IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 February 2023, accepted 26 February 2023, date of publication 6 March 2023, date of current version 10 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3253719*

RESEARCH ARTICLE

# Deep Learning Models for Single-Channel Speech Enhancement on Drones

**DMITRII MUKHUTDINOV, ASHISH ALEX, ANDREA CAVALLARO, AND LIN WANG**
Centre for Intelligent Sensing, Queen Mary University of London, E1 4NS London, U.K.

Corresponding author: Lin Wang (lin.wang@qmul.ac.uk)

This work was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K007491/1; in part by the ARTEMIS-JU and the U.K. Technology Strategy Board (Innovate U.K.) through the COPCAMS Project under Grant 332913; in part by the Pilot Research Project of EPSRC UK Acoustics Network Plus under Grant EP/V007866/1; and in part by the use of time on Tier 2 HPC Facility JADE2, funded by EPSRC under Grant EP/T022205/1.

**ABSTRACT** Speech enhancement for drone audition is made challenging by the strong ego-noise from the rotating motors and propellers, which leads to extremely low signal-to-noise ratios (e.g. SNR $< -15$ dB) at onboard microphones. In this paper, we extensively assess the ability of single-channel deep learning approaches to ego-noise reduction on drones. We train twelve representative deep neural network (DNN) models, covering three operation domains (time-frequency magnitude domain, time-frequency complex domain and end-to-end time domain) and three distinct architectures (sequential, encoder-decoder and generative). We critically discuss and compare the performance of these models in extremely low-SNR scenarios, ranging from $-30$ to 0 dB. We show that time-frequency complex domain and UNet encoder-decoder architectures outperform other approaches on speech enhancement measures while providing a good trade-off with other criteria, such as model size, computation complexity and context length. The best-performing model is a UNet model operating in the time-frequency complex domain, which, at input SNR $-15$ dB, improves ESTOI from 0.1 to 0.4, PESQ from 1.0 to 1.9 and SI-SDR from $-15$ dB to 3.7 dB. Based on the insights drawn from these findings, we discuss future research in drone ego-noise reduction.

**INDEX TERMS** Deep learning, drone audition, ego-noise reduction, single-channel, speech enhancement.

## I. INTRODUCTION

Drone audition enables a flying robot to understand the surrounding acoustic environment from the sound captured by one or multiple onboard microphones [1]. The combination of acoustic signal analysis and the mobility of the drone benefits a wide range of applications including search and rescue, surveillance and monitoring, aerial filming, and human-drone interaction. However, the audition capability of the drone is severely limited by the strong ego-noise from the rotating motors and propellers. The motors and propellers are located much closer to onboard microphones than target sound sources on the ground or in the air, leading to extremely low signal-to-noise ratios at onboard microphones, e.g. SNR $< -15$ dB [2]. The spectrum of the ego-noise varies with the flight behaviour of the drone. Recovering the signal of interest from severely corrupted drone audio recording is a very challenging task.

Multi-channel microphone array approaches have been widely employed to improve the acoustic sensing performance on drones [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. While promising results have been reported, multi-channel techniques require customized microphone array hardware, which is large and heavy for mini-drones [21], [22], [23]. The spatial filtering performance of microphone-array techniques degrades remarkably in dynamic scenarios with moving microphones or sound sources. Developing efficient single-channel ego-noise reduction algorithms would encourage a wider application of drone audition. However, traditional signal processing approaches perform limitedly for single-channel noise reduction in low-SNR and nonstationary scenarios [24], [25], [26], [27].

In recent years, deep learning has revolutionized sound and speech processing [28]. Given a sufficient amount of training

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

data, deep neural networks (DNN) can learn to predict the clean speech from the noisy input, providing better speech enhancement performance than traditional signal processing approaches, especially when the noise is nonstationary and is well represented in the training data [29]. However, while deep learning based speech enhancement has been intensively investigated, a large body of research is targeting daily environments with a relatively high input SNR, e.g. above −5 dB [29]. The extension to the specific drone ego-noise reduction problem is still in the infant stage [30], [31]. We aim to understand how general DNN models perform when applied to the specific speech enhancement problem on drones.

We train twelve representative DNN models, which include *FC* [31] and *SMoLnet* [30], which were originally proposed for drone ego noise reduction, and ten general speech-enhancement models. The ten models were selected among the best performing with a variety of architectures and cover a range of typical designs for the time and time-frequency domains. These include three TF-complex masking models (*DCUNet* [32], *DCCRN* [33] and *PHASEN* [34]); two time-domain UNet mapping models (*WaveUNet* [35] and *Demucs* [36]), three time-domain TasNet masking models (*ConvTasNet* [37], *DPRNN* [38], *DPTNet* [39]), and two generative models (the time-domain mapping *SEGAN* model [40] and the TF-magnitude masking *VAE* model [41]). We train and evaluate these models in extremely low-SNR scenarios, ranging from −30 dB to 0 dB. We systematically analyze and extensively compare the performance of these DNN models with a list of objective measures, including speech enhancement performance, model size, computational complexity, and context length.

The contribution of the work is summarized in two folds.

- This is the first work that comprehensively evaluates the performance of deep learning models to single-channel ego-noise reduction on drones. The survey covers a wide range of representative models from three operation domains (TF-magnitude, TF-complex, and end-to-end time domain) and three distinct architectures (sequential, encoder-coder and generative). The experiment covers an extensive list of objective performance measures in extremely low-SNR scenarios with recorded drone ego-noise.
- The comparison results in extremely low-SNR scenarios suggest great potential from models working in the TF-complex domain and the UNet encoder-decoder architecture. Specifically, TF-complex models offer the best trade-off between speech enhancement performance and model size, followed by time-domain models and TF-magnitude models. Among the three types of architectures, encoder-decoder models achieve the best trade-off between speech enhancement performance and model size, followed by sequential models and generative models. These observations provide significant insights to future research in drone audition.

TF-complex models offer the best trade-off between speech enhancement performance and model size, followed by time-domain models, while TF-magnitude models perform the worst. TF-magnitude models require the shortest context length to make inference, while TF-complex and time-domain models require much larger context length. A better speech enhancement performance typically requires higher computational complexity and larger context length. Overall, TF-complex models show the highest potential for the ego-noise reduction problem.

Among the three types of architectures, encoder-decoder models achieve the best trade-off between speech enhancement performance and model size, followed by sequential models, while generative models perform the worst. Meanwhile, encoder-decoder models have the highest computational complexity. For encoder-decoder architectures, UNet models tend to outperform TasNet models with better speech enhancement performance, smaller model size and less computational complexity. Overall UNet models show the highest potential for the ego-noise reduction problem.

The paper is organized as follows. Section II surveys related work. Sections III-V present the principles of the twelve selected models in three categories: TF-magnitude, TF-complex, and time-domain. Sections VI and VII describe model training and evaluation results, respectively. Finally, conclusions are drawn in Section VIII.

## II. RELATED WORK

### A. DNN FOR GENERAL SPEECH ENHANCEMENT

DNN models for speech enhancement can be broadly categorized based on the operation domain (time and time-frequency (TF) domain) and the training target (mapping and masking). *TF-magnitude* approaches process the magnitude of the noisy signal in the time-frequency domain while keeping the phase unchanged. Mapping models map the noisy spectrum directly to clean magnitude spectra estimate [42], [43], [44], while masking models predict time-frequency masks which yield a clean spectrum estimate when applied to the noisy spectrum, e.g. the ideal binary mask (IBM) [45] or the ideal ratio mask (IRM) [46]. *TF-complex* approaches process both the magnitude and phase information of the noisy signal in the time-frequency domain. Mapping models either estimate the complex spectrum directly [30], or estimate the magnitude and phase information separately [34], [47], [48], while masking models estimate a complex ideal ratio mask (cIRM) [30], [32], [33]. *Time-domain* approaches focus on end-to-end processing and thus involve mapping models only, which predict clean speech waveforms directly from the noisy input in the time domain [49]. Due to the significance of the phase information for speech reconstruction in low-SNR scenarios [50], TF-complex and time-domain approaches typically outperform TF-magnitude ones at the cost of more complicated architectures and higher computational complexity.

Alternatively, the DNN models can be categorized based on the architecture, i.e. sequential, encoder-decoder, and

**TABLE 1.** Summary of the twelve models selected for the comparative study.

| Domain | Ref | Model | Architecture | | | | | Loss | Output | Model Size [$M$ parameters] |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Seq. | Enc.-Dec. | | Generative | | | | |
| | | | | UNet | TasNet | GAN | VAE | | | |
| Time-frequency Magnitude | [31] | Baseline | • | | | | | M-MSE | SMM | 10.51 |
| | [41] | VAE | | | | | • | M-MSE + KL | | 6.72 |
| Time-frequency Complex | [30] | SMoLnet | • | | | | | S-MSE | Complex spectra | 0.22 |
| | [32] | DCUNet | | • | | | | SI-SDR | cIRM | 3.53 |
| | [33] | DCCRN | | • | | | | | | 3.67 |
| | [34] | PHASEN | • | | | | | S-MSE | Phase + SMM | 8.59 |
| Time-domain | [35] | WaveUNet | | • | | | | MSE | Waveform | 10.13 |
| | [36] | Demucs | | • | | | | $L_1$ | | 18.87 |
| | [37] | ConvTasNet | | | • | | | SI-SDR | | 4.98 |
| | [38] | DPRNN | | | • | | | | | 3.64 |
| | [39] | DPTNet | | | • | | | | | 8.52 |
| | [40] | SEGAN | | • | | • | | $L_1$ + LSGAN | | 97.48 |

