

HIERARCHICAL GRAPH-CODING DIFFUSION MODEL WITH ADAPTIVE INFORMATION BOTTLENECK FOR MULTICHANNEL SPEECH ENHANCEMENT

Anonymous authors

Paper under double-blind review

ABSTRACT

Diffusion models have achieved strong performance in multichannel speech enhancement, especially in unseen noisy scenarios. However, most existing diffusion methods rely on globally consistent guidance applied either to the output or uniformly across denoiser layers, which fails to provide layer-specific adaptation and introduces redundancy, thereby constraining denoising performance. To address these challenges, we propose a novel hierarchical graph-coding diffusion model with adaptive information bottleneck (HG-Diff-IB) for multichannel speech enhancement. Specifically, we introduce a hierarchical alignment method to align graph-coding with the denoiser at different depths, together with a layer-wise graph-coding modulation mechanism that injects graph information into intermediate features, enabling layer-specific guidance of diffusion feature distributions. Furthermore, we introduce an adaptive information bottleneck that dynamically adjusts the feature compression according to the estimated SNR, effectively balancing noise suppression and target feature preservation. Experimental results demonstrate that our proposed method outperforms baselines in various evaluation metrics.

1 INTRODUCTION

Multichannel speech enhancement has witnessed significant progress in recent years, driven by the synergy of deep learning innovations. By leveraging the spatiotemporal correlations across multiple microphone arrays, it has become a powerful approach for enhancing speech quality in dynamic acoustic environments (Hao et al., 2022; Chau et al., 2024). This advancement not only enables more robust noise suppression, but also delivers better generalization capabilities across diverse real-world scenarios, such as in smart speakers, mobile phones, and hearing aids, where non-stationary noise and interference are prevalent (Doclo et al., 2015; Sainath et al., 2017; Saryuddin Assaqtly et al., 2020).

Recently, diffusion models have shown remarkable potential in speech enhancement (Lu et al., 2021; Gonzalez et al., 2024), achieving high-quality reconstruction even under challenging acoustic conditions. By leveraging iterative denoising, these models provide stronger robustness compared to other approaches (Elgiriye withana & Kodikara, 2024). However, despite these advances, existing diffusion-based SE methods still face two critical limitations.

Firstly, most approaches lack layer-specific guidance, limiting their ability to fully exploit hierarchical representations in denoise process. For example, G-DiffuMSE (Yu et al., 2025) adjust the sampling mean in each denoise step by introducing STGCN reconstruction loss and adversarial loss. NADiffuSE (Wang et al., 2023) and FUSE (Yang et al., 2024) introduce condition to guide each layer of the diffusion model, but it provides the same guidance for every layer, resulting in a lack of targeted adaptation.

Secondly, the adopted guidance features often contain substantial redundancy, introducing irrelevant information that may mislead the denoising process. For instance, DOSE (Tai et al., 2023) and CDiffuSE (Lu et al., 2022) directly concatenates noisy speech into the input to guide denoising,

Through its layer-specific alignment mechanism, this hierarchical alignment enables graph-coding features to be smoothly embedded into the corresponding layers of the denoiser, rather than being crudely fed into a single layer. This integration ensures that at each stage of the entire diffusion process, the denoiser can leverage local detail cues to enhance signal precision and rely on global reconstruction hints to maintain the consistency of speech structure, thereby preventing the generation of distorted speech.

2.2 LAYER-WISE GRAPH-CODING MODULATION

Based on the hierarchical alignment method established in Sec.2.1, we further introduce a layer-wise graph-coding modulation mechanism inspired by (Hudson et al., 2024), which directly injects graph-coding information into the intermediate features of the denoiser to achieve adjustment for the denoiser feature distribution.

