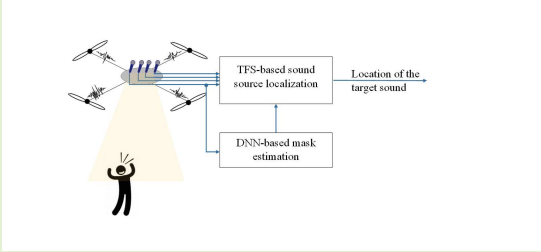
Deep learning assisted sound source localization from a flying drone

Lin Wang, Andrea Cavallaro

***Abstract*— Sound source localization from a flying drone is a challenging task due to the strong ego-noise from rotating motors and propellers as well as the movement of the drone and the sound sources. To address this challenge, we propose a deep learning-based framework that integrates single-channel noise reduction and multi-channel source localization. In this framework we suppress the ego-noise and estimate a time-frequency soft ratio mask with a single-channel deep neural network (DNN). Then we design two downstream multi-channel source localization algorithms, based on Steered Response Power (SRP-DNN) and Time-Frequency Spatial filtering (TFS-DNN). The main novelty lies in the proposed TFS-DNN approach, which estimates the presence probability of the target sound at individual time-frequency bins by combining the DNN-inferred soft ratio mask and the instantaneous direction of arrival of the sound received by the microphone array. The time-frequency presence probability of the target sound is then used to design a set of spatial filters to construct a spatial likelihood map for source localization. By jointly exploiting spectral and spatial information, TFS-DNN robustly processes signals in short segments (e.g. 0.5 seconds) in dynamic and low signal-noise-ratio scenarios (e.g. SNR -20 dB). Results on real and simulated data in a variety of scenarios (static sources, moving sources and moving drones) indicate the advantage of TFS-DNN over competing methods, including SRP-DNN and the state-of-the-art time-frequency spatial filtering.**



***Index Terms*— Deep neural network, drone audition, ego-noise reduction, microphone arraysound source localization**

I. Introduction

Drone audition has been attracting a growing interest because of the increasing availability of agile multi-rotor aerial vehicles [1], [2]. Drone audition techniques enable a drone to localize, enhance, and understand sound sources in the surrounding environment through the analysis of the sound captured by airborne microphones for application in search and rescue [3]–[6], monitoring [7]–[9], recreational filming [10], human-drone interaction and robot autonomy [11], [12]. The performance of acoustic sensing from drones degrades significantly mainly due to two challenges: the strong ego-noise from rotating motors and propellers, which leads to extremely low signal-to-noise ratio (SNR), and the movement of the drone and sound sources, which leads to highly dynamic acoustic transmission paths [13], [14].

Drone-based sound source localization approaches are uni-modal or multi-modal. Uni-modal approaches, such as spatial likelihood methods and spatial filtering methods, use the microphone signal only. Spatial likelihood methods are based on traditional source localization algorithms, such as steered response power (SRP) and steered response power with phase transform (SRP-PHAT) [3], [15]–[18], and multiple signal classification (MUSIC) [9], [19], [20]. While being widely used for ground robot audition, the performance of these algorithms typically degrades with drone platforms due to the strong ego-noise and hence the low SNR [21]. Spatial filtering methods were recently proposed for drone sound processing [22]–[25]: spatial filters steered at multiple candidate locations and sound source localization is based on comparing the spatial filtering output (e.g. Kurtosis) at these locations. By designing the spatial filter in the time-frequency domain, the spatial filtering method can suppress the ego-noise effectively and thus works robustly in low-SNR scenarios. While promising results have been reported, the spatial filtering method typically requires a certain amount of audio data to estimate the statistical information of the sound and design the spatial filters. This leads to degraded performance when signals are processed in short segments, and thus it is not efficient when dealing with moving drones and sound sources. Since the locations of the motors and propellers are fixed, some work proposed to minimize the influence of the ego-noise by mounting the microphone array far from the drone body using an extension pole [5], [26] or rope [16]. While these approaches reduce the effect of the ego-noise, the requirement of additional hardware reduces the versatility of the drone and hence of drone audition applications.

Multi-modal approaches use additional sensors mounted on drones to improve the sound source localization performance. Since the ego-noise is generated by the motors and propellers, speed sensors can be used to monitor the motor rotation speed

Manuscript received: September 15, 2022
The authors are with Centre for Intelligent Sensing, Queen Mary University of London, London, UK (e-mail: {lin.wang, a.cavallaro}@qmul.ac.uk)

and predict the ego-noise. The predicted ego-noise is further incorporated into existing source localization algorithms to improve the robustness to ego-noise [26], [27]. Computer vision algorithms can exploit onboard cameras to localize candidate sound sources (e.g. humans) and thus guide where to steer spatial filters for sound enhancement [28]–[30].

Deep learning, which has revolutionized audio and speech processing [31]–[33], has also found applications in drone audition, in particular for sound enhancement. When appropriate training data are available, deep neural networks (DNN) can learn to predict the clean speech from noisy recordings in low-SNR scenarios (e.g. using fully connected DNN [34] and SMoLNet [35]). In addition to promising results for speech enhancement, recent years saw the emergence of deep learning applied to sound source localization. DNN sound enhancement is typically a pre-processing step for traditional source localization algorithms [36]. While a DNN can also be trained to predict the location of the sound source directly from the multi-channel microphone signal, the performance typically drops significantly in low-SNR scenarios [37].

In this paper, we propose a deep learning-based sound source localization framework that integrates single-channel ego-noise suppression and multi-channel source localization. The proposed approach addresses the challenges of strong ego-noise and can deal with the movement of drones and sound sources. First, we estimate the soft ratio mask at individual time-frequency bins from the noisy signal using single-channel DNN models, and define a baseline sound source localization system, namely SRP-DNN. Since a single-channel DNN can enhance the noisy signal robustly in case of sensor and source movement, SRP-DNN achieves better performance than the original SRP-PHAT algorithm in low-SNR scenarios. Second, we propose a TFS-DNN sound source localization pipeline that incorporates the DNN-estimated soft ratio mask into the state-of-the-art time-frequency spatial filtering (TFS) approach [23]. By jointly exploiting the spectral and spatial information, TFS-DNN can better estimate the presence probability of the target sound at individual time-frequency bins and thus provide better performance when processing sounds in short segments and low-SNRs. We evaluate the performance in different scenarios with static sources, moving sources, and moving drones (Fig. 1). We also investigate the impact of various DNN ego-noise suppression models on the proposed pipeline.

The paper is organized as follows. After formulating the problem in Section II, we introduce in Section III the DNN model for the estimation of the time-frequency mask. In Section IV we propose the two DNN-based sound source localization algorithms. Section V covers the experimental results and analysis. Finally, conclusions are drawn in Section VI.

II. Problem Definition

Let a microphone array mounted on a quadcopter platform consist of I microphones placed in an arbitrary shape (e.g. a circular array in Fig. 1(a)-(c) or a cubic array in Fig. 1(d)).

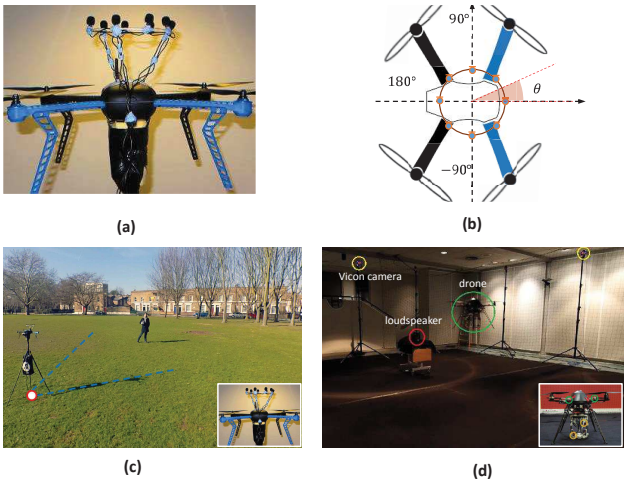


Fig. 1. Drone audition hardware. (a)(b) The 2D coordinate system for a circular array mounted on a drone. (c) A static drone with a moving speaker (image from AVQ [29]). (d) A flying drone with a static speaker (image from DREGON [38]).

