Deep Learning Assisted Time-Frequency Processing for Speech Enhancement on Drones

Lin Wang and Andrea Cavallaro

***Abstract*—This article fills the gap between the growing interest in signal processing based on Deep Neural Networks (DNN) and the new application of enhancing speech captured by microphones on a drone. In this context, the quality of the target sound is degraded significantly by the strong ego-noise from the rotating motors and propellers. We present the first work that integrates single-channel and multi-channel DNN-based approaches for speech enhancement on drones. We employ a DNN to estimate the ideal ratio masks at individual time-frequency bins, which are subsequently used to design three potential speech enhancement systems, namely single-channel ego-noise reduction (DNN-S), multi-channel beamforming (DNN-BF), and multi-channel time-frequency spatial filtering (DNN-TF). The main novelty lies in the proposed DNN-TF algorithm, which infers the noise-dominance probabilities at individual time-frequency bins from the DNN-estimated soft masks, and then incorporates them into a time-frequency spatial filtering framework for ego-noise reduction. By jointly exploiting the direction of arrival of the target sound, the time-frequency sparsity of the acoustic signals (speech and ego-noise) and the time-frequency noise-dominance probability, DNN-TF can suppress the ego-noise effectively in scenarios with very low signal-to-noise ratios (e.g. SNR lower than −15 dB), especially when the direction of the target sound is close to that of a source of the ego-noise. Experiments with real and simulated data show the advantage of DNN-TF over competing methods, including DNN-S, DNN-BF and the state-of-the-art time-frequency spatial filtering.**

***Index Terms*—Deep neural network (DNN), drone, ego-noise reduction, microphone array.**

I. Introduction

THE processing of audio signals captured by small drones has attracted increasing interests in recent years for applications such as search and rescue, recreational filming, and human-robot interaction [1]–[6]. A microphone array mounted on a drone can be used for acoustic sensing tasks, such as recording and localizing sound sources [7], [8]. However, since the microphones are mounted close to the rotors and propellers, which produce strong ego-noise [9], [10], the signals have typically very low signal-to-noise ratios (e.g. SNR lower than −15 dB). While a few signal processing algorithms have been proposed for sound enhancement [10]–[17] and source localization [18]–[22], their performance is still unsatisfactory because of ego-noise nonstationarities [23].

Deep learning has revolutionized sound and speech enhancement [24]–[27]. When a sufficient amount of training data is available, deep neural networks (DNN) can learn to predict clean speech signals from a noisy recording. DNN-based approaches can be separated into single-channel and multi-channel techniques. *Single-channel approaches*, which include fully-connected [28], [29], convolutional [30], [31] and recurrent neural networks [32]–[34], learn the mapping between noisy signals and their corresponding clean signal [28], [35] or estimate the time-frequency masks of the clean signal [36]–[39]. While traditional unsupervised single-channel noise reduction approaches are still favorable in many applications due to their robustness and low computational complexity, the noise reduction performance of supervised DNN approaches is generally better, especially with nonstationary noise and when the noise is represented in the distribution of the training set data [8]. *Multi-channel approaches* typically use time-frequency masks estimated by the DNN model to construct a spatial filter to enhance the target sound [40]–[45]. Extensions of this idea [46], [47] estimate the coefficients of the filter directly from the multi-channel data, which however require a large amount of training data simulated in a variety of scenarios. While significant progress has been made in this domain, the application of DNN-based speech enhancement on drone platforms has not been reported yet.

In this paper, we present the first work that applies single-channel and multi-channel DNN-based approaches for speech enhancement on drones. Our contribution is twofold. First, we employ a DNN to estimate the ideal ratio mask at individual time-frequency bins from the noisy signal, and implement two baseline speech enhancement approaches: single-channel ego-noise reduction (DNN-S) and multi-channel beamforming speech enhancement (DNN-BF). Experimental results demonstrate that while both baseline approaches perform well for non-stationary noise, the performance drops remarkably in low-SNR scenarios. Second, we propose a method that combines the DNN-estimated ideal ratio mask with a time-frequency spatial filtering approach [10]. By jointly exploiting the time-frequency sparsity of the acoustic signals (speech and ego-noise) and the noise-presence probability inferred from the DNN-estimated mask, the proposed method outperforms state-of-the-art approaches, especially when the direction of the target sound is close to that of a source of ego-noise.

Manuscript received March 31, 2020; revised July 3, 2020; accepted July 27, 2020. Date of publication August 24, 2020; date of current version November 23, 2021. This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K007491/1, and by the ARTEMIS-JU and the UK Technology Strategy Board (Innovate UK) through the COPCAMS Project under Grant 332913. Recommended for publication by Associate Editor Dr. Stefano Squartini.

The authors are with the Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, U.K. (e-mail: lin.wang@qmul.ac.uk; a.cavallaro@qmul.ac.uk).

Digital Object Identifier 10.1109/TETCI.2020.3014934

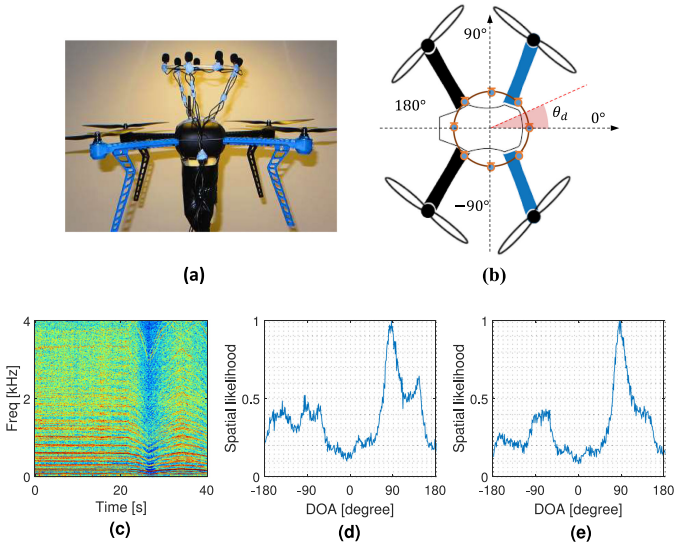


Fig. 1. (a) The drone with the microphone array and (b) the corresponding 2D coordinate system. (c) Spectrogram of the ego-noise. (d)–(e) Spatial likelihood of the ego-noise for 0–20s and 20-40s, respectively.

The paper is organized as follows. After defining the problem in Section II, Section III introduces the DNN model for the estimation of the time-frequency mask. In Section IV we propose the single-channel and multi-channel DNN algorithms. Section V covers the experimental results and analysis, and Section VI discusses the advantages and disadvantages of the proposed algorithms. Finally, conclusions are drawn in Section VII.

II. Problem Definition

Let a target sound source in the far field emit sound with direction of arrival (DOA) θ_d. Let $\boldsymbol{r}_m = [r_{mx}, r_{my}]^{\mathrm{T}}$ be the position of the m-th microphone (the superscript $(\cdot)^{\mathrm{T}}$ denotes the transpose operator). Let $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_I]$ represent the locations of the I microphones in the array. The microphone signal, $\boldsymbol{x}(n) = [x_1(n), \ldots, x_I(n)]^{\mathrm{T}}$, contains the target sound, $\boldsymbol{s}(n) = [s_1(n), \ldots, s_I(n)]^{\mathrm{T}}$, as well as the ego-noise, $\boldsymbol{v}(n) = [v_1(n), \ldots, v_I(n)]^{\mathrm{T}}$, i.e.

$$\boldsymbol{x}(n) = \boldsymbol{s}(n) + \boldsymbol{v}(n), \tag{1}$$

or, after applying a short-time Fourier transform (STFT):

$$\boldsymbol{X}(k,l) = \boldsymbol{S}(k,l) + \boldsymbol{V}(k,l), \tag{2}$$

where k and l denote the frequency and frame indices, respectively. Let K and L be the total number of frequency bins and time frames, respectively.

We assume that $\boldsymbol{R}$ and θ_d are known. Given $\boldsymbol{x}(n)$, the goal is to extract the target sound from the noisy recording. The main challenge is to deal with the very low SNR at the microphones (that can be lower than -15 dB), which most speech enhancement algorithms cannot cope with.

