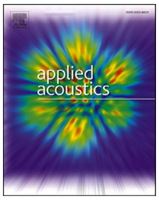
Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust



Drone audition: Audio signal enhancement from drone embedded microphones using multichannel Wiener filtering and Gaussian-mixture based post-filtering ☆

Wageesha N. Manamperi [a,*], Thushara D. Abhayapala [a], Prasanga N. Samarasinghe [a], Jihui (Aimee) Zhang [a,b]

[a] Audio & Acoustic Signal Processing Group, The Australian National University, Canberra, 2601, ACT, Australia
[b] Institute of Sound and Vibration Research, University of Southampton, Southampton, SO17 1BJ, United Kingdom

ARTICLE INFO

Keywords:
Drone noise reduction
Embedded microphones
Gaussian mixture model
Signal enhancement
Wiener filter

ABSTRACT

In this paper, we consider the problem of recovering desired sound source signals from on-board microphone recordings on a noisy drone. Enhancement of source signal degraded by drone noise is considered to be a difficult task due to the strong noise generated from its motors and propellers causing an extremely low signal-to-drone noise ratio (S$\overline{\text{D}}$NR). We propose a solution (i) by combining the widely known multichannel Wiener filter (MWF) to remove drone noise from microphone recordings, and (ii) further reduction of residual noise using a Gaussian mixture model (GMM) based dual-stage parametric Wiener filter (WF). The method exploits known statistics of motor current-specific drone noise. This combination of techniques to the specific context of signal enhancement for drone audition is applicable to irregular microphone arrays embedded on a drone enabling realistic integration to most drones. We demonstrate the validity of the proposed framework with extensive real data through (i) experimental recordings from two different drone acoustics datasets and (ii) outdoor measurements from a hovering drone for a bioacoustic application. The results confirm improved performance in terms of S$\overline{\text{D}}$NR, speech quality (PESQ), and intelligibility (STOI) at very low S$\overline{\text{D}}$NR (up to −30 dB) and show a strong potential for signal enhancement applications using noisy drones.

1. Introduction

1.1. Motivation and background

Drones have received considerable attention in recent years as they emerge with many potential applications in a wide range of areas. Signal enhancement using a drone mounted microphone array enables services mainly in search and rescue missions, wildlife monitoring and video capturing for media and filming industries [1]. However, signal enhancement using a drone is challenging due to its emission of significant noise, e.g., drone motors and propellers cause a highly adverse noisy environment and degrade the quality and intelligibility of the recorded signals. Since the microphones are closer to the drone noise sources compared to the desired sound source on the ground, resulting in an extremely low signal-to-drone noise ratio (S$\overline{\text{D}}$NR) (defined as the power ratio between the source signal and the drone noise) level.[1] This paper addresses the problem of audio signal enhancement at very low S$\overline{\text{D}}$NR conditions (up to −30 dB) for drone mounted or embedded microphone array platforms. Unlike in the conventional speech enhancement in a room scenario [2], it is challenging (depends on the application [3]) to enhance the desired sound source signal on a drone audition as it is difficult to suppress (i) loud, (ii) non-stationary drone noise, and (iii) on-board microphones are nearer to the noise sources than the target sound source.

We review in brief the literature on signal enhancement methods developed for drone on-board microphone arrays. However, in the last few decades, signal enhancement algorithms in noisy environments other than drones, have been widely studied and is still an active field of research. Among the existing methods, beamforming is the standard approach to multichannel speech enhancement [4]. The multichannel Wiener filter (MWF) has been used in speech enhancement on drone [5–7]. In [5], the authors presented a framework for speech enhance-

☆ This work was supported by Australian Research Council (ARC) Discovery Project grant DP200100693.

* Corresponding author.
E-mail address: wageesha.manamperi@anu.edu.au (W.N. Manamperi).

[1] Note that this S$\overline{\text{D}}$NR is not the signal difference-to-noise ratio.

https://doi.org/10.1016/j.apacoust.2023.109818
Received 19 May 2023; Received in revised form 13 December 2023; Accepted 14 December 2023

Available online 22 December 2023

ment using a beamformer with post-filtering. A minimum variance distortionless response (MVDR) adaptive beamformer is coupled to a Wiener post-filtering scheme to extract the target speech. This approach uses a set of MVDR beamformers to estimate the power spectral densities (PSD) of the target speech and drone motors and propellers noise which are used to calculate the parameters of the Wiener filter (WF) for extracting the target speech. The work in [6], compared the performance of the variants of MWF algorithms with an interfering noise source and a single propeller-motor combination. The work in [8], presented a beamforming based spectral distance response algorithm for both localization and enhancement of the target source. This method proposes a diagonal unloading (DU) beamformer for obtaining the target source direction and reports the better signal enhancement capability of the DU beamformer compared to the conventional beamforming methods.

Another method for speech enhancement uses time-frequency (TF) spatial filtering algorithms [9–12]. In [9], the authors proposed a drone noise reduction framework that combines blind source separation, TF spatial filtering, and single channel spectral post-filtering to jointly enhance the target sound. There also exists a video-assisted speech enhancement method together with the TF spatial filters to extract the speech from the visually informed directions [10,11].

Lately, to improve the speech enhancement performance at very low SNR conditions (e.g., $\overline{\text{SDNR}}$ lower than −15 dB), supervised learning methods using deep neural networks (DNNs) are introduced [12–16]. In [12], the authors demonstrated single channel and multichannel integrated TF spatial filtering approaches for speech enhancement on drones. In [13,14], the authors proposed to use multi-sensory information of the drone motors and propellers to accurately estimate drone noise PSD together with microphone signal for speech enhancement. In [13], results are evaluated considering a single motor propeller combination. In [14], a multi-sensory source enhancement framework was proposed for in-flight configuration. The method in [15] presented a partially-shared deep neural network with a small amount of training data. However, this study is not intended to improve the sound quality. The work in [16], proposed a convolutional neural network-based complex spectrogram enhancement method that removes the drone noise. While these techniques show improved signal enhancement results in low $\overline{\text{SDNR}}$ levels (≥ −25 dB), require standard microphone array, e.g., linear, circular, spherical arrays, or specific microphone array configurations mounted below the drone or on a beam attached horizontally to the drone with a set of directional microphones [5,8–12,15,13,14,17,18]. However, in practical drone-based application scenarios, the size of the microphone array and the location of microphones are often restricted. Prior methods for signal enhancement required the transfer functions or assumed free-field sound propagation assumption to obtain the transfer functions, making them difficult to consider as potential methods for signal enhancement using on-board microphones on a drone.

1.2. Approach

The primary aim of this paper is to enhance the audio signal recorded from a drone on-board microphone array when the drone hovers above a constant height from a sound source on the ground. In this paper, we use an irregular microphone array embedded/on-board on a drone that differentiates from a standard microphone array or specific array configurations e.g., mounted below the drone or on a beam attached horizontally to the drone, as in [5,7–15] and demonstrate the audio signal enhancement for (i) speech, and (ii) bird calls.

We present a multichannel signal enhancement framework using Wiener filtering approaches. In particular, we obtain (i) drone noise reduction using the widely known MWF [19–21], and (ii) signal enhancement using Gaussian mixture model (GMM) based dual-stage parametric WF [22]. We use the MWF (similar to [19–21] but not restricting to a preferred microphone channel) for a significant drone noise suppression and a recently proposed GMM WF for single channel speech enhancement from [22] as a post-filter to obtain the enhanced sound source signal. With an accurate estimate of the drone noise PSD, MWF derived from [19–21] suppresses the drone noise more effectively at very low $\overline{\text{SDNR}}$ conditions (≤ −10 dB).

Typically, the spectrum of drone noise is composed of tonal or harmonic and broadband components. In general, the drone noise profile depends on both intrinsic characteristics such as current through the motors or the motor speed, phase difference between propeller blades and flight modes, and extrinsic characteristics such as the pressure, humidity and wind speed conditions [23]. In [24], we demonstrated that drone harmonic noise is proportional to the motor current. Therefore, we use non-acoustical information (as in [13,14]) such as drone motor current, to estimate the drone noise correlation matrix at the MWF. This formulation enables us to more accurately estimate the noise correlation matrix using pre-recorded motor current-specific drone noise recordings. Here, we assume the short-term stationarity of the drone noise when the drone operates on a hovering manoeuvre. This assumption, however, is often not feasible in practical applications specifically in the presence of external wind. In such cases, the drone noise correlation matrix can be obtained recursively with a few seconds of multichannel recording. In this paper, we demonstrate the performance of both cases using experimental recordings.

After suppressing the drone noise from the multichannel recording, we further remove the residual drone noise at each microphone channel from a GMM WF. A clean source signal dataset is used to model the source PSD in the training phase of GMM WF. On the other hand, a few seconds of residual drone noise-only recording from the output of the MWF obtained using a signal activity detector is trained to model the residual drone noise PSD. The calculated GMM mean vectors of both source and residual drone noise at the training phase, are used to estimate the source and noise PSDs from the output of the MWF using dual-stage Wiener filtering.

1.3. Contributions

The *Multichannel Signal Enhancement Framework*: In Section 3, we present the proposed framework using MWF with GMM WF for a drone embedded microphone array at very low $\overline{\text{SDNR}}$ conditions (≥ −30 dB). The novelty of this work lies in the combination of these two techniques in the application of signal enhancement for drones. This algorithm is low in complexity and has a highly applicable outcome for drone applications with any microphone array configuration.

The *Experimental Configurations*: In Section 4, we explain the experimental setups, evaluation protocol, and our drone acoustic dataset. More specifically, the motor current-specific drone noise dataset when the drone hovers in a stable flight mode.

The *Results and Discussion*: In Section 5, we provide the evaluation of the proposed multichannel framework on (i) our motor current-specific drone noise dataset, and (ii) motor speed-specific drone noise dataset (DREGON in [25]) for speech enhancement. We compare the performance in terms of noise reduction, speech quality, and intelligibility. Additionally, we evaluate the proposed method in an outdoor environment for signal enhancement, in particular, bird calls for bioacoustic applications. Moreover, we discuss the performance of the baselines: MVDR [26] and MWF in [25] over our drone acoustic dataset.