**KEY**
Seq.: Sequential; Disc.: Discriminative; SMM: spectral magnitude mask; [c]IRM: [Complex] Ideal ratio mask;
M-MSE: Mean squared error of the mask; S-MSE: Mean squared error of the speech signal.
KL: Kullback-Leibler divergence; SI-SDR: Scale-invariant signal-to-distortion ratio; LSGAN: Least-squares generative adversarial network loss;
VAE: Variational autoencoder; SMoLnet: Small Model on low-SNR; DCUNet: Deep complex U-net; DCCRN: Deep complex conv. recurrent network;
PHASEN: Phase and harmonics-aware speech enhancement network; Demucs: Deep extractor for music sources;
ConvTasNet: Convolutional time-domain audio separation network; DPRNN: Dual-path RNN; DPTNet: Dual-path Transformer network;
SEGAN: Speech enhancement generative adversarial network.

generative. The *sequential* architecture consists of multiple layers that process the input sequentially. This includes traditional DNN models, such as fully-connected [44], convolutional [51], and recurrent ones [52]. The *encoder-decoder* architecture includes distinct encoder and decoder blocks, which are usually symmetric to each other, and (optionally) a middle processing block. The encoder usually produces a compressed representation of the input, while the decoder decompresses the given representation and produces the output. UNet and TasNet are two well-known instances in this category, with the former widely deployed in TF-complex and time-domain approaches [53] and the latter providing benchmark performance for speech separation [54]. In comparison to sequential and encoder-decoder models, which aim to learn a discriminative transformation between noisy and the clean signal, the *generative* category is motivated by a different purpose that aims at learning the underlying distribution of the training data to generate new data points from the learned distribution with certain variations. Variational autoencoders (VAE) and generative adversarial networks (GAN) are two well-known generative approaches, where VAE aims at maximizing the lower bound of the data log-likelihood [41], [42] and GAN aims at achieving an equilibrium between the generator and discriminator [40], [55].

### B. DNN FOR EGO-NOISE REDUCTION

Drone ego-noise removal can be treated a special speech enhancement problem targeting at a specific type of noise. The drone ego-noise mainly consists of harmonic components, from the mechanical sound of the rotating motors, and full-band components, from the rotating propellers cutting the air. The pitch of the harmonic components is proportional to the rotating speed of the motors and is varying dynamically corresponding to the flight behaviour of the drone. The ego-noise is very strong and dominant in the microphone signal: the recording of a person talking aloud in front of the drone in a distance of 2 to 6 meters typically has SNR in the range of [-25, −10] dB [2]. The nonstationariy and the extremely low SNR make drone ego-noise removal a very challenging task. Fig. 1 illustrates an example of stationary and nonstationary ego-noise. A more detailed analysis on the spatial and spectral properties of the ego-noise was given in our previous work [2], [6].

To the best of our knowledge, only two DNN models have been specifically proposed for speech enhancement on drones [30], [31]. One work [31] employed a sequential architecture of fully connected DNN that performs single-channel ego-noise removal in the TF-magnitude domain. The estimated TF masks can be further incorporated into a multichannel spatial filtering framework. Another work [30] proposed a DNN model called *SMoLnet*, which follows a sequential architecture with dilated convolutional blocks in the frequency dimension, which can better capture the long-range harmonic correlations of the ego-noise. The SMoLnet model has three variants, TF-magnitude mapping, TF-complex masking and TF-complex mapping, where the last one gives the best ego-noise reduction performance. While the fully-connected model (which we call *Baseline*) has much lower computational complexity, the SMoLnet

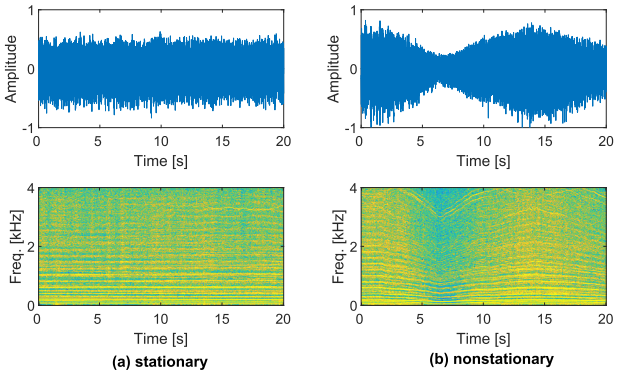



**FIGURE 1.** Time-domain waveform and time-frequency spectrum of the ego-noise. (a) Stationary ego-noise with a constant motor rotating speed. (b) Nonstationary ego-noise with a varying motor rotating speech.

model has a more compact architecture and also reports better results for ego-noise reduction.

We select twelve state-of-the-art DNN models for inclusion of our comparative study. These models cover the three working domains ( magnitude time-frequency domain, complex time-frequency domain, and time domain) and three types of architectures (sequential, encoder-decoder, and generative). Table 1 summarizes these models and their features.

- Two DNN models that were originally proposed for ego-noise removal: the fully-connected DNN [31], which is selected as the *Baseline*, and the *SMoLnet* model, which works in the complex time-frequency domain [30].
- Three masking models working in the complex time-frequency domain: using complex arithmetic (*DCUNet* [32]), *DCCRN* [33]) and real arithmetic (*PHASEN* [34]).
- Two UNet mapping models working in the time domain: *WaveUNet* [35] and *Demucs* [36].
- Three TasNet mapping models working in the time domain: *ConvTasNet* [37], *DPRNN* [38], *DPTNet* [39]).
- Two generative models: the time-domain mapping *SEGAN* model [40] and the magnitude TF-domain *VAE* model [41].

We employed several criteria when selecting DNN models for comparison. The first criteria was to cover a range of commonly encountered design patterns both for time- and TF-domain models for speech enhancement and/or separation. The second was to select among the well-established and best-performing models in each architectural subcategory. For example, WaveUNet and SEGAN are widely used for comparison in state-of-the-art works (e.g. [33], [34], [36]); DCUNet, DCCRN and Demucs are among the top-performing speech enhancement models on the Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) [56] and Deep Noise Suppression (DNS) Challenge [57]; PHASEN is the model designed to capture harmonic noise frequencies, which is likely to be useful for drone ego-noise removal; ConvTasNet, DPRNN and DPTNet are recognised state-of-the-art models for speech separation.

## III. TIME-FREQUENCY MAGNITUDE METHODS

Let $x(n) = s(n) + v(n)$ be the noisy signal, consisting of the mixture of a clean speech signal $s(n)$, and the drone noise $v(n)$, where $n$ is the sample index in the time domain. Let the signal in the time-frequency domain be represented as $X(k,l) = S(k,l) + V(k,l)$, where $k$ and $l$ are the frequency and frame indices, respectively. Time-frequency magnitude methods first estimate the magnitude of the clean speech $|\hat{S}(k,l)|$ from the noisy magnitude $|X(k,l)|$, and then reconstruct the spectrum of the clean speech using the noisy phase as

$$\hat{S}(k,l) = |\hat{S}(k,l)|e^{j\angle X(k,l)}, \tag{1}$$

where $\angle\cdot$ denotes the angle of a complex number.

The magnitude of the clean speech can be estimated with a mapping approach or a masking approach. In the first case, the DNN estimates a nonlinear function $f_M(\cdot)$ that can map the noisy magnitude to the clean magnitude:

$$|\hat{S}(k,l)| = f_M(|X(k,l)|). \tag{2}$$

In the second case, the DNN estimates a time-frequency mask $M(k,l)$ that can be used to recover the clean magnitude:

$$|\hat{S}(k,l)| = |X(k,l)|M(k,l). \tag{3}$$

$M(k,l)$ can be a spectral magnitude mask, which is defined as $\text{SMM}(k,l) = \min\left(\frac{|S(k,l)|}{|X(k,l)|}, 1\right)$, or an ideal ratio mask (IRM), which is defined as $\text{IRM}(k,l) = \frac{|S(k,l)|^2}{|S(k,l)|^2+|V(k,l)|^2}$. We consider two masking approaches: Baseline [31] and VAE [41].

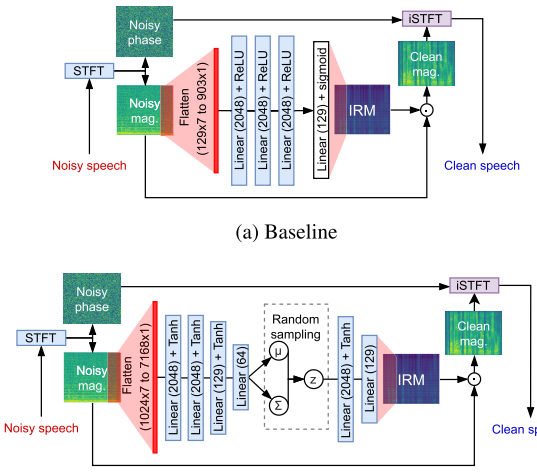



**FIGURE 2.** Time-frequency magnitude models selected for the study: (a) Baseline and (b) VAE. KEY –⊙: multiplication.

### A. BASELINE

Baseline [31] uses a fully connected DNN with an input layer, three hidden layers of size 2048 with ReLU activation, and an output layer with Sigmoid activation (Fig. 2). The signal is

processed frame-by-frame in the normalized log-magnitude domain, which is noted as $\tilde{X}(k, l)$. The input at the $l$-th frame is a spectrogram with a context window of radius $\Delta_L$:

$$\boldsymbol{X}_{\text{in}}(l) = \tilde{X}(1 : K, l - \Delta_L : l + \Delta_L). \tag{4}$$

The output of the network is the SMM estimated at the $l$-th frame:

$$\boldsymbol{M}_{out}(l) = M(1 : K, l). \tag{5}$$

The model is trained to minimize the mean-square error (MSE) loss between the estimated and the true SMM:

$$\mathcal{L}_{\text{baseline}} = \sum_{k,l} (M(k, l) - \text{SMM}(k, l))^2. \tag{6}$$

The network uses short-time Fourier transform (STFT) of size 256 with half overlap and sets context radius $\Delta_L = 3$, which results in an input vector of size 903 and an output vector of size 129.

