An Efficient Dilated Convolutional Neural Network for UAV Noise Reduction at Low Input SNR

Zhi-Wei Tan, Anh H. T. Nguyen, and Andy W. H. Khong
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
E-mail: zhiwei001@e.ntu.edu.sg, nguyenhta@ntu.edu.sg, andykhong@ntu.edu.sg

***Abstract*—Acoustic applications on a multi-rotor unmanned aerial vehicle (UAV) have been hindered by its low input signal-to-noise ratio (SNR). Such low SNR condition poses prominent challenges for beamforming algorithms, statistical methods, and existing mask-based deep learning algorithms. We propose the small model on low SNR (SMoLnet), a compact convolutional neural network (CNN) to suppress UAV noise in noisy speech signals recorded off a microphone array mounted on the UAV. The proposed SMoLnet employs a large analysis window to achieve high spectral resolution since the loud UAV noise exhibits a narrow-band harmonic pattern. In the proposed SMoLnet model, exponentially-increasing dilated convolution layers were adopted to capture the global relationship across the frequency dimension. Furthermore, we performed direct spectral mapping between noisy and clean complex spectrogram to cater to the low SNR scenario. Simulation results show that the proposed SMoLnet outperforms existing dilation-based models in terms of speech quality and objective speech intelligibility metrics for UAV noise reduction. In addition, the proposed SMoLnet requires fewer parameters and achieves lower latency than the compared models.**

I. Introduction

Unmanned aerial vehicle (UAV) has gain popularity given its aerial manoeuvrability and its decreased cost. Light-weight acoustic, radar and image sensors mounted on the UAV extends its applications in many industries. However, acoustic applications on the UAV faces many challenges due to the rotor and environmental noise generated during flight, resulting in significantly low input signal-to-noise ratio (SNR). In particular, for surveillance applications, the sounds-of-interest may be far away from the acoustic sensors, which further reduces the SNR. The problem poses more challenges when intelligibility of a speech signal is required. Many methods for UAV denoising based on classical signal processing techniques such as spectral subtraction [1], adaptive filtering [2], [3], beamforming [4], [5] and blind source separation [6] have not reported nor yield reasonable speech intelligibility or quality performance under such low SNR scenario.

In recent years, supervised deep-learning speech enhancement algorithms have gained significant traction due to the availability of large datasets and higher computing power. These methods assume that the relationship between the noisy and target (clean) signal can be directly learned from observed data. These data-driven approaches, in general, estimate a time-frequency mask [7] or approximate the target spectral components directly [8]. Since they do not rely on explicit statistical assumption, deep learning approaches such as deep auto-encoder (DAE) [9] and deep neural network (DNN) [10] have been shown to outperform classical signal processing approaches in speech enhancement. In our proposed method, we employ the convolutional neural network (CNN) as it has been shown to achieve better speech enhancement performance than DNN [8], [11]–[13] at a lower computation cost [14] . It is useful to note that, for low input SNR below -12 dB, CNN has not been well explored in existing literature.

Exponentially-increasing dilated convolution layers in CNN was first proposed in images [15] and subsequently applied to speech in the time-domain [16]. This structure enables CNN to efficiently increase its number of connectivity to the input space without resolution loss as opposed to pooling [15]. Recently, the gated residual network (GRN) with dilated convolutions [13] and VoiceFilter (VF) [17] have extended such exponentially-increasing dilated CNN structure to the time-frequency domain. However, these models utilise a small short-time Fourier transform (STFT) [18] window, which results in the lack of spectral resolution [19]. We argue that this resolution is crucial to denoise the UAV noise, which consists of strong narrow-band harmonic components [20].

In light of the above, we propose an efficient, compact and fully convolutional network, namely the small model on low SNR (SMoLnet) for UAV noise reduction. As opposed to other dilated denoising models, SMoLnet utilises a large STFT window and employ exponentially-increasing dilated convolution layers to model the relationships between widely-separated high-resolution frequency band. Furthermore, the proposed model employs fewer number of parameters and can achieve lower inference latency while producing higher noise reduction when compared to other existing dilated denoising models. The organisation for the remainder of the paper is as follows: In Section II, we describe the problem formulation, the proposed SMoLnet model, and the training targets. Section III provides the simulation results, and Section IV presents the conclusion.

This work was supported within the STE-NTU Corporate Lab with funding support from ST Engineering and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme (Ref. MRP14) at Nanyang Technological University, Singapore.