2. Signal model and problem formulation

Consider a drone with Q microphones embedded/mounted on a drone, hovering at a constant height above the ground. Let $p_q(f,t)$, $x_q(f,t)$ and $v_q(f,t)$ be the short-time Fourier transform (STFT) domain received noisy source signal, its source content and drone noise content, respectively, at the q^{th} microphone, where t and f are the frame number and frequency index, respectively. The received signal vector at the microphones for a single sound source on the ground can be written as

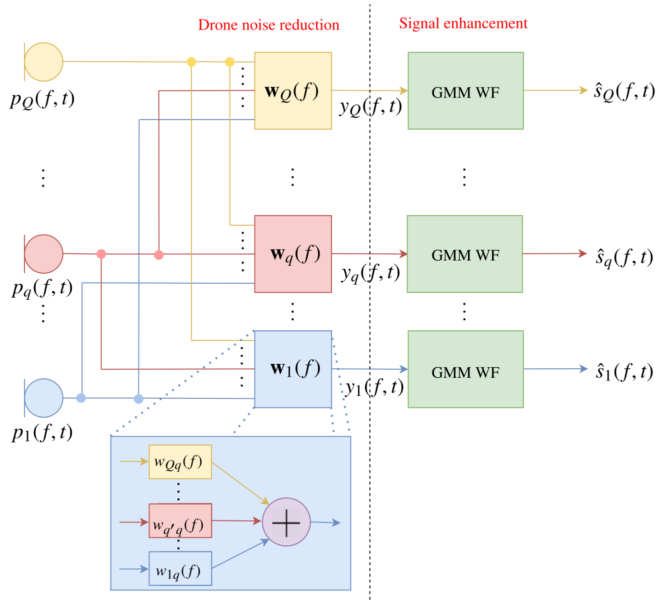


Fig. 1. A block diagram of the proposed multichannel framework.

$$\mathbf{p}(f,t) = \mathbf{x}(f,t) + \mathbf{v}(f,t), \tag{1}$$

where $\mathbf{p}(f,t) \triangleq [p_1(f,t), \cdots, p_Q(f,t)]^T$, $\mathbf{x}(f,t) \triangleq [x_1(f,t), \cdots, x_Q(f,t)]^T$, $\mathbf{v}(f,t) \triangleq [v_1(f,t), \cdots, v_Q(f,t)]^T$ and $(\cdot)^T$ is the non-conjugate transposition. The source component as $\mathbf{x}(f,t)$ can be written

$$\mathbf{x}(f,t) = \mathbf{d}(f)s(f,t), \tag{2}$$

where $\mathbf{d}(f) \triangleq [d_1(f), \cdots, d_Q(f)]^T$, $d_q(f)$ is the acoustic transfer function between the sound source to the q^{th} microphone and $s(f,t)$ is the signal at the source.

We define: (i) the noisy source correlation matrix as $\mathbf{\Phi}_{pp}(f,t) \triangleq E\{\mathbf{p}(f,t)\mathbf{p}^H(f,t)\}$; (ii) the clean sound source correlation matrix as $\mathbf{\Phi}_{xx}(f,t) \triangleq E\{\mathbf{x}(f,t)\mathbf{x}^H(f,t)\}$; (iii) the drone noise correlation matrix as $\mathbf{\Phi}_{vv}(f,t) \triangleq E\{\mathbf{v}(f,t)\mathbf{v}^H(f,t)\}$, where $E\{\cdot\}$ is the statistical expectation operator. We assume drone noise signals to be uncorrelated with the sound source signal. Therefore, using (1)

$$\mathbf{\Phi}_{pp}(f,t) = \mathbf{\Phi}_{xx}(f,t) + \mathbf{\Phi}_{vv}(f,t). \tag{3}$$

Generally, audio signal enhancement in adverse environments is considered to be a difficult problem. For simplicity, here, we consider a scenario where the drone operates at a certain height from the ground in hovering manoeuvre with constant current flow through the motors allowing us to estimate $\mathbf{\Phi}_{vv}(f,t)$ when there is no sound source present.

In this paper, our goal is to extract the desired sound source signal $s(f,t)$ from the received microphone signals $\mathbf{p}(f,t)$.

3. Signal enhancement framework

In this section, we first present a general method to significantly remove drone noise from the multichannel recordings using the multichannel Wiener filter (MWF) and we then propose a post-filtering approach using the Gaussian mixture model (GMM) Wiener filter (WF) to further suppress the interfering drone noise from each microphone channel.

The proposed solution in Fig. 1 is comprised of two components:

(i) MWF, and
(ii) GMM WF.

We discuss each component in separate subsections. We begin with feeding the received microphone signals $\mathbf{p}(f,t)$ through the MWF for a significant drone noise reduction which we discuss in subsection 3.1. Next, we filter the outputs of MWF $y_q(f,t)$, for $q = 1, \ldots, Q$ through a set of GMM WF for signal enhancement to obtain Q estimation of original sound source signal $s(f,t)$, as $\hat{s}_q(f,t)$, for $q = 1, \ldots, Q$. The detailed method is given in subsection 3.2.

3.1. Drone noise reduction with MWF

The output of the MWF is calculated as

$$\mathbf{y}(f,t) = \mathbf{W}^H(f)\mathbf{p}(f,t), \tag{4}$$

where $\mathbf{y}(f,t) \triangleq [y_1(f,t), \cdots, y_Q(f,t)]^T$, $\mathbf{W}(f)$ denotes the $Q \times Q$ filter weight matrix. Unlike the conventional MWF solution, in drone audition applications, the selection of the reference channel for filtering out noise from the other channels at the MWF is not reasonable as the drone motors and propellers are near-field noise sources and this causes all the channels to capture more noise than the sound source signal of interest. Here, the MWF block aims to remove drone noise from each of the microphones signal $p_q(f,t)$ to obtain $y_q(f,t)$ as an estimate of the source component $x_q(f,t)$, where $q = 1, \ldots, Q$.

Let $w_{q'q}(f)$ be the weight applied to the q'^{th} microphone signal $p_{q'}(f,t)$ to obtain $y_q(f,t)$. Thus,

$$y_q(f,t) = \mathbf{w}_q^H(f)\mathbf{p}(f,t), \tag{5}$$

where $\mathbf{w}_q(f) = [w_{1q}(f), \ldots, w_{q'q}(f), \ldots, w_{Qq}(f)]^T$. To find optimal weights $\mathbf{w}_q(f)$, we minimize the following mean squared error criterion

$$J_q(f,t) = E\{\|x_q(f,t) - y_q(f,t)\|_2^2\}, \tag{6}$$

where $\|\cdot\|_2$ is the ℓ_2 norm. The solution to (6) derives the conventional form of MWF [27,28] as

$$\mathbf{w}_q(f) = \mathbf{\Phi}_{pp}^{-1}(f,t)E\{\mathbf{p}(f,t)x_q^*(f,t)\}. \tag{7}$$

By combining all weights vectors $\mathbf{w}_q(f)$ for $q = 1, \ldots, Q$ in (7) together, we obtain

$$\mathbf{W}(f) = \mathbf{\Phi}_{pp}^{-1}(f,t)\mathbf{\Phi}_{xx}(f,t), \tag{8}$$

where $\mathbf{W}(f) = [\mathbf{w}_1(f), \ldots, \mathbf{w}_Q(f)]$. Using (3) and (8), this can be rewritten as

$$\mathbf{W}(f) = \mathbf{I}_Q - \mathbf{\Phi}_{pp}^{-1}(f,t)\mathbf{\Phi}_{vv}(f,t), \tag{9}$$

where $\mathbf{I}_Q$ denotes the identity matrix of size $Q \times Q$. Given the very low $\bar{\text{SDNR}}$ conditions, we reformulate (9) to obtain the optimal weight matrix using the inverse of $\mathbf{\Phi}_{vv}(f,t)$ similar to the work in [19–21].[2] For that, we expand the inverse of $\mathbf{\Phi}_{pp}(f,t)$ in (3). Given that power spectral density (PSD) of source signal $\Phi_{ss}(f,t) = E\{|s(f,t)|^2\}$, we use (2) to write

$$\mathbf{\Phi}_{xx}(f) = \Phi_{ss}(f)\mathbf{d}(f)\mathbf{d}^H(f), \tag{10}$$

and substitute in (3) to get

$$\mathbf{\Phi}_{pp}(f,t) = \Phi_{ss}(f,t)\mathbf{d}(f)\mathbf{d}^H(f) + \mathbf{\Phi}_{vv}(f,t). \tag{11}$$

Using Sherman–Morrison formula [29] and further simplifications (a more detailed steps of this derivation can be found in Appendix A), we obtain

$$\mathbf{\Phi}_{pp}^{-1}(f,t) = \mathbf{\Phi}_{vv}^{-1}(f,t) - \frac{\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{\Phi}_{xx}(f,t)\mathbf{\Phi}_{vv}^{-1}(f,t)}{1 + \Phi_{ss}(f,t)\mathbf{d}^H(f)\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{d}(f)}. \tag{12}$$

[2] Note that when considering the inverse $\mathbf{\Phi}_{vv}(f,t)$ drone noise correlation matrix, there is a reduction of the drone noise from the mixture correlation matrix $\mathbf{\Phi}_{pp}(f,t)$. We discuss this difference in MWF output results in Section 5.4.

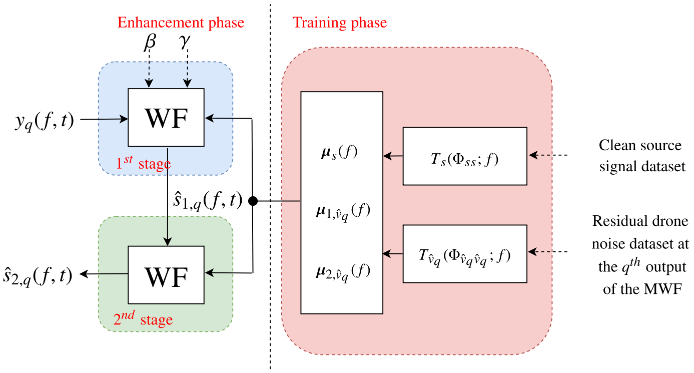

Enhancement phase
Training phase
β
γ
$y_q(f,t)$
WF
1^{st} stage
$\hat{s}_{1,q}(f,t)$
WF
2^{nd} stage
$\hat{s}_{2,q}(f,t)$
$\boldsymbol{\mu}_s(f)$
$\boldsymbol{\mu}_{1,\hat{v}_q}(f)$
$\boldsymbol{\mu}_{2,\hat{v}_q}(f)$
$T_s(\Phi_{ss};f)$
$T_{\hat{v}_q}(\Phi_{\hat{v}_q\hat{v}_q};f)$
Clean source signal dataset
Residual drone noise dataset at the q^{th} output of the MWF

Fig. 2. Illustration of the single channel GMM WF at the q^{th} output of the MWF. The source and residual drone noise PSDs are obtained using (17) and (18).

By using the cyclic property of $tr[\cdot]$ and noting $\Phi_{ss}(f,t)$ is a complex scalar, we get

$$tr[\boldsymbol{\Phi}_{vv}^{-1}(f,t)\boldsymbol{\Phi}_{xx}(f,t)] = \Phi_{ss}(f,t)\mathbf{d}^H(f)\boldsymbol{\Phi}_{vv}^{-1}(f,t)\mathbf{d}(f), \tag{13}$$

where $tr[\cdot]$ denotes the trace operator. Finally, using (13), (12) in (9) together with (3), we express the optimum filter weight that needs to be applied to each microphone outputs of a drone to remove drone noise as

$$\mathbf{W}(f) = \frac{\boldsymbol{\Phi}_{vv}^{-1}(f,t)\boldsymbol{\Phi}_{pp}(f,t) - \mathbf{I}_Q}{1 + tr[\boldsymbol{\Phi}_{vv}^{-1}(f,t)\boldsymbol{\Phi}_{pp}(f,t)] - Q}. \tag{14}$$

A similar derivation can also be found in [19] with respect to a reference microphone channel. Note that the optimal weight applied to the q^{th} microphone channel $\mathbf{w}_q(f)$ for $q = 1,\dots,Q$ to obtain the drone noise suppressed signal as in (5) is given by the q^{th} column of $\mathbf{W}(f)$ matrix.