### B. VAE

Similarly to Baseline, VAE (Fig. 2b) operates on normalized log-magnitude of the noisy spectra, processes the input spectrogram frame-by-frame using the context window of radius $\Delta_L = 3$ (Eq. 4) and outputs the SMM estimate at the given frame (Eq. 5). The same STFT parameters are used as for Baseline (size 256, half overlap). We train the variational autoencoder as a denoiser, using the noisy data as input and the clean speech as a target (this is an adaption of the method in [41]). The encoder maps the noisy input $\boldsymbol{X}_{\text{in}}(l)$ into the latent space, which can be characterized as a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The decoder maps the latent variable, which is randomly sampled from the latent space to the SMM mask. The encoder and decoder are feed-forward DNNs. The encoder has three hidden layers of sizes (2048, 2048, 196) with hyperbolic tangent (*tanh*) activations and a linear output layer of size 64. The latent variable $z$ thus has dimension 64. The decoder has one hidden layer of size 2048 with *tanh* activation and a linear output layer of size 129.

The loss function for model training is composed of the MSE loss between true and estimated SMM and the KL divergence between the distribution of the latent space and a standard normal distribution, which can be formulated as

$$\mathcal{L}_{\text{VAE}} = \sum_{k,l} (M(k, l) - \text{SMM}(k, l))^2 + \text{KL}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(0, \boldsymbol{I})). \tag{7}$$

During training, we use a reparametrization technique to make the gradient descent backpropagation possible despite the random sampling between the encoder and decoder [41].

## IV. TIME-FREQUENCY COMPLEX METHODS

TF-complex methods process the whole complex spectrum of the noisy signal $X(k, l)$ to estimate the complex spectrum of the clean signal $\hat{S}(k, l)$ with mapping, masking or hybrid approaches. In the first case, the DNN estimates a nonlinear function $f_C(\cdot)$ that can map the noisy spectrum to the clean spectrum:

$$\hat{S}(k, l) = f_C(X(k, l)). \tag{8}$$

In the second case, the DNN estimates a complex-valued time-frequency mask $M_C(k, l)$ that can be used to recover the clean spectrum from the noisy spectrum:

$$\hat{S}(k, l) = X(k, l) M_C(k, l). \tag{9}$$

Most commonly, given $X = X_r + jX_i$ and $S = S_r + jS_i$, $M_C(k, l)$ is a complex ideal ratio mask (cIRM) defined as $M_{\text{cIRM}} = \frac{X_r S_r + X_i S_i}{X_r^2 + X_i^2} + j\frac{X_r S_i + X_i S_r}{X_r^2 + X_i^2}$, where $(k, l)$ is omitted for brevity. In the third case, the hybrid approach estimates the magnitude IRM and the phase separately, and combines them into the complex spectrum. To operate in the time-frequency complex domain, the models can be implemented as real-valued and complex-valued ones. Real-valued models treat the complex input as two-channel real data, either using the cartesian (real and imaginary parts) coordinates [30], [34] or the polar (amplitude and angle) coordinates [58], [59]. Complex-valued models perform complex arithmetic on complex tensors directly [60].

We consider a mapping approach (SMoLnet [30]), two masking approaches (DCUNet [32] and DCCRN [33]), and a hybrid approach (PHASEN [34]).

### A. SMoLnet

SMoLnet is a mapping approach that estimates the complex spectrum of the clean speech directly from the noisy spectrum [30].[1] SMoLnet is a real-valued network that treats the real and imaginary components as separate channels. The network consists of fourteen convolutional layers: 10 dilated layers, 3 non-dilated layers and an output layer (Fig. 3b). The input STFT is calculated using a window length of 2048 and a half overlap. The 10 dilated layers aggregate information across the frequency dimension: they have kernel size (3, 1) and dilation factor $2(d - 1)$, where $d \in [1, 10]$ denotes the depth of the layer. The 3 non-dilated layers aggregate information across both time and frequency dimensions, with a kernel size (3, 3). All these 13 layers have 64 filters with ReLU activation. The output layer is a convolutional layer with two filters with kernel size (1, 1): one for real and one for imaginary channel. The input and output spectrograms of the model are represented as

$$T_{\text{in}}(k, l) = \begin{bmatrix} X_r(k, l) \\ X_i(k, l) \end{bmatrix}, \; T_{\text{out}}(k, l) = \begin{bmatrix} \hat{S}_r(k, l) \\ \hat{S}_i(k, l) \end{bmatrix}. \tag{10}$$

The model parameters are optimized by minimizing the MSE between the output $T_{\text{out}}(k, l)$ and the ground-truth spectrum $T_S(k, l) = \begin{bmatrix} S_r(k, l) \\ S_i(k, l) \end{bmatrix}$, i.e.

$$\mathcal{L}_{\text{SMoLnet}} = \sum_{k,l} |T_{\text{out}}(k, l) - T_{\text{s}}(k, l)|^2, \tag{11}$$

[1] We use SMoLnet-TCS, which performs the best among the three variations of SMoLnet [30].

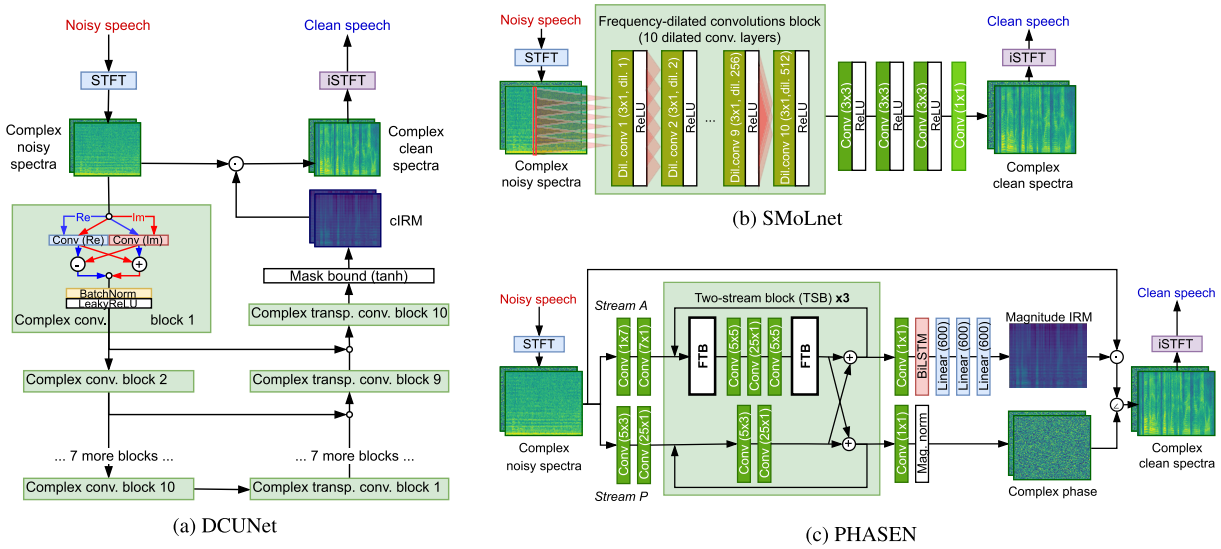



**FIGURE 3.** Time-frequency complex models selected for the study: (a) DCUNet, (b) SMoLnet, and (c) PHASEN. DCCRN, which is similar to DCUNet but has an LSTM block in the middle, is not shown. KEY – FTB: frequency transformation block; $\odot$: multiplication; $\oplus$: addition; $\angle$: magnitude and phase combination; $\circ$: concatenation.

where the 2-element vectors $T_{\text{out}}(k, l)$ and $T_S(k, l)$ are interpreted as complex numbers.

### B. DCUNet AND DCCRN

DCUNet [32] and DCCRN [33] are masking approaches that estimate the complex cIRM from the noisy spectrum. Both models are based upon UNet [53], which is a convolutional encoder-decoder structure with skip connections (Fig. 3). Both models are complex-valued networks that perform arithmetic operations directly in the complex domain [32].

DCUNet obtains the spectrogram via STFT with window 64 ms and skip size 16 ms. We use DCUNet-10, which contains 10 complex convolutional layers in the encoder and decoder parts, respectively. Each convolutional layer is followed by a batch normalization layer and a Leaky ReLU activation. The architecture of DCCRN is very similar to DCUNet. It contains 8 complex convolutional layers in the encoder and decoder, respectively. One key difference is that DCUNet connects the encoder and decoder directly, while DCCRN includes a two-layer LSTM block between the encoder and decoder, which aims to model the temporal dependencies between audio frames.

While targeting at estimating the complex cIRM, DCUNet and DCCRN are optimized using the time-domain loss function between the clean and the reconstructed speech. In the original papers, DCUNet [32] employs the weighted SDR loss while DCCRN [33] employs the SI-SDR loss. In this paper, we use SI-SDR for both models. For instance, the loss function of DCUNet is given by

$$\mathcal{L}_{\text{DCUNET}} = \text{SI-SDR}(s(n), \hat{s}(n)), \tag{12}$$

where for true and estimated clean signals $s$ and $\hat{s}$ the SI-SDR [61] is defined as $\text{SI-SDR} = 10\log_{10}\left(\frac{\left\|\frac{\hat{s}^T s}{\|s\|^2} s\right\|^2}{\left\|\frac{\hat{s}^T s}{\|s\|^2} s - \hat{s}\right\|^2}\right)$.