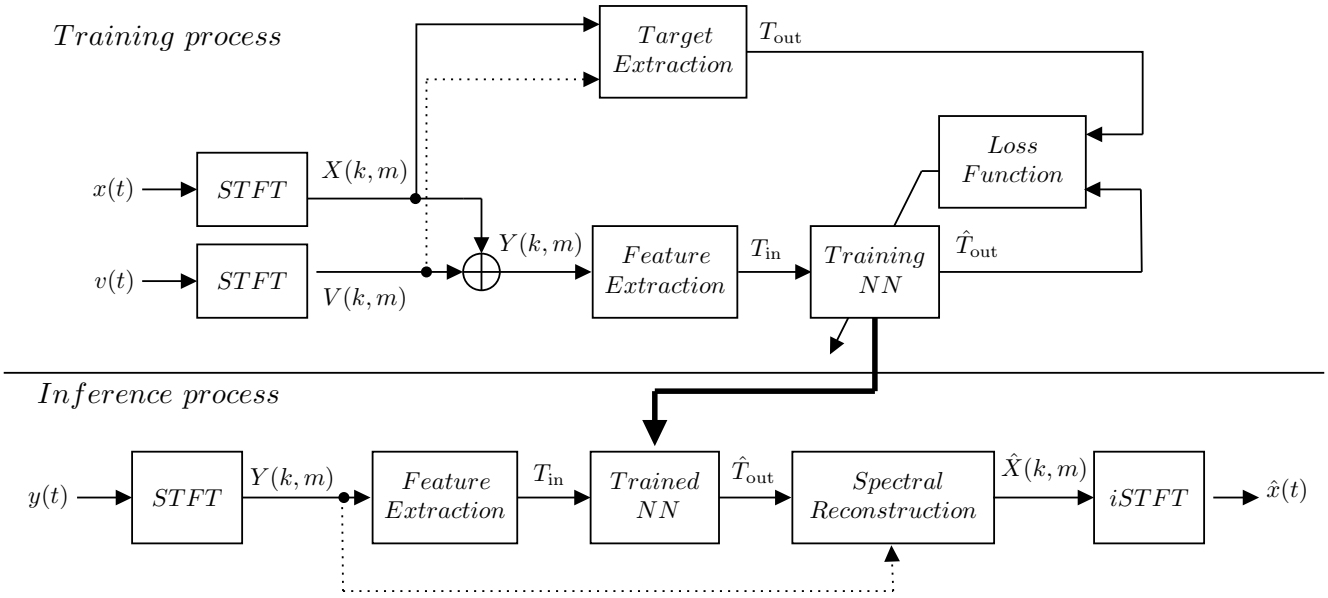

Fig. 1: The block diagram for the training (top) and inference (bottom) process of supervised deep-learning-based speech enhancement. The dotted arrows indicate the optional inputs.

II. UAV NOISE REDUCTION WITH DILATED CONVOLUTIONAL NEURAL NETWORK

A. Problem formulation

We consider the speech enhancement problem for noisy speech signals which are corrupted by UAV rotor noise and background noise. The received (noisy) signal can be expressed as

$$y(t) = x(t) + v_{\text{uav}}(t) + v_{\text{bg}}(t), \tag{1}$$

where t is the time index, $x(t)$ is the desired clean speech, $v_{\text{uav}}(t)$ is the UAV noise, and $v_{\text{bg}}(t)$ is the background noise. Using STFT, (1) can be expressed in the time-frequency domain by

$$\begin{aligned} Y(k,m) &= X(k,m) + V_{\text{uav}}(k,m) + V_{\text{bg}}(k,m) \\ &= X(k,m) + V(k,m), \end{aligned} \tag{2}$$

where k is the frequency index, m is the frame index, $Y(k,m)$ is the noisy speech, $X(k,m)$ is the clean speech, $V_{\text{uav}}(k,m)$ is the UAV noise, $V_{\text{bg}}(k,m)$ is the background noise, and $V(k,m)$ is the total noise.

In deep-learning based speech enhancement, a neural network will be trained to recover $X(k,m)$ from $Y(k,m)$. In this approach, there are three main design factors: the network architecture, input and output features, and loss function. Figure 1 shows the block diagrams of the training and inference process. During training, the input feature T_{in} is extracted from $Y(k,m)$ and being fed into a neural network that approximates the training target T_{out} with an estimate $\widehat{T}_{\text{out}}$. Here, the training target T_{out} is typically a spectral component of $X(k,m)$ or a time-frequency mask that is subsequently employed to extract $X(k,m)$. The variables T_{out} and $\widehat{T}_{\text{out}}$ are then used to compute the loss such that its gradient can be utilised to optimize the neural network [21] to achieve better prediction. During the inference process, the output of the trained neural network is used to reconstruct the denoised spectrogram $\widehat{X}(k,m)$. Finally, $\widehat{X}(k,m)$ is converted back to the time domain using inverse STFT (iSTFT) to obtained the denoised signal $\widehat{x}(t)$.

Although the proposed approach assumes knowledge of the clean signal $x(t)$ during training, no such assumption has been made during inference. This may result in a modest reduction in enhancement performance due to generalization error [22]. However, the deep learning methods for speech enhancement have been empirically shown to generalized well in practice [8], [12], [13], [23].

B. Proposed model for UAV noise reduction

The UAV noise $V_{\text{uav}}(k,m)$ typically consists of narrow-band components caused by the rotation of the rotors and the broadband components caused by air resistance to the rotor blades [20]. Figure 2 depicts how the spectrogram varies with various window length for a noisy speech signal recorded off a UAV operating at approximately 18 m away from a male subject. It can be seen that the narrow-band components of the UAV noise is harmonic and more detrimental to speech signal than the broadband components. More importantly, it is beneficial to employ a long STFT analysis window to achieve high frequency resolution [19]. In our pilot study, it was found that a window length of 2048 or higher is more suitable for the suppression of UAV noise than shorter ones.