Without loss of generality, we consider a 2D coordinate system (Fig. 1(b)) and denote the locations of the microphones as $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m, \ldots, \boldsymbol{r}_I]$, where $\boldsymbol{r}_m = [r_{mx}, r_{my}]^{\mathrm{T}}$ is the position of the m-th microphone, and the superscript $(\cdot)^{\mathrm{T}}$ denotes the transpose operator.

A target source in the far field emits sound with a time-varying direction of arrival (DOA), $\theta_d(n)$, with respect to the microphone array, where n is the time index. The time-varying DOA is caused by the movement of the sound source (Fig. 1(c)) and/or the movement of the microphones with the flying drone (Fig. 1(d)).

The signal from the array, $\boldsymbol{x}(n) = [x_1(n), \ldots, x_I(n)]^{\mathrm{T}}$, captures the target sound, $\boldsymbol{s}(n) = [s_1(n), \ldots, s_I(n)]^{\mathrm{T}}$, as well as the noise, $\boldsymbol{v}(n) = [v_1(n), \ldots, v_I(n)]^{\mathrm{T}}$. The noise term is dominated by the drone ego-noise, and might also contain other components such as the wind noise (from the propellers and the environment) and microphone self-noise in practice. The ego-noise mainly consists of multiple narrow-band harmonic components, which are caused by the rotating rotors, and broadband noise, which is caused by the propellers cutting air [14]. Since the motors and propellers are close to onboard microphones, the ego-noise is typically stronger than the target sound and the other noise components. We express the microphone signal in the time domain as

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

and in the time-frequency domain as

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

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.

Given $\boldsymbol{x}(n)$ and $\boldsymbol{R}$, the goal is to estimate the time-varying DOA $\theta_d(n)$ of the sound source despite the very low SNR (that can be lower than -15 dB) and the movement of the drone or sound sources, which creates a challenging dynamic acoustic transmitting path.

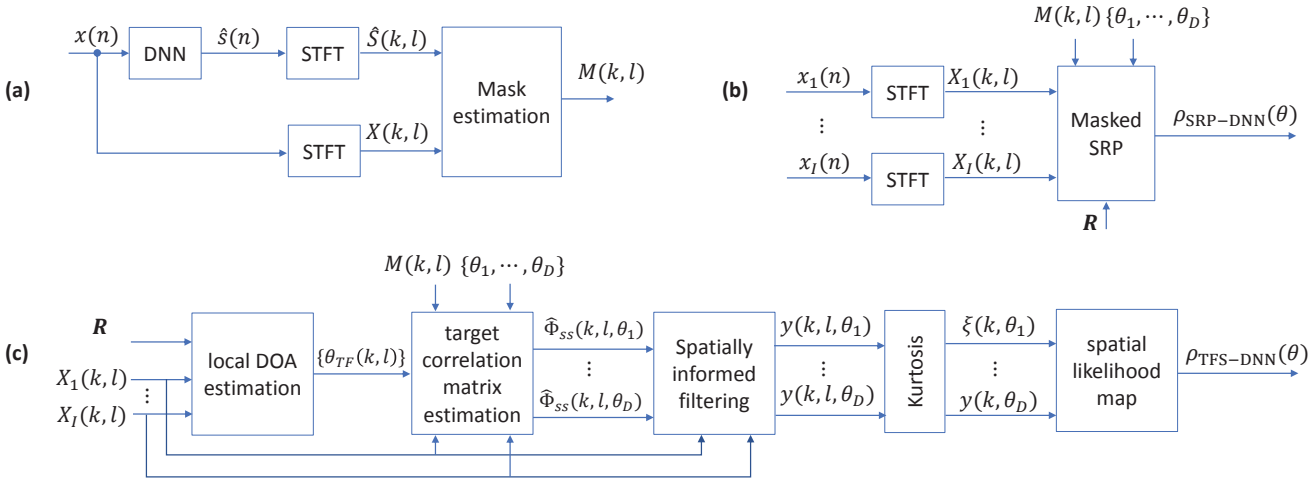


Fig. 2. The proposed deep learning assisted sound source localization framework. (a) Single-channel sound enhancement DNN model that estimates a soft ratio mask, which is then integrated in two multi-channel sound source localization pipelines: (b) SRP-DNN and (c) TFS-DNN. SRP-DNN employs time-frequency weighting when computing the spatial likelihood map. TFS-DNN employs time-frequency weighting when computing the target correlation matrix and the spatial filter. The Kurotis value of the spatial filtering output is used to indicate the spatial likelihood of the target sound.

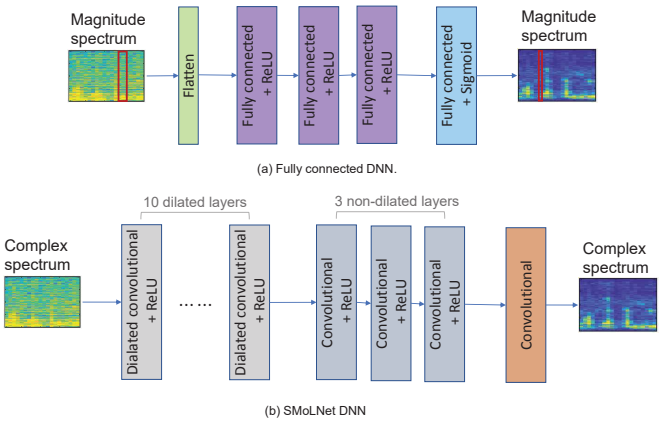


Fig. 3. Two DNN architectures for single-channel drone noise reduction. (a) Fully-connected DNN. (b) SMoLNet DNN.

III. Mask Estimation

Fig. 2 shows the proposed framework, which consists of two steps. In the first step (Fig. 2(a)), we employ a single-channel DNN model to enhance the noisy recording and estimate a soft ratio mask in the time-frequency domain. Based on the estimated mask, in the second step, we propose two sound source localization algorithms, which are referred to as SRP-DNN (Fig. 2(b)) and TFS-DNN (Fig. 2(c)). We describe mask estimation in this section and then source localization in the next section.

A. DNN sound enhancement

Single-channel DNN models are suitable for preprocessing the noisy sound captured on flying drones, due to their robustness to sensor and source movement. We employ single-channel DNN models to enhance the noisy recording, after which we estimate the soft ratio mask. We consider two existing approaches that were originally proposed for ego-noise reduction and speech enhancement on drones: fully connected (FC) DNN [34] and SMoLNet DNN [35]. Fig. 3 depicts the architectures of the two DNN models.

The first DNN model (FC) operates in the time-frequency magnitude domain to estimate the magnitude of the clean speech [34]. After converting the time-domain signal into the short-time Fourier transform (STFT) domain, the DNN model takes as input seven neighboring magnitude frames and generates as output one frame of estimated magnitude. As shown in Fig. 3(a), the DNN architecture simply consists of one flatten layer, three fully-connected layers, and one output layer. The flatten layer converts the neighboring STFT frames into an input vector; the three fully connected layers each consist of 2048 neurons; the output layers generate one STFT frame of magnitude estimation. Finally, the time-domain speech is reconstructed by combining the estimated magnitude and the original noisy phase.

The second DNN model (SMoLNet) operates in the time-frequency complex domain to estimate the complex spectrum of the clean speech [35]. SMoLNet treats the real and imaginary components of the complex spectrum as separate channels. As shown in Fig. 3(b), the DNN architecture consists of ten dilated convolutional layers, three non-dilated convolutional layers, and one output layer. The ten dilated layers aggregate information across the frequency dimension: they have kernel size (3, 1) and dilation factor 2(d-1), where d denotes the depth of the dilated layer. The three non-dilated layers aggregate information across both time and frequency dimensions, with a kernel size (3, 3). The output layer consists of a convolutional layer with two kernel sizes (1, 1), corresponding to the real and imaginary components of the complex spectrum. SMoLNet takes as input the whole STFT complex spectrum, and generates a complex spectrum of the same size, which can be converted to the time domain as clean speech estimation.