Specifically, given the i -th graph-coding $\mathcal{F}_{\phi,i}(xt)$ and the $\mathcal{HA}(i)$ -th layer latent feature of the denoiser $\theta_{\mathcal{HA}(i)}(x_t, t)$, the affine-modulated latent feature is given by:

$$\theta'_{\mathcal{HA}(i)}(x_t, t) = z_s \cdot \text{AdaIN}(\theta_{\mathcal{HA}(i)}(x_t, t)) + z_b \quad (2)$$

Here, z_s and z_b denote the scale and bias parameters derived from a linear projection of $\mathcal{F}_{\phi,i}(x_t)$, which directly modulate the feature distribution. The adaptive instance normalization $\text{AdaIN}(\cdot)$ normalizes each channel independently and is commonly defined as (Huang & Belongie, 2017):

$$\text{AdaIN}(\theta_{\mathcal{HA}(i)}(x_t, t)) = \frac{\theta_{\mathcal{HA}(i)}(x_t, t) - \mu}{\sigma} \quad (3)$$

Here, μ and σ are the mean and standard deviation of the feature $\theta_{\mathcal{HA}(i)}(x_t, t)$, computed independently for each channel.

Speech signals exhibit strong correlation and regularity across multiple channels, while the statistical characteristics of noise are typically more random. AdaIN can normalize the joint features of multiple channels to highlight the consistent patterns of speech features, suppress the random fluctuations of noise, and thereby enable the model to more easily distinguish between speech and noise components.

2.3 INFORMATION BOTTLENECK FOR OPTIMIZATION

2.3.1 INFORMATION BOTTLENECK WITH SNR ADAPTATION

To effectively suppress noise while preserving target features, we introduce an adaptive information bottleneck that dynamically regulates the feature compression according to the estimated SNR. It ensures that the model retains sufficient information under high SNR conditions, while enforcing stronger compression to suppress noise under low SNR conditions.

Formally, based on the general form of the information bottleneck loss (Hu et al., 2024), the adaptive information bottleneck loss is defined as:

$$L_{IB} = -I(Z; Y) + \beta_{\text{adapt}} I(Z; X) \quad (4)$$

where $I(\cdot; \cdot)$ denotes mutual information, β_{adapt} controls the tradeoff, large values results in a highly compressed representation, X , Y and Z denote the input, output, and latent representation, respectively.

The adaptive parameter β_{adapt} is computed directly from the temporal similarity of the input STFT features. Let $\mathbf{x}_t \in \mathbb{R}^{C \times T}$ be the signal of the t -th denoise step, and $W^Q, W^K \in \mathbb{R}^{T \times d_k}$ learned projection matrices. Then β_{adapt} is obtained as

$$\beta_{\text{adapt}} = \text{softmax} \left(\frac{W^Q \mathbf{x}_t \cdot W^K \mathbf{x}_t^\top}{\sqrt{d_k}} \right) \quad (5)$$

In this way, the adaptive information bottleneck allows the model to flexibly regulate feature compression based on real-time SNR, improving both noise reduction and information preservation in diverse SNR conditions.

2.3.2 COLLABORATIVE OPTIMIZATION FOR GRAPH AND DIFFUSION

To prevent potential bias caused by using graph-coding to guide the denoiser alone, we propose a cooperative optimization strategy between the Graph and Diffusion modules. At each denoising step t , the intermediate latent feature of the diffusion model is first updated via layer-wise hierarchical graph-coding modulation, according to Eq.2. Then, the graph network is optimized using the loss:

$$L_{IB} = \|\mathcal{F}_\phi(x_t) - x_{0,t}\|_2^2 + \beta_{\text{adapt}} I(Z; X) \quad (6)$$

where $x_{0,t} = \frac{1}{(1-m_t)\sqrt{\alpha_t}}(x_t - m_t\sqrt{\alpha_t}y - \sqrt{\delta_t}\epsilon_\theta(x_t))$ is the estimated target at step t by denoiser.

Instead of using a single set of graph-coding features to guide all steps, the graph-coding features are dynamically updated with step t . Even if a temporary deviation occurs at a certain step, the graph network optimization process in subsequent steps can quickly correct it, thereby preventing the accumulation of deviations.

3 EXPERIMENT

3.1 EXPERIMENTAL SETUP

Datasets: For dataset construction, we synthesize 6,000 six-channel recordings using the speech from DNS-Challenge Dubey et al. (2022) and noise from ESC50 (Piczak, 2015) dataset via the gpuRIR toolkit (Diaz-Guerra et al., 2021). These recordings simulate general speech-noise scenarios in a 5m×4m×3m room, with a six-channel microphone array placed at the center of the room. For evaluation, we use a test set consisting of 108 samples from DNS-Challenge combined with noise from FSD50K (Fonseca et al., 2021), which includes unseen noise types and covers SNRs ranging from -5 dB to 10 dB.