In the proposed SMoLnet, we divide the noisy input $y(t)$ sampled at 16 kHz into equal segments of 0.64 seconds (corresponding to 10,240 samples per segment). For a STFT window of 2048 samples, the input and output dimensions have 1025 frequencies and 9 frames. To model the dependencies among

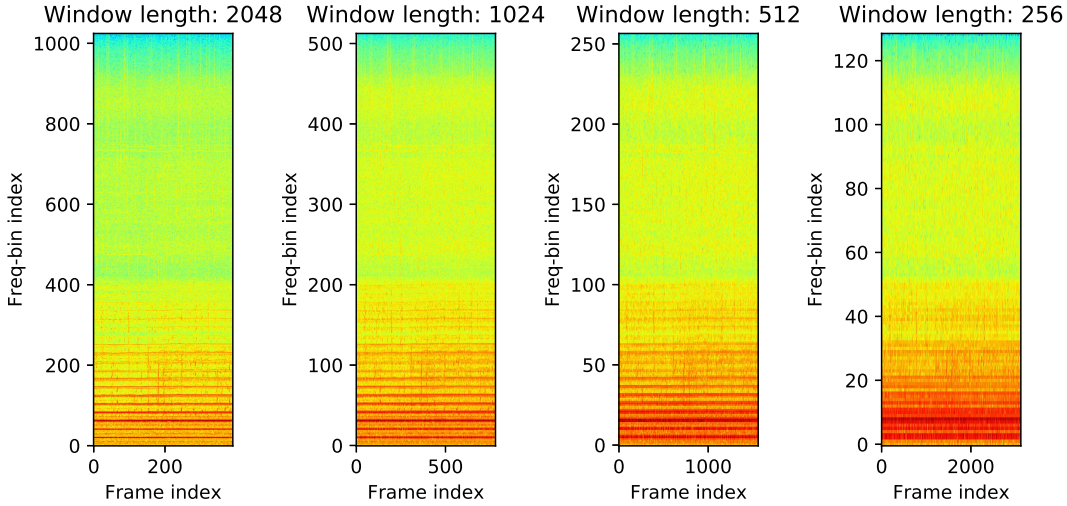

Fig. 2: Spectrograms with respect to different STFT window length for twenty seconds of noisy speech recorded off a male subject who was 18 m away from the operating UAV.

TABLE I: The proposed fully convolutional SMoLnet.

No	Filter Size	No. Filters	Dilation	Dim	Output dim
1	(3, 1)	64	(1,1)	Freq.	(1025,9,64)
2	(3, 1)	64	(2,1)	Freq.	(1025,9,64)
3	(3, 1)	64	(4,1)	Freq.	(1025,9,64)
4	(3, 1)	64	(8,1)	Freq.	(1025,9,64)
5	(3, 1)	64	(16,1)	Freq.	(1025,9,64)
6	(3, 1)	64	(32,1)	Freq.	(1025,9,64)
7	(3, 1)	64	(64,1)	Freq.	(1025,9,64)
8	(3, 1)	64	(128,1)	Freq.	(1025,9,64)
9	(3, 1)	64	(256,1)	Freq.	(1025,9,64)
10	(3, 1)	64	(512,1)	Freq.	(1025,9,64)
11	(3, 3)	64	(1,1)	Freq.-Time	(1025,9,64)
12	(3, 3)	64	(1,1)	Freq.-Time	(1025,9,64)
13	(3, 3)	64	(1,1)	Freq.-Time	(1025,9,64)
14	Output Layer				

time-frequency bins of such dimensions, we proposed a CNN as shown in Table I, where the output layer depends on the training target selected. Each convolutional layer performs the following computation [15], [16]

$$O(k,m,c) = \sum_{i,j,l} W_c(i,j,l) I(k + id_{\mathrm{k}}, m + jd_{\mathrm{m}}, l) + B_c, \quad (3)$$

followed by batch normalization [24] and rectified linear units (ReLU) [25]. In (3), O is the output of the dilated convolution, I is the input, W_c is the cth filter, B_c is the bias of the cth filter, d_{k} is the dilation factor along frequency dimension, and d_{m} is the dilation factor along time dimension. The variable i, j, l are the summation indices along the frequency, time and channel dimension, respectively.

The proposed SMoLnet has fourteen layers. The first ten layers have d_{k} doubled every layer so that the receptive field of the later layers can cover the entire frequency dimension. On the other hand, Layers 11, 12, and 13 model the relationship among time-frequency bins jointly. These layers do not undergo any dilation since the number of frames are small when using a long STFT window. Figure 3 illustrates the receptive fields of two CNNs with filter size of 3, one with dilation of 1 and another with dilation factor that doubles every layer. Generally, the receptive field R for L convolution layers with filter size of s grows linearly with L and is given by

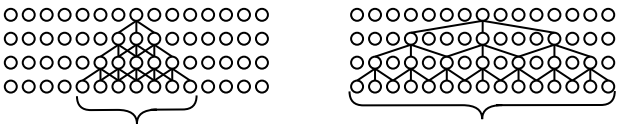

Fig. 3: Receptive field of a CNN with filters size 3 with dilation of 1 (left) and with dilation that doubles every layers (right).

$$R = L(s-1) + 1. \quad (4)$$

On the other hand, the receptive field for the CNN with dilation factor that doubles every layer grows exponentially with L [15], [16] is given by

$$R = 2^L(s-1) - s + 2. \quad (5)$$

In our case, it can be seen that ten dilated layers are sufficient to cover 1025 frequency bins.