Fig. 1(a)–(b) show the 8-microphone circular array ($I = 8$), mounted on the top of a 3D IRIS quadcopter [9], and the coordinate system. The diameter of the array is 0.2 m. Fig. 1(c) shows the time-frequency spectrum of a segment of ego-noise recorded with one of the microphones. The first half of this segment contains stationary noise generated by rotors and propellers operating at a constant speed. The second half of the segment contains nonstationary noise generated by rotors and propellers operating with a time-varying speed. The ego-noise mainly consists of multiple narrowband harmonic components, which are caused by the rotating rotors, and broadband noise, which is caused by the propellers cutting air. The nonstationarity of the ego-noise is caused by variations of the fundamental frequency of the harmonic component with the changes of the rotation speed of the motors. Fig. 1(d)–(e) depict the spatial likelihood functions for two ego-noise segments 0–20s and 20-40s, respectively. The spatial likelihood is computed by building a histogram of the local DOA estimates at individual time-frequency bins (cf. Eq. (19)) and then normalizing it with the highest frequency count [8]. It can be observed that the shape of the spatial likelihood function varies for the two segments. The DOA of the ego-noise is spread widely, but has roughly four peaks, corresponding to the four rotors/propellers.

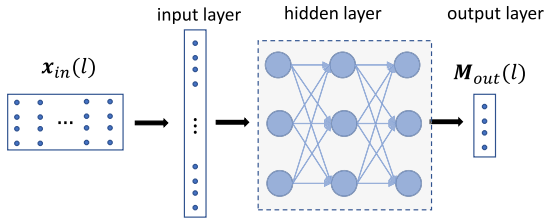


Fig. 2. The DNN architecture for estimating the ideal ratio mask. The input layer corresponds to $(2\Delta_L + 1)$ frames of log spectrogram flattened into a vector. The output layer corresponds to one frame of the ideal ratio mask.

III. Mask Estimation

The ideal ratio mask (IRM) is the soft ratio between the clean speech component and the noisy signal at each time-frequency bin. The ratio also indicates the speech-presence probability at each time-frequency bin. DNNs have been widely used to estimate the ideal ratio mask for speech enhancement [38]. We first present how to estimate the ideal ratio mask with DNNs, and then describe how the mask is used for single-channel and multi-channel speech enhancement.

Let us consider the signal captured at one microphone only:

$$X(k,l) = S(k,l) + V(k,l), \tag{3}$$

where X, S, and V are the single-channel microphone signal, the clean speech, and noise component, respectively. The ideal ratio mask for each time-frequency bin is defined as

$$\mathrm{IRM}(k,l) = \min\left(\frac{|S(k,l)|}{|X(k,l)|}, 1\right), \tag{4}$$

where $|\cdot|$ denotes the absolute value. $\mathrm{IRM}(k,l) \in [0,1]$.

To estimate the ideal ratio mask, we employ a feed-forward neural network that comprises an input layer, three hidden layers with rectifying linear units (ReLUs) and an output layer with sigmoid units (see Fig. 2) [28]. A dropout layer is added at each hidden layer to increase the robustness of the model. The sigmoid unit maps an input into the range of $[0,1]$, which typically corresponds to a soft mask.

The DNN is trained with input the log magnitude spectrogram of the noisy signal. We compute the log magnitude as

$$\check{X}(k,l) = \log|X(k,l)|, \tag{5}$$

and then normalize it to zero mean and unit variance, i.e.

$$\widetilde{X}(k,l) = \frac{\check{X}(k,l) - \check{m}(k)}{\check{\sigma}(k)}, \tag{6}$$

where $\check{m}(k)$ and $\check{\sigma}(k)$ denote the mean and standard deviation of $\check{X}$ at the k-th bin, which are pre-computed from the training dataset. The spectrogram has a smaller scale after log scaling and normalization, which is desirable for better convergence in model training [28].

The spectrogram is fed to the DNN in a frame-wise manner. Let us take the l-th frame as an example. The input $\boldsymbol{X}_{in}(l)$ consists of neighbouring frames $[l-\Delta_L : l+\Delta_L]$ in the whole frequency band, i.e.

$$\boldsymbol{X}_{in}(l) = \widetilde{X}(1:K, l-\Delta_L : l+\Delta_L), \tag{7}$$

where $\Delta_L = 3$ is a predefined constant for STFT with frame length 32 ms (256 points) and half overlap. The $2\Delta_L + 1$ frames are flattened into a vector of length $(2\Delta_L+1)K$ before entering the neural network.

The output of the DNN is the ideal ratio mask estimated at the l-th frame, which is a vector of size K and is represented as

$$\boldsymbol{M}_{\text{out}}(l) = M(1:K, l). \tag{8}$$

Cascading the estimation at all the L frames, we get the mask $\boldsymbol{M} = \{M(k,l)\}$, which approximates the true ideal ratio mask.

The model parameters are computed in a supervised manner by minimizing the squared error between the estimated and the true ideal ratio mask, i.e.

$$J = \sum_{k=1}^{K}\sum_{l=1}^{L} |M(k,l) - \text{IRM}(k,l)|^2 . \tag{9}$$

A detailed configuration of the training and testing dataset is described in Section V-B.

IV. DNN-Assisted Speech Enhancement

In this section, we apply the estimated soft masks to three speech enhancement systems: DNN-S, DNN-BF and DNN-TF. DNN-S and DNN-BF are two baseline methods that use the soft masks to estimate the speech-presence probability at individual time-frequency bins, and to design a single-channel spectral filter and a multi-channel beamformer, respectively. DNN-TF is the proposed multi-channel method, which uses the soft masks to estimate the noise-dominance probability at individual time-frequency bins, which are incorporated into the time-frequency spatial filtering algorithm (TF), a state-of-the-art algorithm for drone sound processing [8], [10].

A. DNN-S

Fig. 3 depicts the single-channel speech enhancement pipeline (DNN-S) that uses the DNN model (Fig. 2) to estimate the ideal ratio mask $M(k,l)$ from the log spectrogram of the noisy signal. The whole procedure is straightforward. The clean speech can be recovered from the noisy signal as

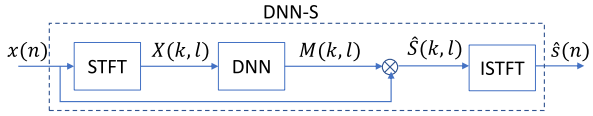


Fig. 3. Block diagram of the DNN-assisted single-channel speech enhancement method (DNN-S).

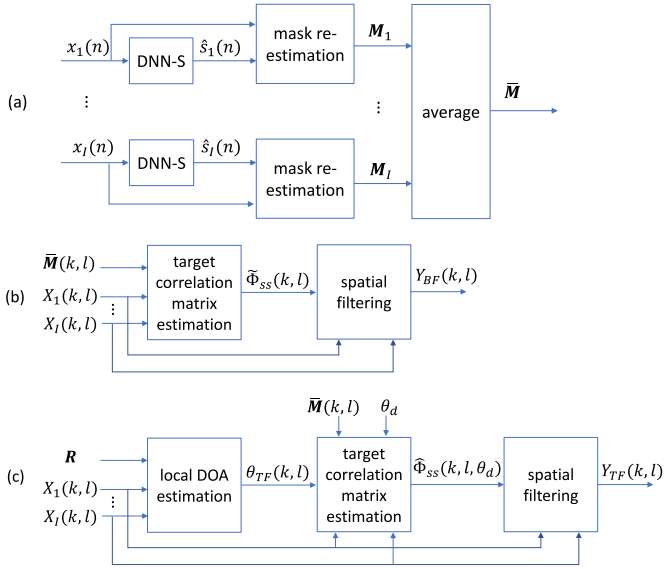


Fig. 4. DNN-assisted multi-channel speech enhancement with I microphones. (a) Ideal ratio mask estimation from multi-channel signals. (b) DNN-BF: DNN-assisted beamforming. (c) DNN-TF: DNN-assisted time-frequency filtering.

$$\hat{S}(k,l) = M(k,l)|X(k,l)|\mathrm{e}^{j\angle X(k,l)} = M(k,l)X(k,l), \tag{10}$$

where $\angle\cdot$ denotes the angle of a complex-valued number. After applying inverse STFT (ISTFT), we obtain the enhanced signal in the time domain.

