Parameterisation, Localisation, and Enhancement of Spatial Soundfield from Drone On-Board Microphones

Wageesha Nilmini Manamperi

MSc, B.Sc. Engineering (Hons) University of Moratuwa, Sri Lanka

October 2023

A thesis submitted for the degree of Doctor of Philosophy
of The Australian National University

School of Engineering
College of Engineering, Computing and Cybernetics
The Australian National University

To my beloved parents and beautiful souls around me

Declaration

The contents of this thesis are the results of original research and have not been submitted for a higher degree to any other university or institution. Much of this work has either been published or submitted for publications as journal papers and conference proceedings. Following is a list of these papers.

Journal Publications

- **W. N. Manamperi**, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone Audition: Sound Source Localization using On-Board Microphones," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 30, pp. 508 – 519, January 2022.

- **W. N. Manamperi**, T. D. Abhayapala, P. N. Samarasinghe, and J. Zhang, "Drone Audition: Audio Signal Enhancement from Drone Embedded Microphones using Multichannel Wiener Filtering and Gaussian-Mixture based Post-Filtering," in *Applied Acoustics*, vol. 216, pp. 109818, January 2024.

- **W. N. Manamperi**, T. D. Abhayapala, L. I. Birnie, P. N. Samarasinghe, and J. Zhang, " Drone Audition: On Measurements and Modelling of Drone-Related Transfer Functions," submitted to *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, February 2023.

Conference Proceedings

- **W. N. Manamperi**, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Estimating Drone Motor Related Acoustic Transfer Function: A Preliminary

Investigation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 156–160, Auckland, New Zealand, December 2020.

- **W. N. Manamperi**, P. N. Samarasinghe, T. D. Abhayapala, and J. Zhang, "GMM based Multi-stage Wiener Filtering for Low SNR Speech Enhancement," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1-5, Bamberg, Germany, September 2022.

- **W. N. Manamperi**, T. D. Abhayapala, P. N. Samarasinghe, and J. Zhang, "Drone Audition: Improved GMM Wiener Filtering Approach for Signal Enhancement", in *INTERNOISE and NOISE-CON Congress and Conference Proceedings*, Greater Tokyo, Japan, August 2023 (Peer-reviewed by the session chair).

- **W. N. Manamperi**, T. D. Abhayapala and P. Holmberg, "Microphone Aligned Continuous Wearable Device-related Transfer Function: Efficient Modeling and Measurements," in *Hands-free Speech Communication and Microphone Arrays (HSCMA) Satellite Workshop* (Accepted), April 2024.

- **W. N. Manamperi**, P. N. Samarasinghe, J. Zhang and T. D. Abhayapala, "Drone Audition: Analysis of the Preservation of Spatial Cues in Multichannel Wiener Filtering", in *Quiet Drones Conference* (Accepted), September 2024.

The following work is also results during my Ph.D. study, but not included in this thesis:

- **W. N. Manamperi**, T. D. Abhayapala, J. Zhang and Y. Hioka, "Drone audition: Use of Interior Response of a Commercial Loudspeaker for Making Drones Talk", in *Quiet Drones Conference* (Accepted), September 2024.

The research work presented in this thesis has been performed jointly with Prof. Thushara D. Abhayapala, Dr. Prasanga N. Samarasinghe, Dr. Aimee Zhang, and Eng. Paul Holmberg Approximately 80% of this work is my own.

Wageesha Nilmini Manamperi
College of Engineering, Computing and Cybernetics
The Australian National University
Canberra ACT 2600
October 2023

Acknowledgments

Without the support of a number of individuals and organisations, this work would not have been completed. I would like to acknowledge and thank each of the following:

- First and foremost, my supervisor, Prof. Thushara Abhayapala, for his understanding, patience, unconditional support and professional guidance on this research work. He trained me to be an independent researcher and provided the opportunity to learn from him. Many thanks to my co-supervisors, Dr. Prasanga Samarasinghe and Dr. Aimee Zhang, for their constructive guidance, insights and direct feedback on my research work. Appreciate their care throughout the years.

- My industry mentor, Eng. Paul Holmberg at Dolby Laboratories, Australia, for kindly welcoming me to his research group during my internship and for his collaboration on some of this work. I express my gratitude to the Australian Mathematical Sciences Institute for additional support during my internship.

- My amazing academic mentor, Dr. Xiaolin Wang, for her inspiration, invaluable guidance on both research and teaching responsibilities and for helping me to become the best version of myself.

- The Australian National University, for the PhD scholarship, the Vice Chancellor's travel grant and the Network for Early Career Teachers, Academics and Researchers mentoring program.

- My fellow students in the Audio and Acoustic Signal Processing Group, especially my seniors Yonggang, Lachlan, Huiyuan, Fahim, Fei and Felix for

their inspiration. I give extra thanks to Lachlan for his constant support and for being there to lend a helping hand with my experiments. I cherish Amy for her true friendship.

- My wonderful housemates in the Graduate House, especially Prof. Peter Kanowski, Achalie, Rebbecca, Audrey, Rabbia, Jovita, Robert and Cindy for their kindness, upstanding and for being there throughout all the struggles of a Ph.D..

- My family, especially my parents for their overwhelming love, emotional support, and endless sacrifices for my happiness. To *Thaththa*, for his encouragement on my academic pursuit. To *Amma*, for her unparalleled love, and for being there to lend an ear on my research discussions.

- Finally, all the beautiful souls around me, for their faith and pure love on my happily ever after.

Abstract

This thesis explores the acoustic signal processing on drones which can enhance the drone audition functions for search and rescue missions, wildlife monitoring, and video recording for media and filming industries. With the recent emergence of drone-based applications and services, microphone arrays are beginning to be mounted to drones. However, audio capturing using drone-mounted microphone arrays leads to very low signal-to-drone noise ratio ($\text{S}\overline{\text{D}}\text{NR}$) recordings (below negative 10 dB) due to the drone-generated motor and propeller noise signals. Hence, this causes a highly adverse noisy environment and degrades the quality and intelligibility of the recorded audio signals. Therefore, there is a growing interest in deriving new signal processing algorithms for the future deployment of the drone audition.

To this end, we develop novel methods for spatial soundfield modelling, and spatial audio processing using a microphone array embedded into the body of drones for sound source localisation and audio signal enhancement. To a lesser extent, drone noise characterisation is also studied.

We propose modelling of the drone noise *harmonics* as current to noise transfer function by taking the significant *harmonics* of the drone noise signature. The model is derived from the experimental measurements of a single motor propeller combination, and validated using extensive error analysis.

A cross-correlation-based DOA estimation method with drone noise angular spectrum subtraction is derived for acoustic source localisation. Our approach is to exploit the underlying spatial characteristics of drone noise to mitigate the noise emission by drones from the microphone recordings. The outcome resulting from this approach enables direction estimation of a single and simultaneously active multiple sound sources. Compared to the baseline approaches, the proposed localisation method is shown to be more robust to extreme $\text{S}\overline{\text{D}}\text{NR}$ conditions (up to negative 30 dB).

A framework of an audio signal enhancement using Wiener filtering approaches is developed with a substantial reduction of drone noise, and an improved signal quality and intelligibility of the target source signal. Application of this method is illustrated with extensive real data through experimental measurements for both (i) speech, and (ii) bird calls.

An efficient parameterisation of the spatial acoustic response to an on-board microphone of the drone due to a sound source, called the *drone related transfer function*, is introduced using spherical harmonics. This parameterisation is derived from the circular harmonic coefficients using a discrete set of measurements by multiple measurement points on circular arrays. The proposed approach models the scattering and diffraction response of a drone to an incident soundfield that can be used to enhance drone audition functions.

Notations and Symbols

Abbreviations and Acronyms

DOA	Direction of Arrival
DRTF	Drone Related Transfer Function
GMM	Gaussian Mixture Model
MWF	Multichannel Wiener Filter
PSD	Power Spectral Density
S$\overline{\text{D}}$NR	Signal to drone noise ratio
TDOA	Time Difference of Arrival
WF	Wiener Filter

Mathematical operations

$\lceil \cdot \rceil$	Ceiling operator
$[\cdot]^*$	Complex conjugate of a matrix
$[\cdot]^T$	Transpose of a matrix
$[\cdot]^H$	Complex conjugate transpose of a matrix
$\lvert \cdot \rvert$	Euclidean norm of a vector
$tr[\cdot]$	Trace operator
$\boldsymbol{A}^{-1}$	Matrix inverse
$\boldsymbol{A}^{\dagger}$	Matrix psuedoinverse: $\boldsymbol{A}^{\dagger} = [\boldsymbol{A}^H \boldsymbol{A}]^{-1} \boldsymbol{A}^H$
$*$	Linear convolution
$\boldsymbol{x} \cdot \boldsymbol{y}$	Dot product between two vectors
$\text{E}\{\cdot\}$	Expectation operator
i	$\sqrt{-1}$

Contents

Declaration iii

Acknowledgements vii

Abstract ix

Notations and Symbols xi

List of Figures xix

List of Tables xxvii

1 Introduction 1

1.1 Motivation and Scope 1

1.2 Problem Description and Proposed Approaches 4

1.3 Overview of Thesis 5

2 Literature Review and Background Theory 11

2.1 Acoustic Signature of Drone Noise 12

2.1.1 Noise measurements 12

2.1.2 Multichannel sound acquisition systems on drones 13

2.1.3 Drone acoustic datasets 14

2.1.4 Drone noise suppression 16

2.1.5 Research gaps 17

2.2 Source Localisation with Drone Noise Cancellation 18

2.2.1 Subspace based methods 19

2.2.2 TDOA-based methods 19
2.2.3 Beamforming-based methods 20
2.2.4 Spatial filtering 20
2.2.5 Supervised-learning-based methods 21
2.2.6 Research gaps 21
2.3 Removing Drone Noise for Source Signal Enhancement 21
2.3.1 Beamforming-based methods 22
2.3.2 Spatial filtering 23
2.3.3 Supervised-learning-based approaches 23
2.3.4 Research gaps 24
2.4 Head Related Transfer Function 24
2.5 Background Theory on Spherical Harmonic Decomposition of a Soundfield 27
2.5.1 Coordinate systems 27
2.5.2 Reduced wave equation 28
2.5.3 Spherical harmonic expansion of the soundfield 29
2.6 Summary 30

3 Characterisation of the Drone Motor-Related Acoustic Transfer Function: Preliminary Investigation 32
3.1 Introduction 33
3.2 Problem Formulation 34
3.3 Experimental Analysis 35
3.3.1 Measurements 35
3.3.2 Results 37
3.4 Current to Noise Transfer Function 40
3.4.1 Harmonic frequency 41
3.4.2 Harmonic amplitude 41
3.5 Estimation of the Harmonic Frequency Coefficients 43
3.6 Summary and Contributions 46
3.7 Related Publications 47

4 Acoustic Source Localisation with Drone Noise Angular Spectrum Subtraction 49
4.1 Introduction . . . 50
4.2 Problem Formulation . . . 51
4.3 Cross-correlation based DOA Estimation . . . 53
4.3.1 The instantaneous angular spectrum . . . 54
4.3.2 Noise angular spectrum subtraction to remove drone noise . 57
4.4 Experimental Verification . . . 58
4.4.1 Methodology . . . 58
4.4.2 Results for single source localisation . . . 62
4.4.3 Results for multi-source localisation . . . 67
4.5 Summary and Contributions . . . 76
4.6 Related Publications . . . 77

5 Drone Noise Suppression with Wiener Filtering for Multichannel Audio Signal Enhancement 79
5.1 Introduction . . . 80
5.2 Problem Formulation . . . 82
5.3 Multi-microphone Signal Enhancement Framework using Wiener Filtering . . . 84
5.3.1 Drone noise suppression with MWF . . . 85
5.3.2 Audio signal enhancement with GMM WF . . . 87
5.4 Experimental Study . . . 91
5.4.1 Algorithm properties . . . 92
5.4.2 Evaluations on outdoor measurements . . . 94
5.4.3 Evaluations on measured impulse responses . . . 96
5.4.4 Evaluations on DREGON dataset . . . 105
5.4.5 Comparison with baseline . . . 107
5.4.6 Discussion . . . 109
5.5 Interfering Source Removal Process using Improved GMM WF . . . 111
5.5.1 Problem reformulation . . . 111
5.5.2 Source enhancement using the improved GMM WF . . . 112
5.5.3 Experimental verification . . . 114

5.6 Summary and Contributions 118
5.7 Related Publications 119

6 Efficient Measurements and Parameterisation of the Drone Related Transfer Function Model 121
6.1 Introduction 122
6.2 Problem Formulation 126
6.3 Parameterisation of the DRTF 127
6.3.1 Spherical harmonic representation of the DRTF 127
6.4 Estimation of the DRTF Coefficients using Multiple Circular Arrays 130
6.4.1 DRTF decomposition by multiple circles 131
6.4.2 Extracting DRTF coefficients 132
6.4.3 Reconstruction error 133
6.5 Validation of the Proposed Parameterisation and Measurement Approach using a Rigid Sphere 134
6.5.1 Simulating the soundfield on a rigid sphere 134
6.5.2 Reconstruction performance of the proposed parameterisation 135
6.6 DRTF Measurements 141
6.6.1 Experimental methodology 142
6.6.2 Post-processing to extract the DRTFs 145
6.6.3 Experimental representation of the DRTF 146
6.7 Parameterisation of the Microphone Aligned DRTF 152
6.7.1 Extraction of microphone-aligned DRTF coefficients 152
6.7.2 Simulation study 155
6.7.3 Application in speech enhancement 159
6.8 Summary and Contributions 161
6.9 Related Publications 162

7 Conclusions and Future Work 164
7.1 Conclusions 164
7.2 Future Work 165

Appendix A Analysis of Preservation of Spatial Cues in MWF 169
A.1 Introduction 396
A.2 The Relative Transfer Function 169
A.3 Use of ReTF to Analyse the Preservation of Spatial Cues 170
A.4 Performance Measure 171
A.5 Experimental Validation 172
A.5.1 Measured impulse responses 172
A.5.2 Real-life environments 173
A.6 Application in Acoustic Localisation and Conclusion 174

Bibliography 177

List of Figures

1.1 Human-UAV interactions: Making drones talk and listen. 2
1.2 Illustration of the drone audition system addressed in this thesis. . . 3
1.3 Outline of this thesis. 6

2.1 The spectrum of the recorded drone noise at microphone 1 for motor speed of 80 turns/s from DREGON dataset. 15
2.2 Signal energy of the drone noise from Recording 6,9,12,14 and 15 at different altitudes at microphone channel 1 from AIRA-UAS dataset. 16
2.3 Illustration of a direct HRTF measurement setup. 25
2.4 Illustration of the spherical coordinate system. 27

3.1 'Black Box' model: Drone motor-related acoustic transfer function for a given microphone takes motor current i and outputs the signal received by the microphone $v(t, i)$. 34
3.2 Experimental setup with a dual microphone module attached to the body of the drone closer to the selected motor. 36
3.3 Time-frequency plots of the recorded drone motor noise at different motor current ratings. 37
3.4 Frequency spectrum of the recorded drone motor noise at 1000 mA and 100 mA with background noise. 38
3.5 Harmonic frequency against input motor current. 39
3.6 Real part of the harmonics amplitude, $\Re(a_n)$ against input motor current, i. 39
3.7 Imaginary part of the harmonics amplitude, $\Im(a_n)$ against input motor current, i. 40