The training targets, their corresponding output layers, and the cost function are described in the following:

1) Training targets and spectral reconstruction: Three training targets are considered in this work. They are target magnitude spectrum (TMS) [26], target complex spectrum (TCS) [8], and complex ideal ratio mask (cIRM) [7]. Other target masks are not considered since they generally perform worse than cIRM [7]. It is also worth noting that masking

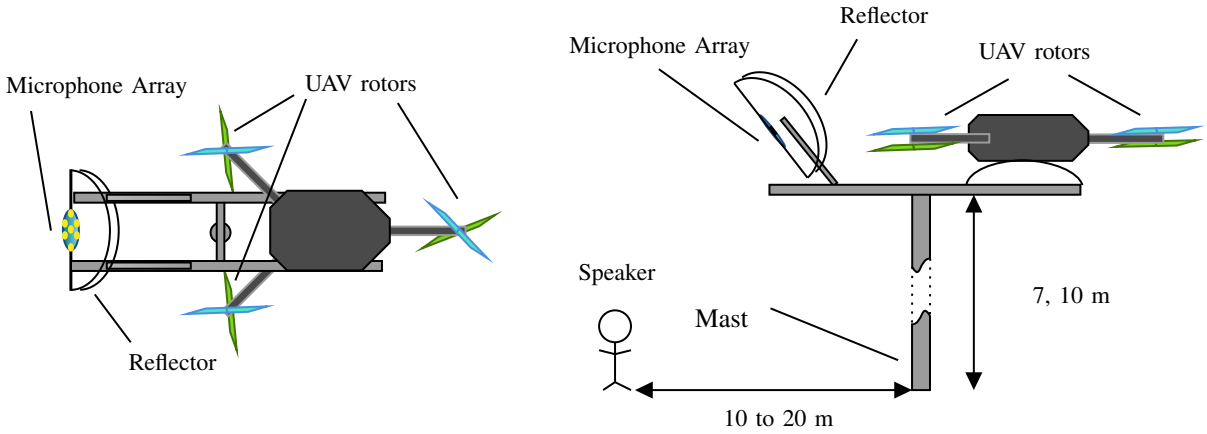

Fig. 4: The top-down (left) and front-view (right) of the UAV configuration.

Fig. 5: The UAV configuration raised on a mast at a height of approximately 10 m in an open space with traffic noise and ambient noise.

based methods are not suitable for significantly low SNR because the signal-of-interest is totally dominated by noise.

For the TMS, the model estimates the target magnitude spectrogram from the noisy one

$$T_{\text{in}}^{\text{TMS}} = |Y(k,m)|,\ T_{\text{out}}^{\text{TMS}} = |X(k,m)|, \tag{6}$$

where $|.|$ is the absolute operand. Similar to [13], [17] denoised signal is reconstructed by combining the denoised magnitude with the noisy phase

$$\widehat{X}^{\text{TMS}}(k,m) = \widehat{T}_{\text{out}}^{\text{TMS}} \exp\left(\jmath \angle Y(k,m)\right), \tag{7}$$

where $\angle$ is the phase operand and $\jmath = \sqrt{-1}$.

For TCS, the model estimates the target complex spectrum from the noisy complex spectrum. Since the input and output are complex, they are converted to real-imaginary composites

$$T_{\text{in}}^{\text{TCS}} = \begin{bmatrix} Y_{\Re}(k,m) \\ Y_{\Im}(k,m) \end{bmatrix},\ T_{\text{out}}^{\text{TCS}} = \begin{bmatrix} X_{\Re}(k,m) \\ X_{\Im}(k,m) \end{bmatrix}, \tag{8}$$

for the real-valued neural network. Here the real and imaginary components are stacked along the filter dimension. The denoised spectrogram is then reconstructed as follows

$$\widehat{X}^{\text{TCS}}(k,m) = \widehat{X}_{\Re}^{\text{TCS}}(k,m) + \jmath \widehat{X}_{\Im}^{\text{TCS}}(k,m). \tag{9}$$

TABLE II: The output layer of the proposed fully convolutional SMoLnet model.