B. DNN-BF

The ideal ratio mask estimated by DNN cannot be used directly to construct the spatial filter, due to the FFT length mismatch between the two. For instance, in DNN the FFT length is usually 32 ms (256 points) while in a spatial filter the FFT length is longer than 128 ms (1024 points) in order to capture the transmission delay and reverberation. We thus employ a synthesis-reanalysis procedure, as shown in Fig. 4(a), to estimate a ratio mask that is useful for spatial filtering.

Suppose the DNN-estimated mask at the i-th microphone channel is $M_i(k_1,l_1)$, where k_1 and l_1 are the frequency and frame indices corresponding to the FFT length N_1 (e.g. 256). The clean signal is estimated from the noisy signal in the frequency domain as

$$\hat{S}_i(k_1,l_1) = M_i(k_1,l_1)X_i(k_1,l_1), \tag{11}$$

and is converted into the time domain as $\hat{S}_i(n)$. We convert the signal $X_i(n)$ and $\hat{S}_i(n)$ back to the STFT domain with FFT length N_2 (e.g. 1024) as $S_i(k_2,l_2)$. The ratio mask is

re-estimated as

$$M_i(k_2, l_2) = \frac{|\hat{S}_i(k_2, l_2)|}{|X_i(k_2, l_2)|}, \tag{12}$$

where k_2 and l_2 are the frequency and frame indices corresponding to the FFT length N_2.

We compute the mask as the average across the channels (e.g. $I = 8$):

$$\bar{M}(k_2, l_2) = \frac{1}{I}\sum_{i=1}^{I} M_i(k_2, l_2). \tag{13}$$

Without introducing ambiguities, from now we replace (k_2, l_2) with (k, l).

Fig. 4(b) shows how we construct a spatial filter based on the average ratio mask $\bar{M}(k, l)$. Given the mask indicating the speech-presence probability at each time-frequency bin, we estimate the correlation matrix of the target speech signal as

$$\tilde{\boldsymbol{\Phi}}_{ss}(k, l) = \frac{1}{L}\sum_{l=1}^{L} \bar{M}^2(k, l)\boldsymbol{X}(k, l)\boldsymbol{X}^{\mathrm{H}}(k, l), \tag{14}$$

where $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose. We then compute the correlation matrix of the noisy signal as

$$\boldsymbol{\Phi}_{xx}(k, l) = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{X}(k, l)\boldsymbol{X}^{\mathrm{H}}(k, l). \tag{15}$$

Finally, we construct a standard multichannel Wiener filter (MWF) as [49]

$$\boldsymbol{W}_{\text{BF}}(k, l) = \boldsymbol{\Phi}_{xx}^{-1}(k, l)\tilde{\boldsymbol{\Phi}}_{ss1}(k, l), \tag{16}$$

where $\tilde{\boldsymbol{\Phi}}_{ss1}(k, l)$ represents the first column of $\tilde{\boldsymbol{\Phi}}_{ss}(k, l)$.

The target signal is extracted as

$$Y_{\text{BF}}(k, l) = \boldsymbol{W}_{\text{BF}}^{\mathrm{H}}(k, l)\boldsymbol{X}(k, l). \tag{17}$$

DNN-BF formulates a beamformer based on the ideal ratio mask estimated by the DNN. Unlike traditional beamforming techniques, DNN-BF does not require the knowledge of the target DOA or the voice activity information to estimate the target correlation matrix, and thus is more flexible.

C. DNN-TF

In [10] we proposed a time-frequency (TF) processing method that, after estimating the instantaneous DOA at each time-frequency bin, estimates – given the target DOA – the correlation matrix of the target signal and constructs the spatial filter. The TF algorithm works effectively for sound processing on drones since it can well exploit the time-frequency sparsity of the speech signal and the ego-noise. However, a drawback of TF is that when the target signal arrives from a direction close to that of the ego-noise, the time-frequency bins from the ego-noise will be included in computing the target correlation matrix [8]. This leads to noise residuals in the spatial filtering output, thus degrading the noise suppression performance. In extremely low-SNR scenarios the noise-presence inferred from the estimated ideal ratio mask tends to be more reliable than the speech-presence. If we can identify the time-frequency bins that are occupied by the ego-noise and exclude them from computing the target correlation matrix, the performance of the spatial filter will be improved. Based on these considerations, we propose the DNN-TF algorithm, as illustrated in Fig. 4(c). The proposed algorithm is composed of three steps: instantaneous DOA estimation, target correlation matrix computation, and spatial filtering.

1) Instantaneous DOA Estimation: Given the microphone signal $\boldsymbol{X}(k, l)$ and location $\boldsymbol{R}$, the DOA of the sound at each time-frequency bin can be estimated by building a local generalized cross correlation (GCC) function [48]

$$\gamma_{\text{TF}}(k, l, \theta) = \Re\left\{\sum_{\substack{m_1, m_2 = 1 \\ m_1 \neq m_2}}^{I} \frac{X_{m_1}(k, l)X_{m_2}^*(k, l)}{|X_{m_1}(k, l)X_{m_2}(k, l)|} \mathrm{e}^{j2\pi f_k \tau(m_1, m_2, \theta)}\right\}, \tag{18}$$

where f_k denotes the frequency at the k-th bin, the superscript $(\cdot)^*$ denotes the complex conjugation, and the operator $\Re\{\cdot\}$ denotes the real component of the argument. The term $\tau(m_1, m_2, \theta) = \frac{\|\boldsymbol{r}_{m_2} - \boldsymbol{r}_\theta\| - \|\boldsymbol{r}_{m_1} - \boldsymbol{r}_\theta\|}{c}$ denotes the delay between two microphones m_1 and m_2 with respect to the sound coming from θ, where c is the velocity of sound and $\boldsymbol{r}_\theta$ is the location of the far-field sound source from direction θ. The local DOA of the sound at the (k, l)-th bin is determined as

$$\theta_{\text{TF}}(k, l) = \underset{\theta \in (-180^\circ, 180^\circ]}{\arg\max}\ \gamma_{\text{TF}}(k, l, \theta). \tag{19}$$

The harmonic components of the ego-noise and the human speech tend to occupy different time-frequency bins. This time-frequency sparsity makes it possible to estimate the instantaneous DOA at individual time-frequency bins even in low-SNR scenarios.

2) Target Correlation Matrix Computation: The instantaneous DOA estimation at individual time-frequency bins can be used to formulate a spatial filtering pointing at the target direction θ_d. We first measure the closeness of each time-frequency bin (k, l) to θ_d. Assuming the distribution of the DOA estimates to be Gaussian with mean θ_d and standard deviation σ_d, the closeness measure is defined as

$$C_d(k, l, \theta_d) = \exp\left(-\frac{(\theta_{\text{TF}}(k, l) - \theta_d)^2}{2\sigma_d^2}\right), \tag{20}$$

where the scalar $C_d(\cdot) \in [0, 1]$. The higher $C_d(\cdot)$, the higher the confidence that the sound at the (k, l)-th bin arrives from direction θ_d. Since the target sound arrives from θ_d, the confidence $C_d(\cdot)$ can also be interpreted as speech-presence probability. In [10], C_d is used to calculate the target correlation matrix as

$$\check{\boldsymbol{\Phi}}_{ss}(k, l, \theta_d) = \frac{1}{L}\sum_{l=1}^{L} C_d^2(k, l, \theta_d)\boldsymbol{X}^{\mathrm{H}}(k, l)\boldsymbol{X}(k, l), \tag{21}$$

where $C_d(\cdot)$ denotes the contribution of the (k, l)-th bin to the correlation matrix. As shown in (16), this target correlation matrix is crucial to formulate a multi-channel beamformer.

Due to time-frequency sparsity of the acoustic signal, Eqs. (20)–(21) can compute the speech-presence probability and the target correlation matrix robustly in low-SNR scenarios. However, when the target sound comes from a similar direction as the ego-noise, the noise component will be accounted as the target speech, thus degrading the accuracy of the target correlation matrix estimation. To mitigate this problem, we use DNN-estimated soft masks to detect the time-frequency bins dominated by the ego-noise and then remove them from computing the target correlation matrix. The details are given below.