A detailed training procedure for the two DNN models will be given in Sec. V. The FC model has a simple architecture and lower computational complexity, while the SMoLNet model has a compact architecture and better ego-noise suppression performance by operating in the time-frequency complex domain (see Sec. V-D). We consider both DNN models in our proposed framework in order to investigate the impact of various DNN models on the performance of the system.

B. Mask estimation

Fig. 2(a) depicts the detailed steps of mask estimation. Let us consider the signal captured at one microphone as $x(n) = s(n)+v(n)$ in the time domain and $X(k,l) = S(k,l)+V(k,l)$ in the STFT domain.

Suppose we already trained a single-channel DNN model (either FC or SMoLNet) that can enhance the microphone signal as

$$\hat{s}(n) = \text{DNN}(x(n)). \tag{3}$$

We convert the noisy and the enhanced signal into the STFT domain as $\hat{S}(k,l)$ and $X(k,l)$, and estimate the signal-noise ratio at each time-frequency bin as

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

Lying in the range $[0,1]$, this ratio can be used to indicate the speech-presence probability at each time-frequency bin.

To help understanding, Fig. 4(a) illustrates a sample of mask estimation result, including the noisy signal $X(k,l)$, clean speech $S(k,l)$, DNN enhanced speech $\hat{S}(k,l)$ and the estimated mask $M(k,l)$. We use SMoLNet as the DNN sound enhancement model. For input SNR -5 dB, the DNN model can improve the SNR effectively by about 11 dB. The estimated mask $M(k,l)$ appears consistent with the energy distribution of the clean speech $S(k,l)$ in the time-frequency domain, which implies that the mask is good to measure the speech presence probability.

IV. Sound source localization

In this section, we show how we incorporate the DNN-estimated soft ratio mask into two sound source localization frameworks, namely SRP-DNN and TFS-DNN.

A. SRP-DNN for sound source localization

The traditional SRP-PHAT algorithm localizes a sound source by computing a spatial likelihood function from the microphone array signal as [39]

$$\boxed{\rho_{\text{SRP-PHAT}}(\theta) = \Re\left\{\sum_{k,l}\sum_{\substack{m_1,m_2=1\\ m_1\neq m_2}}^{I}\frac{X^*_{m_1}(k,l)X_{m_2}(k,l)}{|X_{m_1}(k,l)X_{m_2}(k,l)|}\mathrm{e}^{-j2\pi f_k\tau(m_1,m_2,\theta)}\right\}} \tag{5}$$

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 $\tau(m_1,m_2,\theta)$ denotes the delay between two microphones m_1 and m_2 with respect to the sound coming from a candidate direction $\theta \in \{\theta_1,\cdots,\theta_D\}$. SRP-PHAT in (5) is essentially a time-frequency implementation using GCC-PHAT functions [40]. Alternatively, it can be implemented as delay-an-sum beamforming with phase transform. Here we use the time-frequency implementation as it is more straightforward to extend to the SRP-DNN algorithm, as explained subsequently.

In (5), all the time-frequency bins are treated equally when computing the spatial likelihood map. If the time-frequency bins are weighted with the presence probability of the target sound, the computation of the spatial likelihood map will be more robust to noise [40], [41]. Based on this idea, we propose the SRP-DNN algorithm that computes the spatial likelihood map as

$$\boxed{\rho_{\text{SRP-DNN}}(\theta) = \Re\left\{\sum_{k,l}M(k,l)\sum_{\substack{m_1,m_2=1\\ m_1\neq m_2}}^{I}\frac{X^*_{m_1}(k,l)X_{m_2}(k,l)\mathrm{e}^{-j2\pi f_k\tau(m_1,m_2,\theta)}}{|X_{m_1}(k,l)X_{m_2}(k,l)|}\right\}} \tag{6}$$

where $M(k,l)$ is the DNN-estimated ratio mask (4), indicating the speech presence probability at each time-frequency bin.

Finally, from the spatial likelihood function, the location of the sound source is estimated as the location of the highest peak, i.e.

$$\theta_{\text{SRP-DNN}} = \underset{\theta\in\{\theta_1,\cdots,\theta_D\}}{\arg\max}\ \{\rho_{\text{SRP-DNN}}(\theta)\} \tag{7}$$

B. TFS-DNN for sound source localization

The time-frequency spatial filtering (TFS) algorithm is a recent algorithm for sound source localization on drones [22]. The basic idea is to formulate a set of spatial filters pointing at candidate directions, $\theta \in \{\theta_1,\cdots,\theta_D\}$, and use the kurtosis of the spatial filtering outputs to indicate the spatial likelihood of the target sound. In this section, we propose a new algorithm, TFS-DNN, that incorporates the DNN-estimated ratio mask into the time-frequency spatial filtering algorithm to further improve the source localization performance. The detailed computation procedure is given below.

Given the microphone signal $\boldsymbol{x}(k,l)$ and the microphone locations $\boldsymbol{R}$, we first estimate the instantaneous DOA of the sound at each time-frequency bin. This is achieved by computing a local spatial likelihood function as

$$\gamma_{\text{TF}}(k,l,\theta) = \Re\left\{\sum_{\substack{m_1,m_2=1\\ m_1\neq m_2}}^{I}\frac{X_{m_1}(k,l)X^*_{m_2}(k,l)}{|X_{m_1}(k,l)X_{m_2}(k,l)|}\mathrm{e}^{j2\pi f_k\tau(m_1,m_2,\theta)}\right\}, \tag{8}$$

from which the DOA of the sound at each time-frequency bin is computed as

$$\theta_{\text{TF}}(k,l) = \underset{\theta\in(-180^\circ,180^\circ]}{\arg\max}\ \gamma_{\text{TF}}(k,l,\theta). \tag{9}$$

Next, we define a confidence measure at each time-frequency bin that the target speech comes from the direction θ, i.e.

$$\boxed{\tilde{c}_d(k,l,\theta) = c_d(k,l,\theta)M(k,l).} \tag{10}$$

The confidence consists of two parts. The first term $c_d(k,l,\theta)$ measures the closeness of each time-frequency bin (k,l) to the direction θ, and is defined as [22]

$$c_d(k,l,\theta) = \exp\left(-\frac{(\theta_{\text{TF}}(k,l)-\theta)^2}{2\sigma^2}\right), \tag{11}$$

where we assume the DOA estimate to be Gaussian-distributed with mean θ and standard deviation σ. Lying in the range $[0,1]$, the higher $c_d(\cdot)$, the closer of the local DOA to the direction θ. The second term $M(k,l)$ is the ratio mask estimated by the DNN model, which indicates the speech presence probability at each time-frequency bin. Combining the spatial information and the spectral information, the confidence measure $\tilde{c}_d(k,l,\theta)$ can better indicate the probability that the sound at the (k,l)-th bin arrives from the direction θ.