Here, the optimal filter weights of the MWF depend on the second order statistics of the microphone recordings and drone noise signal, i.e., correlation matrices $\boldsymbol{\Phi}_{pp}(f,t)$ and $\boldsymbol{\Phi}_{vv}(f,t)$: the first one can be easily estimated during the periods when both source and drone are active (directly from the multichannel recordings) while the second one can be estimated using either (i) sample estimates from the motor current-specific pre-recorded drone noise data, or (ii) recursively updates during the noise-only intervals from a multichannel recording when the drone hovers in a stable flight.

More importantly, in (14), there is no requirement of **prior knowledge of the drone-related transfer functions or free-field estimation**, but only estimates of the statistics of the noise, received signal, and the number of microphones in the array.

3.2. Audio signal enhancement with GMM WF

After applying the MWF to the mixture signals, the outputs typically consist of some residual drone noise denoted here as $\hat{v}_q(f,t)$, for $q = 1,\dots,Q$. To further enhance the output signals of the MWF, we apply a single channel GMM based dual-stage parametric WF from [22] to each output of MWF $\mathbf{y}(f,t)$, for $q = 1,\dots,Q$ (see Fig. 1).

The proposed method in [22] has a *training phase* and an *enhancement phase* as shown in Fig. 2. In the training phase, GMM mean vectors of the clean source signal and residual drone noise PSDs are extracted. Those GMM mean vectors are fed through the dual-stage WF in the enhancement phase.

3.2.1. GMM model representation

Let $T_s(\Phi_{ss};f)$ be the probability density function (PDF) of $\Phi_{ss}(f,t)$, and let $T_{\hat{v}_q}(\Phi_{\hat{v}_q\hat{v}_q};f)$ be the PDF of $\Phi_{\hat{v}_q\hat{v}_q}(f,t)$, which takes as a random variable over time frames. Using GMM, we represent the above PDFs as

$$T_s(\Phi_{ss};f) = \sum_{k=1}^{K_s} C_{sk}(f)\mathcal{N}(\mu_{sk}(f),\Sigma_{sk}(f)), \tag{15}$$

and

$$T_{\hat{v}_q}(\Phi_{\hat{v}_q\hat{v}_q};f) = \sum_{k=1}^{K_{\hat{v}_q}} C_{\hat{v}_qk}(f)\mathcal{N}(\mu_{\hat{v}_qk}(f),\Sigma_{\hat{v}_qk}(f)), \tag{16}$$

where $\mathcal{N}$ denotes the Gaussian distribution, K_i, $C_{ik}(f)$, $\mu_{ik}(f)$ and $\Sigma_{ik}(f)$ represent the number of GMM components, the contribution, the mean value, and the variance of k^{th} GMM component at the f^{th} frequency for $i \in \{s,\hat{v}_q\}$, for source and the residual drone noise at the q^{th} output of the MWF, respectively.[3]

3.2.2. Average power spectra

As shown in Fig. 2, we perform dual-stage source enhancement by taking a set of residual drone noise GMM mean vectors (as defined by stacking the mean values of each GMM component for a single f^{th} frequency bin into a vector) at each stage $n = 1,2$. Thus, first-stage consists of the high energy residual drone noise GMM mean vectors, and the rest in the subsequent stage.

We interpret the GMM model representation as in [30] by a weighted sum of GMM mean values over a number of GMM components. Then, PSDs can be expressed as

$$\Phi_{ss}(f,t) = \boldsymbol{\mu}_s^T(f)\boldsymbol{\alpha}_s(t), \tag{17}$$

$$\Phi_{n,\hat{v}_q\hat{v}_q}(f,t) = \boldsymbol{\mu}_{n,\hat{v}_q}^T(f)\boldsymbol{\alpha}_{n,\hat{v}_q}(t), \tag{18}$$

where the GMM mean vectors are given by $\boldsymbol{\mu}_s(f) = [\mu_{s1}(f),\dots,\mu_{sK_s}(f)]^T$, and $\boldsymbol{\mu}_{n,\hat{v}_q}(f) = [\mu_{n,\hat{v}_q1}(f),\dots,\mu_{n,\hat{v}_qK_{n,\hat{v}_q}}(f)]^T$. The power coefficient vectors are obtained by $\boldsymbol{\alpha}_s(t) = [\alpha_{s1}(t),\dots,\alpha_{sK_s}(t)]^T$, and $\boldsymbol{\alpha}_{n,\hat{v}_q}(t) = [\alpha_{n,\hat{v}_q1}(t),\dots,\alpha_{n,\hat{v}_qK_{n,\hat{v}_q}}(t)]^T$. Note that $K_{n,\hat{v}_q}$ denotes $K_{\hat{v}_q}$ at stage n, for $n = 1,2$.

3.2.3. Parameter estimation

Assuming the source signal and residual drone noise are uncorrelated, we obtain the PSD of $y_q(f,t)$ as the weighted sum of the GMM means vectors of both source signal and residual drone noise as

$$\Phi_{y_qy_q}(f,t) = \boldsymbol{\mu}_s^T(f)\boldsymbol{\alpha}_s(t) + \boldsymbol{\mu}_{n,\hat{v}_q}^T(f)\boldsymbol{\alpha}_{n,\hat{v}_q}(t). \tag{19}$$

By arranging (19) for each frequency bin f, we obtain a set of linear equations in a matrix form for a given time frame t. It can be solved to find the power coefficients of the current time frame t at stage n as

$$\begin{bmatrix} \boldsymbol{\alpha}_s(t) \\ \boldsymbol{\alpha}_{n,\hat{v}_q}(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_s\,\boldsymbol{U}_{n,\hat{v}_q} \end{bmatrix}^{\ddagger} \boldsymbol{\Phi}_{y_qy_q}(t), \tag{20}$$

[3] In practice, these parameters are trained separately, using a clean source signal dataset and the residual drone noise-only signal at the q^{th} output of the MWF through the expectation-maximization (EM) algorithm [30]. Note that we omit $\mathbf{w}_q^H(f)\mathbf{d}(f)$ (in (2)) term for simplicity.

where $\boldsymbol{U}_s = [\boldsymbol{\mu}_s^T(f_1), \dots, \boldsymbol{\mu}_s^T(f_F)]$, $\boldsymbol{U}_{n,\hat{v}_q} = [\boldsymbol{\mu}^T_{n,\hat{v}_q 1}(f_1), \dots, \boldsymbol{\mu}^T_{n,\hat{v}_q K_{\hat{v}}^{(n)}}(f_F)]$, and $\boldsymbol{\Phi}_{y_q y_q}(t) = [\Phi_{y_q y_q}(f_1,t), \dots, \Phi_{y_q y_q}(f_F,t)]^T$, and $(\cdot)^\ddagger$ is the Moore-Penrose inverse.

3.2.4. Reconstruction of the source

We reconstruct source-only PSD and the residual drone noise-only PSD of the current time frame using (17) and (18), respectively. Both source and residual drone noise spectrums are smoothed to reduce their temporal fluctuations. Therefore, we estimate the above PSDs recursively as

$$\Phi_{ss}(f,t) = \eta\Phi_{ss}(f,t) + (1-\eta)\Phi_{ss}(f,t-1), \tag{21}$$

and

$$\Phi_{n,\hat{v}_q\hat{v}_q}(f,t) = \eta\Phi_{n,\hat{v}_q\hat{v}_q}(f,t) + (1-\eta)\Phi n,\hat{v}_q\hat{v}_q(f,t-1), \tag{22}$$

where η denotes the forgetting factor.

3.2.5. GMM dual-stage source enhancement

The STFT of enhanced source signal at the q^{th} output of the MWF at stage 1 resulting as

$$\hat{s}_{1,q}(f,t) = \left(\frac{\Phi_{ss}(f,t)}{\Phi_{ss}(f,t) + \beta\Phi_{1,\hat{v}_q\hat{v}_q}(f,t)}\right)^{\gamma} y_q(f,t), \tag{23}$$

where β denotes the over- or underestimation factor and γ denotes the power exponent for trading off source signal distortion for noise suppression [31]. Note that β can be fixed or frequency dependent on the noise characteristics. However, in this work, we mainly focus on fixed values of β for more aggressive drone noise reduction. Several forms of spectral enhancement can be obtained based on γ [32].

Following the WF procedure in the first stage (23), we can filter out the $\hat{s}_{1,q}(f,t)$ at the second stage to obtain the total framework/stage 2 output, derived as

$$\hat{s}_q(f,t) = \left(\frac{\Phi_{ss}(f,t)}{\Phi_{ss}(f,t) + \beta\Phi_{2,\hat{v}_q\hat{v}_q}(f,t)}\right)^{\gamma} \hat{s}_{1,q}(f,t). \tag{24}$$

Currently, the output of this algorithm is a multichannel signal. The most simple way to select the output signal is to take the average out of the enhanced source signals. Note that the combination of the enhanced signals for an optimal output is left for future work.

4. Experimental configurations

The evaluation of the proposed method on experimental data is provided under the following three analysis scenarios:

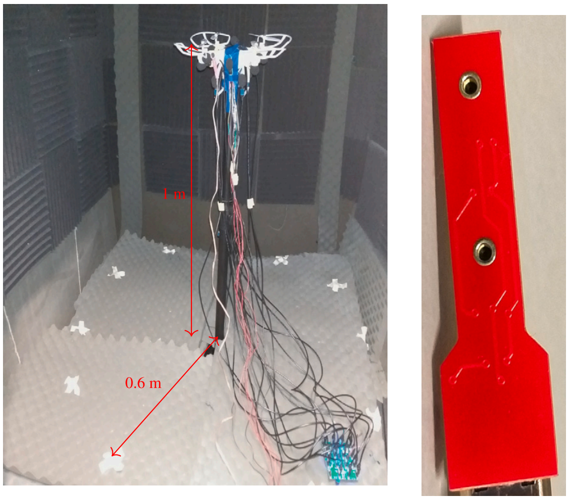
(i) Experimental recordings from outdoor measurements;
(ii) Experimental recordings from measured impulse responses (IRs);
(iii) Experimental recordings from DREGON dataset in [25].

4.1. Experimental setup and recordings

4.1.1. Outdoor measurements

We validate our proposed framework using a real-life application for wildlife monitoring. These applications target learning animal/bird sounds using flying drones. The experiment uses a drone to acquire bioacoustic data and enhances the signal using the proposed solution. We used a Phantom 3 DJI drone to obtain the measurements outdoors at the Australian National University (ANU). We used a ReSpeaker six-channel circular microphone array attached to the top of the drone for the recordings with 16 kHz sampling. Here, we only consider a scenario of the bird calls of an Australian Raven as the target sound source.