3.8 Harmonic strength as a function of current (i) and frequency (ω_n) for $n \in [1, 2, \ldots, 10]$. 42
3.9 MSE estimator of $\widehat{\omega}_{\mathbf{n}}$ for $n \in [1, 2, \ldots, 10]$ with respect to the order of the polynomial M. 44
3.10 Harmonics frequency, ω_n, $n = 1, \ldots, 4$ mapped with the motor current, i. Markers are actual data whereas lines are polynomial fit. 45
3.11 Harmonics frequency, ω_n, $n = 5, \ldots, 10$ mapped with the motor current, i. Markers are actual data whereas lines are polynomial fit and dotted lines are 95% confidence interval. 46

4.1 Problem setup of an irregular microphone array on a drone in the presence of multiple sound sources. 52
4.2 Illustration of the computation of the instantaneous angular spectrum, $\Psi(n, k, \tau_j)$ from the recordings. 56
4.3 Experimental setup inside a semi-anechoic chamber, where the loudspeaker placements are shown by the white markers on the floor. . . 58
4.4 The drone-embedded with 15 dual microphone modules. 59
4.5 Instantaneous angular spectra over (θ, ϕ) for the proposed method at low S$\overline{\text{D}}$NR levels for a sound source position at $(\theta_\ell = -44°, \phi_\ell = 0°)$. 62
4.6 Performance comparison over (θ, ϕ) at S$\overline{\text{D}}$NR level of -15 dB for a sound source at $(\theta_\ell = -44°, \phi_\ell = 0°)$. 63
4.7 Performance comparison over (θ, ϕ) at S$\overline{\text{D}}$NR level is -20 dB with a sound source position at $(\theta_\ell = -44°, \phi_\ell = 0°)$. 64
4.8 Performance comparison over (θ, ϕ) at S$\overline{\text{D}}$NR level is -25 dB with a sound source position at $(\theta_\ell = -44°, \phi_\ell = 0°)$. 65
4.9 Performance comparison over (θ, ϕ) at S$\overline{\text{D}}$NR level is -30 dB with a sound source position at $(\theta_\ell = -44°, \phi_\ell = 0°)$. 66
4.10 Instantaneous angular spectrum over (θ, ϕ) for the proposed method where the azimuth angular difference between the two adjacent sound sources in (a,d) 30°, (b,e) 60°, and (c,f) 90°, and the S$\overline{\text{D}}$NR level of in (a-c) -15 dB, and (d-f) -10 dB. 68

4.11 Instantaneous angular spectra over (θ, ϕ) for the proposed method at S$\overline{\text{D}}$NR level of -30 dB and azimuth angular difference between the two adjacent sound sources in (a) 60°, (b) 90°, and (c) 120°. . . 69

4.12 MUSIC spectrum over (θ, ϕ) for azimuth angular difference between the two adjacent sound sources in (a) 30°, (b) 60°, and (c) 90°. The S$\overline{\text{D}}$NR level is -10 dB. 70

4.13 Performance comparison over (θ, ϕ) at the S$\overline{\text{D}}$NR level of -15 dB with the sound sources positions at $(\theta_1 = -44°, \phi_1 = 0°)$ and $(\theta_2 = -44°, \phi_2 = 90°)$. 71

4.14 Performance comparison over (θ, ϕ) at the S$\overline{\text{D}}$NR level of -20 dB with the sound sources positions at $(\theta_1 = -44°, \phi_1 = 0°)$ and $(\theta_2 = -44°, \phi_2 = 90°)$. 72

4.15 Instantaneous angular spectra over (θ, ϕ) for the proposed method where the azimuth angular difference between the two adjacent sound sources is 90°, and the S$\overline{\text{D}}$NR level difference w.r.t to -25 dB in (a) 5 dB, (b) 10 dB, (c) 15 dB, and (d) 20 dB. Note that the red triangle in (d) indicates a false peak detection. 73

4.16 MUSIC spectra over (θ, ϕ) for azimuth angular difference between the two adjacent sound sources is 120°, and the S$\overline{\text{D}}$NR level difference w.r.t to -20 dB in (a) 0 dB, and (b) 15 dB. 75

5.1 A block diagram of the proposed multichannel audio signal enhancement framework. 84

5.2 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 (5.19) and (5.20). 88

5.3 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). . . 96

5.4 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 $\overline{\text{SDNR}}$ level is at -30 dB whereas the input $\overline{\text{SDNR}}$ level at channel 11 is -24 dB. 98

5.5 Performance evaluation of the proposed framework at the MWF. (a) Average $\overline{\text{SDNR}}$ improvement, (b) average PESQ scores, and (c) average STOI scores (in percent). 99

5.6 Spectrogram of (a) the MWF output, (b) stage I, and (c) stage II of the GMM WF at channel 11. 100

5.7 GMM mean vector plots of residual drone noise for $K_{\widehat{v}_q} = 13$ for $q = 11$. The selection of the set of mean vectors $K_{n,\widehat{v}_q}$ sets to (a) 7 for $n = 1$, and (b) 6 for $n = 2$. 101

5.8 PESQ scores at channel q where $q \in [7, 12]$ with very low average input $\overline{\text{SDNR}}$ levels. 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 at stage 1 and 2, respectively. 102

5.9 STOI scores (in percent) at channel q where $q \in [7, 12]$ with very low average input $\overline{\text{SDNR}}$ levels. 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. 103

5.10 $\overline{\text{SDNR}}$ improvement in the GMM WF at channel q where $q \in [7, 12]$ with very low average input $\overline{\text{SDNR}}$ levels. Dash-dotted and solid lines represent the output at the GMM WF in stage 1 and 2, respectively. 104

5.11 Performance of the MWF outputs for the DREGON dataset. (a) Average $\overline{\text{SDNR}}$ improvement, (b) average PESQ scores, and (c) average STOI scores (in percent). 106

5.12 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 levels. Dotted line represents the input at the GMM WF whereas the dash-dotted and solid lines represent the output at the GMM WF in stages 1 and 2, respectively. 107

5.13 Comparison with baseline methods at average input S$\overline{\text{D}}$NR 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 from typical MWF for channel $q \in \{7, 9, 11\}$, and (h-j) the output at MWF derived in the proposed method for channel $q \in \{7, 9, 11\}$. 108

5.14 Spectrogram plots at $q = 3$ microphone for the outdoor measurements: (a) input at the MWF, (b) output at the MWF, (c) stage one, and (d) stage two output at the improved GMM WF. 115

5.15 Spectrogram plots at $q = 11$ microphone: (a) the clean speech, (b) speech and traffic noise only, (c) input at the MWF, (d) output at the MWF, (e) stage one, and (f) stage two output at the improved GMM WF. 116

5.16 Performance evaluation of the proposed framework with the improved GMM WF at different average input S$\overline{\text{D}}$NR levels at $q = 11$ microphone. Row 1 (in blue): Input SINR of -10 dB, Row 2 (in red): Input SINR of -5 dB, Row 3 (in black): Input SINR of 0 dB. 117

6.1 The spherical coordinates system in DRTF measurement for drone-embedded microphones. 127

6.2 Example of multiple circular array structure for DRTF measurement. 130

6.3 Magnitude of the DRTF of $q = 3$ microphone on the rigid sphere at $f = 750$ Hz with truncation order $N = 1$ ((a), (b) and (c)) and $f = 1125$ Hz with $N = 1$ ((d), (e) and (f)). 137

6.4 Magnitude of the DRTF of $q = 3$ microphone on the rigid sphere at high frequency $f \in \{1500, 2250, 3000\}$ Hz for theory ((a), (b) and (c)) and simulated measurement ((d), (e) and (f)). 138

6.5 The simulated measurement: DRTF reconstruction error performance the proposed parameter estimation method (a) $\epsilon^{(q)}$, (b) $\epsilon^{(q)}_{\text{mag}}$, and (c) $\epsilon^{(q)}_{\text{phase}}$, computed as in (6.16), (6.17), and (6.18), respectively, at the frequency range of [0.2, 5] kHz for the q^{th} microphone where $q \in \{1, 2, 3, 4\}$ 139
6.6 The simulated DRTF reconstruction error performance through the proposed parameter estimation method at the frequency range of [0.2, 5] kHz for the $q = 3$ microphone over different sets of measurements. 141
6.7 A more practical setup for the proposed DRTF measurements using a turntable and a height adjustable loudspeaker. 142
6.8 Experimental setup using direct DRTF measurements. The dotted circles show the multiple circular loudspeaker arrays. 143
6.9 The drone-embedded with 15 dual-microphones. 144
6.10 Microphone placements of $q \in \{1, 2\}$ on the drone with different radii based on its structure. 147
6.11 Magnitude of the DRTFs for *rot-microphone* ((a), (b) and (c)) and *top-microphone* ((d), (e) and (f)). 148
6.12 The experimental DRTF reconstruction error performance through the proposed parameter estimation method across the measurement points (a) $\epsilon^{(q)}$, (b) $\epsilon^{(q)}_{\text{mag}}$, and (c) $\epsilon^{(q)}_{\text{phase}}$, computed as in (6.16), (6.17), and (6.18), respectively, at the frequency range of [0.2, 2] kHz for the q^{th} microphone where $q \in \{1, 2\}$ 150
6.13 The experimental DRTF reconstruction error performance through the proposed parameter estimation method with respect to the ground truth (a) $\epsilon^{(q)}$, (b) $\epsilon^{(q)}_{\text{mag}}$, and (c) $\epsilon^{(q)}_{\text{phase}}$, computed as in (6.16), (6.17), and (6.18), respectively, at the frequency range of [0.2, 2] kHz for the q^{th} microphone where $q \in \{1, 2\}$ 151
6.14 The approximate microphone positions on the AR glasses are shown with the corresponding channel number. Channels 5 and 6 are binaural microphones. 155
6.15 Impulse response measurement grid of the EasyCom dataset. . . . 156

6.16 The magnitude of the WDRTF of $q = 2$ microphone at 1125 Hz over θ and ϕ using 2^{nd} order $d_{nm}^{(q)}$ using our WDRTF parameter estimation method. 157
6.17 The microphone-aligned WDRTF measurement reconstruction error performance through the proposed method at the frequency range of $[0.2,\ 5]$ kHz of $q = 2$ microphone. 158
6.18 The recording setup is shown along with the participant ID numbers, and the concentric rings indicate approximate loudspeaker locations around the room. 159
6.19 Spectrograms of beamforming examples. (a,c,e) Scenario 1: one speaker is active at a given time and (b,d,f) scenario 2: two talkers are active simultaneously. 160

A.1 The distance function: Hermitian angle Θ between the input and output ReTF vectors over frequencies at very low S$\overline{\text{D}}$NR levels. . . 172
A.2 Hermitian angle Θ between the input and output ReTF vectors averaged over frequency bins at very low S$\overline{\text{D}}$NR levels. 173
A.3 The distance function: Hermitian angle Θ between the input and output ReTF vectors over frequencies. 174

List of Tables

2.1 Existing multichannel sound acquisition systems on drones. P: Portable recorder, V: intelligent Voice interface; S: Sound card 14

4.1 The prediction error for single source localisation. Note that the lowest prediction error for a given position is highlighted. 67

4.2 The prediction errors of the proposed method for multi-source localisation under various $\overline{\text{SDNR}}$ conditions. Note that the prediction error convention is given as (first source/second source). 68

4.3 The prediction errors for the sound sources position at ($\theta_1 = -44^\circ, \phi_1 = 0^\circ$) and ($\theta_2 = -44^\circ, \phi_2 = 90^\circ$). Note that the prediction error convention is given as (first source/second source). 72

4.4 The prediction errors of the proposed method for adjacent sound sources in multi-source localisation scenario with varying $\overline{\text{SDNR}}$ level difference. Note that the prediction error convention is given as (first source/second source). 74

4.5 The prediction errors of the MUSIC algorithm for adjacent sound sources in multi-source localisation scenario with varying $\overline{\text{SDNR}}$ level difference. Note that the prediction error convention is given as (first source/second source). 75

6.1 Design formulation of the DRTF measurements for a rigid sphere using multiple circles. 136

6.2 Microphone placement on the rigid sphere. 136

6.3 Loudspeaker locations of the DRTF measurement setups using multiple circular arrays. 140

6.4 Experimental settings of the DRTF measurements using 25 circles to determine DRTF coefficients. 145
6.5 List of number of truncation orders for the choice of radii on the drone with $r_0 = 3.5$ cm, $r_1 = 12$ cm, and $r_2 = 22.3$ cm. 147
6.6 Loudspeaker locations of the WDRTF measurement setups using multiple circular arrays. 156

A.1 The prediction errors Δ over extreme S$\overline{\text{D}}$NR levels. 175

Chapter 1

Introduction

1.1 Motivation and Scope

In recent years, unmanned aerial vehicles (UAVs) or drones have received significant attention as they emerge with many potential applications in a wide range of areas. This includes activities in commercial, recreational, and military spaces. For example, in the application of consumer package delivery [1], emergency rescue missions [2], firefighting [3], surveillance [4], video recording/filming [5, 6], wildlife management [7], agriculture, and health sector industry. Companies such as Amazon, Wing, UPS, Zipline and Domino are the key players in this industry.

For these applications, advancement in measurements, and sensing technologies are integrated with drone platforms. Among them, the most common sensors include global positioning systems (GPS), inertial measurement units (IMU), accelerometer, gyroscope, barometric pressure, and light detection and ranging (lidar) sensors together with a digital/optical or infrared camera, and a MEMS microphone [8]. While drones utlise multiple modalities for communication and information capture, video is of predominant importance as audio recordings. Despite its advantages of visual information, the vision can be occluded due to adverse lighting and environmental conditions, and hence the audio information may act as the sole contributor for sensing sound sources. Of course, this needs to be achieved with suitable methods for drone noise reduction from on-board microphone signals making acoustic scene analysis feasible. Thus, there is a significant interest in audio

signal processing to deliver better drone-based services [9].

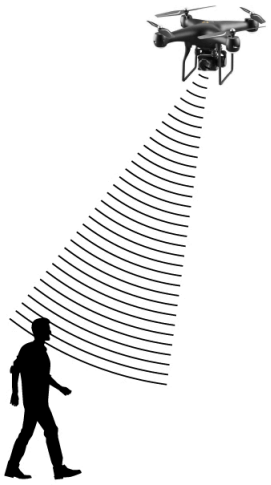


Figure 1.1: Human-UAV interactions: Making drones talk and listen.

Today, human-UAV interaction attracts considerable interest in delivering the aforementioned applications and services (as shown in Fig. 1.1). This raises the requirement of having both listening and speaking capabilities for the drones by including more microphones and loudspeakers to mitigate self-generated noise from drones. Processing of audio recordings from microphone arrays embedded to the drones on real-world scenes has been referred to as drone audition [10], whose function includes sound source localisation [11–14], source separation [15], source tracking [16, 17], and signal enhancement [18, 19]. There also exist some works utilising both audio-video based measurements for scene analysis such as sound source localisation [19, 20], and speech enhancement [19, 20].