Next, we calculate an $I \times I$ target correlation matrix of the direction θ as

$$\boxed{\mathbf{\Phi}_{ss}(k,l,\theta) = \frac{1}{L}\sum_{l=1}^{L}\tilde{c}_d(k,l,\theta)\boldsymbol{x}^{\text{H}}(k,l)\boldsymbol{x}(k,l),} \tag{12}$$

where $\tilde{c}_d(k,l,\theta)$ indicates the contribution of each time-frequency bin to the target correlation matrix. With this target correlation matrix, we can formulate a spatial filter pointing at the direction θ. We use a standard Multi-channel Wiener filter (MWF) that is defined as [42]

$$\boldsymbol{w}_{\text{TF}}(k,l,\theta) = \mathbf{\Phi}_{xx}^{-1}(k,l)\boldsymbol{\phi}_{ss1}(k,l,\theta). \tag{13}$$

The MWF filter can be estimated directly from the microphone signal: $\boldsymbol{\phi}_{ss1}(k,l,\theta)$ is the first column of $\mathbf{\Phi}_{ss}(k,l,\theta)$, and $\mathbf{\Phi}_{xx}(k,l)$ is the correlation matrix of the microphone signal, which can be estimated directly using $\mathbf{\Phi}_{xx}(k,l) = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{x}(k,l)^{\text{H}}\boldsymbol{x}(k,l)$. The sound coming from the direction θ is extracted as

$$y_{\text{TF}}(k,l,\theta) = \boldsymbol{w}_{\text{TF}}^{\text{H}}(k,l,\theta)\boldsymbol{x}(k,l). \tag{14}$$

We calculate the kurtosis value $\xi(k,\theta)$ of the time sequence in each frequency bin:

$$\xi(k,\theta) = \mathcal{K}(\tilde{\boldsymbol{y}}_{\text{TF}}(k,\theta)), \tag{15}$$

where $\tilde{\boldsymbol{y}}_{\text{TF}}(k,\theta)$ denotes the time sequence $|y_{\text{TF}}(k,:,\theta)|$ and $\mathcal{K}(\cdot)$ denotes the kurtosis value of the sequence.

Repeating this procedure for each $\theta \in \{\theta_1,\cdots,\theta_D\}$ and averaging the whole frequency band, we obtain a spatial likelihood function as

$$\rho(\theta) = \frac{1}{K}\sum_{k=1}^{K}\xi(k,\theta). \tag{16}$$

The location of the sound source is then estimated as the location with the highest peak, i.e.

$$\hat{\theta} = \underset{\theta\in\{\theta_1,\cdots,\theta_D\}}{\arg\max}\ \{\rho(\theta)\} \tag{17}$$

TFS and TFS-DNN share exactly the same computation procedure, with the main difference lying in the computation of the confidence measure (10). In TFS, the confidence measure is computed solely as $c_d(k,l,\theta)$ [22], i.e.

$$\tilde{c}_d^{\text{TFS}}(k,l,\theta) = c_d(k,l,\theta). \tag{18}$$

In TFS-DNN, the confidence measure is computed as the product of $c_d(k,l,\theta)$ and $M(k,l)$, where $c_d(k,l,\theta)$ is based on the spatial information and $M(k,l)$ is based on the spectral information of the acoustic signals. By exploiting both spectral and spatial information, TFS-DNN can compute the confidence measure more precisely, thus promising better source localization performance.

C. Discussion

We show intermediate processing results in Fig. 4 to help understand the proposed method. We use the same setup as in Sec. V-D, where the sound arrives from the direction $\theta_d = 20°$, and is captured by the circular microphone array. Fig. 4(a) depicts the mask estimation results at input SNR -5 dB. As already discussed in the last paragraph in Sec. III-B, the estimated mask $M(k,l)$ provides a good indication of speech presence probability at each time-frequency bin. In the last subfigure, we compare the spatial likelihood functions obtained by SRP-PHAT and SRP-DNN[1]. SRP-DNN obtains a correct estimation at 20°, while SRP-PHAT produces a wrong estimation at -90°, which corresponds to the location of one ego-noise source and deviates from the ground truth.

Fig. 4(b) compares the intermediate processing results obtained by TFS and TFS-DNN for the same segment at input SNR -5 dB. The left two columns represent TFS results while the right two columns represent TFS-DNN. In the first row, we present the instantaneous DOA estimation results by the TFS algorithm, i.e. $\theta_{\text{TF}}(k,l)$. The right subfigure depicts the histogram of θ_{TF} at individual time-frequency bins. From the histogram we can observe two peaks at 20° and -90°, corresponding to the target sound source and one ego-noise source, respectively. In the left subfigure we use dark color to indicate the time-frequency bins that have instantaneous DOA 20°, i.e. $\theta_{\text{TF}}(k,l) = 20°$. It can be observed that the majority of these bins correspond to the speech component.

In the second row of Fig. 4(b), we present, for both methods (TFS and TFS-DNN), the confidence measure $\tilde{c}_d$ and the spatial filtering output Y for a candidate direction $\theta = 20°$, which corresponds to the DOA of the target sound source. For TFS, the confidence measure $\tilde{c}_d(k,l,20°) = c_d(k,l,20°)$ presents high values at time-frequency bins that are dominated by both speech and noise, and the spatial filtering output $Y(k,l,20°)$ contains both speech and residual noise. For TFS-DNN, the confidence measure $\tilde{c}_d(k,l,20°) = c_d(k,l,20°)M(k,l)$ presents high values at time-frequency bins that are dominated by speech only, and the spatial filtering output $Y(k,l,20°)$ contains less residual noise. The output SNR by TFS and TFS-DNN are 8.5 dB and 14.9 dB, where TFS-DNN can better enhance the target sound from 20°.

[1]For ease of comparison, the spatial likelihood function is normalized to [0, 1] corresponding to its minimum and maximum value.

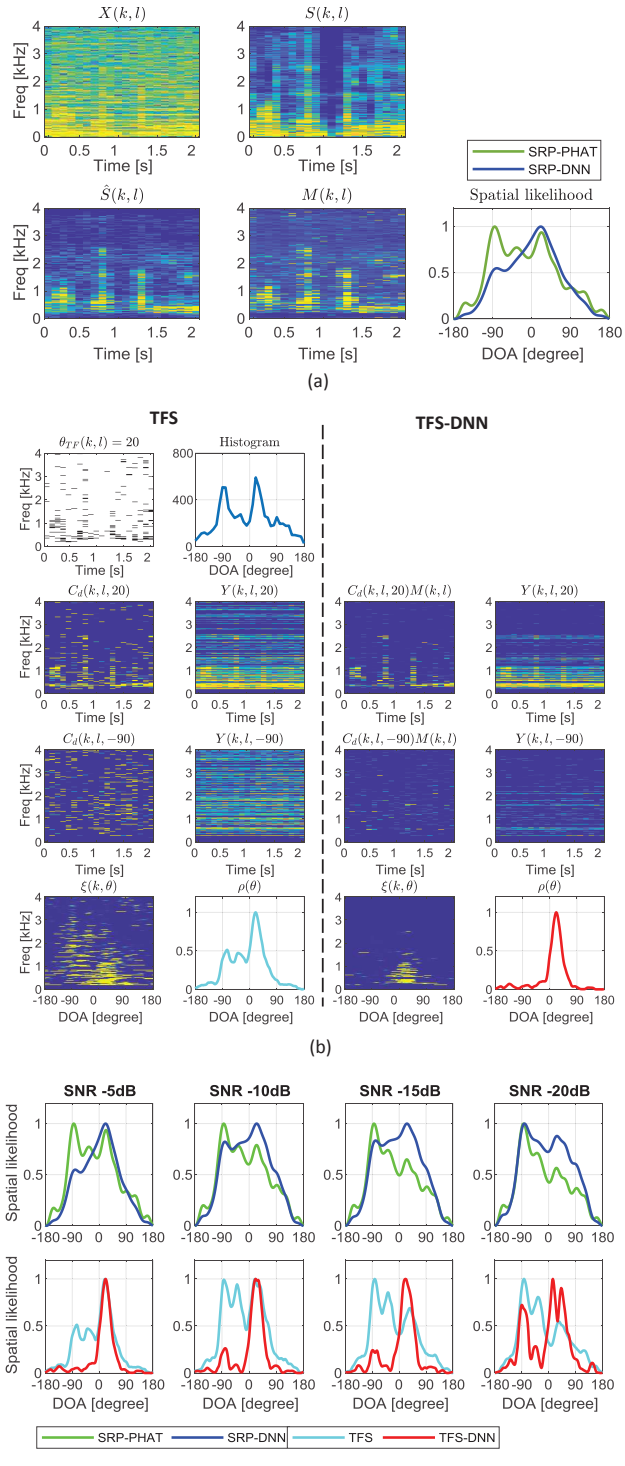


Fig. 4. Sample intermediate and final results of the proposed method. (a) SRP-PHAT vs SRP-DNN at input SNR -5 dB. (b) TFS vs TFS-DNN at input SNR -5 dB. (c) Sound source localization by SRP-PHAT, SRP-DNN, TFS, and TFS-DNN at an input SNR varying from -20 dB to -5 dB. We use SMoLNet for DNN sound enhancement. The DOA of the target sound is 20°.

