MASKING AND INPAINTING: A TWO-STAGE SPEECH ENHANCEMENT APPROACH FOR LOW SNR AND NON-STATIONARY NOISE

Xiang Hao[1,2], *Xiangdong Su*[1*], *Shixue Wen*[2], *Zhiyu Wang*[1], *Yiqian Pan*[2], *Feilong Bao*[1], *Wei Chen*[2]

[1]Inner Mongolia Key Laboratory of Mongolian Information Processing Technology,
College of Computer Science, Inner Mongolia Univeristy, Hohhot, China
[2]AI Interaction Division, Sogou Inc., Beijing, China

ABSTRACT

Currently, low signal-to-noise ratio (SNR) and non-stationary noise cause severe performance degradation for most of speech enhancement models. For better speech enhancement at the above scenarios, this paper proposes a two-stage approach that consists of binary masking and spectrogram inpainting. In the binary masking stage, we first obtain binary mask by hardening soft mask and then use it to remove time-frequency points that are dominated by severe noise. In the spectrogram inpainting stage, we use a CNN with partial convolution to perform inpainting on the masked spectrogram from the previous stage. We compared our approach with two powerful baselines, including Wave-U-Net and CRN, on a low SNR dataset containing lots of non-stationary noises. The experimental results show that our approach outperformed the baselines and achieved the state-of-the-art performance. A demonstration can be found at "https://github.com/haoxiangsnr/Masking-and-Inpainting."

***Index Terms*—** speech enhancement, low signal-to-noise ratio, non-stationary noise, spectrogram inpainting

1. INTRODUCTION

Low signal-to-noise ratio (SNR) and non-stationary noise are two typical problems in speech enhancement. Low SNR means that effective speech account for a small proportion in the noisy speech, and even it has been entirely suppressed by noise. Non-stationary noise mean that its distribution is random and difficult to predict in advance, among whom impact noise is an important representative. These two problems often occur at the same time, which brings significant challenges to speech enhancement.

Speech enhancement methods can be divided into traditional methods and deep learning-based methods according to whether deep neural networks were used. Traditional methods are often based on certain ideal assumptions [1], which result in performance degradation at low SNR involving non-stationary noise. For example, based on the assumption that the noise are stationary, wiener filtering [2] degrades significantly when there is a non-stationary noise. In recent years, speech enhancement methods based on deep learning are increasing progressively [3] [4], and have achieved much better performance than traditional methods. Although they do not dependent on any assumptions, they are equally difficult to deal with at low SNR with non-stationary noise.

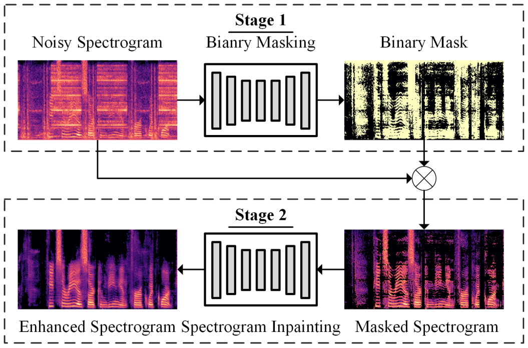


Fig. 1. Illustration of our proposed two-stage approach.

For better speech enhancement at the above scenarios, we propose a two-stage approach consisting of binary masking and spectrogram inpainting, as shown in Fig. 1. The motivations are as follows. When in low SNR condition (below -5dB), the effective speech is almost completely suppressed by the noise. It is difficult to estimate the harmonic structures from the noisy spectrogram. The time-frequency (T-F) points dominated by noise not only provide no useful information for speech enhancement, but also has side effects, which complicate the data distribution of the spectrogram and increase the difficulty of speech enhancement. Therefore, it is necessary to remove the T-F points that are dominated by severe noise in the noisy spectrogram. In the first stage of our approach, we train a binary masking model to achieve this goal.

After the first stage, the T-F points dominated by severe noise are almost completely removed, and the T-F points dominated by effective speech account for the majority of the spectrogram. However, due to the binary masking, there are many "holes" in the noisy spectrogram. Therefore, we design a spectrogram inpainting model to recover the spectrogram in the second stage. This stage is similar to the inpainting task in image processing. The spectrogram inpainting model is to capture the effective context information and complete the missing T-F points. Besides, it can enhance other T-F points that contain weak interference in the spectrogram.