However, the drone on-board microphones mostly capture the loud self-generated noise from the drone motors and propellers, which are closer to the microphone array than sound sources on the ground or elsewhere. Therefore, any sound originating from the ground is obscured by the drone noise 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. This hinders the widespread deployment of drone-based applications as it also contributes to privacy, and noise pollution concerns on exposed communities/public [21].

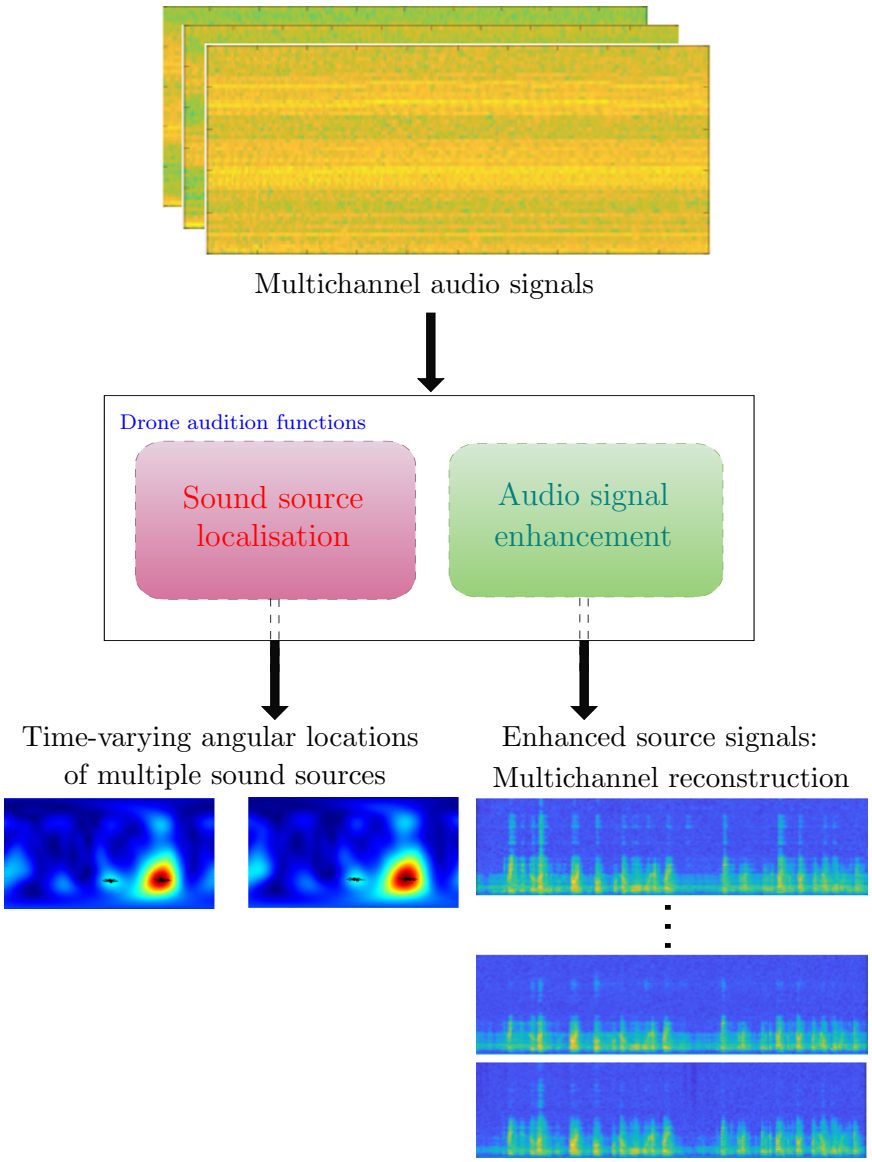


Figure 1.2: Illustration of the drone audition system addressed in this thesis.

While there exists a large number of recent works on audio signal processing for drones ranging from source localisation to signal classification [2, 12, 14, 22–24], the aforementioned challenge of drone related noise reduction is still very much an open problem. The majority of the noise around the drone is produced by itself. As such, there exists a great need to systematically understand and mitigate the noise emission by drones in order to enhance the user experience with drone

audition [25].

In Figure 1.2 we illustrate an example of the drone audition system which takes multichannel audio signals as an input to the system and outputs time-varying angular positions, e.g., elevation and azimuth angles of multiple sound sources for sound source localisation, whereas multichannel reconstruction of the source signals for audio signal enhancement. To accurately estimate the following outcomes for the drone audition functions, the suppression of the drone noise needs to be performed from the multichannel recordings of the drone-embedded microphone array.

Based on the above discussions, we pose the following research question to address in this thesis:

How can drone on-board multiple microphone recordings be processed for future deployment of the drone audition while overcoming the strong self-generated noise?

1.2 Problem Description and Proposed Approaches

We elaborate this research question raised by the thesis into three further subquestions and provide some intended solutions to each of the questions:

(i) How to analyse the acoustic characteristics of drones using drone-embedded microphones, which can then be suppressed using suitable methods?

Solutions: We begin with a preliminary study on modelling the relationship between input motor current and acoustic noise produced by a drone. We operate a single motor propeller combination to learn the drone noise signature using drone-embedded microphones. We conduct our experimental study indoors with varying motor currents and measure the drone noise using a single on-board microphone. We identify multiple tones or *harmonics* in the drone noise spectrum that vary proportionally to the motor current. We then model the current-to-noise transfer function, considering the significant *harmonics*, and estimate its parameters from the experimental recordings.

(ii) How can audio and motor modalities be combined in drone audition for acoustic scene analysis, such as sound source localisation,

and source signal enhancement?

Solutions: Once we learn the acoustic characteristics of drones, we can suppress the drone noise from the microphone recordings. We make use of the prior-measured motor current-specific drone noise with all motors operating at a steady state of manoeuvre to spatially remove the drone noise from the multichannel recordings. We develop two processing algorithms for acoustic scene analysis. We first propose an approach to localise single and multiple sound sources using drone-embedded microphone array with motor current-specific drone noise data. We process the microphone recordings and subtract the learned drone noise characteristics to accurately localise the sound sources. We then develop a framework to enhance the sound source signals from the microphone recordings at very low $\overline{\text{SDNR}}$ levels up to -30 dB. We exploit the known statistics of the motor current-specific drone noise along with an unsupervised learning technique to accurately estimate the drone noise power spectral density (PSD) from the microphone recordings to enhance the target source signals.

(iii) How can drone-related transfer functions (DRTFs) be modelled and parameterised?

Solutions: Applications of drone audition can benefit from the spatial acoustic response of a drone due to a sound source to an on-board microphone of the drone known as the drone-related transfer function (DRTF). The DRTF contains all of the relevant spatial cues including scattering and diffraction. Therefore, we propose a novel approach for measuring and modelling the DRTF with a discrete set of loudspeaker positions by multiple measurement points on circular arrays. We decompose the experimental measurement in the spherical harmonic domain. We also propose using the microphone aligned DRTF for parameterisation, which can further improve the reconstruction performance.

1.3 Overview of Thesis

Motivated by the above problems, this thesis provides some useful tools for spatial soundfield modelling and spatial audio processing using a drone-embedded microphone array. Figure 1.3 illustrates the flow diagram of this thesis. The key contributions to drone audition in this thesis are distributed into four chapters

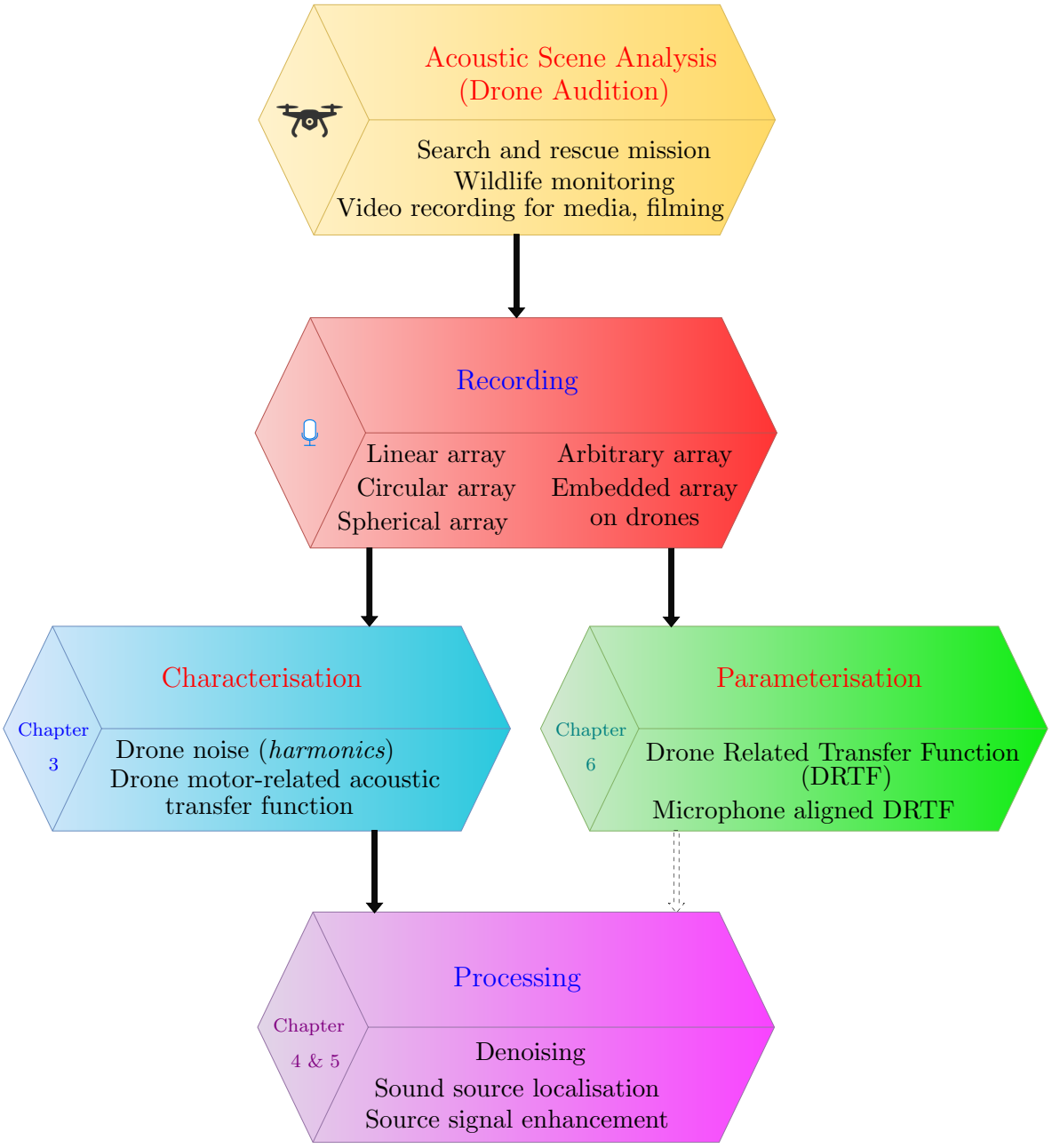


Figure 1.3: Outline of this thesis.

(3-6). Additionally, we begin with a chapter for literature and theoretical review, and end with a chapter on overall conclusions and future work. We summarise the content and contributions of each chapter as:

Chapter 2: Literature Review and Background Theory

This chapter provides a comprehensive literature review of drone audition. A brief review of the head-related transfer function is also given, as well as a brief overview of the background theory related to spherical harmonic analysis for spatial sound, which is used in Chapter 6 of this thesis.

Chapter 3: Characterisation of the Drone Motor-Related Acoustic Transfer Function: Preliminary Investigation

In this chapter, we present our preliminary work on modelling the relationship between input motor current and acoustic noise produced by a drone. This study is conducted indoors for a drone under hovering manoeuvre with a single active motor and propeller. The drone noise is measured by a single on-board microphone. We model the harmonic structure in the drone noise spectrum that varies proportionally to the motor current. We estimate the model parameters and validate this model through a detailed error analysis.

Chapter 4: Acoustic Source Localisation with Drone Noise Angular Spectrum Subtraction

This chapter aims to localise the sound sources using an embedded microphone array on a drone. We present a novel cross-correlation-based DOA estimation method for multiple sound source localisation at extreme S$\overline{\text{D}}$NR levels. This is achieved by using the time-difference-of-arrivals (TDOAs) at different microphone pairs, with drone noise angular spectrum subtraction. The method exploits known characteristics of drone noise such as motor current-specific drone noise for better noise suppression. We subtract the angular spectrum of the current-specific drone noise from the angular spectrum of the microphone recordings before averaging across the microphone pairs. The method can be applied to any microphone array geometry. We demonstrate both single and two simultaneously active sources localisation performance through experimental recordings. We compare our method to the baseline methods.

Chapter 5: Drone Noise Suppression with Wiener Filtering for Multichannel Audio Signal Enhancement

In this Chapter, we propose a solution using Wiener filtering techniques to remove drone noise and enhance the target sound source signal from a drone on-board microphone array recordings. The drone noise suppression is achieved by the multichannel Wiener filter (MWF) and further reduction of the residual noise is obtained using the Gaussian Mixture Model (GMM) Wiener filter (WF). The method exploits the known statistics of motor current-specific drone noise in both MWF and the GMM WF. We validate this audio signal enhancement framework through (i) experimental recordings from two different drone acoustic datasets and (ii) outdoor measurements from a hovering drone for a bioacoustic application. We also assess the sound quality and intelligibility of the enhanced signals. Furthermore, we generalise the proposed drone noise removal process in GMM WF for interfering sources and evaluate the noise reduction and signal enhancement performance using real recordings.

Chapter 6: Efficient Measurements and Parameterisation of the Drone Related Transfer Function Model

This chapter proposes measurements and modelling of the scattering and diffraction of a drone body beyond free-field propagation, as DRTFs. We present a novel method for DRTF measurements using a discrete set of loudspeaker positions by multiple measurement points on circular arrays. The proposed method is a practically viable solution for typical impulse response measurement setups and models the scattering and diffraction response of a drone to an incident soundfield that can be used to enhance drone audition functions. The method is built upon the spherical harmonic representation. An efficient parameterisation of the DRTF is achieved using circular harmonic coefficients. We validate this method using a simple rigid sphere acting analogous to a drone through both theoretical simulation and experimental measurements using an Em32 Eigenmike spherical microphone array. We demonstrate the DRTF representation using a real drone-embedded with microphones using a circular loudspeaker array. Furthermore, we re-define an efficient parameterisation of a continuous DRTF model based on microphone

alignment using spherical harmonic expansions. We evaluate and compare its reconstruction performance over different measurement setups with a head-worn microphone array device using anechoic measurements. We also utilise the derived microphone-aligned DRTF model in an application for beamforming through a real-life environment.

Chapter 7: Conclusion and Future Work

This chapter concludes the thesis and provides possible directions for future research.

Chapter 2

Literature Review and Background Theory

This chapter provides an overview of the up-to-date research progress on drone audition and background knowledge concerning spherical harmonic analysis for spatial soundfield recording. We first review the literature on the acoustic signature of drone noise, and drone audition functions such as sound source localisation, and speech enhancement. We also briefly explore the literature on the Head Related Transfer Functions (HRTFs). We then review the spherical harmonic expansion of a soundfield and its properties.