We first estimate the average ratio mask using multi-channel signals with (13). The time-frequency bins dominated by the ego-noise are then determined as

$$\boxed{M_H(k,l) = \begin{cases} 1, & \bar{M}(k,l) < M_{\text{TH}} \\ 0, & \text{otherwise} \end{cases}} \tag{22}$$

where $M_{\text{TH}} = 0.2$ is a pre-defined threshold. $\boldsymbol{M}_H = \{M_H(k,l)\}$ is a binary matrix indicating the noise-dominance at each time-frequency bin. We refer to $\boldsymbol{M}_H$ as noise-dominance probability.

We then remove the contribution from the noise-dominated time-frequency bins, obtaining a modified confidence matrix with

$$\boxed{\tilde{C}_d(k,l,\theta_d) = C_d(k,l,\theta_d)\left(1 - M_H(k,l)\right).} \tag{23}$$

Finally, the modified target correlation matrix is computed as

$$\boxed{\hat{\boldsymbol{\Phi}}_{ss}(k,l,\theta_d) = \frac{1}{L}\sum_{l=1}^{L} \tilde{C}_d^2(k,l,\theta_d)\boldsymbol{X}^{\text{H}}(k,l)\boldsymbol{X}(k,l).} \tag{24}$$

3) Spatial Filtering: Given the target correlation matrix, we formulate a standard multi-channel Wiener filter, similarly to (16), as

$$\boldsymbol{W}_{\text{TF}}(k,l,\theta_d) = \boldsymbol{\Phi}_{xx}^{-1}(k,l)\hat{\boldsymbol{\Phi}}_{ss1}(k,l,\theta_d), \tag{25}$$

where $\hat{\boldsymbol{\Phi}}_{ss1}(k,l)$ represents the first column of $\hat{\boldsymbol{\Phi}}_{ss}(k,l)$.

The sound arriving from the target direction θ_d is extracted as

$$Y_{\text{TF}}(k,l,\theta_d) = \boldsymbol{W}_{\text{TF}}^{\text{H}}(k,l,\theta_d)\boldsymbol{X}(k,l). \tag{26}$$

The main difference between TF [10] and DNN-TF is the confidence matrix used for computing the target correlation matrix. TF only considers the target DOA and the time-frequency sparsity to compute the confidence matrix $\tilde{C}_d$, using (20)–(21). DNN-TF considers the target DOA, the time-frequency sparsity, as well as the noise-dominance at individual time-frequency bins to compute the confidence matrix C_d, using (22)–(24). In very low-SNR scenarios, the ideal ratio masks estimated by DNN-S tend to have low values, and thus can infer the noise-dominance more robustly than speech-presence. With more accurate target correlation matrix estimation, DNN-TF tends to outperform TF, especially when the target sound comes from a direction close to that of the ego-noise source. A comparison of results obtained with the two confidence matrices, $\tilde{C}_d$ and C_d, will be discussed in the next section (see Fig. 8).

TABLE I
DETAILS OF THE NOISE DATASET

Dataset	ID	Type	Length	Channel
AVQ [51]	n116	s1_seq1: constant (50%)	$120s$	8
	n117	s1_seq2: constant (100%)	$120s$	
	n118	s1_seq1: constant (150%)	$40s$	
	n119	s2_seq1: constant (100%)	$210s$	
	n120	s2_seq2: dynamic	$214s$	
AS [8]	n121	constant (100%)	$130s$	8
	n122	dynamic	$140s$	
Regular noise [54], [55]	n1-n115	Non-speech daily life sound	$<10s$ each	1

V. EXPERIMENTS

A. Experiment Setup

First, we investigate the performance of the single-channel DNN model obtained with different combinations of the training and testing datasets. Then, we investigate the performance of various speech enhancement approaches, namely DNN-S, DNN-BF, DNN-TF, the state-of-the-art time-frequency spatial filtering algorithm (TF) [10], and the single-channel UMMSE noise reduction algorithm [50].

We use two multi-channel drone-sound datasets, AS [8] and AVQ [51], and a single-channel speech and noise dataset. Table I summarizes the details of the noise data used in this paper. AS and AVQ were recorded with the platform shown in Fig. 1, with the drone on a tripod. The multi-channel microphone signal is captured with a Zoom R24 audio recorder at a sampling rate of 44.1 kHz (downsampled to 8 kHz before processing). The speech and ego-noise are recorded separately to allow performance evaluation. The AS dataset consists of real-recorded ego-noise and simulated speech [8]. The ego-noise is recorded in an indoor environment with reverberation time 200 ms. During recording, the drone operates at a constant power, varying from 50% to 150% of the hovering status. The speech component is simulated with the image source model [52] in a space of size 20 m $\times$ 20 m $\times$ 4 m and reverberation time 200 ms. The speech source is placed in the far field (10 m away from the microphone array), emitting sound from a varying DOA from -180° to 180°. The AVQ dataset consists of ego-noise and speech both recorded outdoors [51]. When recording the ego-noise, the drone operates either at a constant power (hovering status) or a time-varying power (from 50% to 150% of the hovering power). When recording speech, a person moves at different locations in front of the drone, with the distance varying from 2 meters to 6 meters. During recording, the person talks while stationary at one location, before moving to the next location. The DOA of the speaker is provided in the dataset. Fig. 11(a) depicts an example of the recording environment.

In addition to multi-channel drone sound recording, we use a single-channel speech and noise dataset. The noise dataset contains 115 types of regular noise in daily life [54], [55]. For speech, we use the TIMIT corpus (630 speakers with 10

TABLE II
DIFFERENT STRATEGIES TO COMBINE TRAINING AND TESTING DATA FOR LEARNING THE DNN MODEL

Model	Noise		Speech		Noisy (training)	
	Train	Test	Train	Test	SNR[dB]	Length
MD1	n1-n115	n121-n122 or n116-n120	TIMIT (TRAIN)	TIMIT (TEST)	[-5, 15]	$60h$
MD2	n1-n115				[-25, -5]	$60h$
MD3	n116-n120				[-5, 15]	$60h$
MD4	n116-n120				[-25, -5]	$60h$
MD5	n116-n120				[-25, 15]	$120h$
MD6	n1-n120				[-25, 15]	$120h$

sentences each), where the training and testing subsets contain, respectively, 4620 and 1680 utterances, each lasting less than 8 seconds [53]. Note that this speech corpus is mainly used to develop and test the single-channel DNN model. AS and AVQ use a completely different speech corpus from TIMIT.

To assess the speech enhancement performance, we use the signal-to-noise ratio (SNR) and the perceptual evaluation of speech quality (PESQ). The SNR is computed assuming the target $\boldsymbol{s}(n)$ and the noise component $\boldsymbol{v}(n)$ at the microphones to be known. Given a spatial filter $\boldsymbol{w}(n)$, which is a time-domain version of $\boldsymbol{W}(k,l)$, the spatial filtering procedure is written as

$$y(n) = \boldsymbol{w}(n) * \boldsymbol{x}(n) = \sum_{p=0}^{L_w-1} \boldsymbol{w}(p)\boldsymbol{x}(n-p)$$

$$= y_s(n) + y_v(n) = \boldsymbol{w}(n) * \boldsymbol{s}(n) + \boldsymbol{w}(n) * \boldsymbol{v}(n), \quad (27)$$

where '$*$' denotes the convolutive filtering procedure and L_w is the length of the filter $\boldsymbol{w}(n)$; $y_s(n)$ and $y_v(n)$ are, respectively, the speech and noise components at the output. The SNR is calculated in the target-sound-active periods $\mathbb{N}_s$ as [56]

$$\text{SNR} = 10\log_{10}\frac{\sum_{n'\in\mathbb{N}_s} y_s^2(n')}{\sum_{n'\in\mathbb{N}_s} y_v^2(n')}. \quad (28)$$

PESQ $\in [0, 4.5]$ is a measure to assess the overall quality of the processed speech relative to the referenced clean speech [57]. The higher the PESQ, the better the speech quality.

B. Single-Channel DNN Model Training

The DNN processes the signal at sampling rate 8 kHz. When computing the STFT, we use a frame length of 32 ms (256 points) and half overlap. We set $\Delta_L = 3$ in (7), and thus the size of the input layer is 903 and that of the output layer is 129. Each of the three hidden layers contains 2048 neurons. The dropout ratio is set as 0.2. The total number of parameters of the DNN is 10,508,417. To train the DNN we use the Keras library [58] and a Tesla V100 GPU with 16 GB memory. We use SGD solver with learning rate 0.01. The mini-batch size is set to 500. The total number of epochs is 50.

