Multi-Domain Processing via Hybrid Denoising Networks for Speech Enhancement

Jang-Hyun Kim[*1,2], **Jaejun Yoo**[*2], **Sanghyuk Chun**[2], **Adrian Kim**[2], **Jung-Woo Ha**[2]
[1]Department of Mathematical Science, Seoul National University
[2]Clova AI Research, NAVER Corp.
[1]blue378@snu.ac.kr
[2]{jaejun.yoo, sanghyuk.c, adrian.kim, jungwoo.ha}@navercorp.com

Abstract

We present a hybrid framework that leverages the trade-off between temporal and frequency precision in audio representations to improve the performance of speech enhancement task. We first show that conventional approaches using specific representations such as raw-audio and spectrograms are each effective at targeting different types of noise. By integrating both approaches, our model can learn multi-scale and multi-domain features, effectively removing noise existing on different regions on the time-frequency space in a complementary way. Experimental results show that the proposed hybrid model yields better performance and robustness than using each model individually.

1 Introduction

The trade-off between temporal and frequency resolution is a well-known phenomenon in the signal processing community, e.g., the window size in discrete Fourier transformation (DFT) [1]. The larger the time segment, the more frequencies are extracted, thus giving us higher frequency resolution in the expense of temporal resolution. Therefore, it is obvious that time-series and time-frequency representations can provide complementary views when investigating a given signal. To the best of our knowledge, however, existing deep learning-based approaches proposed for speech enhancement have only taken either time-series (i.e., raw-audio) [7, 8] or time-frequency representation (i.e., spectrogram) as an input [4, 11, 16]. In this work, we find that models using different audio representations each specialize at tackling specific types of noise, and are also complementary to each other. Grounding on this observation, we propose a hybrid framework which enables the model to learn multi-scale and multi-domain features, dubbed multi-domain processing via hybrid denoising networks (MDPhD). We devise a sequential model integrating two modules of both representations by employing auxiliary loss. Experimental results and ablation studies show that the proposed model can effectively utilize complementary information of time and time-frequency domains. Although our hybridizing strategy is rather straightforward, MDPhD shows better denoising performance than other state of the art (SOTA) algorithms across a variety of noises under multiple measures. Note that the hybrid framework is general and not restricted to the current specific model. The performance can be further improved by employing newly developed models from each domain, by equipping a new loss function, or by designing a better hybridizing strategy.

Our contributions are as follows: 1) We empirically show that the way a model performs denoising depends on its input representation. 2) We propose a hybrid framework that can exploit multi-scale and multi-domain features. To the best of our knowledge, this is the first hybrid approach, effectively utilizing both time and time-frequency domain information. 3) The proposed hybrid model (MDPhD) outperforms SOTA algorithms in the speech enhancement task.

[*]Equal contribution.

2 Model Description

We first describe the objective function and the selected modules that have been reported to show competitive performance using either raw-audio [6] or spectrogram input [3]. Selected models are each used later as components of our proposed hybrid model.

2.1 Objective function

We employ the energy-conserving loss function proposed in [8] which simultaneously considers speech and noise signals. Let the noisy input x consist of clean speech s and noise n. The estimated speech by the model is referred to as $\hat{s}$. Then, our objective function is defined as follows:

$$L(x, s, n, \hat{s}) = \|s - \hat{s}\|_1 + \|n - \hat{n}\|_1, \tag{1}$$

where $\hat{n} = x - \hat{s}$ represents the estimated noise signal and $\| \cdot \|_1$ denotes ℓ_1 norm.

2.2 Hybrid Model

We construct the time domain network based on TasNet [6] which employs one-dimensional dilated convolution to handle long time sequences of raw-audio. TasNet has shown competitive sample quality for speech source separation, which is a similar task to speech enhancement. In our experiments, we used a reduced version of TasNet. With a slight abuse of notation, we refer to the network as "TasNet" for simplicity. For the time-frequency (T-F) domain network, we employ a U-Net structure based on two-dimensional convolutions which has been widely used in various source separation tasks [3, 9]. The T-F domain network aims to learn an ideal ratio mask (IRM) of a noisy spectrogram input [15]. By multiplying the estimated mask to the noisy spectrogram, the model can remove the noise from the time-frequency space.

We hybridize both time and T-F domain networks in a cascaded way (Fig. 1). To make both networks contribute to the denoising task equally well, we devise our model with an auxiliary loss $L(x, s, n, \hat{s}_{i,mid})$ at the intermediate conjunction, where $\hat{s}_{i,mid}$ is the output of the former network. In addition, to let both networks have access to the full data information that is not processed (denoised) by the other, we train the entire model by alternately switching the sequential order of each component. For inference, we can either use a single path or average the results from both paths. Here, we simply average the output of the model, which showed the best performance.