2.1 Acoustic Signature of Drone Noise

Unmanned aerial vehicles (UAVs) or drones are rapidly transforming and/or replacing many activities in commercial, recreational, and military spaces. Compared to a conventional aircraft there is a unique acoustic signature of the noise generated by drones [26, 27].

The acoustic spectrum of a typical drone is composed of tonal noise and broadband noise where tonal noise generates discrete tones characterised as *harmonics* whereas the broadband noise is due to turbulence and the airflow through the propeller blades [28, 29]. It is suggested that the tonal noise also known as discrete frequency noise consists of four primary noise sources: thickness noise, loading noise, blade–vortex interaction (BVI) noise, and the high-speed impulsive (HSI) noise [30], while the broadband noise consists of various distinctive noise sources such as non-deterministic loading noise, turbulence ingestion noise, leading and trailing-edge noise, and blade–wake interaction noise [30]. 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 pressure, humidity, and wind speed conditions [31].

2.1.1 Noise measurements

The drone noise profile based on a full or partial drone or isolated propeller and motor is typically investigated in one of two ways. The first, is by numerical models using computational tools such as Computational Fluid Dynamics (CFD), Vortex methods, Blade Element Theory, Momentum Theory or a combination of them [30, 32–36].

The second method is using experimental approaches inside anechoic or semi-anechoic environments. A significant number of works on drone noise analysis includes in-flight measurements [10, 22, 37–44]. Many works related to drone noise analysis often also include full or partial drones stationary mounted in an anechoic environment [45–47]. Another approach to drone noise analysis discussed isolated drone propellers statically mounted in an anechoic chamber under various flight

conditions such as hovering and forward flight stages at different propeller rotation speeds [48–51]. It is identified that the main sources of drone noise are due to motors and propellers [29]. In [29], an embedded microphone array was used to model drone noise in terms of an equivalent source distribution plus a diffuse field. In another work, drone noise modelling uses aeroacoustic models to predict the drone noise generated from motor and propeller combinations [28,52–55]. However, given prolonged time and computational constraints of the numerical model it is desirable to utilised the real-time experiment to obtain an accurate drone noise profile. We therefore discuss the acoustic signature of drone noise further through experiments.

Most of the previous work has presented drone noise analysis using free-field or ground-based microphone arrays [38, 56, 57]. Such analysis using on-board microphones is rare. Although not directly performing drone noise analysis, some other loosely related works include machine learning-based feature extraction and classification of drone noise for acoustic detection of UAVs [58–61], and passive drone noise reduction via propeller design optimisation methods (lighter propeller, shroud design the use of ducts) [62].

2.1.2 Multichannel sound acquisition systems on drones

A single on-board microphone on a drone is unable to capture spatial information for successful deployment of the drone audition. Therefore, a more advanced approach to soundfield recording from the drone mounted microphones known as an on-board sound acquisition system is required to capture multichannel signals, synchronise, and convert them to multichannel digital signals for further processing. So far, there is only one commercially available multichannel drone sound recording system in which technical information remains undisclosed to the public [70].

Numerous on-board multichannel sound acquisition systems for drones have been presented [11, 14, 19, 22, 23, 63, 64, 66, 67, 69, 71]. These systems are detailed in Table 2.1.

Table 2.1: Existing multichannel sound acquisition systems on drones. P: Portable recorder, V: intelligent Voice interface; S: Sound card

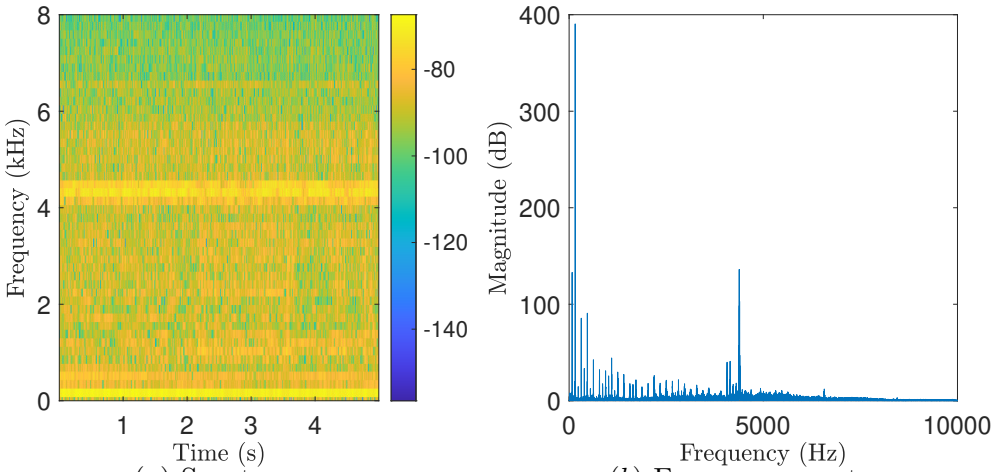
Reference	Number of microphones	Audio interface	Shape of the array	Remarks
[63]	6	Zoom H6	T-shape	P
[64]	8	Zoom H24	Circular	P
[65]	6	Respeaker + Raspberry Pi	Circular	V
[66]	7	UMA-8 array + Raspberry Pi	Circular	V
[67]	8	MiniDSP USBStreamer	Circular	S
[22]	8	8SoundUSB	Cubic	S
[68]	8	8SoundUSB	Circular	S
[14]	12	RASP-ZX	Spherical	S
[23]	8	RASP-24	Circular	S
[11]	16	RASP-ZX	Octagon	S
[69]	8	Bela	Circular	S

2.1.3 Drone acoustic datasets

There exist drone acoustic datasets through on-field measurements considering different intrinsic and extrinsic characteristics [10, 19, 22, 71].

Drone egonoise and localisation (DREGON) dataset [22]

In [22], a UAV-embedded dataset called DREGON is presented with the rotational speed of the propeller for different flight configurations, with or without a target source presence. A quadrotor UAV equipped with an eight-microphone cube-shaped array is used indoors with a motion capture system to obtain the ground truth data of the drone and sound source. The authors observed a relationship between the spectral harmonic components and motor speeds. Figure 2.1(a) and (b) show the recorded drone noise from the on-board microphone channel 1 for an individual motor rotating speed of 80 turns/s in hovering manoeuvre. The spectrum is observed to contain a set of peaks along the harmonics that is proportional to approximately 80 Hz. More prominent harmonics over the drone noise spectrum such as the second harmonic which is equivalent to the blade passing frequency of the propeller are also observed. This confirms that the acoustic characteristics of drones are proportional to the operating motor speed.



(a) Spectrogram (b) Frequency spectrum

Figure 2.1: The spectrum of the recorded drone noise at microphone 1 for motor speed of 80 turns/s from DREGON dataset.

Audio-visual dataset [19]

An audio-visual dataset recorded outdoors from a fixed drone with a camera and an eight-microphone circular array mounted on the top of the drone is presented in [19]. The dataset is comprised of both static and moving sound sources for acoustic scene analysis with multi-modal sensing capabilities.

Acoustic interactions for robot audition - unmanned aerial systems (AIRA-UAS) dataset [71]

In [71], the drone noise signature is recorded from three different flying drones with an eight-microphone circular array mounted on the top of the drone at multiple heights as the AIRA-UAS dataset. Although this dataset provides the drone noise recording at varying altitudes for a hovering drone such as $\{3, 5, 10, 15, 20\}$ m, the interpretation of the data is limited due to the unavailability of the other relevant information such as motor speed data. For example, the recorded signal energy for the microphone 1 at five different altitudes is illustrated in Fig. 2.2. All five plots are observed with a harmonic structure and increased signal energy around the frequency content of 300 to 2 kHz. Overall, it is observed that drone hovering

at 15 m altitude has the highest energy than at an altitude of 20 m. However, all five plots have the same pattern with the increase of the frequency but also have difficulty providing detailed insights into drone noise profiles based on the observations.

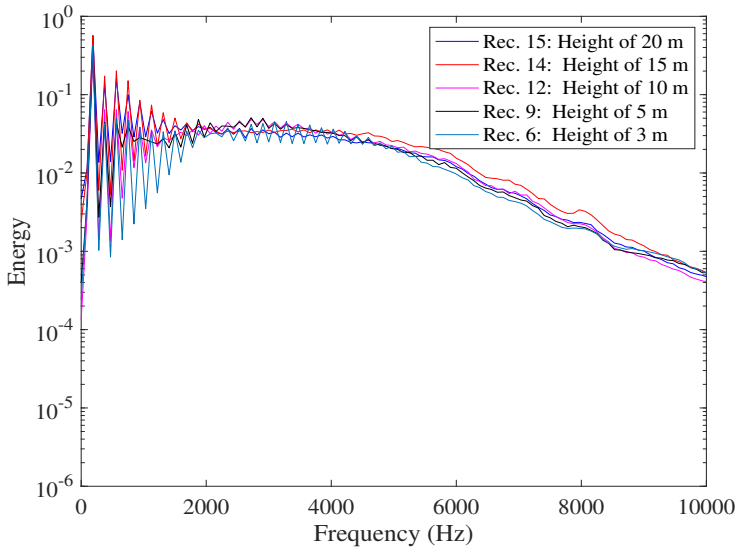


Figure 2.2: Signal energy of the drone noise from Recording 6,9,12,14 and 15 at different altitudes at microphone channel 1 from AIRA-UAS dataset.

2.1.4 Drone noise suppression

The current work on drone noise reduction can be divided into unsupervised and supervised methods. The common forms of unsupervised approaches suppress the drone noise from the microphone array recordings through beamforming, where the recordings are processed to create a sharp directivity pointed toward a specific point of interest [63, 72–76], spatial filtering-based methods which exploit the time-frequency sparsity of the drone noise harmonics [64, 77], blind source separation, where the multichannel recording is separated to the target sound and the drone noise at individual frequency bins using independent component analysis (ICA) [29, 77], and their combination with a post-filtering [78].

The supervised approaches utilise both microphone array recordings and the additional sensors to supervise the status of the motors or the flight configurations

for predicting the drone noise. This includes template-based approaches [79] and correlation matrix-based methods [24], where the operating drone motor speeds are monitored by using a motor-speed sensor. For example, in [24], the predicted drone noise correlation matrix using motor speed data which is proportional to drone noise harmonics, utilises jointly with subspace-based method to localise the target sound source. The use of additional sensors for monitoring can be eliminated by learning the drone noise bases from the pre-recorded noise datasets such as the use of Nonnegative Matrix Factorisation (NMF). There exist NMF-based methods for ground robots [80–83] which use motor data-driven dictionaries to predict the self-generated noise, however, its application for drone noise reduction has not been discussed yet.

The supervised approaches to drone noise reduction also include reference-based methods which use a reference microphone placed close to drone motors to capture the drone noise and then cancel it adaptively jointly with the multichannel recordings [84–86]. Recently favoured are machine learning-based methods for drone noise suppression [66, 87–89]. For example, a denoising autoencoder based on fully convolutional neural networks to cancel drone noise was discussed in [89].

Another category of methods includes spherical sector harmonics-based methods [90,91]. In [90], drone tonal noise reduction method is proposed for a directional spherical sector region. A more detailed survey for the suppression of self-generated noise on both ground robots and drones was discussed in [92].

A brief list of the previous studies explored for the optimum locations of microphones to be mounted on drones was discussed in [64, 72, 74, 93]. Moreover, an accurate estimation of the drone noise correlation matrix for drone flight configuration which is aware of drone motor-propeller noise to improve performance for drone audition with application to speech enhancement is presented in [94].

2.1.5 Research gaps

Even though some drone acoustic datasets are publicly available, their characteristics interpretation is limited due to the unavailability of other relevant information such as motor speed data (except in [22]), making them less useful in drone noise modelling. Among many of the aforementioned drone noise measurement proce-

dures, modelling and analysis of the drone noise from a single drone motor-propeller combination at different operating modes are not well explored. This understanding would be a benefit to suppress the drone noise for drone audition functions such as source localisation (i.e., locating the desired sound source's position), and source enhancement (i.e., enhancing the target source signal).

We review current approaches for acoustic source localisation using drone on-board microphones in the following.

2.2 Source Localisation with Drone Noise Cancellation

The problem of sound source position estimation relative to the microphone array known as source localisation has been widely studied in recent years related to acoustic scene analysis such as signal enhancement [95], source separation [96–98], and source tracking [99–102], as well as for robot audition applications [9, 92, 103–111]. Numerous methods to single and multiple sound source localisation have been introduced including, time-difference-of-arrivals (TDOAs) based methods [112–122], subspace based methods [123–133], beamforming or spatial filtering based method [134–150], maximum likelihood optimisation approaches [151–158], and machine learning approaches [110, 159–168].

Acoustic source localisation plays an essential role in drone audition as the accuracy of the estimated spatial positions of the sound sources affects the subsequent functionality such as source separation, source tracking, and source enhancement. Localisation of sound sources using a drone-embedded microphone array is inherently more challenging than the conventional source localisation techniques in a room scenario as the drone noise sources are louder and nearer to the microphone array than the desired sound source.

A large number of drone-related sound source localisation techniques have been developed over the years to accurately estimate sound source locations from drone on-board microphones. Many contributions [2, 12, 14, 22–24, 169] in this area address sound source localisation using subspace based methods and TDOAs based methods.

2.2.1 Subspace based methods

The most fundamental approach to subspace-based methods is MUSIC (MUltiple SIgnal Classification) [123] which utilises the orthogonality between the signal and noise subspaces to estimate the direction of arrivals (DOAs). In [14], a spherical microphone array on a drone was used to localise the sound sources using MUSIC. However, in MUSIC, the performance degrades as the noise power is higher than the sound sources. Thus, many variants have been proposed to improve the robustness against the drone noise, e.g., generalised eigenvalue decomposition MUSIC (GEVD-MUSIC), which introduces the noise correlation matrix for whitening the noise space using additional sensors [24,170]. In [23], an enhanced version of the GEVD-MUSIC algorithm called incremental GEVD-MUSIC (iGEVD-MUSIC) has been proposed to adapt the dynamic changes in noise using both a microphone array and a Kinect device. Variants of the GEVD-MUSIC and iGEVD-MUSIC methods using generalised singular value decomposition (GSVD-MUSIC) have been introduced to reduce their computational cost [12]. However, it has been observed that in general, GSVD-based MUSIC methods over-estimate the noise correlation matrix and degrade the localisation performance.

2.2.2 TDOA-based methods

TDOA-based approaches find a target sound's incident angular directions with respect to the recording microphones or microphone pairs. The generalised cross-correlation with phase transformation (GCC-PHAT) is the most popular TDOA-based method for DOA estimation [171] due to its robustness against moderate levels of reverberation. Typically, cross-correlation-based methods use the difference of arrival time between two microphone pairs and maximum time-lag yields the location of the sound source. In [2], a cross-correlation-based method along with a particle filter is used to localise the single frequency sound source e.g., a whistle, using the on-board microphone array on the drone. The work in [169], presents a detailed analysis of different TDOA estimation methods and concludes that, overall, GCC-PHAT yields the best performance. However, this analysis was not based on the drone application scenario. In [22], authors compared the performance of DOA estimation methods e.g., SEVD-MUSIC, GEVD-MUSIC, GCC-PHAT, in different

flight conditions and claimed that GCC-PHAT with Wiener filtering obtained better localisation in most scenarios. However, due to a large main lobe width with low frequency content of the sound source or a high amplitude of side lobes with high frequencies of the sound source, GCC-PHAT often struggles to individualise closely spaced sound sources or sources with low sound level [172].