Baselines: Proposed framework are compared with: (1) DM-STGCN-NTA (Zhang, 2024): a framework which employs GCN and spatio-temporal convolution to capture spatial temporal spectral correlations; (2) Diffwave (Kong et al., 2020): a diffusion model with bidirectional dilated conv; (3) DOSE (Tai et al., 2023): a framework which adopts dropout and adaptive prior to address condition collapse; (4) CDiffuSE (Lu et al., 2022): a conditional diffusion which uses the observed noisy speech signal as condition to adapt to non-Gaussian noise; (5) G-DiffuMSE (Yu et al., 2025): a mechanism integrates graph-guided diffusion with noise-conditional modeling for robust multi-channel SE.

Metrics: PESQ (Rix et al., 2001) and STOI (Taal et al., 2011) are used to measure speech quality, whose value ranges are [-0.5, 4.5] and [0, 1], respectively.

For pre-training, we initialize Diffusion with VoiceBank (Veaux et al., 2013)-pretrained weights and fine-tune it on synthetic data for 301 epochs with learning rate setted to 1e-4. And we pre-trained DM-STGCN-NTA on synthetic data for 351 epochs with learning rate setted to 1e-4. During the sampling process, we further update the DM-STGCN-NTA using the optimization strategy mentioned in Sec.2.3 for 10 epochs with a learning rate of 1e-6.

3.2 RESULTS AND ANALYSIS

Table 1: PESQ results of models with different input SNRs

	SNR=-5dB	SNR=0dB	SNR=5dB	SNR=10dB	Avg.
Noisy	1.0338±0.0187	1.0362±0.0202	1.0824±0.0643	1.1122±0.0568	1.0662±0.0558
DM-STGCN-NTA	1.0599±0.0347	1.1465±0.0930	1.2676±0.1320	1.4014±0.1523	1.2189±0.1708
Diffwave	1.0331±0.0178	1.0443±0.0407	1.1104±0.1055	1.1461±0.0838	1.0835±0.0849
DOSE	1.0438±0.0197	1.0987±0.0548	1.1431±0.0587	1.2119±0.0605	1.0944±0.0754
CDiffuSE	1.0882±0.1286	1.1148±0.1425	1.2351±0.1941	1.4474±0.1999	1.2214±0.2207
G-DiffuMSE	1.0617±0.0398	1.1503±0.0995	1.2758±0.1429	1.4010±0.1342	1.2222±0.1701
HG-Diff-IB(ours)	1.1088±0.0566	1.1971±0.0944	1.3171±0.1340	1.4357±0.1258	1.2647±0.1634

3.2.1 MULTI-MODEL COMPARISON UNDER DIFFERENT SNRS

Table.1 presents the PESQ performance of all models across different SNR conditions. Overall, our proposed hierarchical graph-coding diffusion model with adaptive information bottleneck consistently outperforms both discriminative and diffusion-based baselines, achieving robust improvements across all noise levels.

As against the discriminative baselines, our method outperforms DM-STGCN-NTA by an average of 3.76%, as well as 4.61% and 4.41% gain at -5dB and 0dB, which verifies its ability to suppress strong noise.

Compared with diffusion-based baselines, it gains average improvements of 16.72%, 15.56%, 3.55% and 3.48% over Diffwave, DOSE, CDiffuSE and G-DiffuMSE, respectively, which shows the robustness of our method in unseen noisy scenarios. Notably, proposed method shows significant performance gains at -5dB and 0dB: it outperforms Diffwave by 7.32% and 14.64%, DOSE by 6.23% and 8.95%, CDiffuSE by 1.90% and 7.38%, and G-DiffuMSE by 4.43% and 4.07%, confirming robust noise suppression in tough scenarios.

3.2.2 ABLATION STUDY

Table.2 and Fig.2 presents an ablation study to assess the contribution of each component in the proposed method.