The 9 second time domain signals were transformed into the short-time Fourier domain by a 512 point STFT using a 20 ms Hanning window with 50% of overlap to calculate $\boldsymbol{\Phi}_{pp}(f,t)$. On the other hand, $\boldsymbol{\Phi}_{vv}(f,t)$ was estimated using 4–5 seconds of multichannel recordings prior to the bird calls when the drone was stable at the hovering manoeuvre. We note that the wind noise was also included in all the recordings.



(a) Drone mounted microphone array setup. (b) Microphone module.

Fig. 3. Experimental setup. Note that the white markers indicate the loudspeaker placements.

4.1.2. Measured IRs

The performance of the proposed method is evaluated using an experimental setup as shown in Fig. 3(a) for a speech source. The experiment is conducted in a semi-anechoic chamber at ANU with dimensions of [1.45; 1.40; 1.60] m and a room reverberation time of T_{20} of 20 ms. A drone mounted with 15 Micro-Electro-Mechanical System (MEMS) dual-microphone modules[4]: ICS-43432 (as shown in Fig. 3(b)) and fixed on a 1 m tall rigid stand inside the chamber. Hence, there are 30 microphone channels mounted on the drone. Table 1 provides the microphone position on the drone. Note that the microphones are very close to the drone motors and propellers. The microphones are attached underneath each motor, on one side of each motor, on the landing gears, on top of the drone at the origin of the spherical coordinate system, and in between the front and back of the drone arms. We also note that the center of the drone-embedded microphone array and the origin of the spherical coordinate system are aligned.

We placed the sound source 0.6 m from the origin on the ground concentric with the base of the stand as shown in Fig. 3(a). We obtained IR from the source to each microphone channel. It should be noted that the IR recordings incorporate sound scattering by rigid boundaries of the drone structure as well. We mimicked the drone at the hovering manoeuvre by driving all motors using a similar current value. We operated all motors using different current ratings within its operational range and measured the drone noise against different driving currents. The drone operates using four direct current brushed permanent magnet motors. We used the power supply at Constant Current (CC) mode to obtain the root mean square (RMS) value of the motor-current measurements. The drone noise (with all four motors operating) is recorded for approximately 10 s with different motor current levels ranging from 100 mA to 1000 mA with a step of 100 mA. Note that, the current ratings are slightly changed by ±3 mA during the recording. A more detailed version of this drone noise measurement can be found in [23].

[4] Note that the MEMS microphone modules are synchronized by the microphone adapter board.

Table 1
MEMS microphone position on the drone (Cartesian coordinates (x, y, z) in meters).

q	x	y	z	q	x	y	z
1	-0.12	0	-0.02	16	0	-0.09	-0.01
2	-0.11	0	-0.02	17	0	-0.07	-0.09
3	-0.1	0	-0.01	18	0	-0.07	-0.1
4	-0.09	0	-0.01	19	0	0.12	-0.02
5	-0.07	0	-0.09	20	0	0.11	-0.02
6	-0.07	0	-0.1	21	0	0.1	-0.01
7	0.12	0	-0.02	22	0	0.09	-0.01
8	0.11	0	-0.02	23	0	0.07	-0.09
9	0.10	0	-0.01	24	0	0.07	-0.1
10	0.09	0	-0.01	25	0.04	-0.04	0
11	0.07	0	-0.09	26	0.04	-0.04	-0.01
12	0.07	0	-0.10	27	-0.04	0.04	0
13	0	-0.12	-0.02	28	-0.04	0.04	-0.01
14	0	-0.11	-0.02	29	0	0	0
15	0	-0.1	-0.01	30	0	0	-0.01

We generated the received noisy speech signal with IR measurements convolved with the speech signal from the WSJCAM0 database [33] to simulate the sound source signal received at the microphones, and added it together with the current-specific drone noise measurements. We adjusted the S$\overline{\mathrm{D}}$NR of the noisy speech recordings by rescaling the volume of the speech to simulate the real-world scenario which was equivalent to having the speech sources further away from the drone. We simulated our application scenario within the S$\overline{\mathrm{D}}$NR range from -30 dB to -10 dB, with an increment of 5 dB over different motor current ratings. All signals were transformed to the frequency domain by a 512 point STFT using a 20 ms Hanning window with zero padding and 50% of overlap.

4.1.3. DREGON dataset

We analyze the performance of the proposed framework on the DREGON dataset. DREGON dataset has a constellation of 8 microphones in a cubic-shaped structure placed below the motor-propeller plane of the drone. Since the DREGON dataset has a motor speed-specific noise-only dataset, we use that information to learn the drone noise characteristics for our framework. Therefore, we used the noise-free recording of the speech source together with the noise-only recordings of all four motors operating ('allMotors_70.wav') to simulate the extreme S$\overline{\mathrm{D}}$NR conditions (≤ -10 dB). Here, the speech source is placed at the azimuth of 45°, the elevation of $-30°$ and distance of 2.4 meters from the microphone array ('45_ − 30_2.4.wav' in 'DREGON_clean_recordings_speech').

4.2. Algorithm properties

This section provides detailed information on the implementation procedures of the proposed multichannel signal enhancement algorithm.

4.2.1. Noise statistics at the MWF

In this work, we explore drone noise correlation matrix $\mathbf{\Phi}_{vv}(f,t)$ estimation using (i) the multichannel recording on-the-fly (in Section 5.1), and (ii) the pre-recorded drone noise-only measurements using the non-acoustical information, especially, use of pre-recorded motor current-specific drone noise (in Section 5.2) and pre-recorded motor speed-specific drone noise (in Section 5.3).

The main reason for using the non-acoustical information such as motor current-specific drone noise dataset is to learn the noise characteristics of the drone. This improves the tracking of non-stationary drone noise and suppresses the noise from the mixture signal, effectively. Instead of the motor current-specific drone noise data [24,23], we can use motor speed-specific drone noise data as in the DREGON dataset [25].

In the absence of the pre-recorded drone noise data, $\mathbf{\Phi}_{vv}(f,t)$ can be estimated recursively in real-time from a sufficiently long multichannel recording on-the-fly. We note here that we have not provided detailed information on the estimation of the drone noise correlation matrix without using the pre-recorded data. However, in [17], Yen proposes a drone noise covariance matrix estimation method by exploiting the drone mounted microphone array configuration. There are also various approaches in the literature that are proposed in particular to robot-based applications that can be directly adopted to drones [34,35].

We outline this second-order statistic estimation more briefly for a practical scenario. Given T frames of drone noise-only multichannel recordings, the noise correlation matrix is estimated as

$$\mathbf{\Phi}_{vv}(f,t) \triangleq \frac{1}{T}\sum_{t=1}^{T} E\{\mathbf{v}(f,t)\mathbf{v}^H(f,t)\}. \tag{25}$$

Similarly to (25), given T' frames of the multichannel recordings when both source and drone are active, we can directly estimate noisy source correlation matrix as

$$\mathbf{\Phi}_{pp}(f,t) \triangleq \frac{1}{T'}\sum_{t=1}^{T'} E\{\mathbf{p}(f,t)\mathbf{p}^H(f,t)\}. \tag{26}$$

4.2.2. Algorithm parameters of GMM WF

In the GMM WF, the number of GMM components for the desired sound source, e.g., speech, bird calls, and drone noise GMMs are chosen to be $K_s = 6$ and $K_{\hat{v}_q} = 13$ for $q = 1,\dots,Q$, respectively following a trial and error approach proposed in [36]. During the training phase, we use a clean sound source signal to model the GMMs of the target sound source PSDs. In particular, we used a clean bird signal for bird calls in Section 4.1.1, and a clean speech training dataset from the TIMIT database [37] for speech in both Sections 4.1.2 and 4.1.3. While the drone noise is reduced significantly in MWF output $y_q(f,t)$, for $q = 1,\dots,Q$ (as illustrated in Fig. 1), there are some residual drone noise remains as well. To model the GMMs of residual drone noise PSDs, we use a voice activity detector to obtain these remaining noise-only observations. Since the drone operates under the hovering manoeuvre, e.g., fixed motor current, we assume that the drone noise is stationary enough so that the noise-only signals can be estimated during the intervals of the silence of the desired sound source. Here, we note that the residual drone noise for the training phase of GMM WF is always obtained from the output signal at the MWF when the source signal is absent. Hence, it depends on the characteristics of the drone noise and can be done off-line for the pre-recorded measurements. All signals are down sampled to 16 kHz in GMM WF. The training set consists of around 2 minutes of clean bird signal, 2 hours of clean speech, and approximately 6 seconds of motor current-specific residual drone noise data of the particular microphone channel.

In the enhancement phase, to obtain better denoising performance at very low S$\overline{\mathrm{D}}$NR conditions in (23), we use $\beta = 0.5$ and $\gamma = 3.2$ at the first stage ($n = 1$) and set $\beta = 2$ and $\gamma = 1$ i.e., a Wiener gain, at the second stage ($n = 2$) to make the less significant residual drone noise GMM mean vectors to be considerable and minimize the speech distortion introduced by the parametric WF [31]. We set η in (21), and (22) to 0.3. For the experimental setup in Section 4.1.2, we note that two different speech datasets have been used to construct the noisy speech signal of MWF and to model the speech GMMs in the training phase. This means that the experimental setup in Section 4.1.2 is both speaker-independent and speech-content independent.

4.3. Performance measures

For a comprehensive evaluation of the proposed algorithm in both Section 4.1.2 and 4.1.3, we present some useful measures that will help us better understand the quality of the enhanced speech and noise reduction. We evaluate the quality of the enhanced speech by using the Perceptual Evaluation of Speech Quality (PESQ) score [38] and the speech intelligibility by using the Short-Time Objective Intelligibility

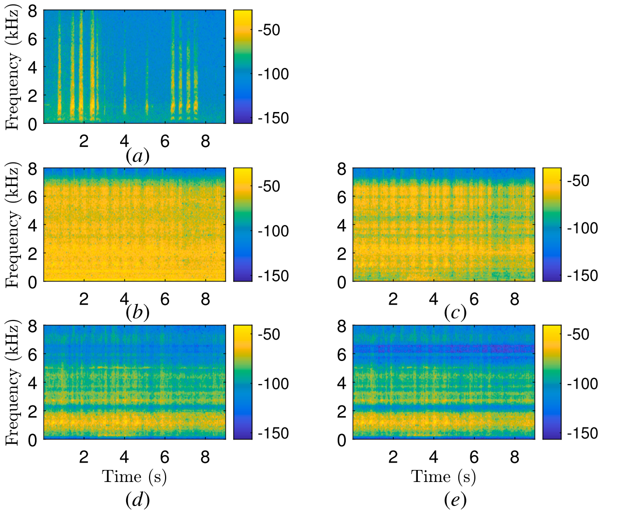


Fig. 4. Spectrogram of (a) the clean bird call, (b) the mixture at channel 3, (c) the output signal of the MWF at channel 3, (d) stage I, and (e) stage II output signal of the GMM WF at channel 3. We note that the clean bird call in (a) used in the *training phase* at the GMM WF is not time aligned with the input at the MWF as shown in (b).