To evaluate our approach, we build a challenging low SNR dataset using TIMIT [5] and 100 Nonspeech Sounds [6], which contains a lot of non-stationary noises. The experimental results show that our approach achieved the state-of-the-art (SOTA) performance compared to the two powerful baselines.

2. METHOD

2.1. Binary masking

The learning objective in the first stage is to determine whether the T-F point is dominated by noise or effective speech information based on a certain threshold, which is similar to ideal binary mask (IBM) [7]:

$$IBM(t,f) = \begin{cases} 1, & SNR(x) \geq T \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In an earlier experiment, we tried to train this ideal binary mask using CNN and LSTM, and we found that the distribution of the final mask was very concentrated. It does not apply to serves as the front-end of the spectrogram inpainting since the spectrogram inpainting needs enough context to better complete the task. In this paper, we use a hardened soft mask, which solves this problem very well. Specifically, we first learn a soft mask. We chose the widely used spectral magnitude mask (SMM) [4] as our learning target:

$$SMM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|} \tag{2}$$

where $|S(t,f)|$ and $|Y(t,f)|$ represent the magnitude spectrogram of clean speech and noisy speech, respectively. As a soft mask, SMM represents the proportion of clean speech in noisy speech. After that, we hardened this soft mask to get binary mask (BM):

$$BM(t,f) = Hard(SMM(t,f)) = \begin{cases} 1, & SMM(t,f) \geq T \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Here, T is a threshold. We use BM to obtain the masked spectrogram.

2.2. Spectrogram inpainting

After the previous stage of processing, we have removed the T-F points that are heavily noise-dominated. In the spectrogram inpainting stage, we compensate for the deleted T-F points based on the context information of the masked spectrogram and further enhance some T-F points that contain weak noise. CNN with the standard convolution has proven to be useful for speech enhancement, but not suitable for the spectrogram inpainting. Specifically, if we retain the bias value in the standard convolution, the resulting spectrogram will be very blurred. On the contrary, if we remove the bias value, the flexibility and fitting ability of the model will be reduced. In this paper, we replace the standard convolution with the partial convolution [8] [9], which solves these problems well. The partial convolution on the feature map covered by the convolution window can be expressed as Eq. 4:

$$\mathrm{x}' = \begin{cases} w^T(X \odot M)\frac{\mathrm{sum}(1)}{\mathrm{sum}(M)} + b, & \mathrm{sum}(M) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where X represents the feature values for the current convolution window, M represents the corresponding binary mask, W represents the weight matrix of the convolution window, and b is the bias value. When the window area covered by the mask M has at least one non-zero value, we perform the convolution to remove the residual weak noise. We do nothing on the window area where the values all are zeros. After each partial convolution, we update the mask according to the value of sum(M) at the corresponding T-F points:

$$m' = \begin{cases} 1, & \mathrm{sum}(M) > 0 \\ x, & \text{otherwise} \end{cases} \tag{5}$$

In conclusion, the model focus only on the existing content and the content derived from it. It is not necessary to use a convolution kernel to learn the mapping between blank space and corresponding clean T-F points, which improves the utilization of convolution kernels.

2.3. Model architecture

The binary masking model and spectrogram inpainting model are both similar to that in work [10], which has achieved impressive results in image-to-image translation [11] and image super-resolution [12]. Each model consists of two downsampling blocks, eight residual blocks, and two upsampling blocks in turn. Except for ReLU in the last upsampling block, the exponential linear unit (ELU) [13] is adopted as the activation function in other blocks. Instance normalization [14] is used after each convolution layer. It is worth noting that the spectrogram inpainting model employs the partial convolution rather than the standard convolution.