Table 2: Ablation study on PESQ and STOI

PESQ					
Input SNR	-5dB	0dB	5dB	10dB	Avg.
Noisy	1.0338	1.0362	1.0824	1.1122	1.0662
Diffusion	1.0617	1.1503	1.2758	1.4010	1.2222
+FiLM Perez et al. (2017)	1.0633	1.1552	1.2861	1.4181	1.2307
+AdaGN Wu & He (2018)	1.0608	1.1521	1.2860	1.4314	1.2326
+AdaIN Huang & Belongie (2017)	1.0628	1.1579	1.2952	1.4332	1.2373
++fixedIB	1.1063	1.1963	1.3166	1.4261	1.2613
++adaptiveIB	1.1088	1.1971	1.3171	1.4357	1.2647
STOI					
Input SNR	-5dB	0dB	5dB	10dB	Avg.
Noisy	0.6539	0.7270	0.7744	0.8385	0.7484
Diffusion	0.6999	0.8015	0.8507	0.8933	0.8114
+FiLM	0.7024	0.8006	0.8484	0.8911	0.8106
+AdaGN	0.6960	0.7999	0.8490	0.8911	0.8090
+AdaIN	0.6972	0.8019	0.8495	0.8923	0.8103
++IB	0.7308	0.8030	0.8444	0.8811	0.8148
++adaptiveIB	0.7305	0.8069	0.8460	0.8814	0.8162

Note: “+” denotes adding a module based on the diffusion module;

“++” is cumulative stacking on top of the previous modules.

FiLM is a general-purpose conditioning method for neural networks that performs feature-wise affine transformations on the network’s features based on conditioning information.

AdaGN is an innovative normalization method that integrates time-step and class embedding information into the group normalization operation.

It is obvious that incorporating layer-wise graph-coding modulation performs better than the diffusion baseline on both metrics. For PESQ, it achieves improvements of 0.70% with FiLM, 0.85% with AdaGN, and 1.24% with AdaIN over original diffusion. In terms of STOI, there are enhancements of 1.49% with FiLM, 1.32% with AdaGN, and 1.15% with AdaIN. These results verify the effectiveness of hierarchical graph-coding modulation and also highlight the superiority of AdaIN.

Furthermore, adding the information bottleneck module further enhances both PESQ and STOI results. For PESQ, improvements are observed particularly under extremely noisy environments, for instance, 4.09% at -5dB and 3.32% at 0dB, verifying its ability to suppress redundant noise information. In terms of STOI, there are also gains, such as 4.43% at -5dB and 2.05% at 0dB.

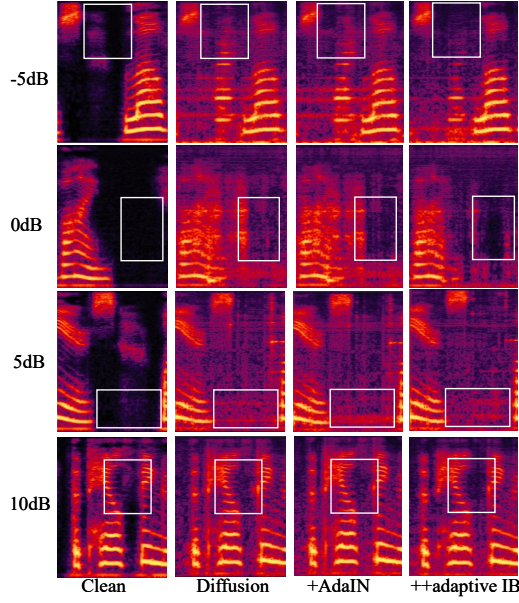


Figure 2: Spectrogram Comparison in Ablation Study

While a fixed information bottleneck causes minor degradation in PESQ and STOI at high SNR, the adaptive IB eliminates this drawback and maintain stable performance on average SNR levels. It can also be observed from the spectrogram shown in Fig.2 that the adaptive IB module can achieve background noise suppression and gain much sparser backgrounds under all SNR conditions.

4 CONCLUSIONS

This paper presented HG-Diff-IB, a hierarchical graph-coding diffusion model with adaptive information bottleneck for multichannel speech enhancement. The proposed method enables precise feature modulation through layer-wise graphcoding and adaptively balances noise suppression with feature preservation. Experimental results show improvements over both discriminative and diffusion-based baselines, especially under challenging scenarios.

ACKNOWLEDGMENTS

The work is supported by National Natural Science Foundation of China(Grant No. 62271039). We appreciate the constructive comments from the anonymous reviewers, which help improve this work.