In the third row of Fig. 4(b), we present, for both methods, the confidence measure $\tilde{c}_d$ and the spatial filtering output Y for a candidate direction $\theta = -90°$, which corresponds to the DOA of one ego-noise source. For TFS, the confidence measure $\tilde{c}_d(k,l,-90°) = c_d(k,l,-90°)$ presents high values at many time-frequency bins that are dominated by noise, and the spatial filtering output $Y(k,l,-90°)$ correspondingly contains the residual noise. For TFS-DNN, the confidence measure $\tilde{c}_d(k,l,-90°) = c_d(k,l,-90°)M(k,l)$ presents only a few high values in the time-domain domain, and the spatial filtering output $Y(k,l,-90°)$ also contains less residual noise. The output SNR by TFS and TFS-DNN are 0.1 dB and 1 dB, respectively, where TFS-DNN can better suppress the noise from -90°.

In the last row of Fig. 4(b), we present, for both methods, the Kurtosis matrix $\xi(k,\theta)$ and the spatial likelihood function ρ. For TFS, the Kurtosis matrix presents high values at many (k,θ) bins and, correspondingly, the spatial likelihood function $\rho(\theta)$ presents peaks not only at 20°, but also at other directions (e.g. -90°). For TFS-DNN, the Kurtosis matrix presents high values only at (k,θ) bins that are close to 20° and, correspondingly, the spatial likelihood function $\rho(\theta)$ presents only one peak at 20°. While from both spatial likelihood functions we can correctly estimate the source location, $\rho_{\text{TFS-DNN}}$ is less affected by the ego-noise.

Fig. 4(c) compares the spatial likelihood functions obtained by the four methods at an input SNR varying from -20 to -5 dB. SRP-PHAT can not estimate the source location correctly for all SNRs. The height of the speech peak at 20° declines continuously with the decreasing SNR. SRP-DNN estimates the sound source location correctly when SNR $\geq$-15 dB. The height of the noise peak at -90° rises with the decrease of SNR, finally surpassing the speech peak at SNR -20 dB. TFS estimates the source location correctly when SNR $\geq$ -10 dB. The height of the speech peak at 20° declines with the decreasing SNR, and is exceeded by the noise peak at SNR -15 dB. TFS-DNN can estimate the source location correctly for all SNRs. The height of the noise peak at -90° rises with the declining SNR, but never surpasses the speech peak at 20°.

Overall, the demonstration in Fig. 4 confirms that the DNN-estimated ratio mask can indicate the speech presence probability at each time-frequency bin. The combination of the ratio mask and the instantaneous DOA at each time-frequency bin can better measure the confidence of a speech source arriving from a target direction. As a result, SRP-DNN and TFS-DNN work more robustly in low-SNR scenarios than their counterparts SRP-PHAT and TFS, respectively.

V. Validation

We consider eights algorithms for the comparison[2]: two baseline approaches, SRP-PHAT and TFS, and six DNN-based approaches: SRP-DNN0, SRP-DNN1, SRP-DNN2, TFS-DNN0, TFS-DNN1 and TFS-DNN2. Here DNN1 and DNN2 are SMoLNet and FC, respectively, while DNN0 is the clean speech reference, which indicates the upper bound of speech enhancement. These eight methods can also be categorized into the SRP family and the TFS family. As set

[2]In [24], we conducted a systematic comparison between TFS and classical sound source localization algorithms, including SRP-PHAT, MUSIC, and histogram-based algorithms, where TFS, by exploiting the time-frequency sparsity of the acustic signal, significantly outperforms these classical ones. We thus consider TFS as the main baseline.

in the original papers [34], [35], DNN1 (SMoLNet) uses an STFT window size of 2048 with half overlap, while DNN2 (FC) uses an STFT window size of 256 with half overlap. For the multi-channel processing in both SRP and TFS, we use an STFT window size 1024 with half overlap. Following [22], we set $\sigma = 10^\circ$ in (11).

A. Datasets

We use three drone sound recording datasets: AS [22], AVQ [29] and DREGON [38]. AS and AVQ provide 8-channel recordings made via a circular microphone array mounted on the top side of a 3DR IRIS quadcopter drone, which is fixed on a tripod (Fig. 1(a)-(b)). The diameter of the circular array is roughly 20 cm. In AS, the recording is made inside a room with sound coming from a loudspeaker at a fixed location. In AVQ, the recording is made outside, with sound from a speaker moving in front of the drone (Fig. 1(c)). AS and AVQ both provide ego-noise recording made when the drone is operating at a constant and varying motor rotating speed. DREGON provides 8-channel recordings made via a cubic microphone array mounted on the bottom side of a MikroKopter drone, which can fly freely (Fig. 1(d)). The side length of the cubic array is roughly 10 cm. In DREGON, the recording is made inside a room, with sound from a loudspeaker at a fixed location while the drone is flying around. For AS, the ground-truth location of the fixed sound source was measured manually. For AVQ, the ground-truth location of the moving source was measured via an external camera mounted on the drone [29]. For DREGON the ground-truth location between the sound source and the moving drone was measured with a Vicon motion tracking system [38]. All the audio samples in the three datasets are resampled at 8 kHz. This helps improve the localization performance as the speech signal typically has significant energy below 4 kHz.

B. DNN Model training

We use the ego-noise in the AVQ dataset in combination with the TIMIT speech corpus [43] for DNN model training. AVQ provides the ego-noise recording with a total duration of 704 seconds, where the drone is operating with either a constant or a varying motoring rotating speed. We use the first channel of the multi-channel recording. TIMIT contains a training subset with 4620 utterances and a testing subset with 1680 utterances. We use 4158 utterances from the training subset for model training, with a total duration of 207 minutes.

For model training, we generate noisy speech by mixing the ego-noise and the TIMIT utterances at different SNRs, uniformly sampled from the [-25, -5] dB. For a specific SNR, the TIMIT utterance is added with a segment of noise randomly cropped from the ego-noise. We generate the training data on the fly in every epoch, which covers 10 iterations of all the speech utterances, corresponding to about 35 hours in total. We use the same training strategies for both FC and SMoLNet, and use their default parameters set in the original papers [34], [35].

C. Performance evaluation

We evaluate the sound source localization performance in three scenarios. $\mathcal{S}1$ considers a static sound source and a static drone, using the AS dataset; $\mathcal{S}2$ considers a moving sound source and a static drone, using the AVQ dataset; $\mathcal{S}3$ considers a moving sound source and a moving drone. In the first two scenarios, the testing data is generated by mixing the speech and the ego-noise at different SNRs varying in the interval [-25, -5] dB with a step size of 5 dB. In the third scenario, the speech and the ego-noise are recorded simultaneously.

In all three scenarios, we evaluate the source localization performance for different processing block sizes B varying from 0.25 to 4 seconds. The localization performance is evaluated per processing segment, where the localization error is defined, given the ground-truth θ_d and the estimation $\hat{\theta}_d$, as

$$e = |\theta_d - \hat{\theta}_d|. \tag{19}$$

For all T processing segments, we define the detection rate as

$$R_d = \frac{T_d}{T}, \tag{20}$$

where T_d is defined as the number of segments with localization error smaller than 10°.

For reference, we also compute SNR and SDR to evaluate the sound enhancement performance of the DNN models and the TFS filters. SNR is defined as the ratio of the speech and the noise after processing, assuming the clean speech and the noise components at the microphones are known in advance [22]. SDR is defined as the scale-invariant signal distortion ratio [44].

D. Localization of static sound sources

We compare the source localization performance of the considered algorithms with a segment of recording of $60s$ from the AS dataset. The drone and the speaker both remain static, and the DOA of the sound is 20° (see the coordinate system in Fig. 1(b)).