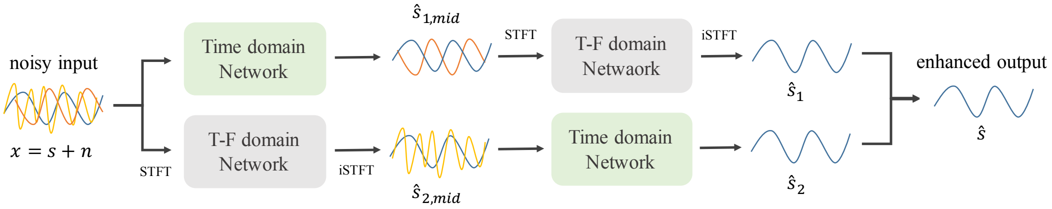

Figure 1: A schematic illustration of the hybrid system (MDPhD). Note that the network of the same domain (same color) shares the parameters. For the time-frequency (T-F) domain network, we convert the time-domain input to a spectrogram using the short time Fourier transform (STFT), whose output is converted back to a waveform using the inverse short time Fourier transform (iSTFT).

The final objective of the hybrid model with auxiliary loss becomes

$$\min_{\theta} \sum_{i=1,2} L(x, s, n, \hat{s}_{i,mid}) + \sum_{i=1,2} L(x, s, n, \hat{s}_i), \tag{2}$$

where θ denotes the network parameter.

3 Experiments

3.1 Data and Experimental Setup

Dataset We used the dataset [13] that has been used in the recent speech enhancement studies [7, 8]. The dataset was produced by synthesizing the clean speech of Voice Bank corpus [14] and the noise data of Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [12]. The training dataset consists of audio from 28 speakers, and the test dataset is composed of the recordings from two speakers. Each speaker's data contains 400 sentences with four noise levels. To deal with signals without voices, we added noise-only data to the training dataset, which is a quarter of the total number. In our experiments, all audio samples recorded at 48kHz were subsampled to 16kHz.

Experimental Setup During training and testing, we split speech waveforms with a sliding window of approximately one second (16384 samples) every 500 ms (50% overlap). To obtain the spectrograms, we used the short time Fourier transform (STFT) of 512 window size and 256 hop size with Hanning window. The output spectrogram is converted back to the time domain using the inverse STFT. For training, we used batch renormalization to cope with a small batch size of 16 and Adam optimizer with the initial learning rate of 2e-4. The learning rate was decayed by half every 100,000 iterations. All of the experiments were performed on NAVER Smart Machine Learning (NSML) platform [5]. For more details, please refer to the supplementary material.

3.2 Experimental Results

Hybrid Framework Validation To show the complementary characteristic of the time and T-F domain networks, we additionally synthesized noisy signals consisting of speech signals from the test dataset and noises which are either babbles (DEMAND), high frequency sinusoidal noise of 1000 $\sim$ 5000 Hz, or both. Note that the networks did not see any of these noises during the training phase. As shown in figure 2, while the spectrogram approach (U-Net) successfully removes high frequency noise that is prominent in the spectrogram, it suffers from dealing with babble noise which is hardly distinguishable from the frequency components of speech signals. On the other hand, the raw-audio approach (TasNet) shows superior results on denoising babble noise, which were even better than that of U-Net with doubled parameter size (Table 1). Note that, however, TasNet fails to remove high frequency noise, which is supposedly hard to capture in the time domain (Fig. 2 red arrow).

Table 1 summarizes these observations along with ablation studies. Our hybrid model (MDPhD) showed the best performance by combining the strength of each model. While the other methods had a noticeable weak domain, MDPhD showed comparable performance across all noise types. Note that MDPhD showed the best performance when the noises are mixed, which is more practical in real world applications. When we only trained a single path of MDPhD, the model failed to fully utilize the complementary information from both domains. Interestingly, we found that the performance of the model tends to follow the characteristics of the network that comes first in order. For example,