REFERENCES

- Hoang Ngoc Chau, Tien Dat Bui, Huu Binh Nguyen, Thanh Thi Hien Duong, and Quoc Cuong Nguyen. A novel approach to multi-channel speech enhancement based on graph neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1133–1144, 2024. doi: 10.1109/TASLP.2024.3352259.
- Chia-Wei Chen, Jen-Cheng Hou, Yu Tsao, et al. Davse: A diffusion-based generative approach for audio-visual speech enhancement. In *3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC)*, 2024.
- David Diaz-Guerra, Antonio Miguel, and Jose R Beltran. gpurir: A python library for room impulse response simulation with gpu acceleration. *Multimedia Tools and Applications*, 80(4):5653–5671, 2021.

- Simon Doclo, Walter Kellermann, Shoji Makino, et al. Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. *IEEE Signal Processing Magazine*, 32(2):18–30, 2015. doi: 10.1109/MSP.2014.2366780.
- Harishchandra Dubey, Vishak Gopal, Ross Cutler, et al. Icassp 2022 deep noise suppression challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9271–9275, 2022.
- Nidula Elgiriye withana and ND Kodikara. A comprehensive review on generative models for speech enhancement. In *2024 4th International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, pp. 236–252. IEEE, 2024.
- Eduardo Fonseca, Xavier Favory, Jordi Pons, et al. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.
- Philippe Gonzalez, Zheng-Hua Tan, Jan Østergaard, et al. Investigating the design space of diffusion models for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- Minghui Hao, Jingjing Yu, and Luyao Zhang. Spatial-temporal graph convolution network for multichannel speech enhancement. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6512–6516, 2022. doi: 10.1109/ICASSP43922.2022.9746054.
- Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, et al. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5325–5344, 2024.
- Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1510–1519, 2017. doi: 10.1109/ICCV.2017.167.
- Drew A Hudson, Daniel Zoran, Mateusz Malinowski, et al. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23115–23127, 2024.
- Zhifeng Kong, Wei Ping, Jiaji Huang, et al. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.
- Yen-Ju Lu, Yu Tsao, and Shinji Watanabe. A study on speech enhancement based on diffusion probabilistic model. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 659–666. IEEE, 2021.
- Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, et al. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7402–7406. Ieee, 2022.
- Ethan Perez, Florian Strub, Harm de Vries, et al. Film: Visual reasoning with a general conditioning layer, 2017. URL <https://arxiv.org/abs/1709.07871>.
- Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.
- Antony W Rix, John G Beerends, Michael P Hollier, et al. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.
- Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, et al. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):965–979, 2017. doi: 10.1109/TASLP.2017.2672401.
- Mohammad Iqbal Saryuddin Assaqty, Ying Gao, Ahmad Musyafa, Weisheng Wen, Quansi Wen, and Noni Juliasari. Independent public video conference network. In *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 142–148, 2020. doi: 10.1109/iSemantic50169.2020.9234229.

- Cees H. Taal, Richard C. Hendriks, Richard Heusdens, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011. doi: 10.1109/TASL.2011.2114881.
- Wenxin Tai, Yue Lei, Fan Zhou, et al. Dose: Diffusion dropout with adaptive prior for speech enhancement. *Advances in Neural Information Processing Systems*, 36:40272–40293, 2023.
- Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, pp. 1–4. IEEE, 2013.
- Wen Wang, Dongchao Yang, Qichen Ye, et al. Nadiffuse: Noise-aware diffusion-based model for speech enhancement. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2416–2423. IEEE, 2023.
- Yuxin Wu and Kaiming He. Group normalization, 2018. URL <https://arxiv.org/abs/1803.08494>.
- Yiyuan Yang, Niki Trigoni, and Andrew Markham. Pre-training feature guided diffusion model for speech enhancement. *arXiv preprint arXiv:2406.07646*, 2024.
- Jingjing Yu, Lirui Wu, Chi Yang, et al. Graph-guided spatial-temporal diffusion model for speech enhancement with microphone array. In *2025 25th International Conference on Digital Signal Processing (DSP)*, pp. 1–5. IEEE, 2025.
- L. Zhang. Research on multichannel speech enhancement in open environment. Master’s thesis, Beijing Jiaotong University, Beijing, China, 2024.