(STOI) score [39] (available at [40]). The range of PESQ score is from −0.5 to 4.5, whereas STOI score is from 0% to 100%.

The level of noise reduction achieved through the MWF and the GMM WF is evaluated through $\overline{\text{SDNR}}$ improvement. Therefore, we define the $\overline{\text{SDNR}}$ as power ratio between the source signal and especially to drone noise [23], i.e.,

$$\overline{\text{SDNR}} = 10\log_{10}\left(\frac{s^2(t)}{v_q^2(t)}\right),$$

where $s(t)$, and $v_q(t)$ are the time domain sound source signal and drone noise signal, respectively. The $\overline{\text{SDNR}}$ improvement is measured as the difference between output $\overline{\text{SDNR}}$ and the input $\overline{\text{SDNR}}$ at each step. The definition of the average input $\overline{\text{SDNR}}$ is found by adding the input $\overline{\text{SDNR}}$ at each channel and averaging it by all channels.

5. Results and discussion

This section presents experimental results of the proposed multichannel signal enhancement framework at extreme $\overline{\text{SDNR}}$ conditions from −30 dB to −10 dB. We also analyze the performance of the baseline methods: MVDR ([26]) and MWF found in [25].

5.1. Experimental recordings from outdoor measurements

Fig. 4(a) provides the spectrogram of the clean bird calls that are used for PSD modeling of the clean bird signal at the GMM WF in the training phase. We observe that the acoustic spectrum of the Raven has high energy in 800–4000 Hz. Fig. 4(b) and (c) show the spectrograms of the input and output signals at the MWF for the channel 3. We observed a great reduction in drone noise, while the wind noise stayed in the MWF output. We note that the clean bird call in Fig. 4(a) used in the *training phase* at the GMM WF is not time aligned with the input at the MWF in Fig. 4(b). The enhanced output signals at the first and the second stage of the GMM WF are illustrated in Fig. 4(d) and (e), respectively. Once more, we observe a considerable reduction in drone noise and additionally, a suppression of wind noise. This suggests that the enhanced signal obtained by (24) (with $n = 2$), is expected to offer good intelligibility for manually identifying the bird species for expertise on bioacoustic.[5] We attach a link to the audio files in Fig. 4.[6]

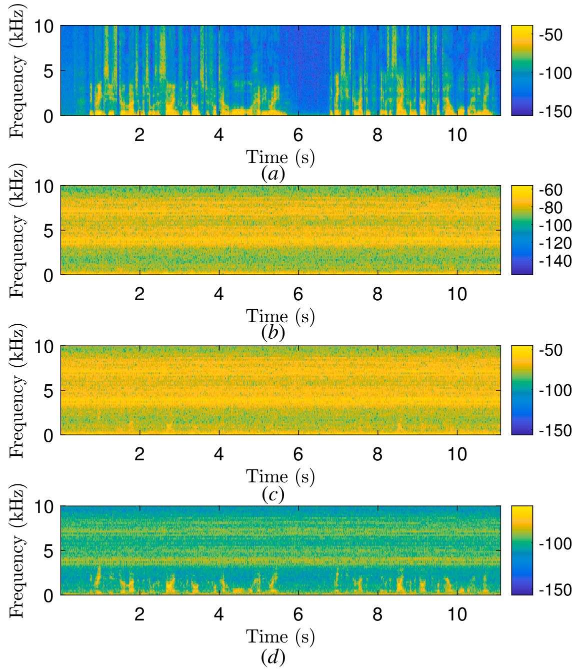


Fig. 5. Spectrogram of (a) the clean speech, (b) the drone noise at channel 11 for motor current of 1020 mA, (c) the mixture at channel 11, and (d) the output signal of the MWF at channel 11. Here the average input $\overline{\text{SDNR}}$ is at −30 dB whereas the input $\overline{\text{SDNR}}$ at channel 11 is −24 dB.

5.2. Experimental recordings from measured IRs

5.2.1. Noise reduction performance with MWF

In this section, we discuss the performance of the MWF for multichannel noise reduction. Fig. 5 shows the time varying spectrograms of the clean speech, drone noise, the mixture signal, and the output signal of the MWF at microphone channel $q = 11$ for average input $\overline{\text{SDNR}}$ of −30 dB. We observe more prominent low and higher order harmonics in the drone noise spectrogram for a motor current rating of 1020 mA (as shown in Fig. 5(b)). More specifically, three segments of frequency range with high power for the given motor current rating. The mixture signal at channel 11 in Fig. 5(c) has a $\overline{\text{SDNR}}$ of −24 dB and still contains the harmonics structure of the drone noise. According to the Fig. 5(d), it is clear that the MWF plays a significant role in reducing the drone noise at the average input $\overline{\text{SDNR}}$ of −30 dB. As it can be seen that the output signal of the MWF well preserves the speech spectra, while a fewer amount of drone noise still remains in the three frequency segments that are observed for the given motor current. The output signal of the MWF has a $\overline{\text{SDNR}}$ improvement of around 21 dB and improves the output $\overline{\text{SDNR}}$ at the channel 11 to −2.6 dB. Moreover, the MWF output signal achieves a PESQ score of 1.99, and a STOI score of 55.6% whereas the mixture signal has a PESQ score of 0.72 and a STOI score of 20.8%. Based on the above plots, we can observe that the output spectrum is denoised using the MWF at average input $\overline{\text{SDNR}}$ of −30 dB. This suggests the MWF achieves better performance in very low $\overline{\text{SDNR}}$ conditions.

[5] Note that for bioacoustic applications, the output of the proposed framework can be fed directly into a machine learning model for automatic identification of the bird species.

[6] https://www.dropbox.com/sh/vup3nsxxuhapov0/AAAANZUu6dq_KtenVGea6JrNa?dl=0.

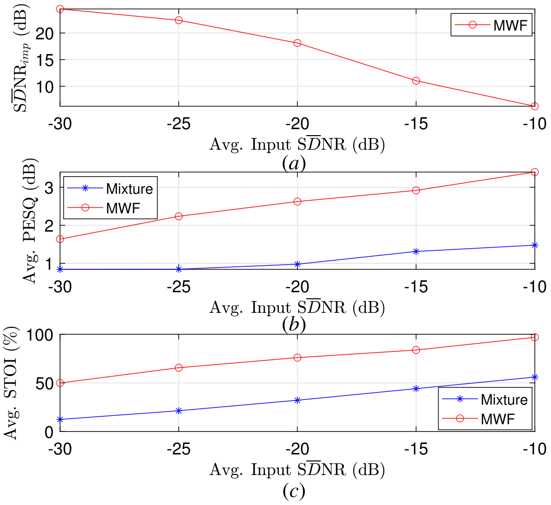


Fig. 6. Performance evaluation at the MWF. (a) Average S$\overline{\text{DNR}}$ improvement, (b) average PESQ scores, and (c) average STOI scores (in percent).

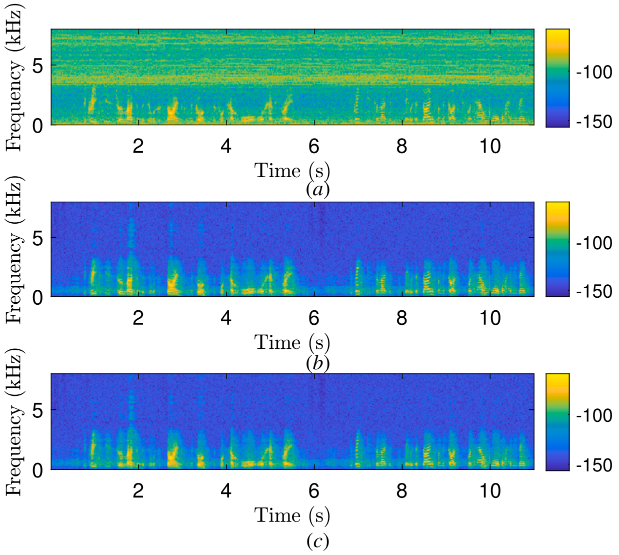


Fig. 7. Spectrogram of (a) the MWF output, (b) stage I, and (c) stage II of the GMM WF at channel 11.

To further understand the drone noise reduction performance at the MWF, we plot the average S$\overline{\text{DNR}}$ improvement, PESQ, and STOI scores over average input S$\overline{\text{DNR}}$ of −30 dB to −10 dB in Fig. 6. The average maximum noise reduction of approximately 25 dB is reached at average input S$\overline{\text{DNR}}$ of −30 dB. A continuous decrease in the noise reduction performance is observed when average input S$\overline{\text{DNR}}$ increases to −10 dB. We observe that MWF yields a great improvement in PESQ and STOI scores over the unprocessed speech. This seems to suggest the MWF improves both S$\overline{\text{DNR}}$ and speech quality without degrading intelligibility. Note that MWF performs well in high S$\overline{\text{DNR}}$ conditions (> −10 dB) as well. However, since we focus on drone application scenarios, we studied the performance in S$\overline{\text{DNR}}$ conditions below −10 dB.

5.2.2. Signal enhancement performance with GMM WF

Fig. 7 shows the time varying spectrograms of the input and outputs of the GMM WF at channel 11. We can observe that the GMM WF removes all of the broadband noise by eliminating most of the wide residual drone noise harmonics from the MWF output using the staging process. Fig. 8 illustrates the selection of the set of GMM mean vectors of the residual drone noise at $q = 11$ microphone in the *training phase*. We observe that different harmonics of the residual drone noise are captured with varying amplitudes for the two stages. Later, in the *enhancement phase*, this helps to obtain a more accurate estimate of the residual drone noise PSD.

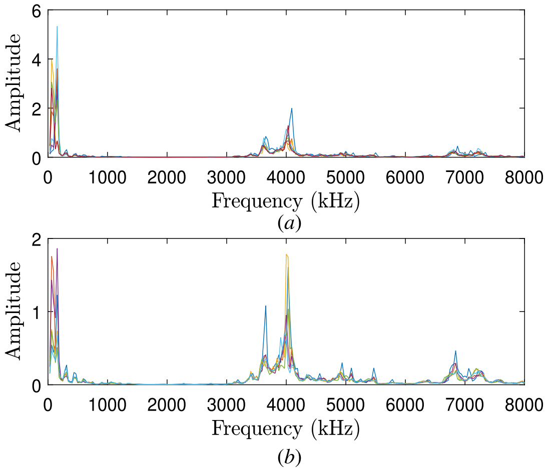


Fig. 8. GMM mean vector plots of residual drone noise for $K_{\hat{v}_q} = 13$ for $q = 11$. The selection of the set of mean vectors $K_{n,\hat{v}_q}$ sets to (a) 7 for $n = 1$, and (b) 6 for $n = 2$.