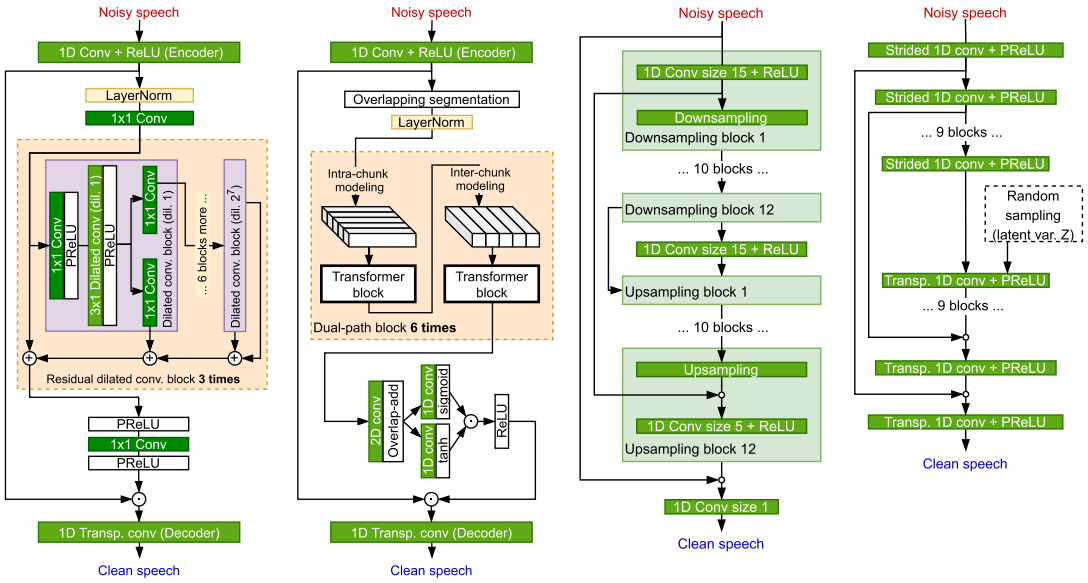
### C. PHASEN

PHASEN is a hybrid approach that estimates magnitude IRM and phase separately to reconstruct the complex spectrum of speech [34]. PHASEN is designed to capture the correlation between harmonic components of the acoustic signal and thus is suitable for ego-noise removal. The architecture is shown in Fig. 3c, which is featured with two-stream blocks (TSB): the amplitude stream (Stream A) in the upper portion estimates the magnitude mask and the phase stream (Stream P) estimates the complex phase. The model contains three TSBs, which are stacked sequentially. In Stream A of each TSB, two frequency transformation blocks (FTBs) are used to capture the harmonic correlation along the frequency dimension. At the end of each TSB, Stream A and Stream P exchange information, which is critical to phase estimation.

PHASEN is a real-valued model, which takes the whole complex STFT spectrogram as input, stacking real and imaginary parts in a way similar to SMoLnet, and outputs two channels the magnitude IRM and the complex phase (real and imaginary parts in two channels, and with the magnitude normalized to 1). The model parameters are optimized by minimizing a two-component MSE loss: power-compressed magnitude spectrum and complex spectrum. This is expressed as

$$\mathcal{L}_{\text{PHASEN}} = \sum_{k,l}(|\hat{S}_c(k,l)| - |S_c(k,l)|)^2 + \sum_{k,l}|\hat{S}_c(k,l) - S_c(k,l)|^2 \tag{13}$$

where $S_c = |S|^{0.3}e^{j\angle S}$ is the ground-truth spectrum with the magnitude compressed with power 0.3 (the same for $\hat{S}_c$).




(a) Conv-TasNet (b) DPTNet (c) WaveUNet (d) SEGAN's generator

**FIGURE 4. Time-domain models selected for the study: (a) Conv-TasNet, (b) DPTNet, (c) WaveUNet and (d) SEGAN. Two models are not shown are: Demucs is similar to WaveUNet, but has an LSTM block in the middle; DPRNN is similar to DPTNet, but uses BiLSTM blocks instead of Transformer blocks. KEY – ⊕: addition; ⊙: multiplication; ∘: concatenation.**

## V. TIME-DOMAIN METHODS

The methods operate on the noisy signal $x(n)$ in the time domain to estimate the clean speech, $\hat{s}(n)$. The DNN estimates a nonlinear function $f_T(\cdot)$ that maps the noisy signal to clean signal:

$$\hat{s}(n) = f_T(x(n)). \tag{14}$$

We consider three architecturally distinct groups, namely the UNet group (WaveUNet [35] and Demucs [36]), the TasNet group (ConvTasNet [37], DPRNN [38] and DPTNet [39]), and GAN (SEGAN [40]).

### A. WaveUNet AND DEMUCS

Similar to DCUNet and DCCRN, WaveUNet and Demucs are UNet-based, but they work with raw waveforms using one-dimensional convolutions. The two models were originally designed for music separation [62], [63] and later adopted for speech enhancement [35], [36]. Fig. 4c depicts the architecture of WaveUNet, where dimensionality reduction/increasing in encoder/decoder is implemented via downsampling/upsampling. A convolutional layer is placed between the encoder and decoder. WaveUNet contains 12 layers in the encoder/decoder part. The number of convolutional filters varies linearly with the depth of each layer with a constant factor $C = 24$. WaveUNet takes the whole noisy audio segment as input and outputs the estimated speech. The model is trained by minimizing the MSE loss:

$$\mathcal{L}_{\text{WaveUNet}} = \sum_n |s(n) - \hat{s}(n)|^2. \tag{15}$$

Demucs shares a similar architecture as WaveUNet. The main difference is that Demucs uses an LSTM layer to connect the encoder and decoder. Demucs contains 5 layers in the encoder/decoder part, and the number of convolutional filters in each layer varies exponentially with a factor of 2. As a result, the size of the Demucs model is almost twice of WaveUNet. Demucs is trained using L1 loss instead of MSE.

### B. ConvTasNet, DPRNN AND DPTNet

The three models were originally designed for speech separation [37], [38], [39] and we adopt them for speech enhancement. The three models are all based on TasNet [54], an encoder-decoder architecture consisting of three components: encoder, masker, and decoder (Fig. 4(a)-(b)). The encoder network performs 1D convolution on the waveform to generate a learnt feature representation $\mathcal{Z}$, which is fed to the masker network to generate a speech mask in the learnt representation space. The mask is element-wise multiplied with $\mathcal{Z}$ to obtain the speech representation, which is passed to the decoder network, which employs 1D transposed convolution to generate a clean speech estimate.

Conv-TasNet uses a masker network which consists of a series of residual one-dimensional dilated convolutional blocks (Fig. 4a). DPTNet's masker features a series of dual-path networks (DPN) (Fig. 4b). Each DPN consists of intra- (DPNintra) and inter- (DPNinter) processing Transformer blocks to model the local and global dependencies between processing chunks. DPRNN shares a similar architecture as DPTNet but uses bidirectional LSTM layers (BiLSTMs) instead of Transformers. All three TasNet networks take the noisy signal in the time domain as input and output the estimated speech. The models are trained by minimizing the SI-SDR loss (Eq. 12).

### C. SEGAN

As one of the first generative adversarial networks applied to speech enhancement, SEGAN [40] consists of two networks: generator ($G$) and discriminator ($D$). The generator is used to estimate the clean speech $\hat{s}(n)$ from the noisy input $x(n)$, i.e. $\hat{s}(n) = G(x(n))$. The discriminator is a binary classifier that aims to distinguish between the true clean signals and the ones produced by the generator: $\begin{cases} [l]D(s,x) = 1 \\ D(G(x),x) = 0 \end{cases}$. The generator aims to fool the discriminator by producing samples which cannot be distinguished from true clean signals, so that $D(G(x),x) = 1$. The two networks are trained alternatively, competing with each other until both converged. The objective functions to be minimized during this joint adversarial learning can be formulated as

$$\begin{cases} \mathcal{L}_G = \mathcal{E}[(D(G(x),x)-1)^2] + \lambda\mathcal{E}[(s-G(x))^2] \\ \mathcal{L}_D = \mathcal{E}[D(G(x),x)^2] + \mathcal{E}[(D(s,x)-1)^2], \end{cases} \tag{16}$$

where $\lambda$ denotes the weight of regularization loss, and $\mathcal{E}$ denotes expectation on the training set.

Fig. 4d depicts the SEGAN generator network, which is a UNet architecture consisting of strided 1D convolutional layers followed by PReLU activations. Both encoder and decoder contain 11 layers, each with kernel size 31 and stride 2. The encoder part only takes noisy signal $x(n)$ as input; the decoder part takes additional input from random samples with prior distribution $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The SEGAN discriminator is similar to the encoder part of the generator and is finalized with a $1 \times 1$ convolutional layer followed by a fully connected layer and sigmoid activation, which enables it to produce a binary decision.

## VI. MODEL TRAINING

We use the clean speech from the TIMIT corpus [64] and the drone ego-noise from the AVQ [11] and AS [2] datasets (Table 2). AVQ and AS contain 8-channel recordings of real drone noise, which were recorded via a microphone array mounted on a drone, which is fixed on a tripod. During recording, the drone either operates at constant power ranging from 50% to 150% of the hovering motor speed or changes the hovering speed dynamically. We only use the first channel of each recording in the experiments. The TIMIT dataset consists of a training subset with 4620 utterances (230 minutes) and a test subset with 1680 utterances (84 minutes). All the audio samples are resampled at 8 kHz. From these datasets, we generate noisy speech for training, validation, and testing.