Targets	No. Filters	Dim	Output dim	Activation Function
TMS	1	Filter	(1025,9,1)	Softplus
TCS	2	Filter	(1025,9,2)	Linear
cIRM	2	Filter	(1025,9,2)	Linear

For cIRM, the model predicts the ratio of desired signal to the noisy signal in the complex domain. The input feature and target are given by

$$T_{\text{in}}^{\text{cIRM}} = T_{\text{in}}^{\text{TCS}},\ T_{\text{out}}^{\text{cIRM}} = \begin{bmatrix} K\tanh(cM_{\Re}(k,m)) \\ K\tanh(cM_{\Im}(k,m)) \end{bmatrix}, \tag{10}$$

where the default variables $K = 10$, $c = 0.1$, while $M_{\Re}(k,m)$ and $M_{\Im}(k,m)$ are the real and imaginary components of the complex ratio, respectively. Here, the target is the compressed version of $\widehat{M}_{\Re}(k,m)$ and $\widehat{M}_{\Im}$ because the uncompressed masks have a range of $(-\infty,\infty)$ and compression gives a numerically stable range of $[-K,K]$. The denoised signal is finally reconstructed by element-wise multiplication

$$\widehat{X}^{\text{cIRM}}(k,m) = \left(\widehat{M}_{\Re}(k,m) + \jmath\widehat{M}_{\Im}(k,m)\right) Y(k,m). \tag{11}$$

2) Output layer: The output layer of the propose model consists of one 1×1 convolution layer with no dilation. This is to weigh the frequency-time spectrum generated by the filter of the previous layer to form the output. For different training targets, the output layer of the proposed model uses different number of filters and activation as tabulated in Table II. Here, the softplus function [27] is employed to ensure that the predicted TMS is always positive.

3) Loss function: We use the mean squared error (MSE) between $\widehat{T}_{\text{out}}$ and T_{out} as our loss function. Due to the low SNR, the model is trained with this loss in double precision. The model is also trainable with root-mean-squared error (RMSE) with lower performance.

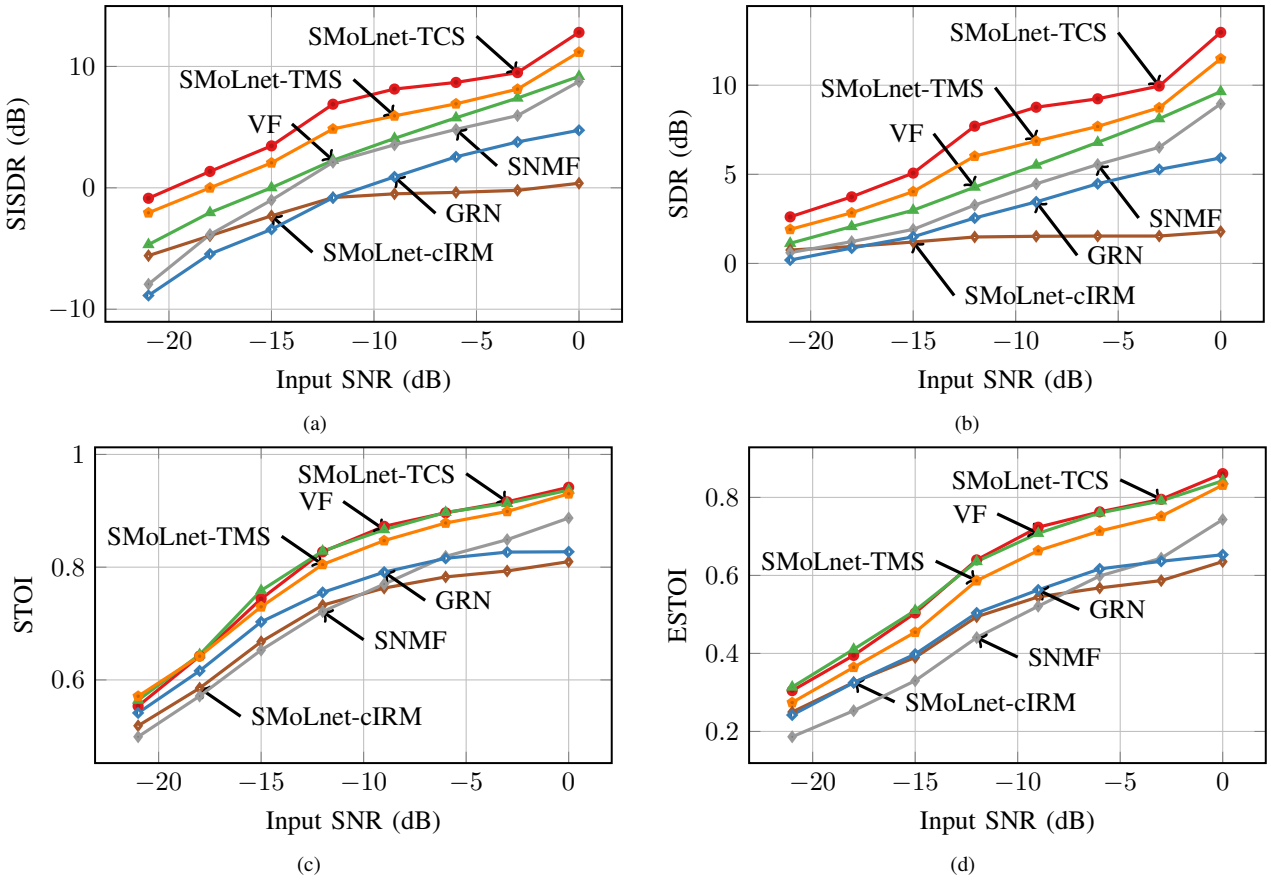

Fig. 6: Noise reduction performance on WSJ0 test-set in term of: a) SISDR (dB), b) SDR (dB), c) STOI, and d) ESTOI.