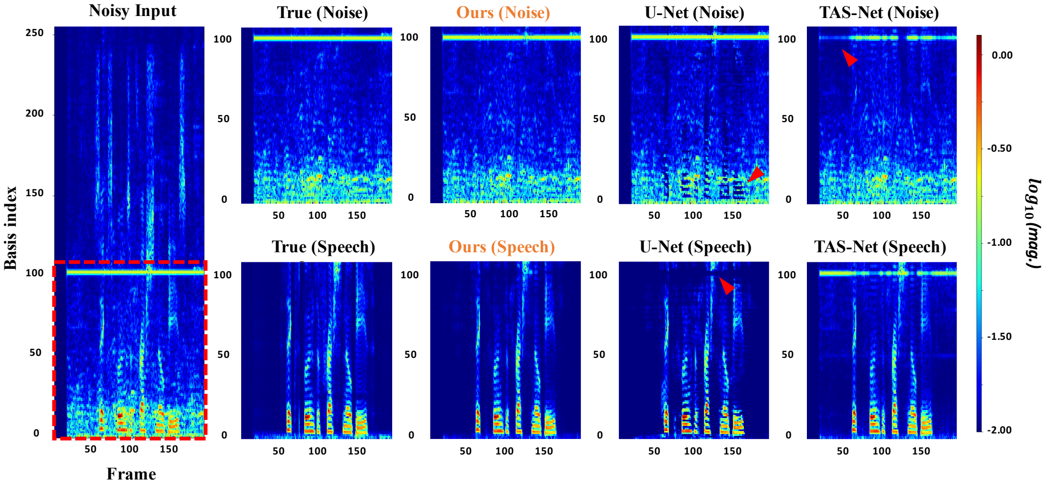

Figure 2: Comparison of denoised results for inputs with babble and high frequency noise. For clarity, the output results of the boxed region (red dotted line) of the noisy input is demonstrated in two perspectives. The top row shows the estimated noise and the bottom row displays the estimated speech signal. Some noticeable distortions of U-Net and TasNet in the spectrogram are marked by red arrows.

Table 1: Ablation study performed on various types of noises (babble, high frequency and a mixture of both) with two signal-to-noise ratios (SNRs) (5 and 10 dB). We evaluated the SNR of each model output in decibel (dB) scale. D and U denotes TasNet using one-dimensional dilated convolution and U-Net, respectively. The number of parameters is noted next to the model (e.g., 1.5 M = 1.5 million). $U \to D$ and $D \to U$ represent single path models without alternately switching training procedure. Our hybrid model is referred to as H (1.5 M + 1.5 M), where the model exploits both pathways ($U \to D$ and $D \to U$) during the training and testing. The best result for each noise type is given in bold style.

	babble		**high freq.**		**babble + high freq.**	
	SNR 5	SNR 10	SNR 5	SNR 10	SNR 5	SNR 10
D (1.5 M)	13.69	16.83	4.86	11.21	11.47	15.23
D^* (3 M)	**14.25**	**17.12**	6.27	11.88	12.74	15.84
U (1.5 M)	10.55	14.51	**17.84**	20.68	11.44	15.41
U^* (3 M)	11.48	15.46	17.60	**21.03**	12.29	16.08
$U \to D$	11.96	15.50	15.08	18.37	12.49	16.01
$D \to U$	14.09	16.97	11.13	17.59	13.42	16.95
H (Ours)	13.81	16.78	15.10	19.09	**14.02**	**17.08**

the $U \to D$ model shares the weakness of U-Net and vice versa. We conjecture that this happens because the latter network cannot reconstruct the information that is already lost from the former network. In addition, we tested various objective functions and confirmed that the complementary nature of the two approaches does not come from a specific choice of the objective function (see the supplementary material).

3.3 Comparison with Other Methods

Using the test dataset, we compared our results to recent studies of speech enhancement field. With the same training and test dataset split, our model showed the best performance quantitatively and qualitatively among the others under various measures [2] (Table 2).

For the qualitative results, please refer to the web demo page (`https://mdphdnet.github.io`), where we have uploaded several denoised examples using the models introduced in the table (except MMSE-GAN whose code is unavailable).

Table 2: Comparison with other methods. Note that, the indicators that are not reported in the original paper are marked as dash (-). The predicted rating of speech distortion (CSIG), background distortion (CBAK) and overall quality (COVL) are reported (from 1 to 5, higher is better). PESQ (from -0.5 to 4.5, higher is better) stands for perceptual evaluation of speech quality and SSNR (higher is better) is segmental SNR. The best result for each measure is given in bold style.

	CSIG	**CBAK**	**COVL**	**PESQ**	**SSNR**
Wiener [10]	3.23	2.68	2.67	2.22	5.07
SEGAN [7]	3.48	2.94	2.80	2.16	7.73
Wavenet [8]	3.62	3.23	2.98	-	-
MMSE-GAN [11]	3.80	3.12	3.14	2.53	-
TasNet (3M)	3.80	3.29	3.18	2.57	9.65
U-Net (3M)	3.65	3.21	3.05	2.48	9.34
MDPhD (3 M + 3 M)	**3.85**	**3.39**	**3.27**	**2.70**	**10.22**

4 Conclusion

We demonstrated that the conventional speech enhancement models have limitations due to using specific representations. Based on this observation, we proposed a hybrid approach that exploits multi-domain features for speech enhancement, dubbed multi-domain processing via hybrid denoising networks (MDPhD). With respect to five metrics, MDPhD achieved the best performance compared to the other concurrent models. Because MDPhD is a general framework, future work may include developing a more elegant way of hybridizing and extending this framework to other signal processing tasks, such as music and speech source separation.