Based on Figs. 7 and 8, we observe that the GMM WF suppresses the drone noise effectively in the frequency regions over 0 to 300 Hz, and 3.2 kHz to 4 kHz. In the first stage, we mainly cancel out the high power residual drone noise GMM mean vectors in between 0 to 300 Hz frequency range as shown in Fig. 7(b). The drone noise in both frequency regions is suppressed using the low power residual drone noise GMM mean vectors at the second stage (as shown in Fig. 7(c)). Overall, the enhanced signal has PESQ of 2.2 and STOI of 84.7% scores. We share a link to the audio files in Fig. 7.[7] We note that the speech enhancement using drone embedded/on-board microphones are different from the smart devices (such as Google Home) as typical signal-to-noise ratios of drone embedded microphones are significantly lower than that of smart devices. It is expected that lower intelligibility due to the heavy drone noise.

In the experimental setup, we placed three dual-microphone modules closer to each motor. This indicates that we can take six microphone channels around a single motor to analyze the drone noise reduction and signal enhancement properties. We use the microphone channels of $q \in [7, 12]$ to analyze the speech enhancement performance at GMM WF in terms of perceptual quality and intelligibility scores. Figs. 9 and 10 illustrate the PESQ and STOI results, respectively. The PESQ scores, shown in Fig. 9, show a better improvement to the MWF processed speech over the average input S$\overline{\text{DNR}}$ levels ranging from −30 dB to −15 dB except channel 12. However, the PESQ scores decrease compared to the MWF output at the S$\overline{\text{DNR}}$ level of −10 dB except for the channel 10. Fig. 10 indicates that all channels ($q \in [7, 12]$) at extremely low average input S$\overline{\text{DNR}}$ levels, e.g., < −15 dB, the proposed method achieves higher STOI scores with respect to the processed speech of MWF. However, STOI scores begin to decrease from −15 dB for all the channels compared to the output at the MWF.

Fig. 11 shows the S$\overline{\text{DNR}}$ improvement at the GMM WF of $q =\in [7, 12]$ microphones. We observe that the S$\overline{\text{DNR}}$ improvement is not very significant until −20 dB. All six channels have a similar S$\overline{\text{DNR}}$ improvement with the increase of the average input S$\overline{\text{DNR}}$ from −15 dB. PESQ and STOI scores are observed to have lower values compared to the MWF output at S$\overline{\text{DNR}}$ levels from −15 dB to −10 dB in Figs. 9 and 10. In general, the parametric WF introduces speech distortion at the cost of

[7] https://www.dropbox.com/sh/6j881piwo5cwi4p/AABp8K5nmylpjp11Ym41aGuta?dl=0.

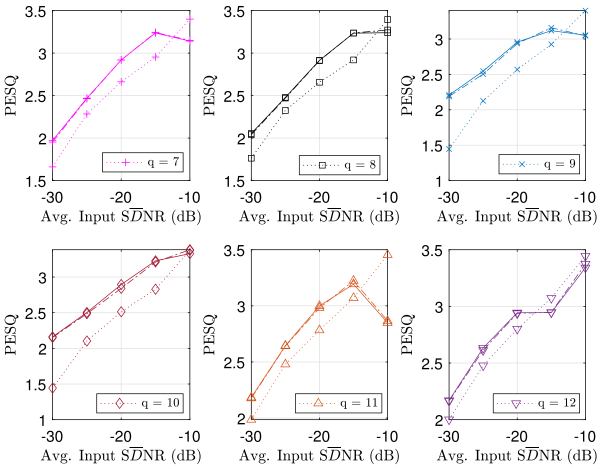


Fig. 9. PESQ scores at channel q where $q \in [7, 12]$ with very low average input S$\overline{\text{D}}$NR conditions. The dotted line represents the input at the GMM WF (output from the MWF) whereas the dash-dotted and solid lines represent the output at the GMM WF in stage 1 and 2, respectively.

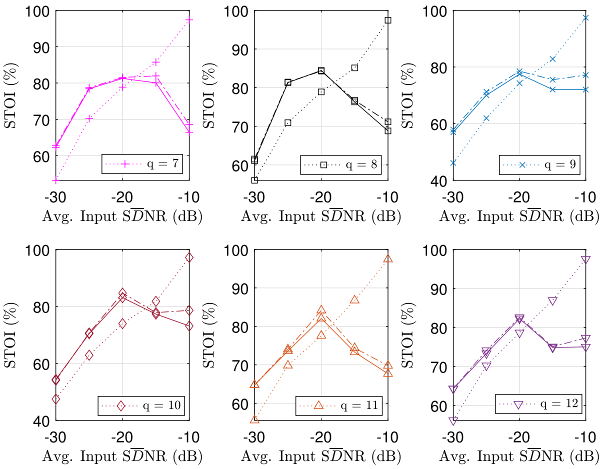


Fig. 10. STOI scores (in percent) at channel q where $q \in [7, 12]$ with very low average input S$\overline{\text{D}}$NR conditions. Dotted line represents the input at the GMM WF (output from the MWF) whereas the dash-dotted and solid lines represent the output at the GMM WF in stage 1 and 2, respectively.

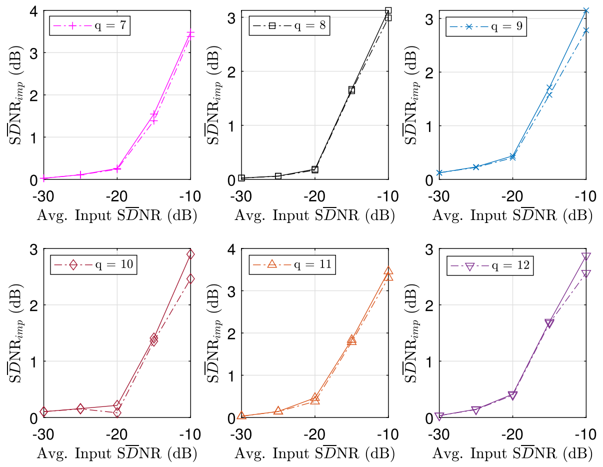


Fig. 11. S$\overline{\text{D}}$NR improvement at the GMM WF at channel q where $q \in [7, 12]$ with very low average input S$\overline{\text{D}}$NR conditions. Dash-dotted and solid lines represent the output at the GMM WF in stages 1 and 2, respectively.

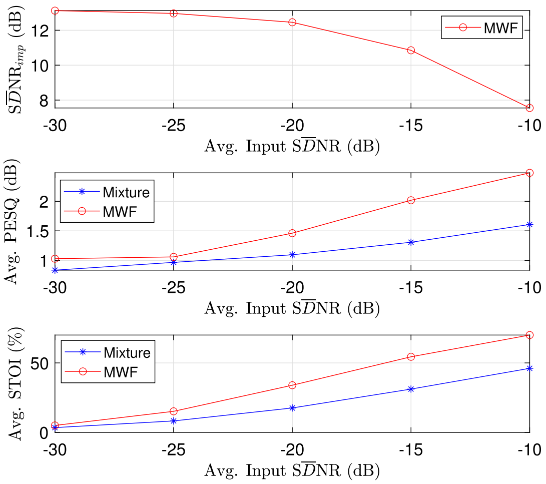


Fig. 12. Performance of the MWF outputs for the DREGON dataset. (a) Average S$\overline{\text{D}}$NR improvement, (b) average PESQ scores, and (c) average STOI scores (in percent).

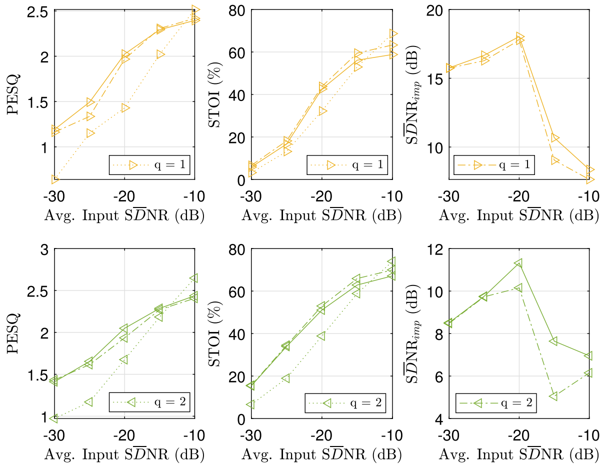


Fig. 13. PESQ, STOI (in percent) scores and S$\overline{\text{D}}$NR improvement at channel q where $q \in [1, 2]$ with very low average input S$\overline{\text{D}}$NR conditions. Dotted line represents the input at the GMM WF (output from the MWF) whereas the dash-dotted and solid lines represent the output at the GMM WF in stage 1 and 2, respectively.

noise reduction. However, for those cases, the processed STOI scores are above 60% and larger than the unprocessed speech for all the S$\overline{\text{D}}$NR levels. Therefore, we expect a limited degradation effect in the processed speech of the MWF at low S$\overline{\text{D}}$NR levels (≤ -10 dB). It seems that the parametric WF should adapt with the changes of the input S$\overline{\text{D}}$NR levels and motor current of the drone, e.g., varying the parameter settings at each stage (with varying β and γ).

5.3. Performance evaluation on DREGON dataset

Fig. 12 illustrates the results at the MWF. The DREGON dataset configuration achieves an average maximum noise reduction of nearly 13 dB at average input S$\overline{\text{D}}$NR level of -30 dB. We can observe that both datasets display the same trend in noise reduction with increasing S$\overline{\text{D}}$NR level. However, the S$\overline{\text{D}}$NR improvement over very low S$\overline{\text{D}}$NR conditions is less compared to our dataset [23]. Moreover, the PESQ and STOI score improvement over the extreme S$\overline{\text{D}}$NR levels are not very considerable. Note that the DREGON microphone array has eight microphones that are mounted below the motor-propeller plane, whereas in our array, we have 30 microphones embedded on the drone structure. It is clear that at the MWF, using a higher number of microphones

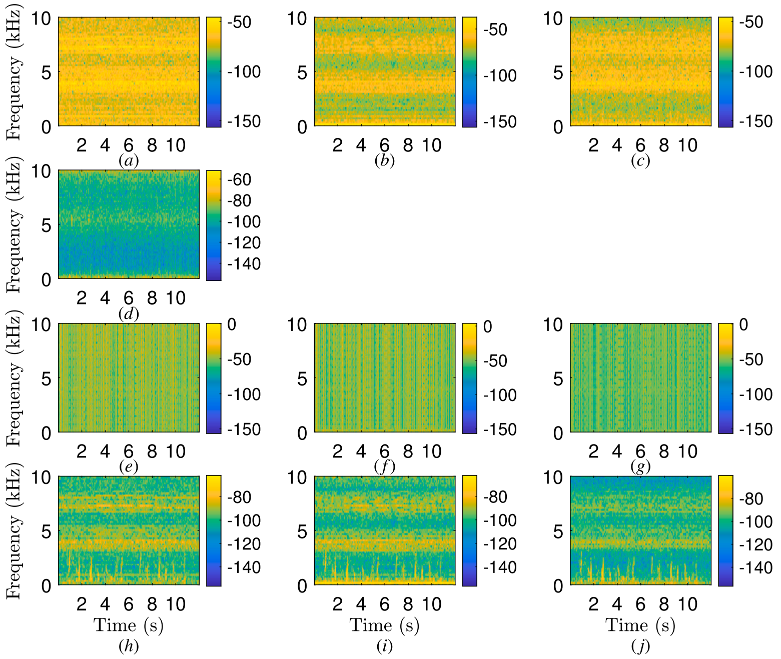


Fig. 14. Comparison with baseline methods at average input $\overline{\text{SDNR}}$ of −30 dB: Spectrograms (a–c) the mixture for channel $q \in \{7,9,11\}$, (d) the MVDR output, (e–g) the output at MWF method in [25] for channel $q \in \{7,9,11\}$, and (h–j) the output at MWF derived in the proposed method for channel $q \in \{7,9,11\}$.

on the drone will capture more propagation paths and add more diversity to the framework. This suggests that a higher number of drone embedded/on-board microphones will result in more noise suppression at the MWF.