C. Experiment setup

We used simulated noisy speech generated by adding clean speech and recorded noise. We recorded the noise which includes the UAV[1] noise, traffic noise and ambient noise using a 7-channel microphone array[2] mounted inside a 13.5-inch parabolic reflector[3] such that the center microphone to the array is at the focal point of the reflector. During the experiment, we note that the UAV noise is dominant. Figure 4 shows the experiment setup where the reflector is about 60 cm away from the center of UAV. The reflector is used to improve input SNR by acting as a sound barrier for the UAV noise and as a physical signal booster for the target speech. This setup allows us to pick up speech signal up to 20 m away from the UAV. For recording, as shown in Fig. 5, the prototype was raised on a mast to an approximate height of 7 and 10 m in an open space. Seventy-eight minutes of noise are recorded. The noise recordings were split to five minutes of validation data, and five minutes of testing data. The remaining were used for training. To obtain more training data, we concatenated the noise along each channel to lengthen each data set by 7-folds as each spatially-separated microphone receives varied noise signature.

III. Simulation Results

For the clean speech, we used speech utterances from the WSJ0-SI84 [28] dataset with eighty-three speakers. Utterances with speaker IDs of 20d, 20e, 20f, and 20g were used for validation, while utterances with speaker IDs of 207, 208, 209, and 20a were used for testing. The remaining speakers in WSJ0-SI84 were used for training. To further evaluate the feasibility of the approach in the presence of unseen data during inference, an additional test set was constructed using the TIMIT dataset [29] which has not used for training nor validation.

The noise and clean speech data were aligned by removing excess data to form approximately 476 minutes of training, 20 minutes of validation and WSJ0 test-set and 9 minutes of TIMIT test-set. They were then transformed into the time-frequency domain using STFT with a sine window and 50% overlap. During training, we scaled the clean speech in each batch of data to a random SNR from -10 dB to -20 dB. This corresponds to the range of SNRs with a subject talking at approximately 10 to 15 m away from the base of the UAV mast. In validating and testing, the clean speech signals were scaled to SNR levels of {-21, -18, -15, -12, -9, -6, -3, 0} dB.

[1]USTAR-Y

[2]https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array

[3]https://kloverproducts.com/shop/klover-mik/klover-mik-16/klover-mik-16-hard-mount/