2.2.3 Beamforming-based methods

There also exist beamforming-based methods for DOA estimation using drones [63, 73, 74, 76]. In [73], a delay-and-sum fixed beamformer-based sound source localisation was presented using an octagonal microphone array on the drone. The author of [76] proposed a drone-based binaural sensing system to locate sound sources using beamforming based on intra-band and inter-measurement beam fusion. In [63, 74], the authors presented an adaptive beamformer called minimum variance distortionless response (MVDR) beamformer-based method with post-filtering techniques for the drone platform. However, this approach assumes the DOA of the sound source is known. The performance of fixed beamformer-based methods that depend on the microphone array geometry, degrades with $\overline{\text{SDNR}}$ condition, whereas adaptive beamformer-based methods perform better, but require prior information related to the drone noise e.g., the noise correlation matrix, to derive the beamformer weights.

2.2.4 Spatial filtering

Another method for DOA estimation uses time-frequency spatial filtering algorithms. The authors of [17, 173], use a spatial likelihood function based on the TDOA of the sound source combined with beamforming and blind source separation techniques. The recordings were made using a circular microphone array which places above the propeller plane to avoid the wind noise that pushes downward. In [17], authors estimated the time-varying DOA of a moving sound source by segmenting the audio recording into blocks and localised the sound source with a time-frequency spatial filter for each block together with a particle filter.

2.2.5 Supervised-learning-based methods

Lately, to improve the localisation performance, most algorithms tend to utilise the knowledge of the fixed noise sources, e.g., propellers and motors, to predict the spatial characteristics of the drone noise. In more recent works, DOA estimation considering learning-based approaches such as deep learning-based methods together with denoising algorithms, especially for the suppression of drone noise harmonics generated from propeller rotation were discussed [174, 175]. In [174], a convolutional neural network with speed-correlated drone noise harmonic cancellation was developed for DOA estimation from the mixture signal. In [175], a UAV-embedded audio recording system was proposed using a denoising autoencoder and post-processing algorithms with restricted peak search. In contrast, a convolutional neural network (CNN) based approach by utilising power-level-based features to sound source localisation using stereo microphones on a drone is proposed in [176]. While supervised-learning-based methods typically have improved performance, they still exhibit drawbacks of requiring large datasets for training and higher computational costs associated with the processing.

2.2.6 Research gaps

In practice, however, the DOA estimation is extremely challenging and often fails at lower S$\overline{\text{D}}$NR conditions (e.g., -30 dB) due to the dynamic nature of the drone noise. In the case of drone audition, it is expected that the source localisation can offer lower complexity solutions with the flexibility of having arbitrary microphones on-board. We also note that exploiting the drone noise characteristics into existing spatial acoustic localisation problems are still not well explored.

2.3 Removing Drone Noise for Source Signal Enhancement

In the last few decades, signal enhancement algorithms in noisy environments other than drones have been widely studied and are still an active field of research. In particular, speech enhancement is an essential task in audio signal processing with

a wide range of applications, including hearing aids, virtual assistants for mobile devices/smart devices [177,178], automatic speech and speaker recognition systems [179–182], teleconferencing [183–185], and robot audition [9, 104, 186–188]. The unwanted environment noise, reverberation, or interfering source removal process for speech enhancement and speech separation is commonly discussed around the "cocktail party problem" [189]. An extensive amount of work has been carried out to tackle this scenario using monaural (single-microphone) and multi-microphone speech enhancement approaches [190]. Common algorithms are spectral subtractive methods [191–197], Wiener filtering based methods [198–201], statistical model-based approaches [202–205], beamforming [95, 206–218], adaptive filtering [219–224], independent component analysis [187, 216, 225–238] and NMF [239–247], and neural networks based methods (NNs) [248–260].

More recent algorithms consider speech enhancement as a supervised learning problem. Supervised learning approaches using deep neural networks (DNNs) can be classified into two categories mainly as Time-Frequency (TF) domain methods and time-domain methods. Generally, DNN-based TF algorithms estimate a real or complex value mask for each TF bin such as Ideal Binary Mask (IBM) or Ideal Ratio Mask (IRM) to obtain the clean speech spectra and show robustness to non-stationary noise sources [248–251]. Time domain DNN-based algorithms such as SEGAN [252], WaveNeT [253], Conv-TasNet [254], TCNN [255] Wavesplit [256], TPARN [261, 262] and their extensions [262, 263] have shown great improvements over TF methods. These techniques recover the original clean speech signals or equivalently, in high (> 0 dB) and moderate (> -5 dB) signal-to-noise (SNR) levels using large models with high complexity and long hours of training and validation data.

Now, we review in brief the literature on signal enhancement methods developed for on-board microphone arrays on drones.

2.3.1 Beamforming-based methods

Among existing methods, beamforming is the standard approach to multichannel speech enhancement [264]. The multichannel Wiener filter (MWF) has been used in speech enhancement on drone [74, 265]. In [74], the authors presented a framework

for speech enhancement using a beamformer with post-filtering. A 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 [265], compared the performance of the variants of MWF algorithms with an interfering noise source and a single propeller-motor combination. The work in [75], presented a beamforming-based spectral distance response algorithm for both localisation and enhancement of the target source. This approach 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.

2.3.2 Spatial filtering

Another method for speech enhancement uses TF spatial filtering algorithms [19, 20, 64, 78, 266]. In [78], the authors proposed a drone noise reduction framework that combines blind source separation, time-frequency spatial filtering, and single channel spectral post-filtering to jointly enhance the target sound at low S$\overline{\text{D}}$NR levels (≥ -15 dB). There also exists a video-assisted speech enhancement method combined with TF spatial filters to extract the speech from visually informed directions [19,20]. In [64], Wang uses a spatial filtering method to localise and enhance a desired sound captured by a drone with different placements of the drone-mounted array to obtain a noiseless sector. The method suggests turning the drone's looking direction to capture the target sound source from the noiseless sector to improve the performance at low S$\overline{\text{D}}$NR levels (≥ -20 dB). However, it only uses a standard microphone array structure on a drone.

2.3.3 Supervised-learning-based approaches

Lately, to improve speech enhancement performance at very low SNR conditions (e.g., S$\overline{\text{D}}$NR lower than -15 dB), supervised learning methods using deep neural networks (DNNs) are introduced [15, 266–269]. In [266], the authors demonstrated single-channel and multichannel integrated TF spatial filtering approaches

for speech enhancement on drones. In [267, 268], 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 [267], results are evaluated considering a single-motor propeller combination. In [268], 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 [269], proposed a CNN-based complex spectrogram enhancement method that removes the drone noise.

2.3.4 Research gaps

While these techniques show improved signal enhancement results in low $\overline{\text{SDNR}}$ levels (≥ -25 dB), some approaches 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 [15, 19, 20, 74, 75, 78, 94, 266–268, 270]. 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 transfer functions or assumed free-field sound propagation to obtain the transfer functions, making them difficult to consider as potential methods for signal enhancement using on-board microphones on a drone.

2.4 Head Related Transfer Function

So far, free-field assumptions or measured impulse responses from specific locations have been used to provide microphone array steering matrices or array transfer functions to audio source localisation and speech enhancement algorithms using drones [12–14, 67, 74, 175, 267, 271]. However, disregard for the drone's acoustic properties such as scattering and diffraction to an incident soundfield limits the performance of those drone audition functions. This motivates us to model an accurate representation of the drone-related transfer function.

People have the ability to perceive spatial locations of sound sources in real-

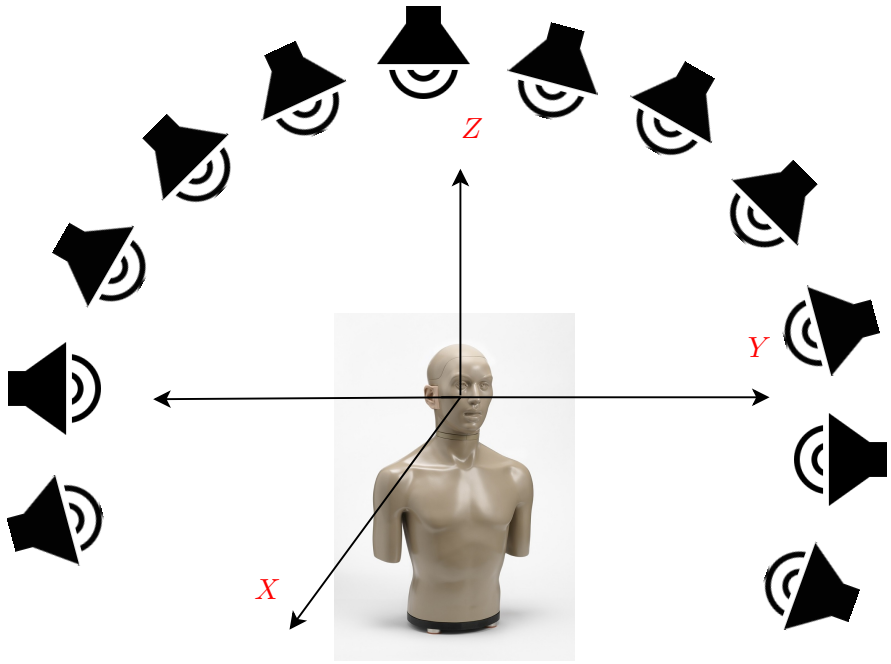


Figure 2.3: Illustration of a direct HRTF measurement setup.

world acoustic scenes. The surrounding spatial soundfield of a listener is modified by the interaction between the sound wave and the listener's body. Specifically, it depends on the diffraction and reflection properties of the individual's body shape. This includes the soundfield reflections from their upper body, head, and outer ear before reaching their eardrum. Measuring the scattering and diffraction of the human head, is known as the Head Related Transfer Function (HRTF) which is the acoustic transfer function from a sound source to a listener's eardrum [272]. HRTFs contain all the relevant spatial cues that are parameterised by both direction and frequency.

HRTF measurements and modeling procedures using a dummy head such as Knowles Electronic Manikin for Acoustic Research (KEMAR), or human subjects have been widely studied in the past few decades [273–277]. In direct HRTF measurements, the microphones are placed inside the listener's ears while impulse response measurements are taken from a discrete set of loudspeakers surrounding the listener's head as shown in Fig. 2.3. It is known that the acoustic reciprocity principle [278] allows the HRTF measurements to be reversed with the source placed

in the ear and microphones around the head [279]. This way we can swap the loudspeaker and the microphone positions in the traditional HRTF measurement setup, and it can still be used in a regular manner in place of the directly measured HRTF [280].

Typically, the HRTF-type measurement process is resource intensive, a set of discrete measurement positions on the sphere are taken and then interpolated (e.g., using spherical harmonics), to estimate the HRTF over all directions [281]. Also, they usually require a spherical loudspeaker array or a single loudspeaker placed at many discrete positions on the sphere. In practice, however, large specialised equipment or feasibility restrictions limit this measurement process [282].

Furthermore, there exists an HRTF-like transfer function also known as the device-related transfer function or direction-dependent response from a microphone array mounted on an arbitrarily-shaped object incorporating the object's scattering, as studied in [104, 283–290]. For example, in [104, 289] parametric transfer function estimation methods using a set of measurements for robot audition is proposed. Transfer function approximation by performing a Fourier series expansion with terminating the approximation at an appropriate order has been explored in [287, 288]. In [284, 286], the device-related transfer functions are modelled in the spherical harmonics domain using a discrete set of points on the surface of a sphere. However, these methods exploit the sphere-like shape objects, e.g., the humanoid robot head, a head-worn device on a spherical head, which facilitates the processing in the Fourier/spherical harmonic domains [104, 284, 286]. Similar to HRTF, *Donley et al.* [291] presented a device-related transfer function dataset by employing the head-to-torso manikin with an augmented reality glasses microphone array while impulse response (IR) measurements are taken from a discrete set of loudspeakers on a sphere surrounding the manikin.

These HRTFs and the device-related transfer functions are used in the application of binaural reproduction [286, 292] using headphones or virtual reality and augmented reality headsets. Likewise, the robot audition systems utilise the HRTF jointly with the interaural phase difference (IPD) and interaural intensity difference (IID) extracted from the binaural microphone recordings to localize the sound sources in a room scenario [104].

2.5 Background Theory on Spherical Harmonic Decomposition of a Soundfield

In this section, we present the background theory of spatial soundfield representation in the spherical harmonic domain, which is used in Chapter 6.

2.5.1 Coordinate systems

We begin with defining the spherical coordinate system as shown in Fig. 2.4 where $\mathcal{O}$ denotes the system's origin. The spherical coordinates of a singular point in space can be represented as (r, θ, ϕ) where $r \in [0, \infty)$ denotes the distance from the origin in meters, $\theta \in [0, \pi]$ denotes the elevation angle downwards from the positive Z-axis, $\phi \in [0, 2\pi)$ denotes the azimuth angle counterclockwise from the positive X-axis and the projection of the point in the XY-plane.

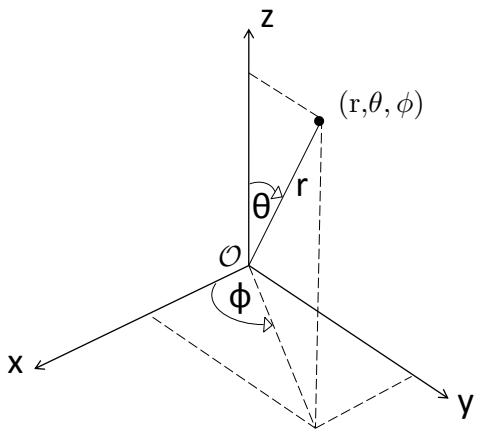


Figure 2.4: Illustration of the spherical coordinate system.

The spherical coordinates are related to the Cartesian coordinates (x, y, z)

through

$$x = r \sin\theta \cos\phi \tag{2.1a}$$
$$y = r \sin\theta \sin\phi \tag{2.1b}$$
$$z = r \cos\theta. \tag{2.1c}$$

Conversely, spherical coordinates can be related to Cartesian coordinates through

$$r = \sqrt{x^2 + y^2 + z^2} \tag{2.2a}$$
$$\theta = \pi/2 - \text{atan2}(z, \sqrt{x^2 + y^2}) \tag{2.2b}$$
$$\phi = \text{atan2}(y, x), \tag{2.2c}$$

where atan2$(\cdot)$ denotes the four-quadrant inverse tangent.

2.5.2 Reduced wave equation

Consider the propagation of sound waves in a homogeneous isotropic medium assumed to be an inviscid fluid (such as air), with stationary initial conditions. Let $p(k, \boldsymbol{x})$ be the Fourier transform of the sound pressure through this medium at a single point in space $\boldsymbol{x} = (r, \theta, \phi)$ where k denotes the *wavenumber*. The *reduced wave equation*, also termed as the *Helmholtz equation* [293],

$$\nabla^2 p(k, \boldsymbol{x}) + k^2 p(k, \boldsymbol{x}) = 0 \tag{2.3}$$

where ∇^2 denotes the Laplacian operator on a chosen coordinate system. The *wavenumber* k is related to the angular frequency ω and wavelength λ and is given by

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda}. \tag{2.4}$$

The quantity c is the propagation speed where at 20° C $c = 343$ m/s in air. We assume the speed of sound c is independent of the frequency variations, implying that the *wavenumber* k is a constant multiple of angular frequency ω. Therefore, we refer to k as "frequency".