Table II shows our different choices of combining speech and noise data for training the DNN. The noisy data is generated by adding the clean speech and the noise at different SNRs. We always use the TIMIT corpus to generate the training and testing clean speech. For training noise, we choose either the regular noise (n1-n115), or the drone ego-noise (n116-n120) from the AVQ dataset, or a combination of both (n1-n120). For testing noise, we choose drone ego-noise either from the AVQ dataset (n116-n120) or from the AS dataset (n121-n122). For n116-n122 each with eight channels, we only use the data from the first channel.

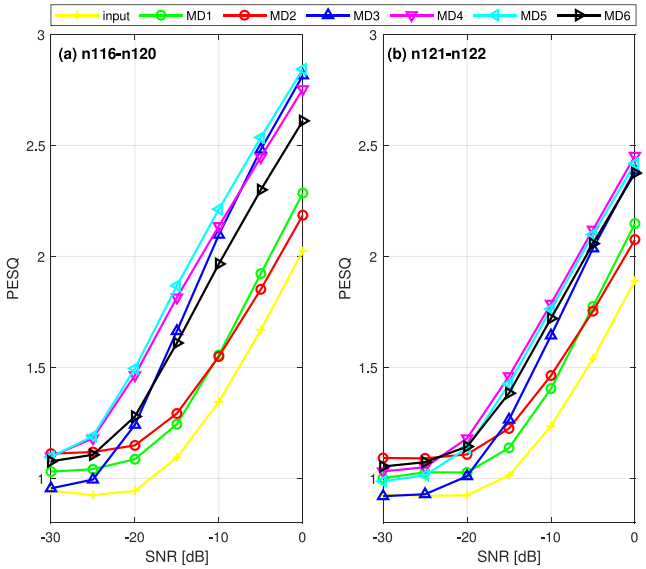


Fig. 5. PESQ values obtained by the DNN models for various testing SNRs. (a) Testing noise n116-n120. (b) Testing noise n121-n122.

When generating the training noisy data, we consider different SNRs, varying from -25 dB to 15 dB, with an increment of 5 dB. For a specific SNR, each of the 4620 utterances in the TIMIT training subset is added with four noises randomly selected from the noise dataset, i.e. we generate 18,480 files with a total duration of about 12 hours at each SNR. When the noise segment (e.g. n116-n120) is longer than the speech segment, we randomly choose the same length of segment from the noise. The combination of different choices leads to 6 DNN models: MD1, . . ., MD6. The testing noisy data is generated similarly to the training data. The difference is that for a specific SNR we combine all the available speech and noise. For instance, given 1680 speech segments and 2 noise segments, we generate 3360 noisy segments.

Fig. 5 shows the average PESQ obtained by the 6 models (MD1, . . ., MD6) for the testing data at different SNRs. The testing data is generated with n116-n120 in Fig. 5(a), and with n121-n122 in Fig. 5(b), respectively. Note that n116-n120 are also used to generate the training data, whereas n121-n122 are unseen to the trained DNN. As shown in Fig. 5, the way to generate the training data impacts the performance of the DNN model considerably. In both Fig. 5(a) and (b), all the 6 models improve the PESQ of input noisy speech to a certain degree. However, the improvement is quite limited in scenarios with low testing SNRs (e.g. < -20 dB).

In Fig. 5(a), which uses n116-n120 as testing noise, MD4-MD7, which are trained using n116-n120, remarkably outperform MD2-MD3, which are trained using n1-n115. This indicates the importance of training noise data for the performance of the DNN model. MD4, which is trained at SNRs $[-25, -5]$ dB, outperforms MD3, which is trained at SNRs $[-5, 15]$ dB, when

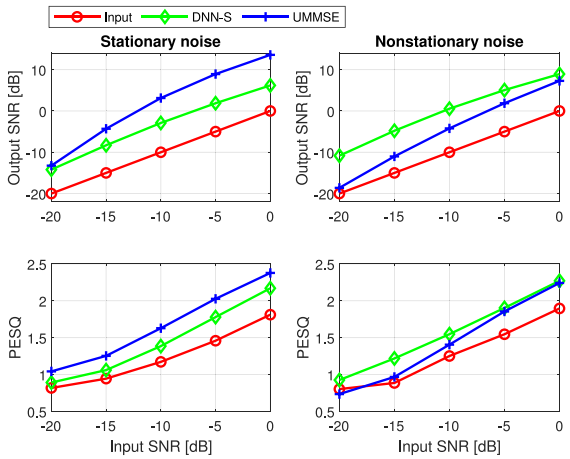


Fig. 6. Speech enhancement performance (SNR and PESQ) by DNN-S and UMMSE for stationary and nonstationary ego-noise at various input SNRs.

the testing SNR is lower than -10 dB. On the other hand, MD3 outperforms MD4 when the testing SNR is higher than -10 dB. MD5, which is trained at SNRs $[-25, 15]$ dB, outperforms both MD3 and MD4, at all testing SNRs. This indicates the importance of training SNR for the performance of the DNN model. MD6, which is trained using all types of noise n1-n120 and all SNRs $[-25, 15]$ dB, performs worse than MD3-MD5 in most testing scenarios. It only outperforms MD3 in a specific testing scenario (SNR $\leq$ -20 dB). This is possibly because the size of the DNN model is not big enough to capture all the variations in the training noise.

In Fig. 5(b), which uses unseen noise n121-n122 as testing noise, the performance of all the models is lower than the corresponding one in Fig. 5(a). This is reasonable as the performance of the DNN model tends to drop for unseen noise. The performance of all the models shows similar variation trends as those observed in Fig. 5(a). However, the difference between MD3-MD6 becomes smaller. It would be interesting to compare the performance of MD4, which is trained at lower SNR $[-25, -5]$ dB, and MD5, which is trained at a wider SNR range $[-25, 15]$ dB. MD5 outperforms MD4 for seen noise in Fig. 5(a), but performs slightly worse than MD4 for unseen noise in Fig. 5(b). This suggests that the generality of the model can be improved.

Based on the observations in Fig. 5, we choose the model MD5, which performs the best in Fig. 5(a) and performs similarly to MD4 and MD6 in Fig. 5(b), for the DNN-S algorithm in the rest of the paper.

We further compare the performance of DNN-S (using MD5) and a traditional unsupervised single-channel UMMSE noise reduction algorithm [50] for stationary and nonstationary ego-noise. To this end, we generate 100 testing files by randomly selecting 100 speech files from the TIMIT testing subset and the same length of stationary (n121) or nonstationary (n122) ego-noise, and mixing them at a varying input SNR from -25 dB to 0 dB, with step 5 dB. We apply the two algorithms to compute the average output SNR and PESQ values. Fig. 6 depicts the experimental results. It can be observed that, for both algorithms, the output SNR and PESQ improve with the increasing input

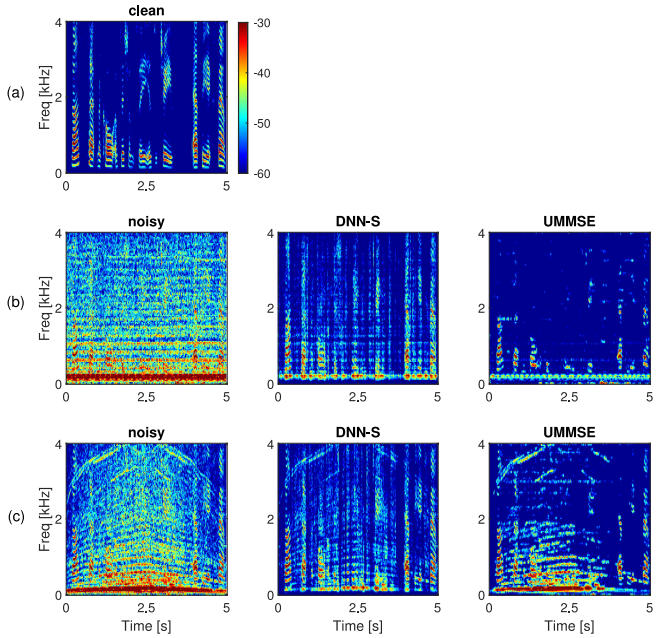


Fig. 7. Spectrograms of the speech enhancement results using DNN-S and UMMSE for stationary and nonstationary ego-noise. (a) Clean speech. (b) Stationary noise. (c) Nonstationary noise. The input SNR is -10 dB.