Fig. 5(a) shows the output SNR and SDR achieved by the DNN1 (SMoLNet) and DNN2 (FC) at an input SNR varying from -25 to -5 dB. Both DNN models can improve the SNR of the noisy input, with DNN1 evidently outperforming DNN2. DNN1 improves the SNR consistently by about 15 dB for all input SNRs, while DNN2 only improves the SNR by 5-10 dB. Both models introduce speech distortion when suppressing the noise, as indicated by the output SDR lower than the output SNR. DNN1 improves the SDR by 5-12 dB at various input SNRs, while DNN2 achieves minor or even negative SDR improvement over the noisy input.

Fig. 5(b) shows the detection rate achieved by the eight considered algorithms at various input SNRs and processing block sizes. For all algorithms, the detection rate naturally improves with the increasing SNR.

For the SRP family, the detection rate of SRP-DNN0 and SRP-DNN1 improves with the increasing block size B, while SRP-PHAT and SRP-DNN2 remain nearly constant. SRP-PHAT fails in most scenarios and SRP-DNN2 outperforms SRP slightly. SRP-DNN1 achieves a much higher detection

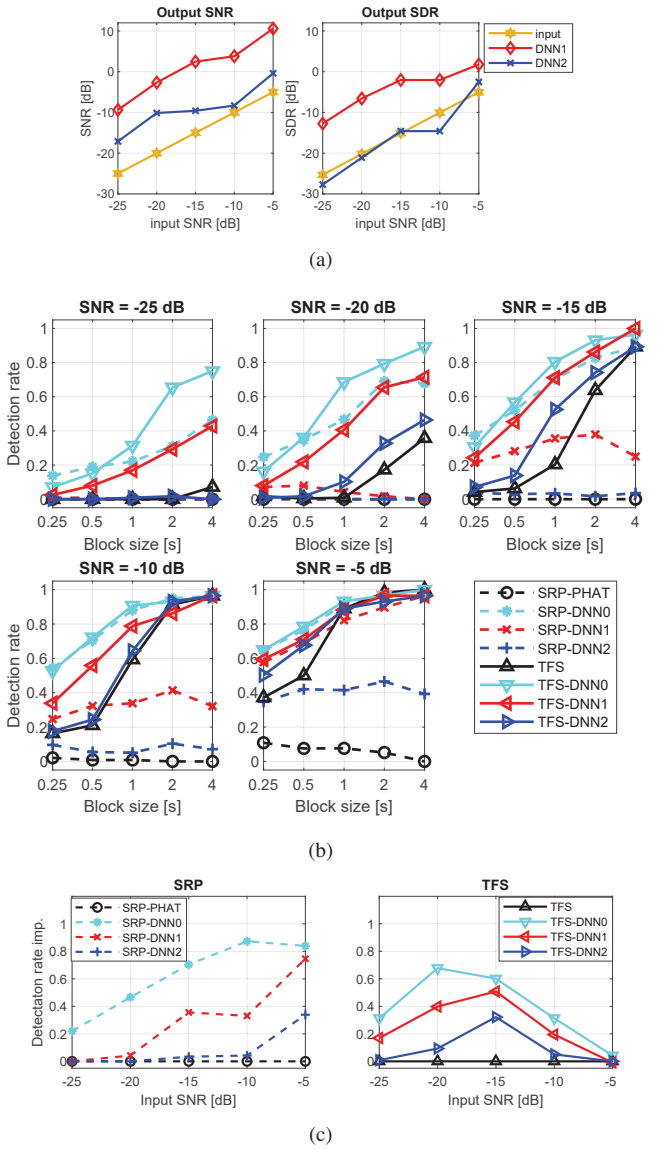


Fig. 5. Sound source localization for static sources (AS dataset). (a) Output SNR and SDR achieved by DNN1 (SMoLNet) and DNN2 (FC). (b) Detection rate achieved by the eight considered methods at various input SNRs and processing block sizes. (c) Detection rate improvement achieved by using DNN over SRP and TFS at block size 1s and various input SNRs.

rate than SRP-DNN2. However, the detection rate of SRP-DNN1 is still not satisfactory at low SNRs, e.g. below 0.4 at SNR -15 dB with all block sizes. SRP-DNN0 provides benchmark performance by using clean speech as a reference.

For the TFS family, the detection rate improves with the increasing block size B, which allows more data to estimate the statistical information of the acoustic signal. TFS fails when SNR $\leq$ -20 dB and block size $B \leq 1s$. However, its performance rises quickly with the increasing block size when $B > 1s$. DNN-based preprocessing can improve the performance of TFS effectively, with the performance of the TFS family ranked as TFS-DNN0 $>$ TFS-DNN1 $>$ TFS-DNN2 $>$ TFS. TFS-DNN0 provides benchmark performance by using clean speech as a reference, while TFS-DNN1 outperforms TFS-DNN2. The performance of TFS-DNN1 is close to the benchmark TFS-DNN0 when SNR $\geq$ -15 dB.

The TFS family remarkably outperforms the SRP family, for instance, when comparing TFS-DNNx and SRP-DNNx (x = 0, 1, 2). The two benchmark algorithms TFS-DNN0 and SRP-DNN0 achieve similar performance when SNR $\geq$ -10 dB, although the former performs better at lower SNRs. Notably, even if a clean speech reference is available, neither of the two benchmarks TFS-DNN0 and SRP-DNN0 achieves perfect source localization results. For instance, TFS-DNN0 achieves a detection rate lower than 0.5 when SNR $<$ -20 dB and $B < 0.5s$.

Fig. 5(c) shows the detection rate improvement by using DNN over SRP and TFS at block size 1s and various input SNRs. It can be observed the impact of DNN is always positive for both SRP and TFS. For SRP, the performance improvement by DNN tends to increase monotonically with the increasing SNR. However, DNN1 improves the performance marginally when SNR$\leq$ -15 dB, and DNN2 also performs limitedly when SNR $\leq$ -20 dB. For TFS, the performance improvement tends to increase first with the increasing SNR and then drops at SNR -15 dB. The drop in the improvement is due to the high detection rate of TFS when SNR $\geq$ -10 dB. The performance of TFS-DNN1 is close to TFS-DNN0 when SNR $\geq$ -10 dB. However, the gap becomes noticeable when SNR $\leq$ -15 dB. Overall, TFS can take better advantage of DNN to improve the source localization performance, especially when the noise suppression performance of DNN is limited in low-SNR scenarios.

E. Localization of moving sound sources

We compare the considered algorithms with a segment of recording of 60s from the AVQ dataset. During recording, a speaker moved in front of the drone, which remained static. We perform sound source localization with a processing block size varying from 0.25 to 4 seconds. The fourth subfigure in Fig. 6(c) depicts the variation of the input SNR to time during the source movement, with an average input SNR of -18.6 dB. It can be observed that the input SNR varies intensely in the range [-25, -10] dB, depending on the location of the sound source (see the coordinate system in Fig. 1(b))[3].

Fig. 6(a) shows the detection rate for the moving sound source by the eight considered algorithms. Interestingly, the observation made in Fig. 6(a) is consistent with ones observed in Fig. 5(b) for a static sound source at SNR -20 dB.

The SRP family performs limitedly in this experiment. SRP-PHAT and SRP-DNN1 fail in all scenarios. SRP-DNN2 improves the detection rate only slightly by 0.2. SRP-DNN0 improves the detection rate significantly, e.g. with a detection rate of 0.8 when $B = 1s$. However, it should be noted that the clean speech reference is not available in practice.

The TFS family achieves much better performance than the SRP family. The detection rate of the four TFS algorithms improves with the increasing B, although the improvement slows or even drops slightly when $B \geq 2s$. The drop in the performance at large B is possibly due to the bigger movement

[3]The video of human moving in front of the drone was also provided in [29].

of the sound source within a large processing segment. TFS and TFS-DNN2 fail when $B \leq 0.5s$, but the performance rises quickly when $B > 0.5s$. TFS-DNN0 and TFS-DNN1 achieve significantly higher detection rates than the other two when $B \leq 1s$. TFS-DNN0 provides benchmark performance by using the clean speech reference. The detection rate of TFS-DNN1 is about 0.2 lower than TFS-DNN0 when $B = 0.25s$, and the gap becomes smaller as B increases.