We use 90% (4158 utterances) of the TIMIT train subset and the five AVQ ego-noise sequences (*n116-n120*) to generate the training set. For each TIMIT utterance, a clean speech segment of length $T$ is randomly cropped from the utterance and a noise segment of the same length is randomly cropped from the ego-noise, which are mixed at a given SNR, which is uniformly sampled from the interval [-25, −5] dB. Given a clean speech signal $s(n)$ and noise $v(n)$, the SNR of the mixture is defined as $\text{SNR} = 10\log_{10}\frac{\sum_{n=1}^{T} s^2(n)}{\sum_{n=1}^{T} v^2(n)}$. The training set is generated on the fly, i.e. the model will see different noisy mixtures every epoch. We define one epoch to be 10 full iterations over clean speech utterances, which corresponds to approximately 35 hours of training data encountered per epoch. We use the same random seed across all our experiments so that each model sees the same pieces of data in the same order during training.

We employ a similar protocol to construct the validation set, using the remaining 10% (462 utterances) of the TIMIT train subset and the five AVQ ego-noise sequences to generate the validation set. One difference is that the noisy mixtures are generated with a fixed setup (not on the fly), and the clean speech and the noise are added at an SNR selected from {-25, −20, −15, −10, −5} dB. In total, we generate 9.5 hours of validation data at the default crop length of 3 seconds.

We use all the 1680 utterances of the TIMIT test subset and the two AS ego-noise sequences (*n121-n122*) to construct the testing set. Each utterance (full length) is mixed with a noise segment randomly cropped from the two ego-noise sequences, at an SNR selected from {-30, −25, −20, −15, −10, −5, 0} dB. Here we use a wider range of testing SNRs to analyze how well the models behave for SNRs outside of the training range. In total, we generate 20 hours of testing data (23520 clips).

**TABLE 2. Drone ego-noise used in the experiments.**

| Dataset | ID | Noise power | Length [s] |
|---|---|---|---|
| AVQ [11] (training) | n116 | constant (50%) | 120 |
| | n117 | constant (100%) | 120 |
| | n118 | constant (150%) | 40 |
| | n119 | constant (100%) | 210 |
| | n120 | dynamic | 214 |
| AS [2] (testing) | n121 | constant (100%) | 130 |
| | n122 | dynamic | 140 |

**TABLE 3. Training parameters of all the models. $T$: crop length [samples]; $\alpha$: initial learning rate; $N_\alpha$: learning rate reduction patience [epochs]; $N_E$: early stopping patience [epochs]; $B$: batch size.**

| Model | $T$ | $\alpha$ | $N_\alpha$ | $N_E$ | $B$ |
|---|---|---|---|---|---|
| Baseline | 24000 | $10^{-4}$ | 5 | 10 | 32 |
| VAE | 24000 | $10^{-4}$ | 5 | 10 | 32 |
| SMoLnet | 10240 | $10^{-4}$ | 3 | 10 | 32 |
| DCUNet | 24000 | $10^{-3}$ | 15 | 30 | 32 |
| DCCRN | 24000 | $10^{-3}$ | 15 | 30 | 32 |
| PHASEN | 24000 | $10^{-3}$ | 15 | 30 | 32 |
| WaveUNet | 16384 | $10^{-3}$ | 15 | 30 | 32 |
| Demucs | 24149 | $3 \times 10^{-4}$ | 15 | 30 | 32 |
| ConvTasNet | 24000 | $10^{-3}$ | 15 | 30 | 8 |
| DPRNN | 24000 | $10^{-3}$ | 15 | 30 | 8 |
| DPTNet | 24000 | $10^{-3}$ | 5 | 10 | 8 |
| SEGAN | 16384 | $2 \times 10^{-4}$ | – | – | 256 |

Table 3 specifies the parameters used by each model during training. We use fixed-length audio clips for ease of mini-batching processing. We set $T = 24000$ samples by default, which corresponds to 3 seconds of 8 kHz audio, because most

utterances in the TIMIT dataset have a duration of around 3 seconds. However, we change the value of $T$ for some models whose architecture and/or original training setup demand it. For instance, both WaveUNet and Demucs require a particular input length to produce the output of the same length after internal down-sampling and up-sampling operations; for SMoLnet, we use the original setup $T = 10240$ [30].

We use Adam optimizer to train all the models, except SEGAN. We employ early stopping to finish training if there is no improvement in validation loss for 10 consecutive epochs. During training, the learning rate is multiplied by 0.1 after no improvement in validation loss for $N_\alpha$ consecutive epochs. The default values for initial learning rate $\alpha$ and $N_\alpha$ are $10^{-3}$ and 5, respectively, but those values also differ for some models as shown in Table 3. We use a batch size of 32 for all models for which GPU memory limits allow it, and use a smaller batch size of 8 where it is impossible.

SEGAN consists of a pair of models which compete during training and the losses observed during their training are prone to oscillations. Approaches like on-plateau learning rate reduction and early stopping are hardly applicable. Hence, we employ a different training setup for SEGAN, as instructed in the original paper [40], with one difference: we do not use a high-frequency preemphasis as a preprocessing step. Both generator and discriminator are trained using RMSprop optimizer with learning rate $2 \times 10^{-4}$, with crop length $T = 16384$ and batch size 256 are used. The model is trained for 300 epochs, where the discriminator and the generator are updated one after another for each batch of data.

## VII. EVALUATION RESULTS

We evaluate the speech enhancement performance of the models using perceptual evaluation of speech quality (PESQ) [65], extended short-time objective intelligibility measure (ESTOI) [66] and scale-invariant signal-to-distortion ratio (SI-SDR) [61]. PESQ is a psychoacoustics-based metric designed to predict the subjective mean opinion score (MOS) of speech. It is one of the most popular metrics when evaluating speech enhancement performance [56], [57]. ESTOI is a modification of STOI [67], which is a widely used metric to predict the intelligibility of the speech [36], [55]. ESTOI was originally shown to correlate well (including better than STOI) with subjective intelligibility tests on low-SNR data [66]. SI-SDR is the signal energy measure equivalent to SNR but invariant to the absolute amplitude of the estimated signal, and unlike the two previous metrics, it is not specific to speech signals. We also compare the size of the model, in terms of the number of trainable parameters, the training time, and the inference speed of the models. Additionally, we analyze the lengths of input audio segments that are used to produce a given sample of output audio (i.e. the context length of a model).

### A. SPEECH ENHANCEMENT

We split the twelve models into four groups: TF-magnitude, TF-complex, time-domain UNet and time-domain TasNet models. We discuss the results per group and then compare the best models from each group.

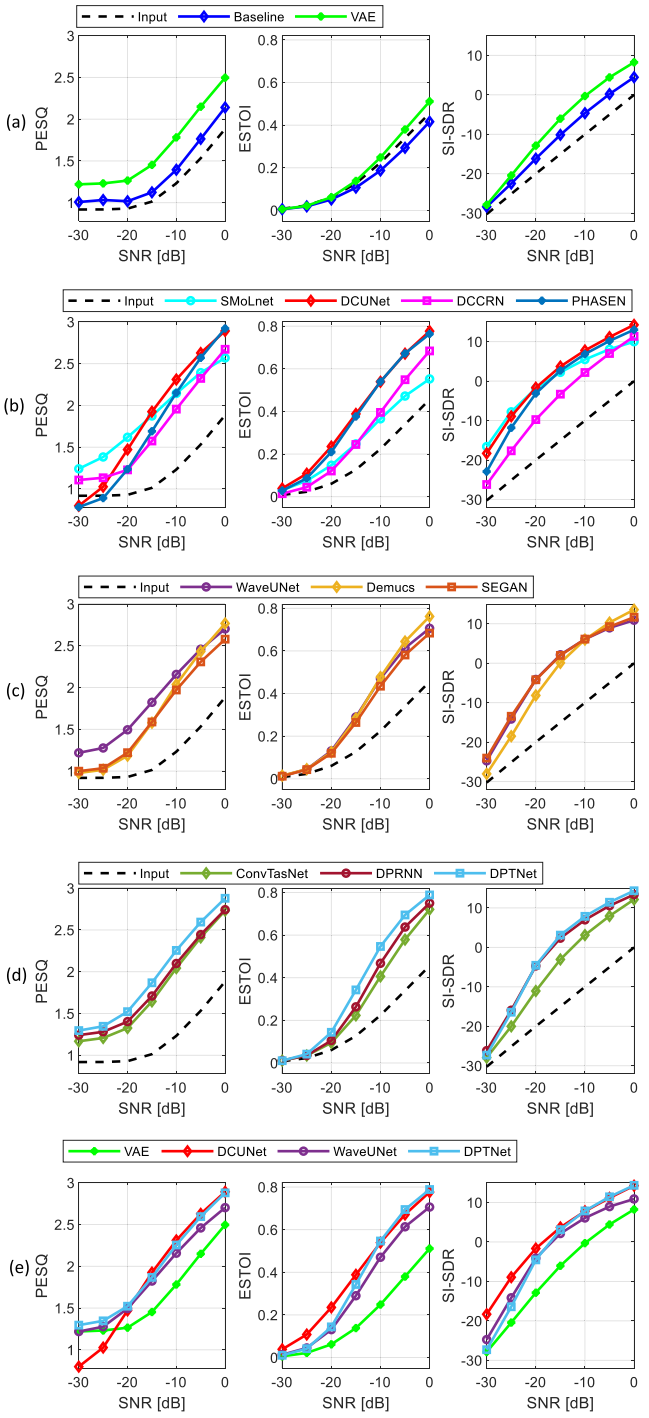



**FIGURE 5.** Speech enhancement performance of all the twelve models, which are split into four groups: (a) TF-magnitude models; (b) TF-complex models; (c) time-domain UNet models; (d) time-domain TasNet models. The leading models in each group are compared in (e).

#### 1) TF-MAGNITUDES MODELS

In Fig. 5(a), VAE shows consistent performance improvement over Baseline. The latter even shows negative improvement

over the noisy input in terms of the ESTOI and SI-SDR measures.