3. EXPERIMENT

3.1. Datasets and metrics

To evaluate our proposed model, we built a low SNR dataset containing a large number of non-stationary noises based on TIMIT and 100 Nonspeech Sounds. To build a training dataset and a validation dataset, we randomly selected 950 clean speeches from the TIMIT training dataset, and used these clean speeches to mixed N18, N19, N21, N25, N30, N48, and N98 (from 100 Nonspeech Sounds, N18-N25: impact noise generated by the machine, N30: siren, N48: chirp, N98: door moving) at -7dB, -3dB, 0dB, 3dB, and 7dB. Eventually, It produced 33,250 noisy speeches. We randomly selected 31,500 noisy speeches as the training dataset and the rest as the validation dataset. To build a test dataset, we randomly selected 100 clean speeches from the TIMIT test dataset. In addition to the noises that appeared in the other two datasets, we mixed N11, N12, N13, N22, N23, N28, N88, and N95 (N11-N13: crowd noise, N22-N28: impact noise generated by the machine, N88: click sound, N95: toothbrushing) at -7dB, -5dB, 0dB, 5dB, and 7dB, resulting in 7,500 noisy speeches as the test dataset. It is worth emphasizing that the test dataset contains a large number of noises that have not appeared in the other two datasets, and all datasets contain non-stationary noises (even impact noises). These conditions make speech enhancement very challenging.

We use STOI [15] and PESQ [16] to measure speech intelligibility and quality, respectively.

3.2. Implementation details

The sampling rate of all speeches is 16,000 Hz. We used a 20 ms Hann window, a 10 ms overlap, and a 20 ms FFT size for STFT. We used the magnitude spectrogram as the input of both the binary masking model and the spectrogram inpainting model. It is worth mentioning that we fixed the number of input frames to 160 frames in our models, which requires that the number of sampling points of the selected clean speeches must be greater than 25,440 points. For the magnitude spectrogram with more than 160 frames, we start taking 160 consecutive frames from a random position at the beginning of each epoch as the input of the models.

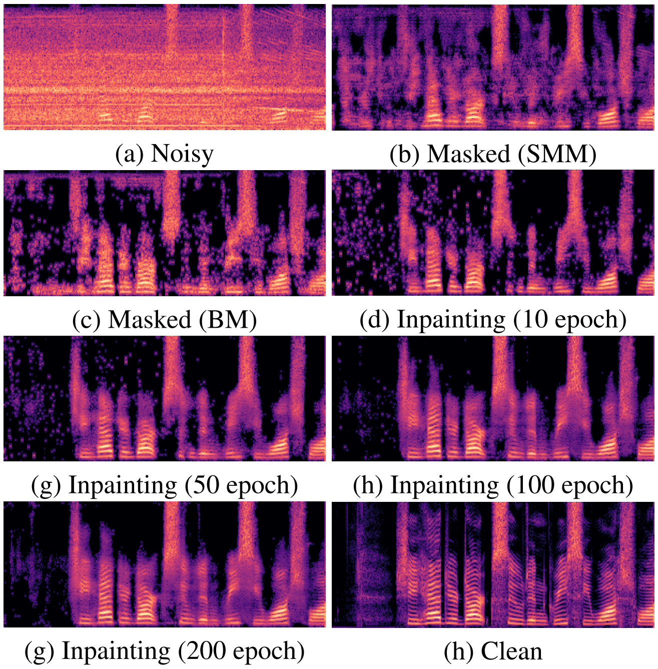
Both models used MSE loss function and Adam optimizer [17] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We set the initial learning rate of the binary masking model to 0.002, and decay by half every 100 epochs. We set the learning rate of the spectrogram inpainting model to a small constant value, which equals to 0.0006. The batch size of both models is 32. We first train the binary masking model until the loss of the validation dataset no longer continues to decrease. Then we fixed the weights of the binary masking model and train the spectrogram inpainting model until convergence. According to the performance of the validation dataset, we set T in Eq. 3 to 0.15.

3.3. Baseline methods

We compared our approach with the two SOTA models. The first model is Wave-U-Net [18], which is an adaptation of the U-Net architecture to the one-dimensional time domain to perform end-to-end audio source separation. We used the implementation on [19] and its default hyperparameters. The second model is CRN [20], which incorporates a convolutional encoder-decoder and LSTM into one framework. We reimplemented it and used the same hyperparameters as the original paper.