Fig. 6(c) compares the ground-truth trajectory and the trajectories estimated by three representative algorithms (TFS, TFS-DNN1 and SRP-DNN1) at various processing block sizes B. It can be observed that SRP-DNN1 performs the worst when estimating the trajectory of the moving sound source, with many errors occurring at all block sizes. In comparison to SRP-DNN1, TFS produces more errors when $B = 0.5s$, but fewer errors when $B \geq 1s$. TF-DNN1 performs the best, with only a few errors at $B = 1s$ and very few errors (detection rate close to 0.9) at $B = 2s$.

F. Localization with a moving drone

We compare the considered algorithms with a segment of recording of $34s$ from the DREGON dataset. During recording, a drone is flying freely in front of a static loudspeaker. The sound is recorded with a cubic microphone array mounted on the bottom side of the drone. The coordinate system was provided in [38]. We use the sequence 'Free Flight Speech Source at High Volume (Room 1)'.[4]. The input SNR is estimated to be -12 dB during recording. The dataset does not provide a clean speech at the microphones; however, an external camera was recording the whole scene thus providing certain clean speech reference that is less affected by the drone noise. We use the sound recording from the external camera as the DNN0 reference.

We perform sound source localization with a processing block size varying from 0.25 to 4 seconds. For ease of comparison, we only focus on the horizontal azimuth, although the dataset provides both azimuth and elevation ground truth. In comparison to Sec. V-E, the source localization task in this experiment is more challenging due to the following reasons. First, a 3D cubic eight-microphone array is used for recording, with maximally four microphones in one plane. While our algorithm can be extended to the 3D case easily [45], the cubic array has less discriminability in the horizontal plane than the circular array used in the AVQ dataset. Second, the relative locations between the sensors and sources are changing more quickly for a moving microphone array than for a moving sound source. Third, the ego-noise in the DREGON dataset is significantly different from the ego-noise used for training the DNN models. The noise suppression performance of the DNN models tends to decrease for unseen noise.

Fig. 6(b) shows the detection rate for the moving drone by the eight considered algorithms. The SRP family works limitedly in this experiment. All the four SRP algorithms, including SRP-DNN0, achieve a detection rate lower than 0.3 for all processing block sizes. In addition, the performance of all the four algorithms does not vary much with the block size B, and even drops when $B > 2s$.

The TFS family achieves much better performance than the SRP family, especially when $B \geq 1s$. The performance of all the four TFS algorithms improves with the increasing B, but drops when $B > 2s$. The drop at large B is due to the big

[4]The video of the drone flying in the room was also provided in [38].

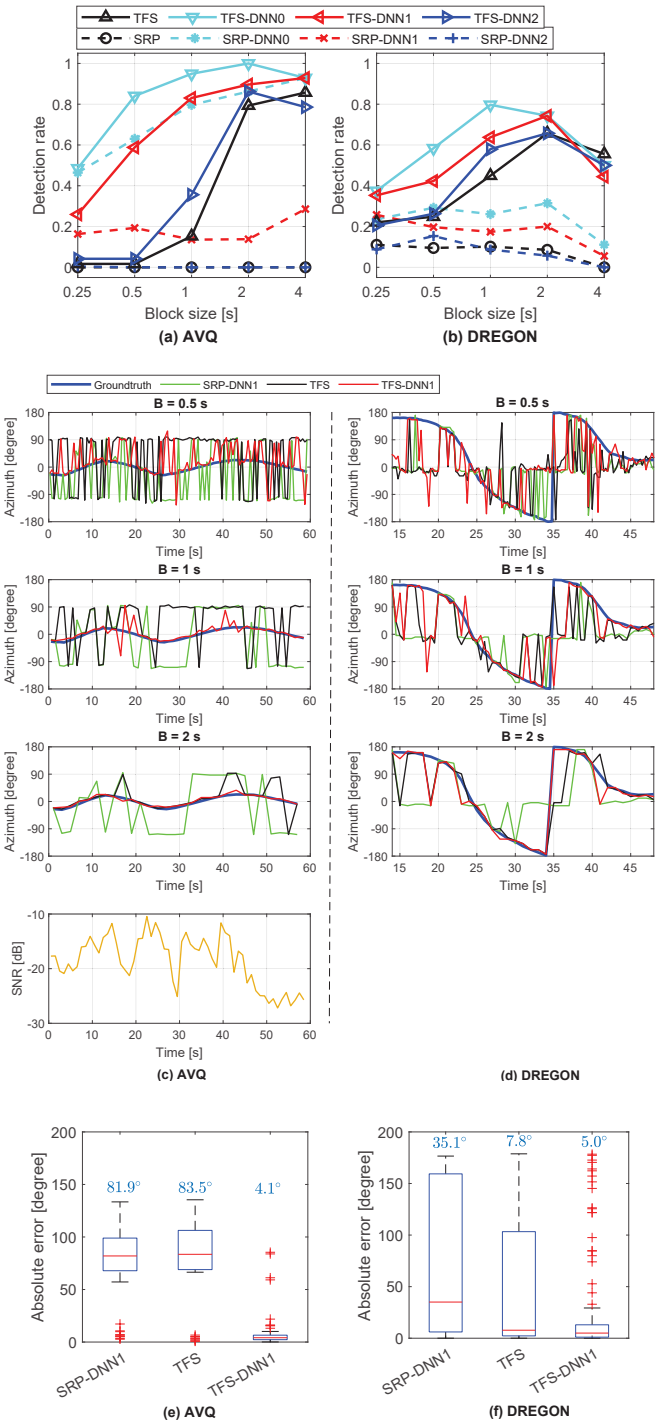


Fig. 6. Sound source localization for moving sound sources (AVQ dataset) and moving drones (DREGON dataset). (a)(b) Detection rate by the considered algorithms at various processing block sizes. (c)(d) Ground-truth and estimated trajectories of the sound source relative to the drone. (e)(f) Box-plot of the absolute localization error for $\boldsymbol{B = 2s}$. The median value is indicated on top of the box.

movement of the drone in a large processing segment (e.g. $4s$). TFS and TFS-DNN2 fail when $B \leq 0.5s$, but the performance rises quickly when $B > 0.5s$. TFS-DNN0 and TFS-DNN1 achieve significantly higher detection rates than the other two when $B \leq 1s$. TFS-DNN0 provides benchmark performance by using a speech reference from the external camera. The detection rate of TFS-DNN1 is lower than TFS-DNN0 in most cases, but close to TFS-DNN0 when $B \geq 2s$. Due to the challenge of this task, the highest detention rate is 0.8, which is achieved by TFS-DNN0 at $B = 1s$, followed by 0.75, which is achieved by TFS-DNN1 at $B = 2s$.

Fig. 6(d) depicts the ground-truth trajectory of the sound source relative to the moving microphone array, as well as the trajectories estimated by three representative methods (TFS, TFS-DNN1 and SRP-DNN1) at various block sizes. It can be observed that the azimuth in Fig. 6(d) is changing more quickly and intensely than the one in Fig. 6(c). SRP-DNN1 performs the worst when estimating the trajectory of the sound source, with many errors occurring at all block sizes B. TFS performs better than SRP-DNN1, with similar errors at $B = 0.5s$, and fewer errors at $B \geq 1s$. TFS-DNN1 performs the best, with minimum errors for all block sizes. When $B = 2s$, TFS-DNN1 can estimate the trajectory of the sound source robustly with a detection rate of 0.75.

Finally, we illustrate in Fig. 6(e) and (f) the boxplot of the absolute localization error as well as the median localization error at $B = 2$. The observations made in Fig. 6(e) and (f) are consistent with the ones made in Fig. 6(a)-(d).