References

[1] Franz Hlawatsch and G. F. Boudreaux-Bartels. Linear and quadratic time-frequency signal representations. *IEEE Signal Processing Magazine*, 9:21–67, 1992.

[2] Yi Hu and Philipos C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:229–238, 2008.

[3] Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-Net convolutional networks. In *International Society for Music Information Retrieval (ISMIR)*, 2017.

[4] Pavan Karjol, M. Ajay Kumar, and Prasanta Kumar Ghosh. Speech enhancement using multiple deep neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[5] Hanjoo Kim, Minkyu Kim, Dongjoo Seo, Jinwoong Kim, Heungseok Park, Soeun Park, Hyunwoo Jo, KyungHyun Kim, Youngil Yang, Youngkwan Kim, Nako Sung, and Jung-Woo Ha. NSML: Meet the MLaaS platform with a real-world case study. *CoRR*, abs/1810.09957, 2018.

[6] Yi Luo and Nima Mesgarani. TasNet: Surpassing ideal time-frequency masking for speech separation. *arXiv preprint arXiv:1809.07454*, 2018.

[7] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: Speech enhancement generative adversarial network. In *Interspeech*, 2017.

[8] Dario Rethage, Jordi Pons, and Xavier Serra. A Wavenet for speech denoising. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[10] Pascal Scalart and Jozue Vieira Filho. Speech enhancement based on a priori signal to noise estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996.

[11] Meet H Soni, Neil Shah, and Hemant A Patil. Time-frequency masking-based speech enhancement using generative adversarial network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[12] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, 133(5):3591–3591, 2013.

[13] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In *ISCA Speech Synthesis Workshop*, 2016.

[14] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *In International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013.

[15] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22:1849–1858, 2014.

[16] Xiao-Lei Zhang and DeLiang Wang. Multi-resolution stacking for speech separation based on boosted dnn. In *Interspeech*, 2015.

A More Experimental Results

To show that the complementary nature of the time domain and T-F domain networks does not come from a specific choice of the objective functions, we trained each network module with various objective functions. Table 3 summarizes the results. We found that the performance was not significantly different.

Table 3: SNR evaluation of models with various objective functions. D and U denote the TasNet (reduced) using one-dimensional dilated convolution and U-Net, respectively. The type of objective functions are noted next to the model name. ℓ_1 represents our baseline objective function. ℓ_2 represents an objective function that substitutes the ℓ_1 term of equation (1) with ℓ_2. SNR indicates an objective function that directly optimizes the SNR. SPEC represents the ℓ_2 distance between a clean speech spectrogram and the estimated spectrogram.

	babble noise			**high frequency band**		
	SNR 5	SNR 10	SNR 15	SNR 5	SNR 10	SNR 15
D-ℓ_1	13.69	16.83	19.57	4.96	11.31	16.54
D-ℓ_2	13.53	16.57	19.26	6.67	12.82	16.94
D-SNR	13.45	16.71	19.51	4.48	10.90	16.38
U-ℓ_1	10.55	14.51	18.11	17.87	20.68	21.92
U-ℓ_2	10.54	14.48	17.97	17.89	20.65	22.32
U-SPEC	10.47	14.38	18.01	19.73	21.47	22.27

B Model Architecture

In this section, we present the detailed configuration of the models we used. In the following figures, each block consists of a convolutional operation, normalization and an activation function. Note that, normalization is not used at the first and the last layer of each model. The operation ⊙ means element-wise multiplication and the preceding layer of this operation uses sigmoid as an activation function.