TABLE III
SPEECH ENHANCEMENT PERFORMANCE USING DNN-S AND UMMSE FOR STATIONARY AND NONSTATIONARY EGO-NOISE. THE RESULT CORRESPONDS TO FIG. 7

	Stationary		**Nonstationary**	
	SNR [dB]	**PESQ**	**SNR** [dB]	**PESQ**
Input	-10	1.44	-10	1.42
DNN-S	-1.2	1.59	-3.3	1.77
UMMSE	5.2	1.87	-6.9	1.46

SNR. DNN-S achieves similar performance (output SNR and PESQ) for stationary and nonstationary noises, while UMMSE shows significantly worse performance for nonstationary noise. UMMSE outperforms DNN-S for stationary noise for all input SNRs. DNN-S outperforms UMMSE for nonstationary noise especially at low input SNRs. This difference in performance, however, becomes less evident as the input SNR increases.

Fig. 7 shows the spectrograms of sample processing results obtained by DNN-S and UMMSE for stationary and nonstationary ego-noise. It can be observed that DNN-S works well for both stationary and nonstationary noise. UMMSE works better for stationary noise but worse for nonstationary noise. The SNR and PESQ values shown in Table III also confirm this observation. Note that while the noise reduction performance for non-stationary noise is worse for UMMSE, this method does need training data and has a much lower computational complexity.

C. Multi-Channel Speech Enhancement

We compare the three multi-channel processing algorithms (DNN-BF, DNN-TF and TF) and the single-channel algorithm

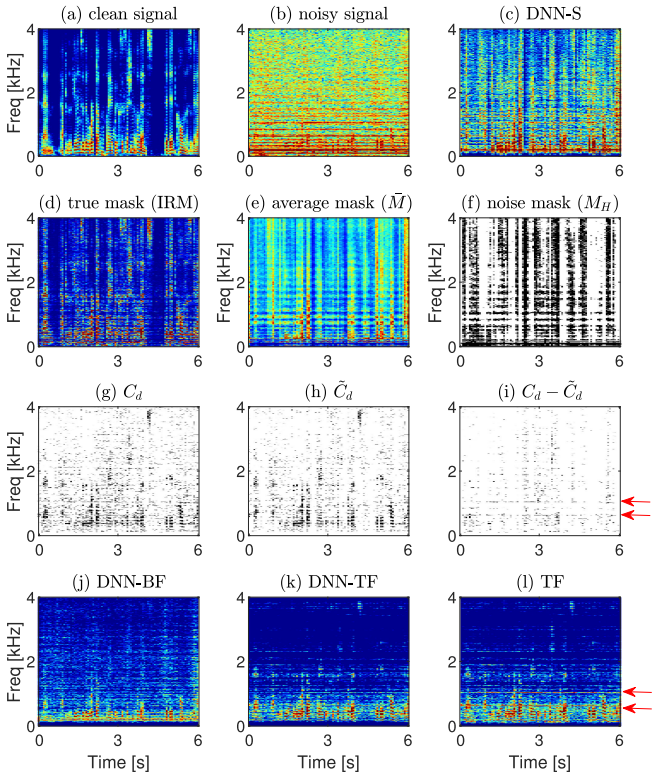


Fig. 8. Intermediate processing results for the proposed algorithms when the DOA of the target source is 70° and the input SNR is −15 dB.

TABLE IV
SNR AND PESQ FOR THE ALGORITHMS UNDER COMPARISON IN FIG. 8

	Input	**DNN-S**	**DNN-BF**	**DNN-TF**	**TF**
SNR [dB]	-15	-6.4	1.6	11.4	8.0
PESQ	1.18	1.47	1.83	2.44	2.38

(DNN-S) with simulated and real-recorded data. For multi-channel algorithm implementation, we set the STFT length to be 1024 with half overlap. We set $\sigma_d = 10°$ in (20), as suggested in [10].

Fig. 8 depicts intermediate processing results of four algorithms (DNN-S, DNN-BF, DNN-TF and TF) for a segment (6 seconds) of multi-channel noisy signal, which is generated by adding a simulated speech with DOA 70° and a real-recorded ego-noise (n121). The input SNR is −15 dB. In this case the target DOA is close to one ego-noise source (cf. Fig. 1(b)). Table IV gives the corresponding SNR and PESQ values obtained by the four algorithms.

The input signal becomes very noisy at input SNR −15 dB. As shown in Fig. 8(b), it is difficult to identify speech-occupied time-frequency bins from the spectrogram. However, as shown in Fig. 8(c), DNN-S is able to partly recover the clean speech from the noisy input signal, improving the SNR from −15 dB to -6.4 dB and the PESQ value from 1.18 to 1.47.

Fig. 8(d)–(f) compare the true ideal ratio mask, the average mask $\bar{M}$ estimated from eight channels in Eq. (13), and the binary noise mask M_H in Eq. (22). While not very precise, the average mask $\bar{M}$ can roughly estimate the speech-presence probability at each time-frequency bin. In this low-SNR scenario, the binary noise mask M_H can identify the noise-dominated time-frequency bins more accurately.

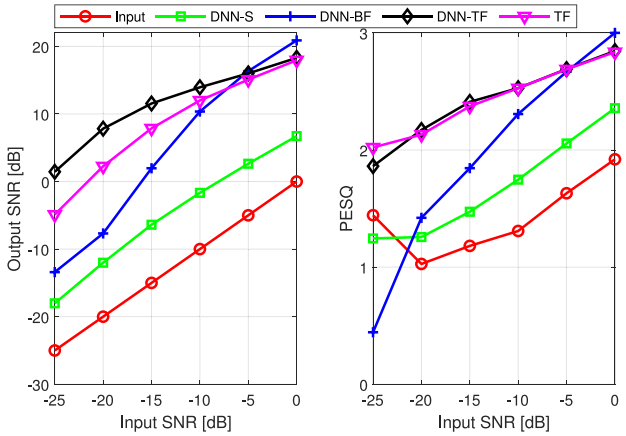


Fig. 9. SNR and PESQ performance for target speech with DOA 70° at various input SNRs.

Fig. 8(g)–(h) illustrate the confidence matrix C_d in Eq. (20) and the modified confidence matrix $\tilde{C}_d$ in Eq. (23), respectively. For ease of comparison, Fig. 8(i) depicts the difference of the two $C_d - \tilde{C}_d$, which clearly shows that some noise-dominated time-frequency bins are removed in the modified confidence matrix. For instance, the time-frequency bins at 750 and 1000 Hz (as indicated by the red arrows) are dominated by noise. These time-frequency bins are erroneously included in C_d, but removed in $\tilde{C}_d$.

Fig. 8(j)–(l) compare the spectrograms obtained by DNN-BF, DNN-TF and TF. DNN-BF is designed solely based on the average mask $\bar{M}$; it outperforms the single-channel DNN-S remarkably, improving the SNR by 8 dB (from -6.4 to 1.6) and PESQ by 0.36 (from 1.47 to 1.83). TF and DNN-TF outperform DNN-BF by exploiting the DOA information of the target signal. DNN-TF can further improve the SNR of DNN-BF by 10 dB (from 1.6 to 11.4) and PESQ by 0.61 (from 1.83 to 2.44). DNN-TF outperforms TF by computing the confidence matrix more accurately and thus suppressing the noise more efficiently. As shown in Fig. 8(k)–(l), TF still retains the noise components at 750 Hz and 1000 Hz, while DNN-TF removes these noise components completely. Consequently, the SNR of DNN-TF is 3.4 dB higher than that of TF (from 8.0 to 11.4).

Fig. 9 shows the SNR and PESQ for the results of the four algorithms on the same speech and noise with various input SNRs. The worst performance is by the single-channel DNN-S. DNN-TF and TF outperform DNN-BF in most cases, except when the input SNR is higher than -5 dB. DNN-TF achieves higher output SNR than TF for all input SNRs. As the input SNR decreases, the difference in performance between DNN-TF and TF increases, thus confirming the advantage of DNN-TF in low-SNR scenarios.