In the following section, we solve the *Helmholtz equation* (2.3) in the spherical coordinate system to obtain the space dependent acoustic pressure $p(k, \boldsymbol{x})$.

2.5.3 Spherical harmonic expansion of the soundfield

The spherical solutions to the *Helmholtz equation* are found by the spherical definition of ∇^2 and applying the separation of variables technique to (2.3). The operator ∇^2 in spherical coordinates is given by

$$\nabla^2(\cdot) = \frac{1}{r^2}\frac{\partial}{\partial r}[r^2\frac{\partial}{\partial r}(\cdot)] + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}[\sin\theta\frac{\partial}{\partial\theta}(\cdot)] + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}(\cdot). \tag{2.5}$$

The solution to the *Helmholtz equation* can be derived as two soundfield categories. We review two cases based on the relative position of the sound sources as *interior soundfield* and *exterior soundfield*.

Interior soundfield

We consider an incoming soundfield, in which the origin is included and the sources are located completely outside the outer boundary. The solution for the interior field is [293]

$$p(k, \boldsymbol{x}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\alpha_{nm}(k)j_n(k|\boldsymbol{x}|)Y_{nm}(\widehat{\boldsymbol{x}}), \tag{2.6}$$

where $\alpha_{nm}(k)$ is the spherical harmonic coefficients of the interior field, $j_n(\cdot)$ is the n^{th} order spherical Bessel function of the first kind, $|\cdot|$ is the norm or Euclidean distance, $\widehat{\boldsymbol{x}} = \boldsymbol{x}/|\boldsymbol{x}|$, and $Y_{nm}(\cdot)$ is the spherical harmonic of order n and degree m, defined by

$$Y_{nm}(\widehat{\boldsymbol{x}}) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}}P_{nm}(\cos\theta)e^{im\phi}, \tag{2.7}$$

where $P_{nm}(\cos\theta)$ is the associated Legendre function, $e \approx 2.71828$ is the Euler's number and $i = \sqrt{-1}$.

Exterior soundfield

We consider an outgoing soundfield, in which the inner boundary encloses all the sources and the outer boundary may extend up to infinity. The solution for the exterior field is

$$p(k, \boldsymbol{x}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \beta_{nm}(k) h_n(k|\boldsymbol{x}|) Y_{nm}(\widehat{\boldsymbol{x}}), \tag{2.8}$$

where $\beta_{nm}(k)$ is the spherical harmonic coefficients of the exterior field, $h_n(\cdot)$ is the n^{th} order spherical Hankel function of the first kind.

2.6 Summary

In this chapter, we have provided the current research progress on acoustic signal processing on drones. More precisely, we discussed the literature on the acoustic characterisation of drones, and acoustic scene analysis using drone-embedded microphone array. We also briefly presented on the HRTF measurements and modelling approaches. Finally, we defined the concepts of decomposing a soundfield in the spherical harmonic domain.

Chapter 3

Characterisation of the Drone Motor-Related Acoustic Transfer Function: Preliminary Investigation

The potential of drone based services is enormous with applications ranging from consumer product delivery to health services. When meeting this demand, one of the major challenges we face is the noise produced by drones, which not only contributes to listener discomfort, but also hinders the device's ability to effectively communicate via audio. Thus, there exists a pressing need for understanding the characteristics of drone related noise, which can then be suppressed using suitable methods. In this chapter, we present a preliminary study on modelling the relationship between input motor current and acoustic noise produced by a drone. We conduct an experimental study indoors for a drone mimicking hovering manoeuvre with a single active motor and propeller. The drone noise was measured by a single on-board microphone. We identify multiple tones or harmonics in the drone noise spectrum that vary proportionally to the motor current. Based on this observation, we define a transfer function between the input current and output noise, and model its harmonic behaviour using a higher order polynomial function. We present a detailed error analysis to validate the model.

3.1 Introduction

Drones are rapidly transforming and/or replacing many activities in the commercial, recreational and military spaces with potential applications of consumer/health package delivery [1], emergency rescue missions [2], and wild-life management [7]. However, the widespread deployment of these applications is not yet accepted by the public due to two main reasons; (i) privacy concerns, and (ii) noise pollution. Addressing the latter by modelling the drone noise signature is the motivation for this chapter.

Drones utilise multiple modalities for communication and information capture, where audio is of predominant importance. Thus, audio signal processing for drone applications has a significant interest. While there exist a large amount of recent work on audio signal processing for drones ranging from source localisation to signal classification [2, 12, 14, 22–24], the aforementioned challenge of drone related noise reduction is still very much an open problem. The majority of the noise around the drone is produced by itself. As such, there exist a great need to systematically understand and mitigate the noise emission by drones in order to enhance the user experience [25].

Most of the previous work have presented drone noise analysis using free-field or ground-based microphone arrays [38, 56]. While such analysis using on-board microphones is rare. In this chapter, we present our preliminary work on deriving a drone related acoustic transfer function that can in return estimate the noise generated for a given flight manoeuvre. Our approach is to start from recorded drone noise, observe their most evident characteristics, and fit a mathematical model(s) to these characteristics. We use an on-board microphone to obtain noise when only a single motor and propeller pair is active while the drone is at a steady state of hovering manoeuvre, and we repeat the measurement for different motor currents. By analysing the recorded noise spectrum for varying currents, our intention is to find a mathematical relationship between the input current and apparent characteristics of the said spectrum. Specifically, we wish to investigate the transfer function between the current through the motor and corresponding sound pressure measured from the drone-embedded microphone, and also parameterise the transfer function for a fixed manoeuvre.

The rest of this chapter is organised as follows. In Section 3.2, we define the problem of modelling the drone noise signature through the motor current measurements. In Section 3.3, we describe our experimental setup with an analysis of the results. In Section 3.4, we introduce the current to noise transfer function with a model to obtain the harmonic frequencies of drone motor noise. In Section 3.5, we present the estimation of the model parameter. In Section 3.6, we provide our concluding remarks.

3.2 Problem Formulation

In this chapter, we aim to analyse the relationship between the current that drives the drone motor and the resulting acoustic signal received from a microphone attached/embedded to the structure of the drone. The drone motor is driven by an electronic speed controller such that it regulates the current through the motor to control the rotational speed of the rotor. Thus, drone noise may be modelled against the input current that drives the motor or the rotational speed of rotors, and we use the former.

Consider the current through the motor as i (constant over time for a fixed manoeuvre), and the noise signal received by the microphone at time t as $v(t, i)$. The drone motor-related acoustic transfer function can be modelled as 'Black Box' model as shown in Fig. 3.1, and can be expressed as

$$v(t, i) = T\{t, i\}, \tag{3.1}$$

where $T\{\cdot\}$ is the drone motor-related acoustic transfer function.

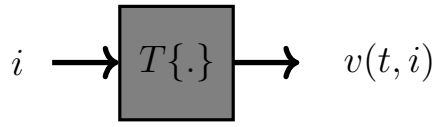


Figure 3.1: 'Black Box' model: Drone motor-related acoustic transfer function for a given microphone takes motor current i and outputs the signal received by the microphone $v(t, i)$.

The main objective of this chapter is to parameterise a drone motor-related acoustic transfer function for each microphone which characterises the acoustics

signature of the drone from an on-board microphone. This parameterisation is based on tonal components of the drone noise, and therefore, we provide a straightforward method to extract the aforementioned coefficients for each microphone.

We wish to conduct noise measurements at different motor current levels and model the acoustic characterisation for a single microphone embedded to the drone for a fixed manoeuvre such as hovering. Note that unlike the motor current which can be directly measured, motor-propeller rotation speed can use for modelling drone motor-related acoustic transfer function.

3.3 Experimental Analysis

In this section, we detail our preliminary measurements that is used to model the drone motor-related acoustic transfer function. We first discuss the experimental setup and configurations. We then present the measurement results and its analysis for a particular microphone attached to the drone. We note that drones can operate in different flight modes and manoeuvres by controlling the throttle, roll, pitch, and yaw. In the following, we mimic a hovering maneuver where a constant current passes through all four motors, which leads to operating them at the same angular speed.

3.3.1 Measurements

Our objective is to find a relationship between the driving current i of the motor and the noise field captured by an on-board microphone. For this experiment, we drive a single motor using a known input current i and acquire the sound pressure signal $v(t, i)$ received by the microphone fixed to the under-body of the drone. We repeat the measurement procedure over different motor current ratings within its operating range.

The experiment is carried out in a quiet room with dimensions of $[5, 4.1, 2.45]$m for the length, width, and height, respectively, and the reverberation time of approximately $T_{60} = 140$ ms. The experimental setup is shown in Fig. 3.2. We mount the drone to a 1 m tall rigid stand to minimise vibrations. Using a bench top power supply, we operate a single motor attached with a two-bladed propeller. We use

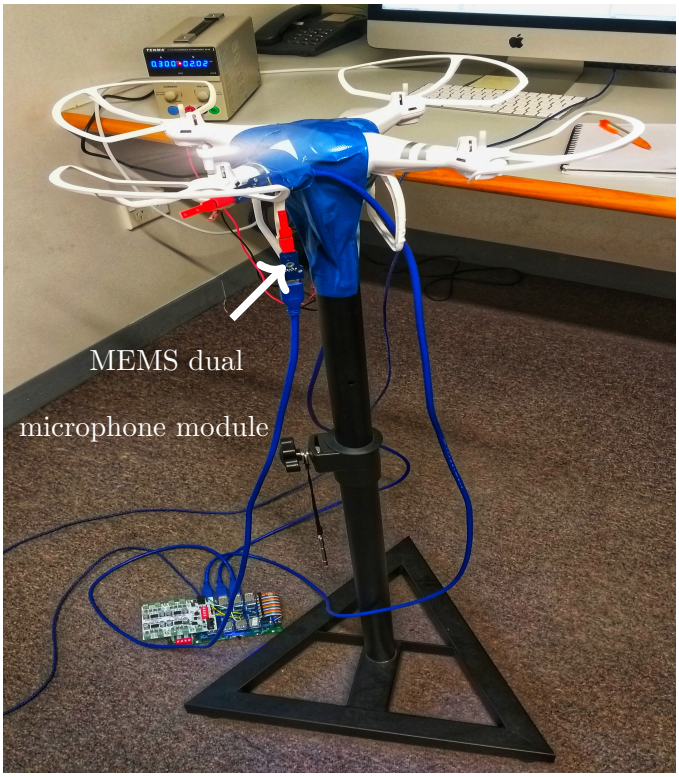
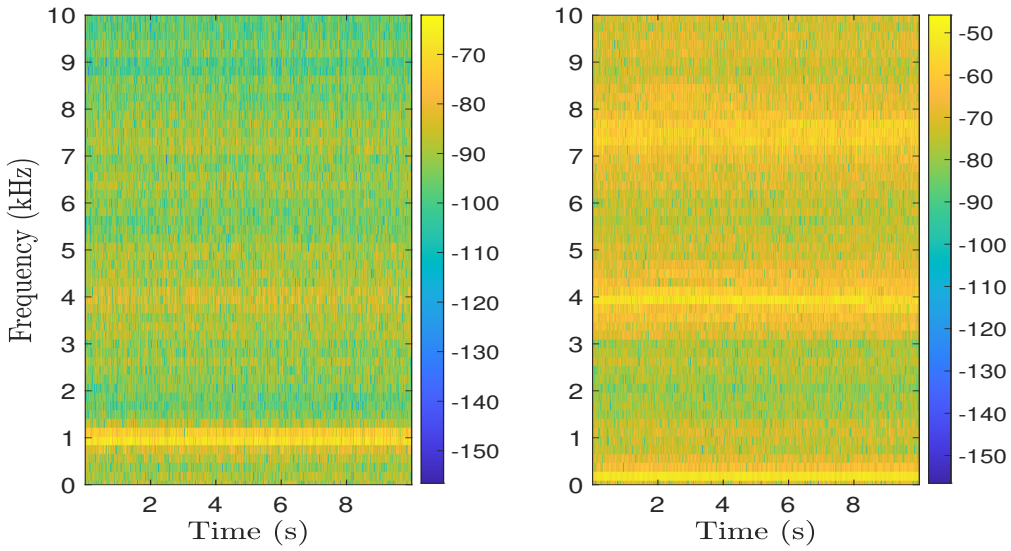


Figure 3.2: Experimental setup with a dual microphone module attached to the body of the drone closer to the selected motor.

a drone with two pairs of clockwise and anti-clockwise propellers with a diameter of 150 mm, and are driven by the direct current brushed permanent magnet motors. Note that the concept set out in this chapter can be used to extend for the brushless direct current motor and brushless alternating current motor as well. The power supply is driven at Constant Current (CC) mode, for each measurement and repeated at 50 mA steps ranging from 100 mA to 1000 mA. At each step, the acoustic measurements are obtained for a duration of 10 seconds using a Micro-Electro-Mechanical System (MEMS) microphone and a USB-powered external audio interface with 48 kHz sampling frequency. The microphone was attached to the landing gear of the drone, close to the motor (as shown in Fig. 3.2).

3.3.2 Results

Here, we present the recorded data for varying input motor currents, and draw insights in to their apparent characteristics.



(*a*) 100 mA (*b*) 1000 mA

Figure 3.3: Time-frequency plots of the recorded drone motor noise at different motor current ratings.

Figures 3.3(a) and (b) present the time-frequency plots of the recorded drone noise at the given microphone for the input motor current of 100 mA, and 1000 mA, respectively. We observe that the drone noise has different spectral characteristics over the frequencies with respect to the operating motor current. Specifically, different segments of energy distribution over the spectrogram at varying motor currents. Additionally, we also observe a tonal structure in each of the spectrogam plots, which we investigate next by analysing the frequency spectrum of the drone noise recordings at different motor current.

Figure 3.4 shows the frequency spectrum of noise recorded by the microphone when only one of the motors is driven by an input current of (i) 100 mA (black), and (ii) 1000 mA (blue). We used a low-pass filter to remove frequencies over 10 kHz. Also, shown in red is the spectrum of background noise measured by the

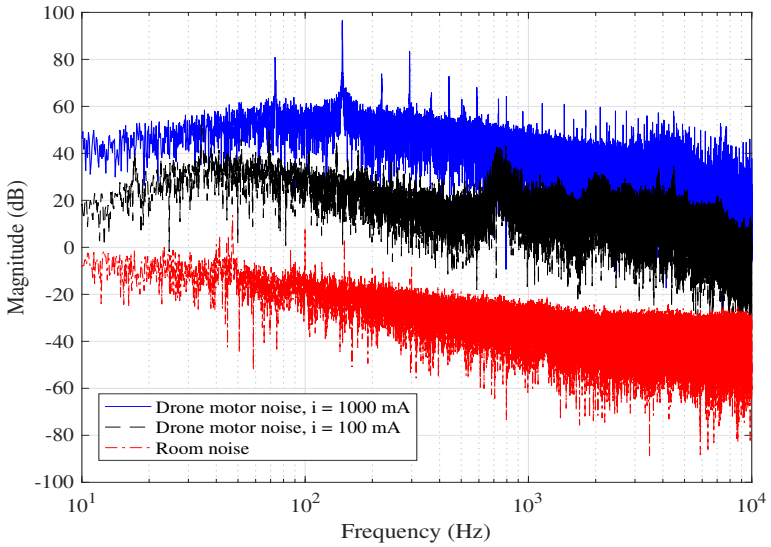


Figure 3.4: Frequency spectrum of the recorded drone motor noise at 1000 mA and 100 mA with background noise.

microphone when the drone was inactive. Note that the background room noise is negligible compared to drone noise. However, we subtract room noise from the acoustic measurements in spectral domain for normalisation. From Fig. 3.4, we can observe a number of peaks in the noise spectrum reflecting multiple tones. We refer to these distinct tones as *harmonics*, because they are almost harmonically related.