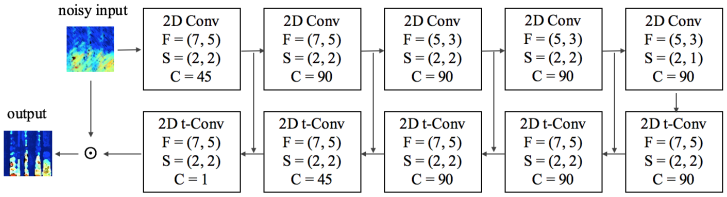
noisy input
2D Conv F = (7, 5) S = (2, 2) C = 45
2D Conv F = (7, 5) S = (2, 2) C = 90
2D Conv F = (5, 3) S = (2, 2) C = 90
2D Conv F = (5, 3) S = (2, 2) C = 90
2D Conv F = (5, 3) S = (2, 1) C = 90
output
⊙
2D t-Conv F = (7, 5) S = (2, 2) C = 1
2D t-Conv F = (7, 5) S = (2, 2) C = 45
2D t-Conv F = (7, 5) S = (2, 2) C = 90
2D t-Conv F = (7, 5) S = (2, 2) C = 90
2D t-Conv F = (7, 5) S = (2, 2) C = 90

Figure 3: U-Net (1.5M) architecture. 2D Conv means a two-dimensional convolution block consisting of a two-dimensional convolution operation with filter size F (height, width), stride size S (height, width) and output channel size C followed by batch renormalization and leaky-RELU activation function. 2D t-Conv means a two-dimensional transposed convolution block. Our baseline models used in experiments process the log-magnitude of the input spectrogram.

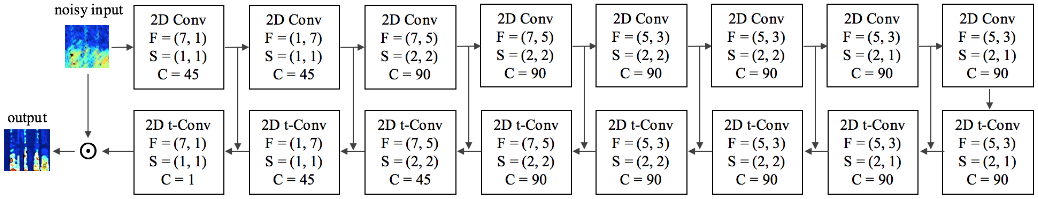
noisy input
2D Conv F = (7, 1) S = (1, 1) C = 45
2D Conv F = (1, 7) S = (1, 1) C = 45
2D Conv F = (7, 5) S = (2, 2) C = 90
2D Conv F = (7, 5) S = (2, 2) C = 90
2D Conv F = (5, 3) S = (2, 2) C = 90
2D Conv F = (5, 3) S = (2, 2) C = 90
2D Conv F = (5, 3) S = (2, 1) C = 90
2D Conv F = (5, 3) S = (2, 1) C = 90
output
⊙
2D t-Conv F = (7, 1) S = (1, 1) C = 1
2D t-Conv F = (1, 7) S = (1, 1) C = 45
2D t-Conv F = (7, 5) S = (2, 2) C = 45
2D t-Conv F = (7, 5) S = (2, 2) C = 90
2D t-Conv F = (5, 3) S = (2, 2) C = 90
2D t-Conv F = (5, 3) S = (2, 2) C = 90
2D t-Conv F = (5, 3) S = (2, 1) C = 90
2D t-Conv F = (5, 3) S = (2, 1) C = 90

Figure 4: U-Net (3M) architecture.

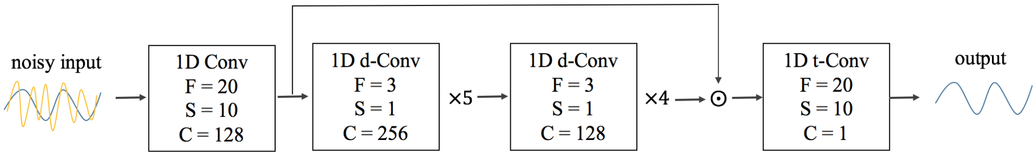

Figure 5: TasNet (1.5M) architecture. 1D Conv means a one-dimensional convolution block and 1D d-Conv stands for a one-dimensional dilated convolution block. The dilation rate of each dilated convolution block is doubled as it goes forward. The convolution operation of the dilation convolution block follows the non-causal method, which takes the value of both ahead and back of the current time step. 1D t-Conv means a one-dimensional transposed convolution block.

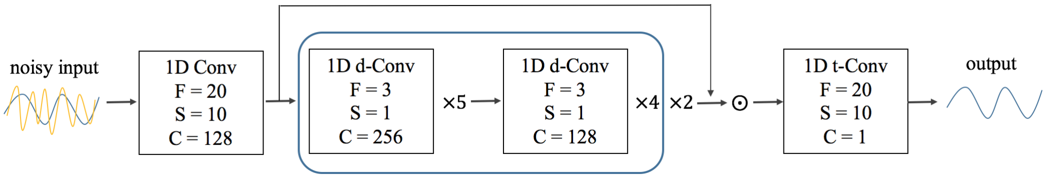

Figure 6: TasNet (3M) architecture.