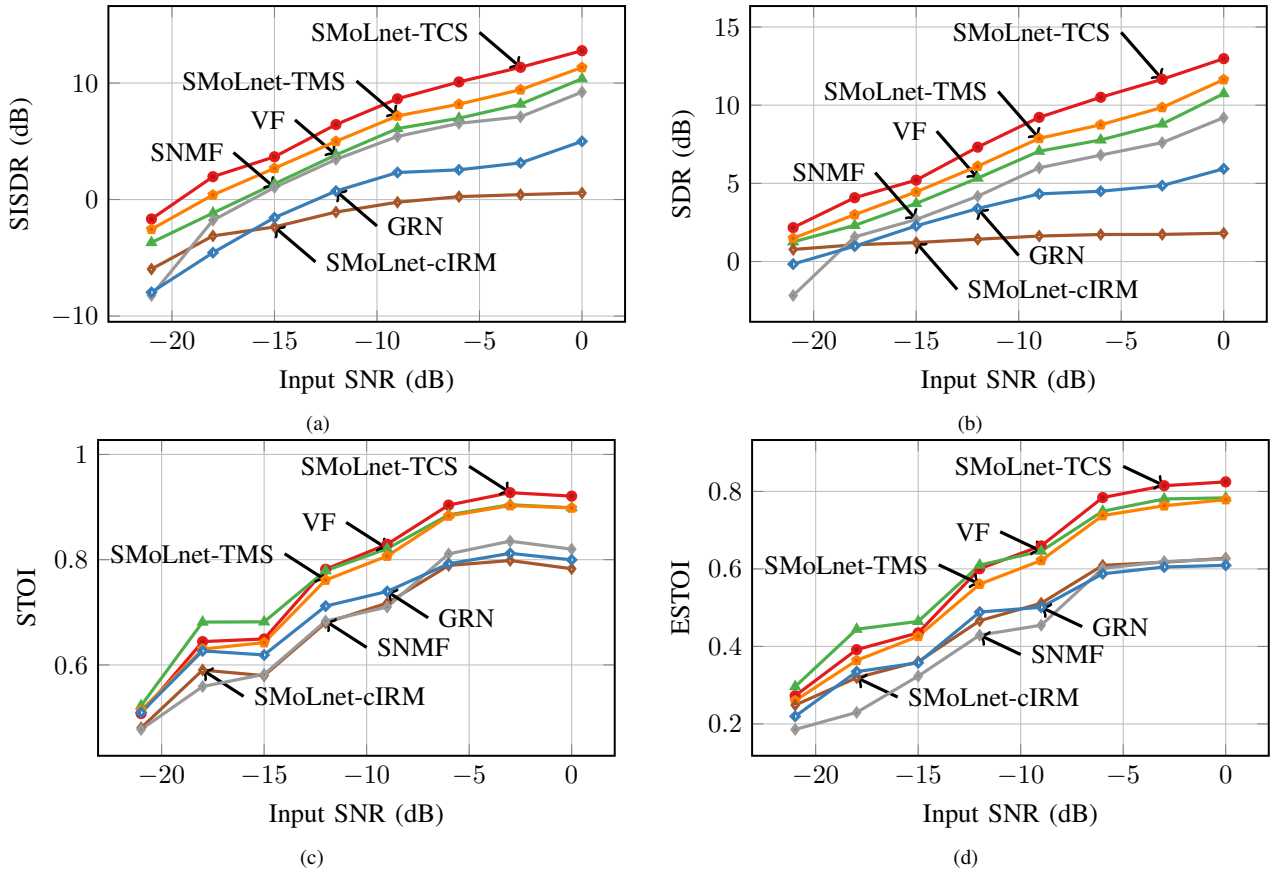

Fig. 7: Noise reduction performance on TIMIT test-set in term of: a) SISDR (dB), b) SDR (dB), c) STOI, and d) ESTOI.

These scalings were necessary to evaluate whether a trained model can generalizes to an input SNR that it was not trained with. After scaling, the speech and noise were synthetically added to form the noisy mixture.

A. *Performance evaluation*

For UAV noise reduction, we compared our proposed SMoLnet with a classical approach, namely, sparse non-negative matrix factorization (SNMF) [30] and two recent deep learning models based on dilated convolution: the GRN [13] and VoiceFilter (VF) [17]. The proposed SMoLnet is trained using TCS, TMS and cIRM features as described in Section II-B1. These variants are denoted as SMoLnet-TCS, SMoLnet-TMS and SMoLnet-cIRM. For performance evaluation, we have employed the signal-distortion-ration (SDR) [31], scale-invariant SDR [32], short-time objective intelligibility (STOI) [33], and extended STOI (ESTOI) [34]. We have also employed the latency and number of parameters to compare the computational requirements.

As opposed to the SMoLnet which employs an STFT window of 2048, GRN and VF employs shorter window length of 320 and 512 samples, respectively, resulting in a large number of frames. As a result, GRN and VF place a higher priority on modelling temporal relation using dilated layers along the time dimension. In addition, for VF, it employs a bi-directional long-short term memory (bi-LSTM). Since different models use different window lengths, for a fair comparison, we split the spectrogram to segments that have the same duration of 0.64 s in the time domain. More specifically, SMoLnet, VF and GRN utilised 9, 39 and 63 time frames respectively.

TABLE III: Average performance over input SNR on WSJ0 test-set.

Model-Target	SDR(dB)	SISDR(dB)	STOI (%)	ESTOI (%)
SMoLnet-TCS	**7.5**	**6.24**	79.9	**62.3**
SMoLnet-TMS	6.2	4.62	78.8	58.0
VF	5.07	2.75	**80.1**	62.1
SNMF	4.06	1.55	72.1	46.5
GRN	3.03	-0.823	73.5	49.2
Noisy Input	-10.5	-10.5	57.2	29.1

TABLE IV: Average performance over input SNR on TIMIT test-set.

Model-Target	SDR(dB)	SISDR(dB)	STOI (%)	ESTOI (%)
SMoLnet-TCS	**7.89**	**6.66**	77.1	**59.8**
SMoLnet-TMS	6.64	5.21	75.5	56.4
VF	5.87	3.99	**77.2**	59.7
SNMF	4.48	2.85	68.5	43.4
GRN	3.26	-0.0409	70.1	46.3
Noisy Input	-10.5	-10.5	60.9	29.2

TABLE V: Latency and parameters of various models.

Model	No. of parameters	Inference Latency(ms)
GRN	2.49 M	26.7
VFCNN	9.04 M	52
SMoLnet	224 k	19.4
SNMF	46 M	-

SMoLnet, VF and GRN were trained with Adam [35] with the batch size fixed at 16 and the learning rate decreased by a factor of 0.5 after three epochs of loss plateau. These models were trained using three learning rates (0.001, 0.0005, 0.0001) and the best performing model on the validation set in terms of average ESTOI was selected for evaluation.

The SNMF was trained and tested using the most suitable configuration discussed in [30]. Specifically, we used KL divergence and sparsity factor of 5 to train randomly initialized basis matrix. During testing, randomly initialized activation matrix are then optimized with the trained basis matrix on the noisy mixture. For fairer comparison with SMoLNet model, we employed an STFT window length of 2048 samples with 9 time frames and utilized a larger basis size of 2500. Due to memory constraint, we only used 20% of randomly selected training samples for training.

Figures 6 and 7 show the denoising performance with different input SNR on WSJ0 test-set and TIMIT test-set, respectively. Correspondingly, the average performance on WSJ0 test-set and TIMIT test-set are shown in Table III and IV. Table V shows the computational metrics. Overall, SMoLnet-TCS achieves the best performance in terms of SDR, SISDR for both test sets and best computational efficiency.