Figure 3.5 illustrates how the variation of first 10 *harmonics* ω_n, $n = 1, \ldots, 10$ as a function of the driving current i of the drone motor. Figures 3.6 and 3.7 plot the real and imaginary parts of the amplitude of each of the first 10 dominant *harmonics*, respectively. We can observe that there is a certain pattern in the above curves, inferring to an underlying relationship between the amplitude and frequencies of the *harmonics* and the driving current of the motor.

In the next section, we will model the experimental results to identify the complex relationship between the noise field *harmonics* and the motor current.

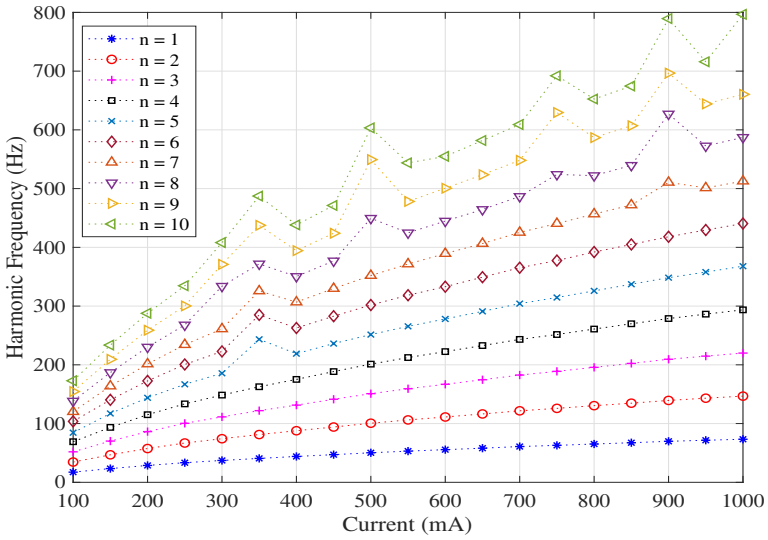


Figure 3.5: Harmonic frequency against input motor current.

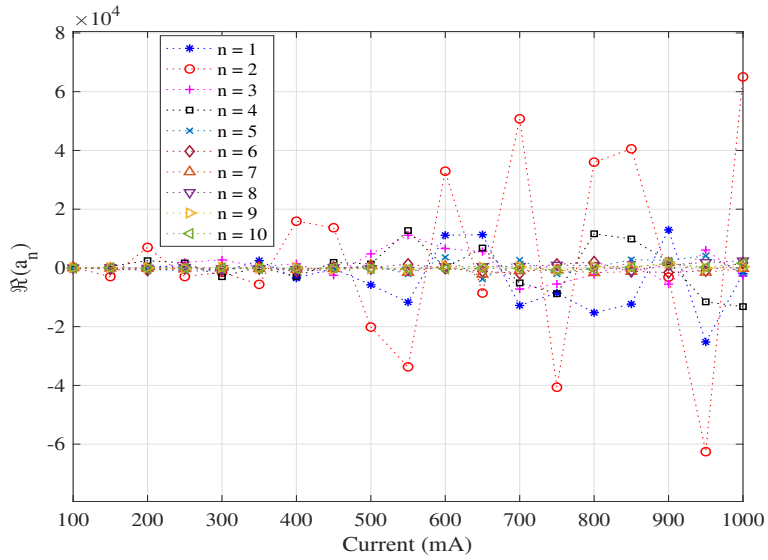


Figure 3.6: Real part of the harmonics amplitude, $\Re(a_n)$ against input motor current, i.

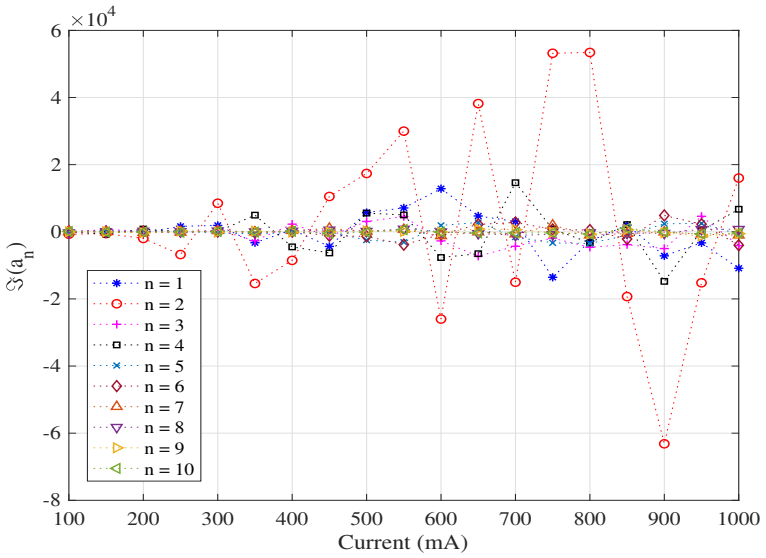


Figure 3.7: Imaginary part of the harmonics amplitude, $\Im(a_n)$ against input motor current, i.

3.4 Current to Noise Transfer Function

In this section, we obtain an expression for the current to noise transfer function, considering only the significant *harmonics* of the received signal.

Let N be the number of *harmonics*, then the received signal can be given by

$$v(t,i) = \sum_{n=-N, n\neq 0}^{N} a_n(i)\, e^{j\omega_n(i)t}, \tag{3.2}$$

where $\omega_n(i)$ is the n^{th} harmonic frequency and $a_n(i)$ is the amplitude of the n^{th} harmonic. Note that

$$\omega_{-n} = -\omega_n, \tag{3.3}$$

$v(t,i)$ is a real signal, and $a_n(i)$ is a complex amplitude with the property,

$$a_{-n} = a_{n^*}. \tag{3.4}$$

As a result, the current to noise transfer function is derived as a finite-order

harmonics governed by motor current i that can be represented as

$$i^{\lambda_n} e^{j\alpha_n it} = a_n(i)\, e^{j\omega_n(i)t}, \tag{3.5}$$

where $\lambda_n = \log_e |a_n(i)| / \log_e i$ and $\alpha_n = (\omega_n(i) + 2\pi\iota)/i$. The scale ι denotes an integer. We note that the current to noise transfer function models the periodic nature of the propeller dynamics that capture the interactions between the propeller blade and air.

3.4.1 Harmonic frequency

Figure 3.5 depicts the behaviour of harmonic frequency against input motor current. We observe that harmonic frequencies, w_n seem to behave as a higher order polynomial of current i. Hence, we may be able to model $w_n(i)$ as

$$\omega_n(i) = \sum_{m=0}^{M} \rho_{nm}\, i^m, \tag{3.6}$$

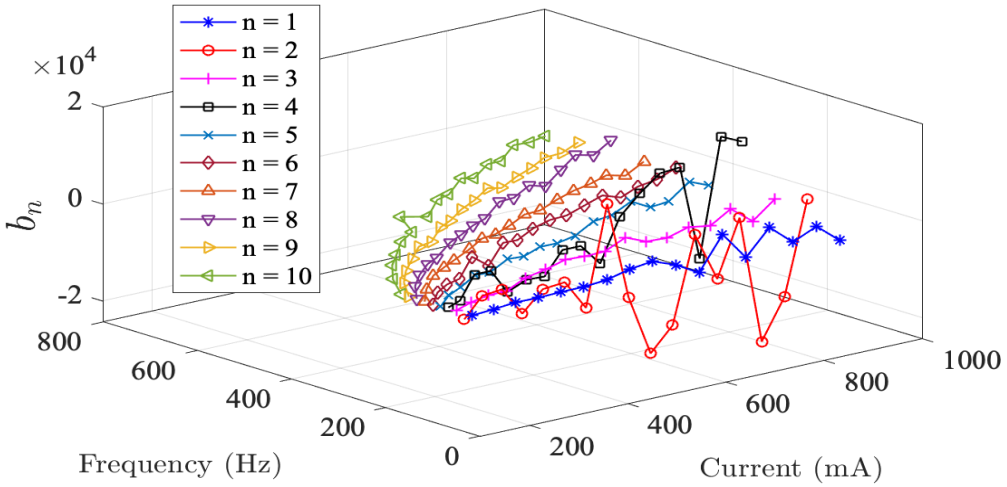
where M is the order of the polynomial, m is the index and ρ_{nm} are the coefficients of the polynomial. We aim to estimate ρ_{nm} to show the relationship between the drone motor driving current and the *harmonics* of the noise.

3.4.2 Harmonic amplitude

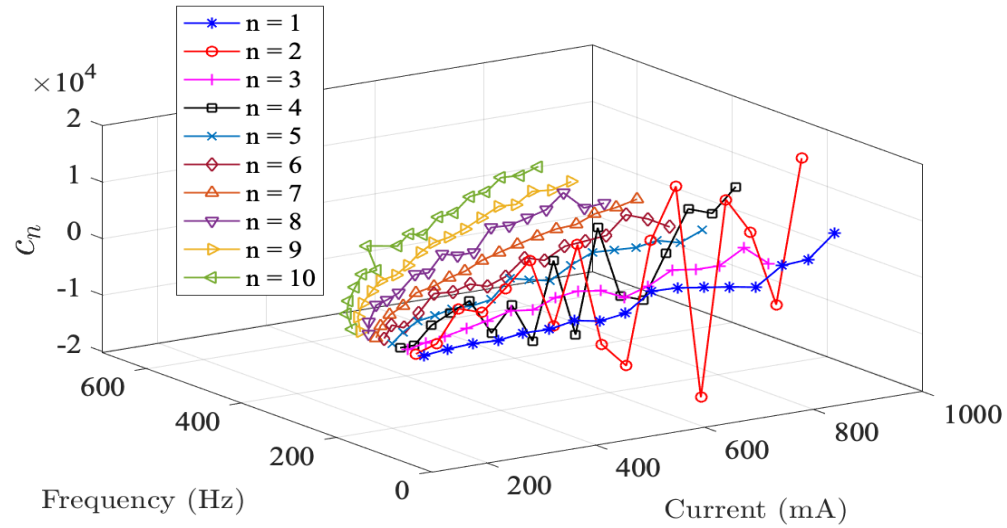
Figures 3.6 and 3.7 illustrate that real and imaginary parts of a_n grow exponentially with current i. Based on the analysis of the acoustic characterisation of the drone in Section 3.3.2, it is indicated that $a_n(\cdot)$ and $\omega_n(\cdot)$ are nonlinear functions. Since $a_n(i)$ is also complex, $v(t, i)$ can be modelled as

$$v(t,i) = \sum_{n=-N, n\neq 0}^{N} b_n(i)\cos(\omega_n(i)t) + jc_n(i)\sin(\omega_n(i)t), \tag{3.7}$$

where $b_n(\cdot)$, and $c_n(\cdot)$ represent the real and imaginary coefficients for the n^{th} harmonic frequency. Investigating the *harmonics* strength, we plot a 3-D representation of the a_n in terms of b_n and c_n as a function of i, and ω_n up to 10 *harmonics*



(a)



(b)

Figure 3.8: Harmonic strength as a function of current (i) and frequency (ω_n) for $n \in [1, 2, \ldots, 10]$.

in Fig. 3.8. It shows that the harmonic strength behaviour against frequency appears to have a negatively underdamped sinusoidal structure allowing us to use the matrix pencil method [294] to model these sum of complex exponentials. The

modelling of the harmonic strength is not derived in this thesis but is suggested for future research directions in Chapter 7.

3.5 Estimation of the Harmonic Frequency Coefficients

Following the above model, we estimate the parameters of the *harmonic* frequency using the higher order polynomials. Here, we provide a way to estimate the proposed model parameters $\rho_{nm} \in \mathbb{R}$ in (3.6) using a least-squares approximation method that minimises the squared error. We evaluate the model accuracy based on the mean squared error (MSE).

Suppose we have Z samples $\{(i_z, \omega_n(i_z))\}_{z=1}^{Z}$ where $i_z, \omega_n(i_z) \in \mathbb{R}$. We modify model (3.6) as

$$\omega_n(i_z) = \sum_{m=0}^{M} \rho_{nm}\, i_z^m + \epsilon_z, \tag{3.8}$$

where ϵ_z is the random error component of the z^{th} case. This can be written in matrix form as

$$\begin{bmatrix} \omega_n(i_1) \\ \omega_n(i_2) \\ \vdots \\ \omega_n(i_Z) \end{bmatrix} = \begin{bmatrix} 1 & i_1 & i_1^2 & \dots & i_1^M \\ 1 & i_2 & i_2^2 & \dots & i_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & i_Z & i_Z^2 & \dots & i_Z^M \end{bmatrix} \begin{bmatrix} \rho_{n0} \\ \rho_{n1} \\ \vdots \\ \rho_{nM} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_M \end{bmatrix}, \tag{3.9}$$

$$\omega_{\mathbf{n}} = \mathbf{X}\ \rho_{\mathbf{n}} + \epsilon, \tag{3.10}$$

where the vector $\omega_{\mathbf{n}} \in \mathbb{R}^Z$ and the matrix $\mathbf{X} \in \mathbb{R}^{Z\times(M+1)}$ are the given data, and $\rho_{\mathbf{n}} \in \mathbb{R}^{(M+1)}$ is the unknown parameter vector. With $Z \geq (M+1)$, the basic least-squares approximation method [295] involves estimating $\rho_{\mathbf{n}}$, as a solution of the optimisation problem such that squared error is minimised.

$$\widehat{\rho}_{\mathbf{n}} = \underset{\rho_{\mathbf{n}}\in\mathbb{R}}{\operatorname{argmin}} ||\mathbf{X}\ \rho_{\mathbf{n}} - \omega_{\mathbf{n}}||^2. \tag{3.11}$$

The least-squares approximation solution to (3.11) is

$$\widehat{\rho}_{\mathbf{n}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\omega_{\mathbf{n}}, \tag{3.12}$$

where $\widehat{\rho}_{\mathbf{n}}$ is an unbiased estimator of $\rho_{\mathbf{n}}$. While the above can be computed via singular value decomposition, for the purpose of this chapter, we use a simple least-squares method to fit the data resulting in a model function given by

$$\widehat{\omega}_{\mathbf{n}} = \mathbf{X}\widehat{\rho}_{\mathbf{n}}. \tag{3.13}$$

Based on (3.11), the best estimate of the model $\widehat{\rho}_{\mathbf{n}}$ can be obtained by minimising the MSE given by

$$\text{MSE} = \frac{1}{Z}\sum_{z=1}^{Z}(\omega_{\mathbf{n}} - \widehat{\omega}_{\mathbf{n}})^2. \tag{3.14}$$