#### 2) TF-COMPLEX MODELS

In Fig. 5(b), DCUNet consistently outperforms all other models, except for PESQ at SNR $\leq -20$ dB. The closest runner-ups are PHASEN, which shows very similar performance overall for ESTOI and SI-SDR, but much lower PESQ than DCUNet. SmoLnet achieves the highest PESQ at SNR $\leq -15$ dB; however, it achieves the lowest ESTOI at most SNRs $\geq -20$ dB. DCCRN performs the worst for all three measures and at most SNRs. Overall, DCUNet ranks as the best-performing model in the TF-complex group.

#### 3) TIME-DOMAIN UNet MODELS

In Fig. 5(c), WaveUNet outperforms other models for all three measures (especially PESQ) at SNR $\leq -10$ dB. Demucs performs slightly better than WaveUNet for the three measures at SNR $\geq -5$ dB. SEGAN performs the worst for PESQ and ESTOI, but performs slightly better than Demucs for SI-SDR at SNR $\leq -10$ dB.

#### 4) TIME-DOMAIN TasNet MODELS

In Fig. 5(d), DPTNet consistently outperforms DPRNN, while ConvTasNet performs the worst for all three measures.

Finally, Fig. 5(e) compares the best-performing models from each group, i.e. VAE, DCUNet, DPTNet, and WaveUNet. VAE is outperformed by other models for all three measures, except at SNR $\leq -20$ dB for the PESQ score. DCUNet outperforms other models, in terms of ESTOI and SI-SDR, especially when the SNR is lower than $-10$ dB. However, DCUNet performs relatively badly in terms of PESQ and is outperformed by other models especially when the SNR is lower than $-20$ dB. The closet runner is DPTNet, which achieves higher PESQ but lower ESTOI and SI-SDR than DCUNet at SNR $\leq -20$ dB. WaveUNet performs slightly worse than DPTNet for all three measures.

Table 4 summarizes the performance improvement averaged for input SNR [-25, $-10$] dB.[2] Overall, the speech enhancement performance of the four groups can be ranked as TF-complex > T-TasNet > T-UNet > TF-magnitude. SMoLnet achieves the highest PESQ improvement while DCUNet achieves the highest ESTOI and SI-SDR improvement. Finally, Fig. 6 illustrates the time-frequency spectrogram of a segment of clean speech, noisy signal at $-15$ dB, and the output by the twelve DNN models.[3]

### B. MODEL SIZE

Fig. 7(a) compares the model size (number of trainable parameters) and the speech enhancement performance. We choose the ESTOI measure because it is closely related to speech intelligibility and we compute the ESTOI improvement averaged at SNR [$-25$, $-10$] dB of each model. The upper-left corner of the plot indicates a better trade-off between speech enhancement and model size. The twelve models are grouped based on their operation domain and architecture: TF-FC, TF-CNN, TF-UNet, T-UNet, T-TasNet, T-VAE and T-GAN. The operation domain is indicated with the color and the architecture is indicated with the marker.

We first compare the models based on their operation domain. TF-complex models show the best trade-off, followed by T-TasNet models and T-UNet models, whereas TF-magnitude models show the worst trade-off. We then

[2]We consider this SNR range as those SNR values are more commonly encountered in the outdoor drone audio recordings which include a human speaking at the distance from 2 to 6 meters from the drone [2].

[3]More audio samples available at http://www.eecs.qmul.ac.uk/~linwang/demo/single-dnn.html

**TABLE 4. Mean performance improvement over the noisy input by each model. The evaluation score is averaged over the SNR range [-25, −10] dB. The twelve models are split into four groups. The highest values per group and measure are underlined. The highest values for each measure are in bold.**

| Group | Model | ΔPESQ | ΔESTOI | ΔSI-SDR [dB] |
|---|---|---|---|---|
| TF-magnitude | Baseline | 0.119 | -0.018 | 4.169 |
| | VAE | 0.409 | 0.009 | 7.621 |
| TF-complex | SMoLnet | **0.731** | 0.100 | 17.043 |
| | DCUNet | 0.660 | **0.209** | **17.751** |
| | DCCRN | 0.450 | 0.093 | 10.395 |
| | PHASEN | 0.469 | 0.194 | 16.173 |
| Time-domain UNet | WaveUNet | 0.664 | 0.125 | 15.020 |
| | Demucs | 0.430 | 0.124 | 12.388 |
| | SEGAN | 0.430 | 0.107 | 15.147 |
| Time-domain TasNet | ConvTasNet | 0.532 | 0.081 | 9.785 |
| | DPRNN | 0.599 | 0.109 | 14.705 |
| | DPTNet | 0.724 | 0.160 | 15.037 |

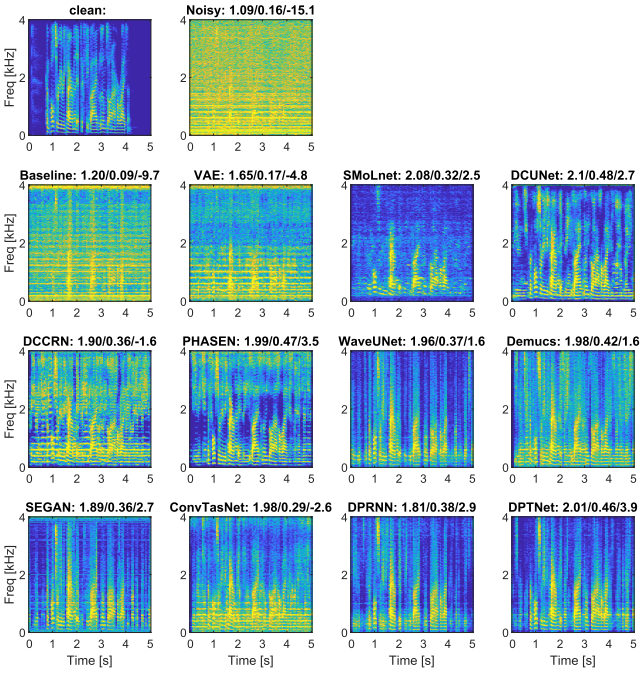



**FIGURE 6. Time-frequency spectrogram of the clean speech, the noisy signal, and the output of the DNN models. The input SNR is −15 dB. The title of each panel indicates 'Model: PESQ/ESTOI/SI-SDR'.**

compare the models based on their architectures. TF-UNet models show the best trade-off, followed by TF-CNN models, T-TasNet models, and T-UNet models, where TF-FC, TF-VAE and T-GAN show the worst trade-off. The two TF-UNet models (DCUNet, DCCRN) show similar model sizes but varying speech enhancement performance. The two TF-CNN models (SMoLnet, PHASEN) vary in both speech enhancement and model size. The three T-TasNet models (ConvTasNet, DPTNet, DPRNN) perform similarly, with minor differences in speech enhancement and model size. The two T-UNet models (WaveUNet, Demucs) show similar speech enhancement performance but varying model sizes. TF-FC (Baseline) and TF-VAE (VAE) perform limited in terms of speech enhancement, whereas T-GAN (SEGAN) has a very large model size.

Among the twelve models, DCUNet achieves the highest ESTOI improvement with the second smallest model size (3.53M), while SMoLnet achieves a medium ESTOI improvement with a very compact model (0.22M). SEGAN, being the biggest model (94.78M), achieves a similar ESTOI improvement as SMoLnet.

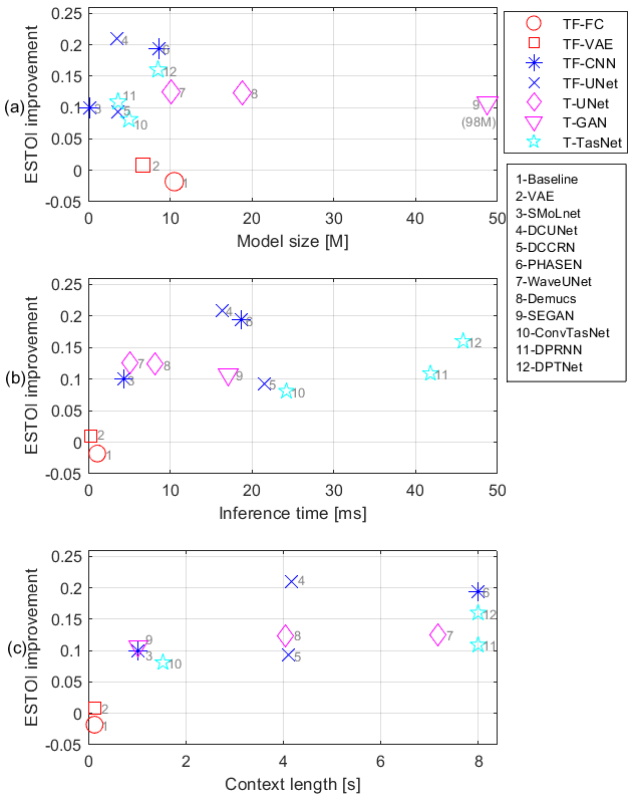



**FIGURE 7. Juxtaposition of each model comparing speech enhancement performance and (a) model size; (b) inference speed; (c) context length.**

### C. TRAINING TIME AND INFERENCE SPEED

All the models are trained on Tesla V100 GPU using the setup specified in Sec. VI. We measure the training time based on the GPU clock time. We also measure the inference speed of each model when processing an audio clip with 16384 samples, which roughly corresponds to 2 seconds at a sampling rate 8 kHz. The inference speed is measured for end-to-end processing, e.g. including the STFT time for time-frequency methods. We use GTX 1080 Ti GPU for the computation, where the mean GPU clock time among 1000 runs is measured. Table 5 shows the training time and inference speed of each model. Fig. 7(b) compares the inference speed and the speech enhancement performance. The upper-left corner of the plot indicates a better trade-off between speech enhancement and model size.