SMoLnet-TCS consistently outperforms SNMF, GRN and VF on SDR and SISDR while maintaining similar STOI and ESTOI compared with VF. More specifically, SMoLnet-TCS achieves higher performance than VF by 2.43 dB in SDR, 3.49 dB in SISDR for the WSJ0 test-set and 2.02 dB in SDR, 2.67 dB in SISDR for the TIMIT test-set while having an STOI and ESTOI difference of $\pm 0.2\%$ when compared with VF on all datasets. The VF model has yields modestly higher STOI and ESTOI performance at low SNR for the TIMIT test-set than SMoLnet-TCS. However, we argue that STOI and ESTOI are ineffective metrics at very low SNR since they do not take phase information into account in their computation; the phase exhibits a significant difference at lower SNR [8] and has shown to be crucial in speech intelligibility studies [36].

When comparing different training targets, SMoLnet-TCS achieves the best performance on all metrics, followed by SMoLnet-TMS, and subsequently SMoLnet-cIRM. The difference in performance between SMoLnet-TCS and SMoLnet-TMS highlights the importance of phase information, which is ignored by TMS, at low SNR. At the same time, SMoLnet-cIRM achieves the worst performance among the three training targets which may be attributed by the fact that this target is less suitable for low SNR.

Table V shows the average interference latency on a GTX1080 Ti graphics processing unit (GPU) to denoise a 0.64 s segment. As SNMF is computed using the central processing unit (CPU), its latency is omitted to avoid an unfair comparison with the other models which employ GPU. In terms of computational complexity, the SMoLnet is approximately 1.38 times and 2.68 times faster than the VF and GRN, respectively, while using one magnitude fewer parameters than both. SMoLnet also uses two magnitude lesser parameters than SNMF.

IV. Conclusions

We proposed an efficient, compact and fully convolutional SMoLnet for UAV noise reduction. The design of stacked dilation convolution along the frequency domain allows higher frequency resolution. The weighted filter on the last layer lowers the computational cost of the network. The target complex spectrum mapping has shown to be effective under low SNR, which is consistent with [8]. Simulation results show that the SMoLnet outperforms a classical approach while requiring two magnitude fewer number of parameters and two existing deep learning models with fully connected or LSTM modules while requiring one magnitude fewer number of parameters and lower latency.

References

[1] A. M. Prodeus, "Performance measures of noise reduction algorithms in voice control channels of uavs," in *Proc. IEEE Int. Conf. Actual Probl. Unmanned Aer. Veh. Dev.*, 2015, pp. 189–192.

[2] S. Yoon, S. Park, E. Youmie, and Y. Sunggeun, "Advanced sound capturing method with adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2015, pp. 26–29.

[3] S. Yoon, S. Park, and S. Yoo, "Two-stage adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2016, pp. 219–222.

[4] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *Proc. IEEE Int. Work. Acoust. Signal Enhanc.*, 2016, pp. 1–5.

[5] B. Yen, Y. Hioka, and B. Mace, "Improving power spectral density estimation of unmanned aerial vehicle rotor noise by learning from non-acoustic information," in *Proc. Int. Work. Acoust. Signal Enhanc.*, 2018, pp. 545–549.

[6] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sens. J.*, vol. 17, no. 8, pp. 2447–2455, Apr. 2017.

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[8] S. W. W. Fu, T. Y. Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE Int. Work. Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[9] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2013.

[10] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," *International Conference on Acoustics, Speech and Signal Processing*, pp. 1562–1566, 2014.

[11] L. Hui, M. Cai, C. Guo, L. He, W. Zhang, and J. Liu, "Convolutional maxout neural networks for speech separation," in *Proc. International Symposium on Signal Processing and Information Technology*, Dec 2015, pp. 24–27.

[12] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2016.

[13] K. Tan, J. Chen, and D. L. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 1, pp. 189–198, Jan. 2019.

[14] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2017.
[15] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv Prepr. arXiv1511.07122*, 2015.
[16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv Prepr. arXiv1609.03499*, pp. 1–15, 2016.
[17] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv Prepr. arXiv1810.04826v4*, 2018.
[18] S. H. Nawab and T. F. Quatieri, "Short-time fourier transform," in *Adv. Top. Signal Process.* Upper Saddle River, New Jersey: Prentice-Hall, 1987, pp. 289–337.
[19] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, Jan 1978.
[20] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2016, pp. 152–158.
[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[22] D. Jakubovitz, R. Giryes, and M. R. D. Rodrigues, "Generalization error in deep learning," *arXiv Prepr. arXiv1810.04826v4*, vol. abs/1808.01174, 2018.
[23] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan 2014.
[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. International Conference on Machine Learning*, 2010, pp. 807–814.
[26] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2013, pp. 436–440.
[27] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," in *Proc. Int. Conf. Artif. Intell. Stat.*, vol. 15, 2011, pp. 315–323.
[28] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguist. Data Consortium, Philadelphia*, 2007.
[29] J. S Garofolo, L. Lamel, W. M Fisher, J. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.
[30] J. L. Roux, F. J. Weninger, and J. R. Hershey, "Sparse NMF half-baked or well done?" Mitsubishi Electric Research Laboratories (MERL), Tech. Rep., 2015.
[31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[32] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR- half-baked or well done?" *arXiv Prepr. arXiv.1811.02508v1*, 2018.
[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[34] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.
[36] K. Paliwal, K. Wjcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465 – 494, 2011.