3.4. Result

3.4.1. Analysis of intermediate results


(a) Noisy

(b) Masked (SMM)

(c) Masked (BM)

(d) Inpainting (10 epoch)

(g) Inpainting (50 epoch)

(h) Inpainting (100 epoch)

(g) Inpainting (200 epoch)

(h) Clean

Fig. 2. Illustration of the intermediate results using our approach.

In Fig. 2, we show the intermediate results of our two-stage approach. We selected a -5dB noisy speech with N25 (non-stationary machine noise) from the test dataset and showed its magnitude spectrogram in (a). In (a), we can only see a small number of harmonic structures in its low frequencies, which is very challenging for speech enhancement. When we use the SMM to mask the noisy speech, we get (b). We noticed that the quality and intelligibility of noisy speech had been greatly improved, but there are still a lot of noise-dominated T-F points. In fact, SMM can be used as an independent training target in many cases, but it does not work well with low SNR that include non-stationary noise. Next, we perform binarization of SMM, and mask the noisy speech to obtain (c). We noticed that a large number of noise-dominated T-F points are enhanced, which are difficult to handle for traditional models. However, the large number of “holes” in the spectrogram has led to a decline in the quality and intelligibility of speech. (d), (e), (f) and (g) show the processes of spectrogram inpainting. In this process, the spectrogram inpainting model continuously fills in the missing parts of the spectrogram and maps the T-F points that still contain weak noise to clean counterparts. Since we use partial convolution, spectrogram inpainting is only done in a neighborhood with effective T-F points to avoid large blurs. In the end, we obtain the spectrogram are very close to the final clean spectrogram.

3.4.2. *Comparison with the baselines*

We compared our proposed approach with baselines based on STOI and PESQ. The results are shown in Table 1. The "Match" and the "Mismatch" represent whether the noise contained in the test speech has appeared in the training dataset. Compared with the other models, our approach achieved the best performance regardless of whether the noises matched, and CRN ranks second. Compared our approach with CRN in the case of noise matching, STOI is improved by 0.062, and PESQ is improved by 0.120. It is well known that achieve good generalization on the low SNR dataset containing non-stationary mismatch noises is very difficult. However, our approach also improves significantly. Compared with CRN in noise mismatching, STOI is improved by 0.023, and PESQ is improved by 0.110.

Model	STOI		PESQ	
	Match	Mismatch	Match	Mismatch
Unprocessed	0.719	0.737	1.614	1.719
Wave-U-Net	0.796	0.811	2.554	2.252
CRN	0.823	0.816	2.651	2.353
Ours	**0.885**	**0.839**	**2.771**	**2.463**

Table 1. Comparison with other SOTA models.

In Fig. 3, we show the result of a noisy speech (N18 noise at -5dB) enhanced with different models. We plot three columns of rectangular wireframes to illustrate the performance of different models. These three columns of rectangular wireframes show some similar phenomena. Let us take the first column as an example. We observed the following experimental phenomena. **(I)** CRN can remove noise very well. However, it treats the harmonic structures completely masked by noise in the noisy speech directly as noise, causing significant disconnection of the spectrogram. **(II)** Wave-U-Net still has a large number of T-F points dominated by noise in the middle and high frequencies. **(III)** Our approach has the best speech enhancement performance. It can enhance the T-F points that are heavily masked by noise. Even if the harmonic structures of the noisy speech are entirely masked by noise, our approach still recovers some of the harmonic structures reasonably. Besides, it does not disconnect the spectrogram. From the remaining two columns, we can observe the phenomena similar to the aboves.

In summary, our approach reasonably estimates more speech harmonic structures from the T-F points that are completely masked by noise and achieves the best performance compared to the baselines.