We first compare the models based on their operation domains. TF-complex models show the best trade-off, followed by T-UNet models and T-TasNet models, whereas TF-magnitude models show the worst trade-off. We further compare the models based on their architectures. The two TF-UNet models (DCUNet, DCCRN) vary in both speech enhancement and inference speed, with DCUNet outperforming in both measures. The two TF-CNN models (SMoLnet, PHASEN) vary in both speech enhancement and model size, with PHASEN achieving better speech enhancement performance and SMoLnet showing faster inference speed. The two T-UNet models (WaveUNet, Demucs) show similar speech enhancement performance but varying inference speeds. SEGAN achieves similar speech enhancement performance but slower inference speed than T-UNet models. The three T-TasNet models (ConvTasNet, DPTNet, DPRNN) vary in both speech enhancement performance and inference speed, with ConvTasNet performing the best in terms of inference speed while DPTNet performs the best in terms of speech enhancement. TF-FC (Baseline) and TF-VAE (VAE) use the least time for inference, but achieve the worst speech enhancement performance. Noticeably, T-TasNet models have smaller model sizes but slower inference speeds, whereas T-UNet models behave inversely. This leads to swapped rank in Fig. 7(a) and (b).

Similar observations can be made for the training time, which is proportional to the inference time. The T-TasNet models take the longest time for training, followed by TF-complex models and T-UNet models, whereas TF-magnitude models take the least time for training. One exception is DCUNet, which takes less inference time than DCCRN, but much longer training time than the latter.

### D. TEMPORAL CONTEXT WINDOW

Table 6 summarizes the maximum lengths of future and past input contexts used by each model to generate any given sample of the output. For Baseline and VAE, the fixed context window radius of 3 STFT frames was used and translated to the number of samples and seconds; models with recurrent or Transformer layers (DCRNN, PHASEN, Demucs, DPRNN, DPTNet) use all the available context (future and/or past, depending on the directionality of recurrent layers); in other cases (SMoLnet, DCUNet, DCRNN, WaveUNet, Demucs, SEGAN, Conv-TasNet), the rule for computing the receptive field of convolutional networks is applied to compute

**TABLE 5.** Training and inference time for each model. The models are trained on Nvidia V100 GPU. The models are tested on GTX 1080 Ti for an audio segment of 2 seconds.

| Model | Training time [h] | Inference time [ms] |
|---|---|---|
| Baseline | 2.5 | 1.1 |
| VAE | 0.7 | 0.3 |
| SMoLnet | 46.2 | 4.3 |
| DCUNet | 140.0 | 16.3 |
| DCCRN | 90.1 | 21.5 |
| PHASEN | 44.4 | 18.7 |
| WaveUNet | 13.5 | 5.1 |
| Demucs | 38.4 | 8.2 |
| SEGAN | 42.5 | 17.1 |
| ConvTasNet | 140.4 | 24.2 |
| DPRNN | 140.9 | 41.8 |
| DPTNet | 140.5 | 45.8 |

**TABLE 6.** Maximum past and future context used by the models.

| Model | Past | | | Future | | |
|---|---|---|---|---|---|---|
| | STFT frames | Samples | Time[s] | STFT frames | Samples | Time[s] |
| Baseline | 3 | 512 | 0.064 | 3 | 512 | 0.064 |
| VAE | 3 | 512 | 0.064 | 3 | 512 | 0.064 |
| SMoLnet | 3 | 4096 | 0.512 | 3 | 4096 | 0.512 |
| DCUNet | 58 | 16640 | 2.080 | 58 | 16640 | 2.080 |
| DCCRN | All available | | | 6 | 896 | 0.112 |
| PHASEN | All available | | | All available | | |
| WaveUNet | — | 28656 | 3.585 | — | 28656 | 3.585 |
| Demucs | — | All available | | — | 298 | 0.037 |
| SEGAN | — | 30784 | 3.848 | — | 30784 | 3.848 |
| ConvTasNet | — | 6128 | 0.766 | — | 6128 | 0.766 |
| DPRNN | — | All available | | — | All available | |
| DPTNet | — | All available | | — | All available | |

the context length [68]. Models with causal LSTM blocks (DCCRN, Demucs) use all the past information available. Models with BiLSTM or Transformer blocks (PHASEN, DRNN, DPTNet) use all the past and future information available. Since the maximum length of the TIMIT audio clips used for evaluation does not exceed 8 seconds, we set the past context length to 4 seconds for models using all the past information available and set the future context length to 4 seconds for models using all the future information available.

Fig. 7(c) compares the context length and the speech enhancement performance. There seems to be no obvious relationship between the context length and the operation domain of the model. TF-magnitude models (Baseline, VAE) have the shortest context windows, but also perform the worst. TF-magnitudes, T-UNet and T-TasNet models all have substantially varying context lengths. However, for models with similar architecture, the speech enhancement performance tends to increase with the context length. For instance, among the three T-TasNet models, DPRNN and DPTNet achieve better speech enhancement performance with a larger context window than ConvTasNet. Among the two TF-CNN models, PHASEN has a larger context window than SMoLnet, and also better speech enhancement performance.

In general, a large amount of context is helpful for speech enhancement performance, but could lead to restrictions on online processing. It is worthwhile noting that there is some ongoing research on casual TF and casual TasNet models that can be deployed for online processing by utilizing the previous state of the models [69].

### E. DISCUSSION

TF-complex models significantly outperform TF-magnitude models by additionally estimating phase information, which is important for speech intelligibility in low-SNR scenarios [70]. TF-complex models also show better speech enhancement performance than time-domain models. The two-dimensional convolutional layers in TF-complex models seem to capture the temporal context more efficiently than one-dimensional convolutional layers of time-domain models or (Bi)LSTM or Transformer layers. The two-dimensional convolutional kernels in TF-complex models can detect speech/noise patterns along the frequency dimensions more efficiently, which helps to separate speech and noise. For instance, both speech and ego-noise have distinctive harmonic structures in the time-frequency domain [6].

Sequential FC models perform limited in low-SNR scenarios due to their simple architecture. Sequential CNN models perform much better than FC models by working in the 2D time-frequency domain. Depending on specific architectures, the sequential CNN models vary in terms of speech enhancement, model size and computational complexity. For instance, SMoLNet achieves medium-level speech enhancement performance with a compact model and low computational complexity, whereas PHASEN achieves better speech enhancement performance with a larger model and higher computational complexity.

Encoder-decoder models, including UNet and TasNet, take the majority of the twelve models, achieving medium- or top-level speech enhancement performance. UNet models are deployed in both TF-domain and time-domain processing. TF-UNet models achieve a good balance between speech enhancement and model size and computational complexity. DCUNet achieves the highest speech enhancement performance among all twelve models. T-UNet models have larger model sizes but less computational complexity than TF-UNet models. This implies the significant computational advantage of one-dimensional convolutions used by time-domain models over two-dimensional UNet architectures and the complex operations in the time-frequency domain. T-TasNet models achieve comparable speech enhancement performance with T-UNet models. T-TasNet models have smaller model sizes than T-UNet models, but more computational complexity. DPRNN and DPTNet are the slowest among the twelve models. This is possibly due to their dual-path processing techniques that introduce redundancy in the data flow. For instance, using the default chunking parameters (chunking with half overlap) essentially results in processing each input audio sample twice. The employment of BiLSTMs (DPRNN)

and Transformers (DPTNet) further increases the computational complexity.

The good performance of UNet architecture for our task might be explained as follows: while the encoder extracts high-level compressed features of the input, the skip connections between encoder and decoder give the decoder access to more high-fidelity information from the input lost in the compressed representations (e.g. small individual variations in the spectrum corresponding to speech in the noise background). This behaviour might be beneficial for speech enhancement in low-SNR scenarios, where high-level compressed features represent global properties of speech or noise (e.g. drone motors' speed) and high-fidelity information corresponds to subtle variations in the spectrum which indicate the presence of the speech (e.g. the phase discrepancies between the noisy and clean signals).

Among the two generative models, VAE has very low computational complexity but performs limitedly for speech enhancement; SEGAN achieves medium-level speech enhancement performance with an extremely large model. Due to special architecture and training strategy, generative models appear more suitable for other applications than ego-noise reduction.

## VIII. CONCLUSION

We evaluated and discussed single-channel DNN approaches to drone ego-noise reduction in extremely low-SNR scenarios. We trained twelve models that cover three operation domains (TF-magnitude, TF-complex and time-domain) and three types of architectures (sequential, encoder-decoder and generative), and evaluated model performance in terms of speech enhancement, model size, model complexity, and context length.

TF-complex models offer the best trade-off between speech enhancement performance and model size, followed by time-domain models, while TF-magnitude models perform the worst. TF-magnitude models require the shortest context length to make inference, while TF-complex and time-domain models require much larger context lengths. A better speech enhancement performance typically requires higher computational complexity and larger context length. Overall, TF-complex models show the highest potential for the ego-noise reduction problem.

Among the three types of architectures, encoder-decoder models achieve the best trade-off between speech enhancement performance and model size, followed by sequential models, while generative models perform the worst. Meanwhile, encoder-decoder models have the highest computational complexity. For encoder-decoder architectures, UNet models tend to outperform TasNet models with better speech enhancement performance, smaller model size and less computational complexity. Overall UNet models show the highest potential for the ego-noise reduction problem.

As a result, DCUNet, a UNet operating in the time-frequency complex domain, achieves the highest ESTOI score, the highest SI-SDR and the fourth PESQ. For instance, at input SNR $-15$ dB, DCUNet improves ESTOI from 0.1 to 0.4, PESQ from 1.0 to 1.9 and SI-SDR from $-15$ dB to 3.7 dB.