Next, we investigate how the speech enhancement performance varies with the target DOA. To this end, we generate testing data with the target DOA varying from -180° to 180°, with increment of 10° and the input SNR varying from -25 dB to 0 dB, with increment of 5 dB. For each combination of DOA

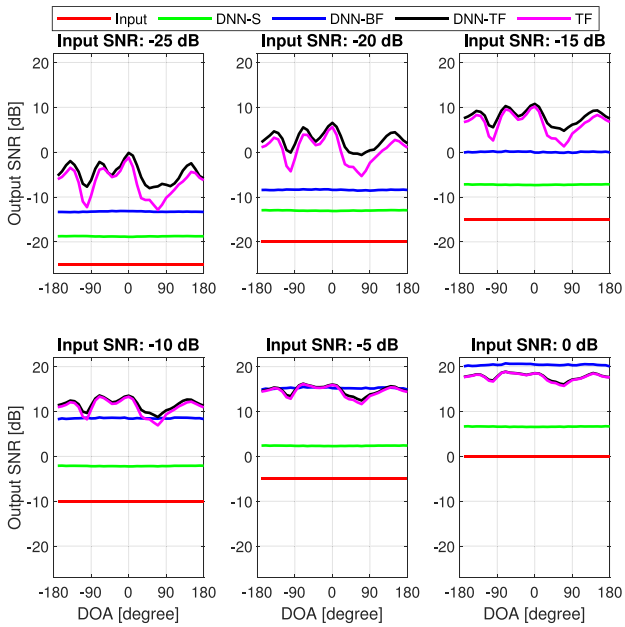


Fig. 10. SNR performance vs. target DOA obtained by the algorithms under comparison when varying the input SNR.

and input SNR, we generate 20 segments (6 seconds long) of testing data, randomly selected from the simulated speech and real-recorded ego-noise (n121 and n122).

Fig. 10 depicts the variation of the output SNR obtained by the four algorithms for varying DOA and input SNR. The performance of TF drops significantly when the target sound comes from a direction close to the ego-noise sources, e.g. at around 90° and -90° (which correspond to two peaks of the spatial likelihood function in Fig. 1(d)). DNN-TF works more robustly than TF to the variation of the target DOA, especially at low SNRs (≤ -10 dB). The two algorithms perform similarly at high SNRs (≥ -5 dB). The performance of DNN-S and DNN-BF does not vary with the DOA. This is because these algorithms rely on single-channel information, which does not vary much with the target DOA. The performance of the two algorithms improves as the input SNR increases. DNN-BF outperforms DNN-TF and TF when the input SNR is higher than -5 dB.

Finally, we investigate the performance of the considered algorithms with real-recorded speech and ego-noise, both from the AVQ dataset [51]. The speech is recorded when a person speaks at four different locations in front of the drone (speaker A at locations ⑦, ④, ①, ③ in Fig. 11(a)). The ego-noise is recorded at the hovering status (n117). Both speech and ego-noise are recorded continuously. The speech and the ego-noise are added without any scaling when generating the noisy signal. The noisy signal is processed by the algorithms block by block, where each block is defined with a non-overlap sliding window of length 6 seconds. Fig. 11(b) shows the block-wise SNR and PESQ results in voice-active periods and Table V shows the average results. All the four algorithms can improve the SNR of the input signal (which is -26.8 dB on average). DNN-TF and TF perform significantly better than DNN-BF and DNN-S.

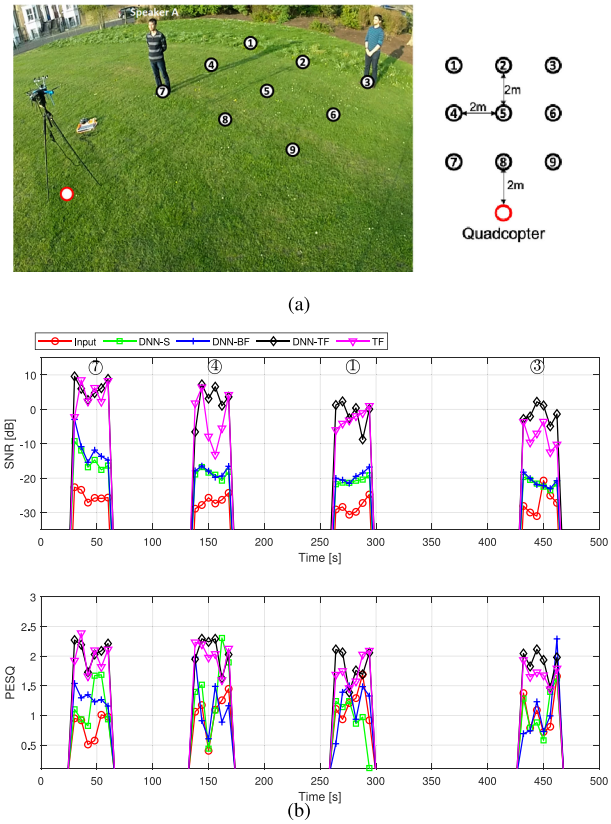


Fig. 11. (a) Recording environment and (b) speech enhancement performance on the AVQ dataset in terms of SNR and PESQ for the algorithms under comparison.

TABLE V
AVERAGE SNR AND PESQ FOR THE ALGORITHMS COMPARED IN FIG. 11

	Input	DNN-S	DNN-BF	DNN-TF	TF
SNR [dB]	-26.8	-18.9	-17.5	1.6	-2.1
PESQ	1.04	1.16	1.20	1.97	1.87

DNN-TF outperforms TF, while DNN-BF slightly outperforms DNN-S. On average, DNN-TF improves the input SNR by 28.4 dB and improves PESQ by 0.93.

VI. DISCUSSION

DNN-S is a single-channel method that employs deep neural networks to estimate the ideal ratio mask at each time-frequency bin. It works well for both stationary and nonstationary noise, and outperforms traditional single-channel noise reduction approaches (e.g. [50]). The types of noise play a crucial role for the supervised approaches and the performance of DNN-S tends to degrade with decreasing input SNR. As a result, DNN-S does not improve the sound quality sufficiently in low-SNR scenarios. However, the estimated ideal ratio mask can be interpreted as the speech-presence probability at each time-frequency bin and can be used to assist multi-channel processing algorithms.

DNN-BF is a multi-channel speech enhancement algorithm that formulates a beamformer based on the ideal ratio mask

estimated by DNN-S. In comparison to traditional beamforming techniques, DNN-BF does not require the knowledge of the target DOA nor the voice activity information to estimate the target correlation matrix, and thus is more flexible. Through multi-channel processing, DNN-BF outperforms DNN-S. A drawback of DNN-BF is that it relies on the accuracy of the mask estimation and thus its performance is limited with low SNR.

DNN-TF can be interpreted as a combination of TF and DNN-S. DNN-TF outperforms DNN-BF by combining the knowledge of the target DOA, the time-frequency sparsity of the acoustic signal and the estimated ideal ratio mask. By exploiting the speech-presence probability, DNN-TF can better estimate the correlation matrix of the target signal, and thus outperforms TF in low-SNR scenarios, especially when the target sound comes from a direction close to that of the ego-noise sources. Unlike DNN-BF, DNN-TF requires the knowledge of the target DOA and the location of the microphones. In practice, the target DOA can be estimated using existing sound source localization algorithms that are dedicated to drone sound processing [18], [19], or be estimated by exploiting additional vision information [13], [51]. Furthermore, DNN-TF can separate multiple speakers talking simultaneously if their DOAs are known. This is challenging for DNN-BF, as it requires to know the TF bins associated with each speaker, which is difficult to obtain in practice.

