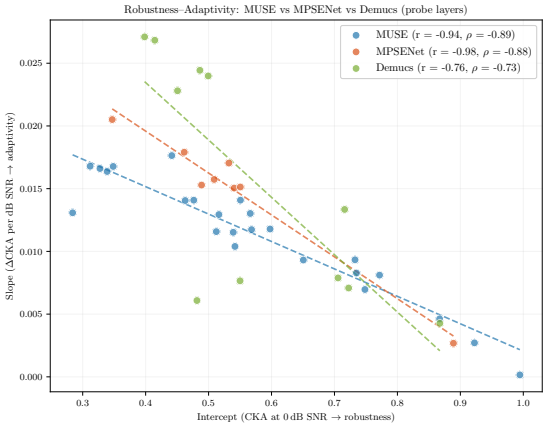

Fig. 6: Robustness (intercept) vs. adaptivity (slope) per probed layer. MUSE ($n = 24$, $r = -0.94$, $p < 10^{-11}$), MP-SENet ($n = 8$, $r = -0.98$, $p < 10^{-4}$), and Demucs ($n = 11$, $r = -0.76$, $p = 0.007$) all exhibit significant negative trends. Spearman correlations: $\rho = -0.89$, $\rho = -0.88$, and $\rho = -0.73$, respectively.

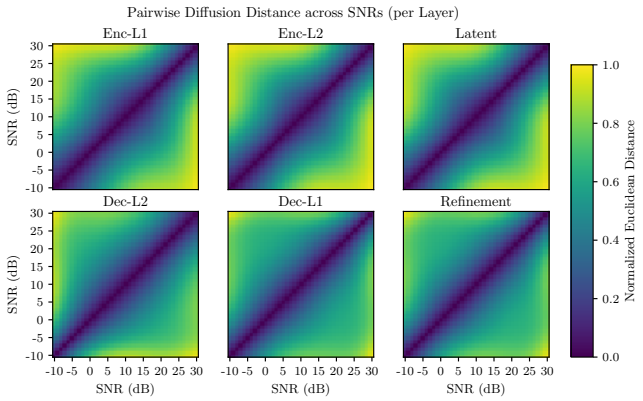

Fig. 7: Pairwise diffusion distances between centroid representations across SNR values (-10 to 30 dB) for each probed block. Encoder blocks show gradual distance gradients; decoder blocks exhibit sharper transitions and larger off-diagonal distances, particularly at low SNR.

4. CONCLUSIONS

We presented a probing study of speech enhancement models combining controlled degradation with CKA-based regression and diffusion-based geometric analysis. Applied to MUSE, the results reveal a consistent depth-dependent organizational pattern: encoder layers maintain stable representations across degradation conditions while decoder layers adapt strongly, with skip-connection boundaries marking the sharpest transitions. This tradeoff is well captured by a linear model ($R^2 > 0.95$), reducing each layer's behavior to a compact, robustness-adaptivity profile. Partial correlation analysis further shows that decoder layers diverging from clean-aligned representations are associated with higher perceptual quality, suggesting the tradeoff has functional significance beyond structural regularity. The same pattern emerges under reverberation, with a compressed dynamic range, and is reproduced independently by MP-SENet ($r = -0.98$) and Demucs ($r = -0.76$) despite

substantial architectural differences, indicating that the tradeoff is induced by the enhancement objective rather than by any particular design choice. Diffusion-based geometric analysis corroborates these findings, confirming that the representational differences are manifold-structured rather than merely scalar. Together, these results complement output-level evaluation by exposing how architectural components divide the enhancement problem internally. Whether this organizational structure persists across larger-scale training corpora and more diverse acoustic conditions remains an open question. Extending the probing methodology to self-supervised and generative SE models, and connecting the observed tradeoff to targeted layer-wise adaptation strategies, are natural next steps.

5. REFERENCES

[1] Ian Tenney, Dipanjan Das, and Ellie Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez, Eds., Florence, Italy, July 2019, pp. 4593–4601, Association for Computational Linguistics.

[2] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein, "SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 6076–6085.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 12449–12460.

[4] Ankita Pasad, Xinjian Zhang, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ICASSP*, 2021, pp. 284–288.

[5] Joao Felipe Santos and Tiago H. Falk, "Investigating the effect of residual and highway connections in speech enhancement models," in *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language*, 2018.

[6] Stéfanos A. Mimilakis, Konstantinos Drossos, Tuomas Virtanen, and Gerald Schuller, "Examining the mapping functions of denoising autoencoders in singing voice separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1019–1030, 2019.

[7] Johannes Heitkaemper, Simon Leglaive, Romain Serizel, and Reinhold Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *Proc. ICASSP*, 2020, pp. 6354–6358.

[8] Sriram Sivasankaran, Emmanuel Vincent, Srikanth Tamilselvam, and Marc Ferras, "Explaining deep learning models for speech enhancement," in *Proc. Interspeech*, 2021, pp. 2816–2820.

[9] Zizhen Lin, Xiaoting Chen, and Junyu Wang, "MUSE: Flexible voiceprint receptive fields and multi-path fusion enhanced taylor transformer for u-net-based speech enhancement," in *Proc. Interspeech*, 2024, pp. 672–676.

[10] International Organization for Standardization, "Acoustics — Measurement of room acoustic parameters — Part 1: Performance spaces," ISO Standard 3382-1:2009, June 2009.

[11] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton, "Similarity of neural network representations revisited," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019, pp. 3519–3529.

[12] Ronald R. Coifman and Stephane Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[13] Boaz Nadler, Stephane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis, "Diffusion maps, spectral clustering and eigenfunctions of fokker–planck operators," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 113–127, 2006.

[14] Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling, "Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *INTERSPEECH 2023*. Aug. 2023, interspeech 2023, p. 3834–3838, ISCA.

[15] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, 2020, pp. 3291–3295.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[17] Cassia Valentini-Botinhao, Xin Wang, Junichi Yamagishi, and Simon King, "Noisy speech database for training speech enhancement algorithms and tts models," in *Proc. Interspeech*, 2016, pp. 503–507.

[18] Pablo Peso Parada, Dushyant Sharma, Jorge Lainez, Daniel Barreda, Toon van Waterschoot, and Patrick A. Naylor, "A single-channel non-intrusive C50 estimator correlated with speech recognition performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 719–732, 2016.

[19] Nicoleta Roman and John Woodruff, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1707–1717, 03 2013.

[20] Marco Jeub, Magnus Schaefer, and Peter Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *2009 16th International Conference on Digital Signal Processing*, pp. 1–5, 2009.

[21] Mandip Goswami, "Rir-mega: a large-scale simulated room impulse response dataset for machine learning and room acoustics modeling," 2025.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[23] Michael H Kutner, "Applied linear statistical models.," 2005.