The work is based on the comparison of twelve DNN models. It would be interesting to include more types of architecture for comparison. For instance, the TF magnitude model is based on a simple fully-connected architecture, it would make sense to use more advanced architecture for predicting the magnitude mask. More recent GAN models [55] or speech enhancement diffusion models [71] could be adapted for the drone ego-noise reduction task.

We recommend three main research directions. First, while the considered deep learning models have promising results at input SNR $-15$ dB, the performance drops quickly when the input SNR is further decreased, which can be lower than $-30$ dB if the human speaker is far from the drone. Developing a better deep learning model that works in this extremely challenging scenario is necessary. Given the good performance of time-frequency UNet architectures in our comparative study, further optimizing on this might lead to favorable solutions.

Second, drone audition applications require real-time and onboard processing, which imposes restrictions on the model size, computational complexity and context length. A better deep learning model with a good balance between speech enhancement performance and other criteria is necessary. For instance, the algorithm's future context length (Table 6) should be not more than a few milliseconds to avoid latency detectable by a human listener (e.g. $\leq 40$ ms by rules of DNS Challenge real-time track [57]). Thus, for models that involve a large future context window (e.g. DCUNet), optimization on the window length is necessary.

Third, the considered deep learning models were evaluated with only one type of drone noise. In real applications, the ego-noise from various types of drones sounds different. Investigating and improving the generality of deep learning models is an important direction.

## REFERENCES

[1] K. Nonami, K. Hoshiba, K. Nakadai, M. Kumon, H. G. Okuno, Y. Tanabe, K. Yonezawa, H. Tokutake, S. Suzuki, K. Yamaguchi, and S. Sunada, "Recent R&D technologies and future prospective of flying robot in tough robotics challenge," in *Disaster Robotics*. Springer, 2019, pp. 77–142.

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

[3] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 3943–3948.

[4] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-board relative bearing estimation for teams of drones using sound," *IEEE Robot. Automat. Lett.*, vol. 1, no. 2, pp. 820–827, Feb. 2016.

[5] 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 Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.

[6] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 152–158.

[7] 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. (ICASSP)*, Mar. 2017, pp. 496–500.

[8] 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.

[9] R. Sanchez-Matilla, L. Wang, and A. Cavallaro, "Multi-modal localization and enhancement of multiple sound sources from a micro aerial vehicle," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1591–1599.

[10] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Tracking a moving sound source from a multi-rotor drone," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 2511–2516.

[11] 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. Robots Syst. (IROS)*, Nov. 2019, pp. 5320–5325.

[12] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Beamforming-based acoustic source localization and enhancement for multirotor UAVs," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 987–991.

[13] 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.

[14] Y. Hioka, M. Kingan, G. Schmid, R. McKay, and K. A. Stol, "Design of an unmanned aerial vehicle mounted system for quiet audio recording," *Appl. Acoust.*, vol. 155, pp. 423–427, Dec. 2019.

[15] B. Yen, Y. Hioka, and B. Mace, "Source enhancement for unmanned aerial vehicle recording using multi-sensory information," in *Proc. Asia–Pacific Signal, Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 850–857.

[16] A. B. A. Qayyum, K. M. N. Hassan, A. Anika, M. F. Shadiq, M. M. Rahman, M. T. Islam, S. A. Imran, S. Hossain, and M. A. Haque, "DOANet: A deep dilated convolutional neural network approach for search and rescue with drone-embedded sound source localization," *EURASIP J. Audio, Speech, Music Process.*, vol. 2020, no. 1, pp. 1–18, Dec. 2020.

[17] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2523–2537, 2020.

[18] L. Wang and A. Cavallaro, "Deep-learning-assisted sound source localization from a flying drone," *IEEE Sensors J.*, vol. 22, no. 21, pp. 20828–20838, Nov. 2022.

[19] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 508–519, 2022.

[20] E. Tengan, T. Dietzen, S. Ruiz Sanchez, M. Alkmim, J. Cardenuto, and T. Van Waterschoot, "Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle," in *Proc. Int. Congr. Acoust.*, 2022, pp. 1–11.

[21] T. Ishiki and M. Kumon, "A microphone array configuration for an auditory quadrotor helicopter system," in *Proc. IEEE Int. Symp. Saf., Secur., Rescue Robot.*, Oct. 2014, pp. 1–6.

[22] T. Ishiki and M. Kumon, "Design model of microphone arrays for multirotor helicopters," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 6143–6148.

[23] M. Clayton, L. Wang, A. McPherson, and A. Cavallaro, "An embedded multichannel sound acquisition system for drone audition," 2021, *arXiv:2101.06795*.

[24] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[25] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[26] B. M. Mahmmod, A. R. Ramli, S. H. Abdulhussain, S. Al-Haddad, and W. A. Jassim, "Low-distortion MMSE speech enhancement estimator based on Laplacian prior," *IEEE Access*, vol. 5, pp. 9866–9881, 2017.

[27] B. M. Mahmmod, A. R. Ramli, T. Baker, F. Al-Obeidat, S. H. Abdulhussain, and W. A. Jassim, "Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials," *IEEE Access*, vol. 7, pp. 103485–103504, 2019.

[28] H. Purwins, B. Li, T. Virtanen, J. Schlüter, 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, May 2019.

[29] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[30] Z.-W. Tan, A. H. T. Nguyen, and A. W. H. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 1885–1892.

[31] L. Wang and A. Cavallaro, "Deep learning assisted time-frequency processing for speech enhancement on drones," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 6, pp. 871–881, Dec. 2021.

[32] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–20.

[33] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," 2020, *arXiv:2008.00264*.

[34] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 9458–9465.

[35] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-U-Net," 2018, *arXiv:1811.11307*.

[36] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, Oct. 2020, pp. 3291–3295.

[37] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[38] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 46–50.

[39] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. InterSpeech*, 2020, pp. 2642–2646.

[40] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," 2017, *arXiv:1703.09452*.

[41] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1788–1800, 2020.

[42] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, Aug. 2013, pp. 436–440.

[43] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4628–4632.

[44] Y. Xu, J. Du, L.-R. R. Dai, and C.-H. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. Audio. Speech. Lang. Process.*, vol. 23, no. 1, pp. 7–19, May 2015.

[45] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA, USA: Kluwer, 2005, pp. 181–197.

[46] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7092–7096.

[47] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "PhaseNet: Discretized phase modeling with deep neural networks for audio source separation," in *Proc. Interspeech*, Sep. 2018, pp. 2713–2717.

[48] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 286–290.

[49] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[50] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

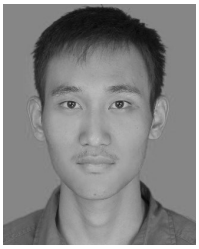
[51] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," 2016, *arXiv:1609.07132*.

[52] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. INTERSPEECH*, Sep. 2012, pp. 22–25.
[53] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[54] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 696–700.
[55] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2031–2041.
[56] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, vol. 19, no. 1. Montreal, QC, Canada: Acoustical Society of America, 2013, Art. no. 035081.
[57] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," 2020, *arXiv:2001.08662*.
[58] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.
[59] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 63–76, Jan. 2019.
[60] L. Drude, B. Raj, and R. Haeb-Umbach, "On the appropriateness of complex-valued neural networks for speech enhancement," in *Proc. Interspeech*, Sep. 2016, pp. 1745–1749.
[61] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—Half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 626–630.
[62] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-net: A multi-scale neural network for end-to-end audio source separation," 2018, *arXiv:1806.03185*.
[63] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," 2019, *arXiv:1909.01174*.
[64] 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, Nat. Inst. Standards Technol. (NIST), Gaithersburg, MD, USA, 1988.
[65] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, May 2001, pp. 749–752.
[66] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[67] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[68] A. Araujo, W. Norris, and J. Sim, "Computing receptive fields of convolutional neural networks," *Distill*, vol. 4, no. 11, p. e21, Nov. 2019.
[69] Y. Liu and D. Wang, "Causal deep CASA for monaural talker-independent speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2109–2118, 2020.
[70] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.
[71] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7402–7406.

**DMITRII MUKHUTDINOV** received the B.Sc. and M.Sc. degrees in informatics and applied mathematics from ITMO University, Saint Petersburg, Russia, in 2017 and 2019, respectively, and the M.Sc. degree in artificial intelligence from The University of Edinburgh, in 2020. He is currently pursuing the Ph.D. degree in computer science with the Centre for Intelligent Sensing, Queen Mary University of London (QMUL). His main research interests include artificial intelligence, signal processing, and audio processing.

**ASHISH ALEX** received the bachelor's degree in electronics and communication from VIT University, India, in 2016, and the master's degree in computer systems and networks from the University of Greenwich, U.K., in 2018. He is currently pursuing the Ph.D. degree with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. His research interests include machine learning, deep learning, and audio signal processing.

**ANDREA CAVALLARO** received the Ph.D. degree in electrical engineering from Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He is currently a Professor of multimedia signal processing and the Founding Director of the Centre for Intelligent Sensing, Queen Mary University of London, a Turing Fellow with The Alan Turing Institute, U.K. National Institute for Data Science and Artificial Intelligence, and a fellow of the International Association for Pattern Recognition. He is the Editor-in-Chief of *Signal Processing: Image Communication*; a Senior Area Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING; the Chair of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee; and an IEEE Signal Processing Society Distinguished Lecturer.



**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 Dalian University of Technology, China, in 2010. From 2011 to 2013, he was an Alexander von Humboldt Fellow with the University of Oldenburg, Germany. From 2014 to 2017, he was a Postdoctoral Researcher with the Queen Mary University of London, U.K. From 2017 to 2018, he was a Postdoctoral Researcher with the University of Sussex, U.K. Since 2018, he has been a Lecturer with the Queen Mary University of London. His research interests include audio–visual signal processing, machine learning, and robotic perception. He is an Associate Editor of IEEE ACCESS and IEEE SENSORS JOURNAL.