Similarly to DNN-BF, several existing works integrate the time-frequency masks and the multichannel beamformer. These works either employ a single-channel DNN model [40]–[42], which exploit the spectral information only, or employ a multi-channel DNN model [43]–[45], which exploit both the spectral and spatial information of the microphone signals. These various types of DNN models can also be used in the proposed method. The single-channel DNN regression model employed in this paper works optimally estimating the time-frequency masks when using a FFT of length 32 ms to take advantage of the short-time stationarity of speech signals [28]. This parameter is consistent with traditional single-channel noise reduction approaches [50]. A spatial filter usually employs a larger FFT length (e.g. 128 ms) to achieve better noise reduction performance. A synthesis-reanalysis procedure is thus required to deal with the inconsistency between the different FFT lengths between the spectral and spatial filters. One interesting future direction would be to develop a more advanced DNN structure that can estimate the time-frequency masks at the same FFT length of the spatial filter.

VII. Conclusion

We presented the first method that employs deep learning for speech enhancement on noisy drone platforms. The proposed method, DNN-TF, can suppress the ego-noise and improve the sound quality by exploiting the knowledge of the target DOA, the time-frequency sparsity of the ego-noise and speech signals, and the DNN-estimated time-frequency ratio mask. DNN-TF outperforms traditional single-channel (DNN-S) and multi-channel deep-learning-based speech enhancement (DNN-BF) approaches in most cases. DNN-TF also outperforms the state-of-the-art TF algorithm in low-SNR scenarios, and provides a more robust performance when the target DOA is close to the ego-noise sources. Future work will investigate the performance of the proposed approach on various drone platforms.

References

[1] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2012, pp. 3288–3293.

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

[3] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2012, pp. 4737–4742.

[4] J. Cacace, R. Caccavale, A. Finzi, and V. Lippiello, "Attentional multimodal interface for multidrone search in the Alps," in *Proc. IEEE Int. Conf. Sys. Man, Cybern.*, Budapest, Hungary, 2016, pp. 1178–1183.

[5] T. Latif, E. Whitmire, T. Novak, and A. Bozkurt, "Sound localization sensors for search and rescue biobots," *IEEE Sensors J.*, vol. 16, no. 10, pp. 3444–3453, May 2016.

[6] K. Hoshiba *et al.*, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, pp. 1–16, Nov. 2017.

[7] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for UAV-embedded sound source localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Madrid, Spain, 2018, pp. 1–8.

[8] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors J.*, vol. 18, no. 11, pp. 4570–4582, Nov. 2018.

[9] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles," in *Proc. Int. Conf. Adv. Video Signal-Based Surveillance*, 2016, pp. 1–7.

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

[11] 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.

[12] R. P. Fernandes, E. C. Santos, A. L. L. Ramos, and J. A. Apolinario, "A first approach to signal enhancement for quadcopters using piezoelectric sensors," in *Proc. Int. Conf. Transformative Sci. Eng. Bus. Social Innov.*, 2015, pp. 536–541.

[13] R. Sanchez-Matilla, L. Wang, and A. Cavallaro, "Multi-modal localization and enhancement of multiple sound sources from a micro aerial vehicle," *Proc. ACM Multimedia*, 2017, pp. 1591–1599.

[14] 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. Workshop Acoust. Signal Enhancement*, 2016, pp. 1–5.

[15] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Beamforming-based acoustic source localization and enhancement for multirotor UAVs," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 987–991.

[16] P. Misra, A. A. Kumar, P. Mohapatra, and P. Balamuralidhar, "Aerial drones with location-sensitive ears," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 154–160, Jul. 2018.

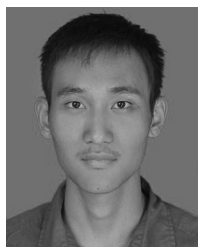
[17] B. Yen, Y. Hioka, and B. Mace, "Improving power spectral density estimation of unmanned aerial vehicle rotor noise by learning from non-acoustic information," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Tokyo, Japan, 2018, pp. 1–5.

[18] L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 1–5.

[19] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Tracking a moving sound source from a multi-rotor drone," *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Madrid, Spain, 2018, pp. 2511–2516.

[20] K. Furukawa *et al.*, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Tokyo, Japan, 2013, pp. 3943–3948.

[21] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-board relative bearing estimation for teams of drones using sound," *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, pp. 820–827, Jan. 2016.
[22] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and L. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2014, pp. 1902–1907.
[23] G. Sinibaldi and L. Marino, "Experimental analysis on the noise of propellers for small UAV," *Appl. Acoust.*, vol. 74, no. 1, pp. 79–88, Jan. 2015.
[24] Z. Zhang, J. Geiger, J. Pohjalainen, A. E. D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 5, pp. 1–28, May 2018.
[25] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
[26] J. C. Hou, S. S. Wang, Y. H. Lai, Y. Tsao, H. W. Chang, and H. M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.
[27] H. Purwins, B. Li, T. Virtanen, J. Schlueter, S. Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, Feb. 2019.
[28] Y. Xu, D. Jun, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2014.
[29] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. InterSpeech*, Lyon, France, 2013, pp. 436–440.
[30] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. InterSpeech*, 2016, pp. 1993–1997.
[31] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2017, pp. 1265–1269.
[32] A. Maas, Q. V. Le, T. M. Oneil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. InterSpeech*, 2012, pp. 22–25.
[33] F. Weninger *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2015, pp. 91–99.
[34] L. Sun, J. Du, L. R. Dai, and C. H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 136–140.
[35] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian wavenet," in *Proc. InterSpeech*, Stockholm, Sweden, 2017, pp. 2013–2017.
[36] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[37] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, Canada, 2013, pp. 7092–7096.
[38] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[39] D. S. Williamson and D. L. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.
[40] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, China, 2016, pp. 196–200.
[41] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. InterSpeech*, 2016, pp. 1981–1985.
[42] T. Nakatani, I. Nobutaka, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 286–290.
[43] L. Pfeifenberger, M. Zohrer, and F. Pernkopf, "DNN-based speech mask estimation for eigenvector beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 66–70.
[44] W. Jiang, F. Wen, and P. Liu, "Robust beamforming for speech recognition using DNN-based time-frequency masks estimation," *IEEE Access*, vol. 6, pp. 52385–52392, 2018.
[45] N. Furnon, R. Serizel, I. Illina, and S. Essid, "DNN-based distributed multichannel mask estimation for speech enhancement in microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Spain, 2020, pp. 4672–4676.
[46] X. Xiao *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, China, 2016, pp. 5745–5749.
[47] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multi-channel robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 271–275.
[48] L. Wang, T. K. Hon, J. D Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1079–1093, Jun. 2016.
[49] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.
[50] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, Apr. 2011.
[51] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Audio-visual sensing from a quadcopter: Dataset and baselines for source localization and sound enhancement," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Macau, China, 2019, pp. 5320–5325.
[52] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[53] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, *Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*, NISTIR 4930, National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA, 1988.
[54] G. Hu, "A corpus of nonspeech sounds," 2005, [Online]. Available: http://web.cse.ohiostate.edu/pnl/corpus/HuNonspeech/HuCorpus.html
[55] Y. Xu, J. Du, Z. Huang, L. Dai, and C. H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *Proc. InterSpeech*, 2015, pp. 1508–1512.
[56] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1493–1508, Sep. 2015.
[57] Y. Hu and P. C. Loizou.,"Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2007.
[58] F. Chollet, "Keras: The Python deep learning library," *Astrophys. Source Code Library*, 2018.



Lin Wang received the B.S. degree in electronic engineering from Tianjin University, China, in 2003; and the Ph.D degree in signal processing from the Dalian University of Technology, China, in 2010. From 2011 to 2013, he has been an Alexander von Humboldt Fellow in University of Oldenburg, Germany. From 2014 to 2017, he has been a Postdoctoral Researcher in Queen Mary University of London, UK. From 2017 to 2018, he has been a Postdoctoral Researcher in the University of Sussex, U.K. Since 2018, he has been a Lecturer in Queen Mary University of London. He is Associate Editor of IEEE ACCESS. His research interests include audio-visual signal processing, machine learning, and robotic perception.

Andrea Cavallaro received the Ph.D. degree in electrical engineering from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He is Professor of Multimedia Signal Processing and the founding Director of the Centre for Intelligent Sensing at the Queen Mary University of London, Turing Fellow at the Alan Turing Institute, the UK National Institute for Data Science and Artificial Intelligence, and Fellow of the International Association for Pattern Recognition. He is Editor-in-Chief of Signal Processing: Image Communication; Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING; Chair of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee; and an IEEE Signal Processing Society Distinguished Lecture.