Fig. 13 shows the speech enhancement performance of DREGON dataset over the channel q where $q \in (1,2)$. Both stages of GMM WF outputs have higher PESQ and STOI scores compared to the output of the MWF over the −30 dB and −15 dB $\overline{\text{SDNR}}$ range. However, at $\overline{\text{SDNR}}$ of −10 dB, we observe that the enhanced signal has slightly lower PESQ and STOI scores than the output signal of the MWF. According to Fig. 13, we notice that the $\overline{\text{SDNR}}$ improvement has reached a peak at −20 dB and decreased with the increase of the $\overline{\text{SDNR}}$ for both channels. It is noticed that our experimental setup in Section 4.1.2 achieves greater performance in terms of speech quality and intelligibility compared to the DREGON dataset configuration. In the experimental validation, we set $\beta < 1$, and $\gamma > 1$ at the first stage to obtain more aggressive noise reduction whereas, in the second stage, we assign $\beta > 1$, and $\gamma = 1$ i.e., a Wiener gain, to overestimate the less significant drone noise for better noise suppression. However, the DREGON setup has a higher combined $\overline{\text{SDNR}}$ improvement through both MWF and GMM WF than our experimental setup. This seems to suggest that there is a trade-off between noise suppression and target signal distortion levels. To this end, the parametric WF needs to fine-tune its parameters to improve the performance for the best compromise.

5.4. Comparison with the baseline methods

We now compare two baseline methods: MVDR [26] and MWF found in [25] for the average input $\overline{\text{SDNR}}$ of −30 dB at different microphone channels of the array for our drone acoustic dataset. Figs. 14(a–c) show the spectrograms of the recorded signal at microphones $q \in \{7,9,11\}$, respectively. As expected, the different characteristics of the drone noise are captured based on the location of the microphones on the drone. We observe that the microphone attached to the bottom of the motor ($q = 7$ microphone in Fig. 14(a)) has more drone noise compared to the microphone far away from the drone motors and propellers such as one on the landing gear ($q = 11$ microphone in Fig. 14(c)). Fig. 14(d) provides the spectrogram of the MVDR output signal. It is noticed that the output signal has spectral energy below 300 Hz and very low energy in the speech frequency band. Note that the MVDR requires prior knowledge of the drone-related transfer functions as well as the direction-of-arrivals of the sound sources. To evaluate, a more general representation of the MWF is obtained by using (3) and (8) as in [25]

$$\mathbf{W}^{(1)}(f) = \mathbf{\Phi}_{pp}^{-1}(f,t)(\mathbf{\Phi}_{pp}(f,t) - \mathbf{\Phi}_{vv}(f,t)). \tag{27}$$

Figs. 14(e–g) illustrate the output spectrograms at the MWF in [25] at microphones 7, 9, and 11, respectively. A high energy of the signal over all frequency bands indicates that the output signals are very noisy and clipped. We observe that the outputs at both MVDR and MWF in [25] are not satisfactory. Figs. 14(h–j) present the MWF in (14) at microphones 7, 9, and 11, respectively. We observe that the large residual noise remains with respect to the frequency range of the drone noise in MWF output at each microphone. This difference in MWFs may be due to the difference of the optimal weight matrix, specifically, $\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{\Phi}_{pp}(f,t)$ vs $\mathbf{\Phi}_{pp}^{-1}(f,t)\mathbf{\Phi}_{vv}(f,t)$ terms in (14) and (27), respectively. This suggests that the optimal filter weights in (14) suppress more drone noise from the multichannel recordings.

5.5. Discussion

We list a few comments on the performance of the proposed signal enhancement algorithm.

- *Microphone array configurations*: The proposed algorithm is independent of the microphone array constellation. We demonstrated the experimental results for circular, cubic, and arbitrary microphone arrays on the drone with different numbers of microphones in each array. However, it was observed more suppression of drone noise with a great improvement in the perceptual quality and intelligibility scores for the embedded array on the drone structure than the compact microphone array mounted on the drone. As expected, more diversity of the drone noise is captured by this arbitrary microphone array on the drone with a large number of microphones ($q = 30$) than a standard microphone array mounted beneath the drone body. This may be due to utilizing the spectral, and spatial information of the drone noise more accurately by placing the microphones on the drone body. Hence, the algorithm suppresses the drone noise more accurately for the drone embedded microphone array configuration with a higher number of microphones.

- *Comparison with baselines*: The output of the baseline methods such as MVDR [26] and MWF found in [25] are not satisfactory on our drone embedded microphone array. Note that we are unable to compare the results with the machine learning based speech enhancement methods as the available drone acoustic database is not large enough for training and leads to the problem of over-fitting.
- *Robustness in real-life environment*: The algorithm is sensitive to noise correlation matrix. We showed audio signal enhancement performance for a hovering drone. The use of recursively updated drone noise, in contrast, to use pre-recorded drone noise data for estimating $\boldsymbol{\Phi}_{vv}(f,t)$ has shown much less satisfactory results as its more challenging. We leave a better estimation of the drone noise correlation matrix in more non-stable flight modes as future work. We observed that the proposed framework achieves a larger noise reduction and improves speech quality while preserving speech intelligibility at extremely low S$\overline{\text{D}}$NR conditions (up to −30 dB) in applications to speech enhancement. The proposed framework is computationally effective compared to DNN-based approaches. This algorithm facilitates dynamic noise adaptation such as motor current-specific drone noise, and ultimately improves the tracking of non-stationary drone noise for audio signal enhancement. **We also note that our method can be applied disregarding the IR measurements.**

6. Conclusion

This paper described a multichannel framework for audio signal enhancement using a drone embedded/mounted arbitrary microphone array platform. The method uses (i) the widely known MWF for a strong drone noise reduction and (ii) enhances the desired sound source signal using GMM WF with the idea of using motor current-specific data to inform the Wiener filtering techniques. This algorithm is simple and facilitates a dynamic noise adaptation, e.g., motor current-specific drone noise, and improves the tracking of non-stationary drone noise for audio signal enhancement. For drone applications where only a single sound source is present, the proposed framework has impressive performance for both speech and bird calls. Extensive experimental study confirmed that better audio signal enhancement was achieved at extremely low S$\overline{\text{D}}$NR (up to −30 dB) from drone embedded microphones.

This work can lead to multiple future research directions. The proposed framework can be verified in more complex outdoor environments, e.g., for moving target sound sources, and different flight configurations of the drone. It can also be investigated to select an optimal output from the multichannel enhanced sound source signals. Moreover, a possible extension of this paper would be to consider beamforming at the MWF output signals as it preserves spatial information.

CRediT authorship contribution statement

Wageesha N. Manamperi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thushara D. Abhayapala:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Prasanga N. Samarasinghe:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Jihui (Aimee) Zhang:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Wageesha N. Manamperi reports financial support was provided by the Australian National University.

Data availability

The audio files are shared in the dropbox folders. We note that the preprint of this manuscript is available on Techrxiv on DOI: 10.36227/techrxiv.20576823.v2.

Acknowledgement

The authors would like to thank the editor and anonymous reviewers for their constructive comments that helped to improve the presentation of this work. The authors also would like to thank Dr. Fei Ma and Mr. Xingyu Chen for their assistance with the outdoor measurements.

Appendix A. Theoretical derivation – MWF

This section presents detailed steps of the theoretical derivation of the MWF in Section 3.1.

Considering Sherman–Morrison formula [29], we expand the inverse of $\boldsymbol{\Phi}_{pp}(f)$ as

$$\boldsymbol{\Phi}_{pp}^{-1}(f)=\boldsymbol{\Phi}_{vv}^{-1}(f)-\frac{\boldsymbol{\Phi}_{vv}^{-1}(f)\Phi_{ss}(f)\mathbf{d}(f)\mathbf{d}^{H}(f)\boldsymbol{\Phi}_{vv}^{-1}(f)}{1+\mathbf{d}^{H}(f)\boldsymbol{\Phi}_{vv}^{-1}(f)\Phi_{ss}(f)\mathbf{d}(f)}. \tag{A.1}$$

Substituting (10) into (A.1) simplifies to (12). Let $\lambda(f)\triangleq tr[\boldsymbol{\Phi}_{vv}^{-1}(f)\boldsymbol{\Phi}_{xx}(f)]$. Using (10), we write

$$\lambda(f)\triangleq tr[\boldsymbol{\Phi}_{vv}^{-1}(f)\Phi_{ss}(f)\mathbf{d}(f)\mathbf{d}^{H}(f)]. \tag{A.2}$$

By using the cyclic property of $tr[\cdot]$ and noting $\Phi_{ss}(f)$ is a complex scalar,

$$\lambda(f)\triangleq\Phi_{ss}(f)\mathbf{d}^{H}(f)\boldsymbol{\Phi}_{vv}^{-1}(f)\mathbf{d}(f). \tag{A.3}$$

Putting (A.2) and (A.3) together, we get (13). Now we substitute (13) in (12) to obtain

$$\boldsymbol{\Phi}_{pp}^{-1}(f)=\boldsymbol{\Phi}_{vv}^{-1}(f)-\frac{\boldsymbol{\Phi}_{vv}^{-1}(f)\boldsymbol{\Phi}_{xx}(f)\boldsymbol{\Phi}_{vv}^{-1}(f)}{1+tr[\boldsymbol{\Phi}_{vv}^{-1}(f)\boldsymbol{\Phi}_{xx}(f)]}. \tag{A.4}$$

We can rewrite (9) using (A.4) as

$$\mathbf{W}(f)=\frac{\boldsymbol{\Phi}_{vv}^{-1}(f)\boldsymbol{\Phi}_{xx}(f)}{1+tr[\boldsymbol{\Phi}_{vv}^{-1}(f)\boldsymbol{\Phi}_{xx}(f)]}. \tag{A.5}$$

Finally, using (3), (A.5) can be expressed explicitly as (14).

References

[1] Basiri M, Schill F, Lima PU, Floreano D. Robust acoustic source localization of emergency signals from micro air vehicles. In: Proc IEEE/RSJ int conf on intell robots and syst; 2012. p. 4737–42.