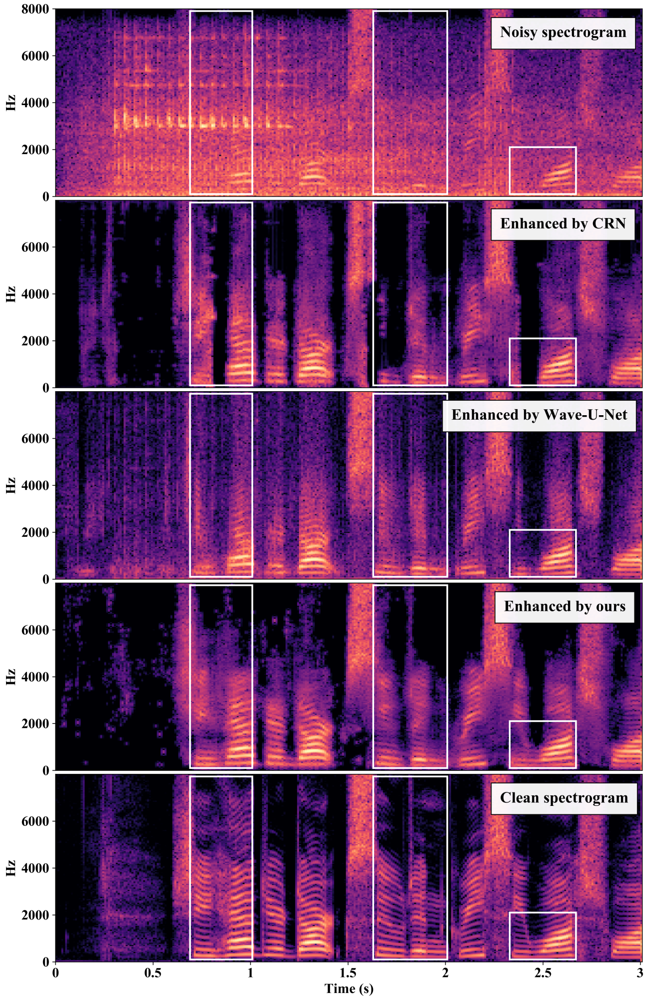


Fig. 3. Demonstrate the result of a noisy speech processed by different models.

4. CONCLUSION

For better speech enhancement in low SNR and non-stationary noise, we propose a two-stage approach that consists of binary masking and spectrogram inpainting. The binary masking stage is used to remove T-F points that are dominated by severe noise, and the spectrogram inpainting stage is used to perform inpainting on the masked spectrogram. We compared our approach with two powerful models on a well-design challenging dataset. The resulting prove that our approach achieves state-of-the-art performance. Besides, we demonstrate the advantages of our approach dealing with these scenarios through two case studies.

5. REFERENCES

[1] Siddala Vihari, A Sreenivasa Murthy, Priyanka Soni, and DC Naik, “Comparison of speech enhancement algorithms,” *Procedia computer science*, vol. 89, pp. 666–676, 2016.

[2] Philipos C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[3] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[4] DeLiang Wang and Jitong Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[5] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, “Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, 1993.

[6] Guoning Hu, “A corpus of nonspeech sounds,” 2010.

[7] DeLiang Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[8] Guilin Liu, Kevin J. Shih, Ting-Chun Wang, Fitsum A. Reda, Karan Sapra, Zhiding Yu, Andrew Tao, and Bryan Catanzaro, “Partial convolution based padding,” in *arXiv preprint arXiv:1811.11718*, 2018.

[9] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *The European Conference on Computer Vision (ECCV)*, 2018.

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[11] Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo, “Contextual-based image inpainting: Infer, match, and translate,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[12] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch, “Enhancenet: Single image super-resolution through automated texture synthesis,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4491–4500.

[13] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” *arXiv preprint arXiv:1511.07289*, 2015.

[14] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” *arXiv preprint arXiv:1607.08022*, 2016.

[15] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings*, pp. 749–752 vol.2.

[17] Diederik Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, 12 2014.

[18] Daniel Stoller, Sebastian Ewert, and Simon Dixon, “Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation,” June 2018, arXiv: 1806.03185.

[19] Daniel Stoller, “Implementation of the Wave-U-Net for audio source separation: f90/Wave-U-Net,” Mar. 2019, original-date: 2018-04-24T16:48:57Z.

[20] Ke Tan and DeLiang Wang, “A convolutional recurrent neural network for real-time speech enhancement.,” in *Interspeech*, 2018, pp. 3229–3233.