VI. Conclusion

We proposed a deep learning-based sound source localization framework that integrates DNN-estimated soft ratio masks into two multi-channel source localization algorithms, namely steered response power SRP-DNN and time-frequency spatial filtering TFS-DNN. In particular, the proposed TFS-DNN algorithm combines both spectral and spatial information to estimate the presence probability of the target sound at individual time-frequency bins, and performs source localization robustly when processing sound signals in short segments and low-SNR scenarios. The source localization performance tends to improve with the increasing processing block size, which can better estimate the statistical information of the sound signal. However, the performance tends to drop if the processing block size is too large, where large movement of the sound source is observed. It was shown that the inclusion of DNN preprocessing can always improve the source localization performance, and the improvement is related to the noise suppression capability of the DNN model. The TFS framework can better exploit the DNN preprocessing results than the traditional SRP framework.

We compared three approaches to estimating the soft ratio mask, including two existing models (DNN1 and DNN2) and a clean speech reference (DNN0). The clean speech indicates the upper bound of speech enhancement and thus provides a benchmark performance of the proposed framework. The benchmark method outperforms the two DNN methods especially in low-SNR scenarios. Our future work would target at this benchmark performance by developing a better (single-channel or multi-channel) DNN model that suppresses the ego-noise more efficiently. Experimental results also indicate that even the benchmark method still struggles when localizing a sound source in extremely adverse scenarios, such as extremely low SNR (e.g. -25 dB) and short processing segments (e.g. $0.25s$). This implies that the performance of the proposed method may degrade when the drone or the sound source is moving quickly. The incorporation of sound source movement and drone motion estimation would help improve the performance in this challenging scenario [25], [46].

Future work will validate the performance of the proposed framework in a more realistic environment with environmental noise, occlusions, and other scene interferences. Drone audition research is still in a very preliminary stage. Additional advances are also needed in hardware design and system integration, dataset collection and sharing, and our method can be combined with other techniques, such as drone motion estimation, moving source tracking, and multimodal fusion.

References

[1] J. Martinez-Carranza and C. Rascon, "A review on auditory perception for unmanned aerial vehicles," *Sensors*, vol. 20, no. 7276, pp. 1-24, 2020.

[2] D. Floreano, R. J. Wood, "Science technology and the future of small autonomous drones," *Nature*, vol. 521, pp. 460-466, May 2015.

[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.*, Vilamoura-Algarve, Portugal, 2012, pp. 4737-4742.

[4] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, and L. Marcenaro, "Audio-based search and rescue with a drone: highlights from the IEEE signal processing cup 2019 student competition," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 138-144, Sep. 2019.

[5] K. Nakadai, M. Kumon, H. G. Okuno, et al., "Development of microphone-array-embedded UAV for search and rescue task," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Vancouver, Canada, 2017, pp. 5985-5990.

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

[7] F. G. Serrenho, J. A. Apolinario, A. L. L. Ramos, and R. P. Fernandes, "Gunshot airborne surveillance with rotary wing UAV-embedded microphone array," *Sensors*, vol.19, no. 4271, pp. 1-26, 2019.

[8] A. Michez, S. Broset, and P. Lejeune, "Ears in the Sky: Potential of Drones for the Bioacoustic Monitoring of Birds and Bats," *Drones*, vol. 5, no. 9, pp. 1-19, 2021.

[9] 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.*, Vilamoura-Algarve, Portugal, 2012, pp. 3288-3293.

[10] 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.*, Las Vegas, USA, 2015, pp. 26-29.

[11] J. R. Cauchard, K. Y. Zhai, and J. A. Landay, "Drone and me: an exploration into natural human-drone interaction," in *Proc. 2015 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing*, Osaka, Japan, 2015, pp. 361-365.

[12] A. Schmidt, H. W. Löllmann, and W. Kellermann, "Acoustic self-awareness of autonomous systems in a world of sounds," *Proceedings of IEEE*, vol. 108, no. 7, pp. 1127-1149, Jul. 2020.

[13] 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.

[14] 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 Surv.*, Colorado Springs, USA, 2016, pp. 1-7.

[15] 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.
[16] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic source localization from multirotor UAVs," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, pp. 8618-8628, 2019.
[17] W. N. Manamperi, T. D. Abhayapala, J. A. Zhang, and P. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 508-519, 2022.
[18] Y. J. Go and J. S. Choi, "An acoustic source localization method using a drone-mounted phased microphone array," *Drones*, vol. 5, no. 75, pp. 1-18, 2021.
[19] 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.*, Chicago, USA, 2014, pp. 1902-1907.
[20] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, and H. G. Okuno, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, pp. 1-16, 2017.
[21] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. IEEE Int. Conf. Acoust, Speech Signal Process.*, Brisbane, Australia, 2015, pp. 5610-5614.
[22] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors J.*, vol. 18, no. 11, pp. 4570-4582, Jun. 2018.
[23] 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.
[24] 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.*, New Orleans, USA, 2017, pp. 1-5.
[25] 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. Robot. Syst.*, Madrid, Spain, 2018, pp. 1-8.
[26] B. Yen, Y. Hioka, G. Schmid, and B. Mace, "Multi-sensory sound source enhancement for unmanned aerial vehicle recordings," *Applied Acoustics*, vol. 189, no. 108590, pp. 1-22, 2022.
[27] 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. Robot. Syst.*, Tokyo, Japan, 2013, pp. 3943-3948.
[28] 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*, Mountain View, USA, 2017, pp. 1591-1599.
[29] 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.
[30] Y. Masuyama, Y. Bando, K. Yatabe, Y. Sasaki, M. Onishi, and Y. Oikawa, "Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Las Vegas, USA, 2020, pp. 4848-4854.
[31] 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. Intelligent Systems and Technology*, vol. 9, no. 5, pp. 1-28, May 2018.
[32] 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.
[33] H. Purwins, B. Li, T. Virtanen, J. Schlueter, S. Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Selected Topics Signal Process.*, vol. 13, no. 2, pp. 206-219, Feb. 2019.
[34] L. Wang and A. Cavallaro, "Deep learning assisted time-frequency processing for speech enhancement on drones," *IEEE Trans. Emerging Topics in Computational Intelligence*, vol. no. 6), pp. 871-881, Dec. 2021.
[35] Z. W. Tan, A. H. Nguyen, and A. W. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *Proc. Asia-Pacific Signal, Inf. Process. Association Annual Summit Conf.*, Lanzhou, China 2019, pp. 1885-1892.
[36] T. Spadini, G. S. I. Aldeia, G. Barreto, et al. "On the application of SEGAN for the attenuation of the ego-noise in the speech sound source localization problem," in *Proc. 2019 Workshop Communication Networks Power Systems*, Brasilia, Brazil, 2019, pp. 1-4.
[37] A. B. A. Qayyum, K. M. N. Hassan, A. Anika, et al. "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, 2020.
[38] 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.
[39] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Honolulu, USA, 2007, pp. 121-124.
[40] L. Wang, T. 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.
[41] S. Thakallapalli, S. V. Gangashetty, and N. Madhu, "NMF-weighted SRP for multi-speaker direction of arrival estimation: robustness to spatial aliasing while exploiting sparsity in the atom-time domain," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, pp. 1-18, 2021.
[42] 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.
[43] 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," National Institute of Standards and Technology (NIST), Gaithersburgh, MD, vol. 107, 1988.
[44] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDRhalf-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, UK, 2019, pp. 626-630.
[45] L. Wang and A. Cavallaro, "Sound source localization and enhancement in 3D space from a flying drone," *Proc. Quiet Drones 2022*, Paris, France, 2022, pp. 1-9.
[46] M. Wakabayashi, H. G. Okuno, and M. Kumon, "Multiple sound source position estimation by drone audition based on data association between sound source localization and identification," *IEEE Robotics Automation Lett.*, vol. 5, no. 2, pp. 782-789, 2020.

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 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, UK. Since 2018, he has been a Lecturer in Queen Mary University of London. He is Associate Editor of IEEE ACCESS and IEEE SENSORS JOURNAL. His research interests include audio-visual signal processing, machine learning, and robotic perception.

Andrea Cavallaro 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 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.