[2] Doclo S, Moonen M. Gsvd-based optimal filtering for single and multimicrophone speech enhancement. IEEE Signal Process Mag 2002;50(9):2230–44.

[3] Marple Jr SL, Carey WM. Digital spectral analysis with applications. Upper Saddle River, NJ: Prentice-Hall, Inc.; 1989.

[4] Van Veen BD, Buckley KM. Beamforming: a versatile approach to spatial filtering. IEEE ASSP Mag 1988;5(2):4–24.

[5] Hioka Y, Kingan M, Schmid G, Stol KA. Speech enhancement using a microphone array mounted on an unmanned aerial vehicle. In: Proc IEEE int workshop on acoust signal enhancement; 2016. p. 1–5.

[6] Li Y, Yen B, Hioka Y. Performance evaluation on multi-channel Wiener filter based speech enhancement for unmanned aerial vehicles recordings. In: Proc INTER-NOISE and NOISE-CON congr and conf, vol. 263. 2021. p. 3584–94.

[7] Deleforge A, Di Carlo D, Strauss M, Serizel R, Marcenaro L. Audio-based search and rescue with a drone: highlights from the IEEE signal processing cup 2019 student competition [SP competitions]. IEEE Signal Process Mag 2019;36(5):138–44.

[8] Salvati D, Drioli C, Ferrin G, Foresti GL. Beamforming-based acoustic source localization and enhancement for multirotor UAVs. In: Proc Eur signal process conf; 2018. p. 987–91.

[9] Wang L, Cavallaro A. A blind source separation framework for ego-noise reduction on multi-rotor drones. IEEE/ACM Trans Audio Speech Lang Process 2020;28:2523–37.

[10] Sanchez-Matilla R, Wang L, Cavallaro A. Multi-modal localization and enhancement of multiple sound sources from a micro aerial vehicle. In: Proc ACM int conf on multimedia; 2017. p. 1591–9.
[11] Wang L, Sanchez-Matilla R, Cavallaro A. Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement. In: Proc IEEE/RSJ int conf on intell robots and syst; 2019. p. 5320–5.
[12] Wang L, Cavallaro A. Deep learning assisted time-frequency processing for speech enhancement on drones. IEEE Trans Emerg Top Comput Intell 2020;5(6):871–81.
[13] Yen B, Hioka Y, Mace B. Source enhancement for unmanned aerial vehicle recording using multi-sensory information. In: Proc Asia-Pacific signal and inf process assoc annu summit and conf; 2020. p. 850–7.
[14] Yen B, Hioka Y, Schmid G, Mace B. Multi-sensory sound source enhancement for unmanned aerial vehicle recordings. Appl Acoust Feb. 2022;189:108590.
[15] Morito T, Sugiyama O, Kojima R, Nakadai K. Partially shared deep neural network in sound source separation and identification using a UAV-embedded microphone array. In: Proc IEEE/RSJ int conf on intell robots and syst; 2016. p. 1299–304.
[16] Kim Y-J, Kim E-G. CNN based complex spectrogram enhancement in multi-rotor UAV environments. J Korea Inst Inf Commun Eng 2020;24(4):459–66.
[17] Yen B, Li Y, Hioka Y. Rotor noise-aware noise covariance matrix estimation for unmanned aerial vehicle audition. IEEE/ACM Trans Audio Speech Lang Process Jun. 2023.
[18] Li Y, Yen B, Hioka Y. Improvement of rotor noise reduction for unmanned aerial vehicle audition by rotor noise PSD informed beamformer design. In: Proc QUIET DRONES second int e-symp on UAV/UAS noise; 2022.
[19] Souden M, Benesty J, Affes S. On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE/ACM Trans Audio Speech Lang Process 2009;18(2):260–76.
[20] Cohen I, Benesty J, Gannot S. Speech processing in modern communication: challenges and perspectives. Springer Science & Business Media; 2009.
[21] Löllmann HW, Barfuss H, Deleforge A, Meier S, Kellermann W. Challenges in acoustic signal enhancement for human-robot communication. In: Proc 11th ITG symp speech comm; 2014. p. 1–4.
[22] Manamperi W, Samarasinghe PN, Abhayapala TD, Zhang J. GMM based multi-stage Wiener filtering for low SNR speech enhancement. In: Proc IEEE int workshop on acoust signal enhancement; 2022. p. 1–5.
[23] Manamperi W, Abhayapala TD, Zhang JA, Samarasinghe PN. Drone audition: sound source localization using on-board microphones. IEEE/ACM Trans Audio Speech Lang Process 2022;30:508–19.
[24] Manamperi W, Abhayapala TD, Zhang J, Samarasinghe PN. Estimating drone motor related acoustic transfer function: a preliminary investigation. In: Proc Asia-Pacific signal and inf process assoc annu summit and conf; 2020. p. 156–60.
[25] Strauss M, Mordel P, Miguet V, Deleforge A. DREGON: dataset and methods for UAV-embedded sound source localization. In: Proc IEEE/RSJ int conf on intell robots and syst; 2018. p. 1–8.
[26] Cox H, Zeskind R, Owen M. Robust adaptive beamforming. IEEE/ACM Trans Audio Speech Lang Process 1987;35(10):1365–76.
[27] Cornelis B, Moonen M, Wouters J. Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors. IEEE/ACM Trans Audio Speech Lang Process 2010;19(5):1368–81.
[28] Doclo S, Klasen TJ, Van den Bogaert T, Wouters J, Moonen M. Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions. In: Proc int workshop on acoust echo noise control; 2006. p. 1–4.
[29] Golub GH, Van Loan CF. Matrix computations. JHU Press; 2013.
[30] Chehresa S, Savoji M. MMSE speech enhancement based on GMM and solving an over-determined system of equations. In: 2011 IEEE 7th int symp intell signal proc; 2011. p. 1–5.
[31] Lim JS, Oppenheim AV. Enhancement and bandwidth compression of noisy speech. Proc IEEE 1979;67(12):1586–604.
[32] Doire CS, Brookes M, Naylor PA, Hicks CM, Betts D, Dmour MA, et al. Single-channel online enhancement of speech corrupted by reverberation and noise. IEEE/ACM Trans Audio Speech Lang Process 2016;25(3):572–87.
[33] Robinson T, Fransen J, Pye D, Foote J, Renals S. WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition. In: Proc IEEE int conf on acoust, speech and signal process; 1995. p. 81–4.
[34] Ince G, Nakadai K, Rodemann T, Hasegawa Y, Tsujino H, Imura J-i. Ego noise suppression of a robot using template subtraction. In: Proc IEEE/RSJ int conf on intell robots and syst; 2009. p. 199–204.
[35] Ito A, Kanayama T, Suzuki M, Makino S. Internal noise suppression for speech recognition by small robots. In: Proc Eur conf on speech comm and technol; 2005.
[36] Chehrehsa S. Single-channel speech enhancement using statistical modelling. Ph.D. thesis. Auckland University of Technology; 2016.
[37] Garofolo JS. TIMIT acoustic phonetic continuous speech corpus. In: Linguistic data consortium; 1993.
[38] ITU-T. Perceptual evaluation of speech quality (pesq): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union, Recommendation P.862 (2001).
[39] Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE/ACM Trans Audio Speech Lang Process 2011;19(7):2125–36.
[40] Taal C. MATLAB code for algorithms. http://www.ceestaal.nl/matlab-code/, 2011.

Wageesha N. Manamperi received the B.Sc. (Hons.) degree in Electronics and Telecommunication Engineering from the University of Moratuwa, Sri Lanka, in 2016 and the M.Sc. degree (Major Component of Research) in Wireless Communication from the University of Moratuwa, Sri Lanka, in 2018. She was a junior lecturer from 2017 to 2019 with the Department of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka. She is currently pursuing her Ph.D. degree in audio and acoustic signal processing at the Australian National University (ANU), Canberra, ACT, Australia. She has been a Ph.D. Research Intern at the Dolby, Australia. Her research interests include spatial sound recordings, sound source localization, signal enhancement, and device-related transfer function.

Thushara D. Abhayapala received the B.E. degree in engineering and the Ph.D. degree in telecommunication engineering from the Australian National University (ANU), Canberra, ACT, Australia, in 1994 and 1999, respectively. He is currently a Professor of Audio and Acoustic Signal Processing with ANU. He held a number of leadership positions, including the Deputy Dean with the ANU College of Engineering and Computer Science from 2015 to 2019, Head of the ANU Research School of Engineering from 2010 to 2014, and the Leader of the Wireless Signal Processing Program with the National ICT Australia, Australia, from 2005 to 2007. His research interests include the areas of spatial audio and acoustic signal processing, and multichannel signal processing. Among many contributions, he is one of the first researchers to use spherical harmonic based Eigen-decomposition in microphone arrays and to propose the concept of spherical microphone arrays, has shown the fundamental limits of spatial sound field reproduction using arrays of loudspeakers and spherical harmonics. This theory is now termed as higher order Ambisonics. He also made seminal contributions to the problem of the multizone sound field reproduction and ANC over space. He worked in industry for two years, before his doctoral study and has active collaboration with several companies. He has supervised 46 Ph.D. students and co-authored more than 300 peer-reviewed papers. He was the Co-chair of IEEE WASPAA 2021. He was an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. From 2011 to 2016, he was the Member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE Signal Processing Society. He is an IEEE Fellow for contributions to the theory of spherical harmonic-based spatial sound field recording, reproduction, and control. He is also a Fellow of Engineers Australia.

Prasanga N. Samarasinghe received the B.E. (Hons.) degree in electronic and electrical engineering from the University of Peradeniya, Peradeniya, Sri Lanka, in 2009, and the Ph.D. degree from Australian National University (ANU), Canberra, ACT, Australia, in 2014. She is currently an Associate Professor with the College of Engineering, Computing and Cybernetics, ANU. Her research interests include spatial sound recording and reproduction, spatial noise cancellation, and array optimization using compressive sensing techniques.

Jihui (Aimee) Zhang received the B.S. degree in measurement, control technology, and instrument from Harbin Institute of Technology, China, in 2011, the M.S. degree in instrumentation engineering from Harbin Institute of Technology, China, in 2013, and the Ph.D. degree in audio and acoustic signal processing from the ANU, Australia, in 2019. She is currently a lecturer (Assistant Professor) in the Institute of Sound and Vibration Research (ISVR), University of Southampton, the United Kingdom. She is also an honorary lecturer in the Audio and Acoustic Signal Processing (AASP) group, the Australian National University (ANU), Australia. From 2018 to 2022, she has been a Postdoctoral Research Fellow and a Research Fellow in the AASP group, ANU, Australia. From May 2018 to Aug.2018, she has been a Research Engineer Intern in SONY, Japan. She is currently a member of the IEEE and Signal Processing Society (SPS). She served as chair of Women in Engineering Affinity Group in IEEE Australian Capital Territory Section 2021-2022. Her research interests include audio signal processing, especially in spatial active noise control, spatial audio solutions for drones, and spatial audio solutions for virtual realities.