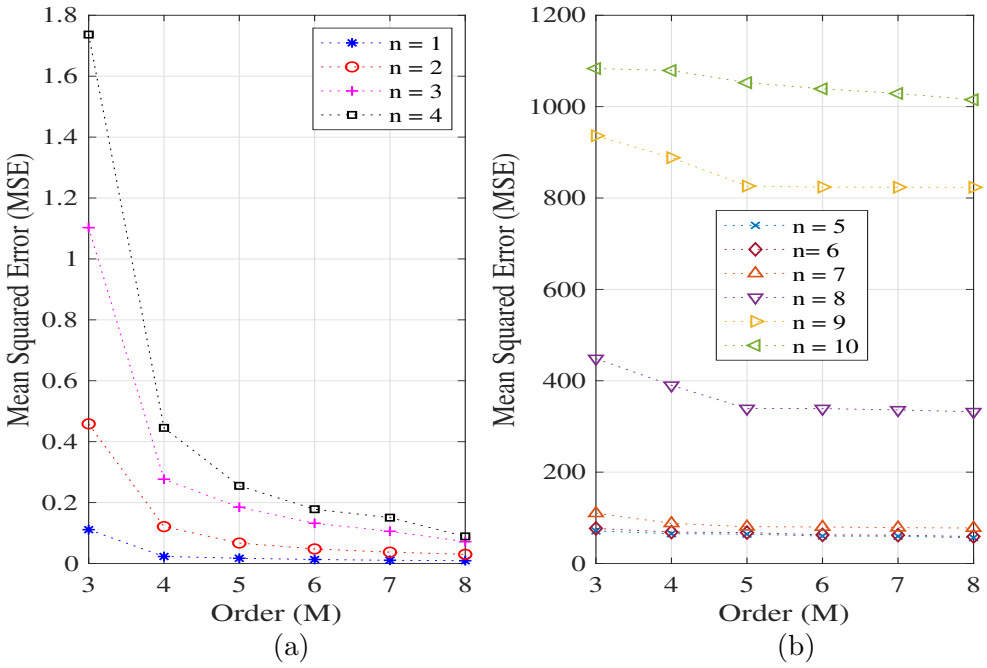


Figure 3.9: MSE estimator of $\widehat{\omega}_{\mathbf{n}}$ for $n \in [1, 2, \ldots, 10]$ with respect to the order of the polynomial M.

Figure 3.9 presents the results after fitting an M^{th} order polynomial where $M \in [3, 4, \ldots, 8]$. From Fig. 3.9(a) we observe that MSE decreases more rapidly with in-

creasing M for $n \in [1, 2, \ldots, 4]$. From Fig. 3.9(b), we observe that MSE estimators are comparably high and become stable with M for $n \in [5, 6, \ldots, 10]$. This may be due to the notches appearing in Fig. 3.5. The reason for these notches is not clear at this stage. It can be due to a measurement error or due to the resonance effect in the room.

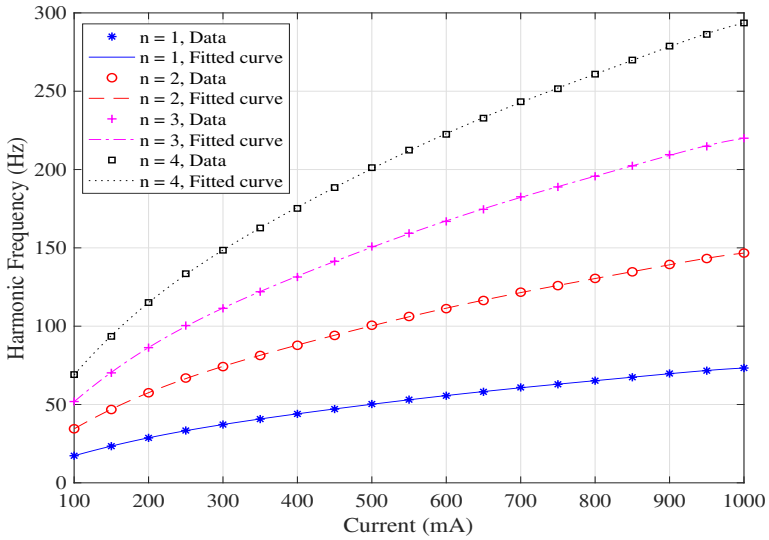


Figure 3.10: Harmonics frequency, ω_n, $n = 1, \ldots, 4$ mapped with the motor current, i. Markers are actual data whereas lines are polynomial fit.

From the above observation, we decide to fit a seventh order $M = 7$ polynomial regression model for the *harmonic* frequency variation over current. Figure 3.10 shows the reconstructed harmonics using the above model for the first four *harmonics*. We observe that the seventh-degree polynomial provides a good approximation for ω_n, where $n = 1, \ldots, 4$ with respect to the measured value. In Fig. 3.11 we show similar results for higher order harmonics from $n \in [5, 6, \ldots, 10]$, which still manage to achieve 95% confidence interval estimates. These results confirm that the harmonic frequencies of drone motor noise against input current can be successfully modelled using higher order polynomials.

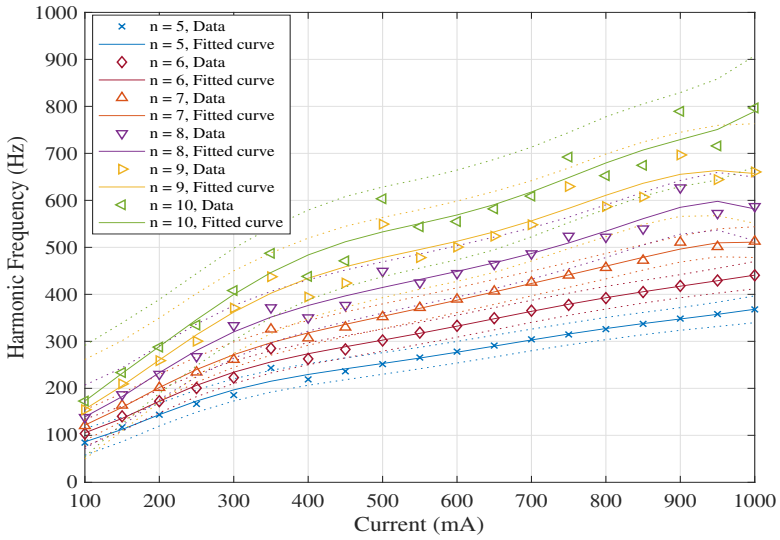


Figure 3.11: Harmonics frequency, ω_n, $n = 5, \ldots, 10$ mapped with the motor current, i. Markers are actual data whereas lines are polynomial fit and dotted lines are 95% confidence interval.

3.6 Summary and Contributions

This chapter presents our preliminary measurements and analysis framework for the acoustic characterisation of the drone noise from an on-board microphone array at fixed maneuver. This work has highlighted the practical advantage of modelling a drone motor-related acoustic transfer function for each microphone through drone noise *harmonics* of a single motor. The chapter outcomes and insights can be itemised as follows:

- We conducted an initial measurement of a drone for a hovering manoeuvre with varying current through a single motor which may be useful for predicting the noise from all four motors.
- We showed that the recorded noise has a harmonic structure proportional to the current driving through the motor.
- Motivated by an apparent harmonic structure observed in the noise measurements, we defined a transfer function between the motor current and the drone noise.

- The frequency behaviour of these harmonics were modelled using higher order polynomials.
- We provided a detailed error analysis validating the model accuracy.

In this chapter, we have investigated the current-specific drone noise measurement procedure and analysis of acoustic noise produced by a commercial drone at steady state of hovering. This chapter heavily relies on the experimental data from a single motor to derive a drone related acoustic transfer function that can in return estimate the noise generated from all four motors for a given manoeuvre. In the next chapter, we show that sound source localisation using a drone-embedded microphone array can be benefited by using motor current-specific measurements to suppress the drone noise in real-world acoustic scenes.

3.7 Related Publications

- **W. N. Manamperi**, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Estimating Drone Motor Related Acoustic Transfer Function: A Preliminary Investigation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 156–160, Auckland, New Zealand, December 2020.

Chapter 4

Acoustic Source Localisation with Drone Noise Angular Spectrum Subtraction

Source localisation is an important function of drone audition. There exist many solutions to localise both single and multiple sound sources other than using drone-embedded microphone arrays. Sound source localisation using a drone is challenging due to its emission of significant noise from drone motors and propellers causing a highly adverse noisy environment. In this chapter, we present a method that suppresses drone noise from multichannel recordings. The method uses real recordings from an on-board microphone array on a drone and the known characteristics of the drone noise. We demonstrate the method's applicability through experimental measurements with prior-measured motor current-specific drone noise data.

4.1 Introduction

Acoustic source localisation is an important function in drone auditions specially when the video is occluded. The problem of sound source localisation using a drone is an emerging area in recent years related to intelligent audio applications for military and civil environments including search and rescue missions [2]. However, the major challenge associated with sound source localisation using drones is to identify the directions of the target sound sources in the presence of noise produced by the drone itself. It is necessary to suppress the noise emitted by drones for enhancing real-world applications [25].

In this chapter, we investigate the problem of sound source localisation using a drone in practical conditions where the *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] can be extremely low (e.g. S$\overline{\text{D}}$NR < -20 dB). We provide an algorithm for estimating the direction of arrival (DOA), both azimuth and elevation of single and multiple sound sources on the ground using an embedded microphone array on the drone. We propose a cross-correlation based DOA estimation method with drone noise angular spectrum subtraction.

We derive the DOAs of the sound sources at different microphone pairs, considering the instantaneous angular spectrum of both recordings (mixture) and drone noise for each time-frequency bin. We exploit the known characteristics of drone acoustics to spatially remove the drone noise to do a better localisation. We learn these drone noise characteristics by using either drone-motor current or the speed-specific drone-noise-only data. This can be done either by modelling the drone noise as discussed in the previous chapter, or from the pre-recorded measurements. We obtain the instantaneous angular spectrum of the motor current-specific drone noise that corresponds to the motor current of the microphone recordings. We suppress the drone noise by subtracting the angular spectrum of the current-specific drone noise from the angular spectrum of the microphone recordings before averaging across the microphone pairs. This way, we preserve the spatial acoustic features for source localisation.

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

We demonstrate the effectiveness of the proposed method with experimental measurements inside a semi-anechoic chamber. We conduct a study on the localisation performance in terms of prediction error over extreme S$\overline{\text{D}}$NR conditions. We also show that two sound sources at certain azimuth angular distances, and with different S$\overline{\text{D}}$NR levels can localise with acceptable accuracy. Further, we compare the performance with state-of-the-art methods: traditional MUSIC algorithm [123], GEVD-MUSIC [24], and GCC-PHAT [171].

We preview the contents of this chapter as follows. In Section 4.2, we formulate the problem of multiple sound source localisation using a drone. In Section 4.3, we describe the proposed method based on noise angular spectrum subtraction. In Section 4.4, we evaluate and analyse the performance of the proposed method in a practical experiment and compare it with state-of-the-art methods. Finally, we summarise and provide conclusions in Section 4.5.

4.2 Problem Formulation

In this section, we formulate the problem of sound source localisation using an on-board microphone array on a drone.

Consider an irregular microphone array of Q microphones attached to a drone, where the position of each microphone $\mathbf{x_q} \equiv (r_q, \theta_q, \phi_q)$, $q = 1, \ldots, Q$ for radial distance r_q, elevation θ_q and azimuth ϕ_q with respect to a suitably chosen origin of a spherical coordinate system. Let there be L active sound sources located in the far-field of the microphone array at $\mathbf{x}_\ell \equiv (r_\ell, \theta_\ell, \phi_\ell)$, for $\ell = 1, \ldots, L$ as shown in Fig. 4.1. Let $d_{\ell,q}(t)$ denotes the drone-related impulse response[2] from the ℓ^{th} source to the q^{th} microphone. The signal received by the q^{th} microphone is given by

$$p_q(t) = \sum_{\ell=1}^{L} d_{\ell,q}(t) * s_\ell(t) + \sum_{m=1}^{M} v_{m,q}(t), \tag{4.1}$$

[2]Here we define the drone-related impulse response as the response of a microphone on the drone to an impulse from a source location. Note that the drone-related impulse response depends on the spatial parameters of the microphone position and source position as well as the acoustic properties such as scattering and diffraction from the drone structure. We propose a method to measure drone-related transfer function in Chapter 6.

where '$*$' represents the convolution operation, $s_\ell(t)$ denotes the speech signal from the ℓ^{th} sound source, $v_{m,q}(t)$ denotes the drone noise due to the m^{th} motor at the q^{th} microphone, and M represents the number of motors on the drone. Note that we consider thermal microphone noise to be negligible.

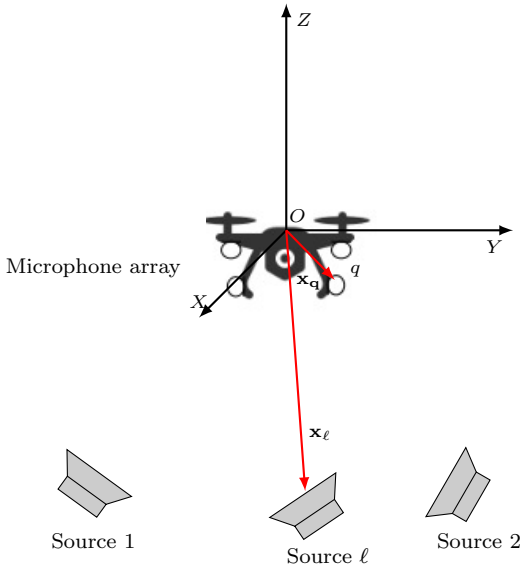


Figure 4.1: Problem setup of an irregular microphone array on a drone in the presence of multiple sound sources.

With the multiplicative transfer function (MTF) approximation [296], the recorded signal of (4.1) is approximated in the short-time Fourier transform (STFT) domain as

$$P_q(n,k) = \sum_{\ell=1}^{L} D_{\ell,q}(k) S_\ell(n,k) + \sum_{m=1}^{M} V_{m,q}(n,k), \tag{4.2}$$

where $S_\ell(n,k)$ and $V_{m,q}(n,k)$ are the STFTs of $s_\ell(t)$ and $v_{m,q}(t)$, respectively, $D_{\ell,q}(k)$ is the Fourier transform of $d_{\ell,q}(t)$, and n and k are the time frame and frequency bin indices, respectively with $n \in \{1,\ldots,N\}$ and $k \in \{1,\ldots,K\}$ where N and K are the number of time frames and frequency bins, respectively.

In addition, noise characteristics varying with the motor current make the source localisation using the on-board microphone signals difficult.

Typically noise characteristics are varied with the motor current making the source localisation using the on-board microphone signals difficult. For our prob-

lem, we assume the drone operates at a certain height from the ground in a stable maneuver (e.g. hovering) with constant current flow through the motors. Note that we assume stationary sound sources.

Our objective is to estimate the direction of L sound sources (θ_ℓ, ϕ_ℓ), $\ell = 1, \ldots, L$ by processing the received microphone signals, $P_q(n,k)$, $q = 1, \ldots, Q$. The main challenge with sound source localisation using drone on-board microphones is the noise from nearby drone motors.

Approximation

We assume that sound sources are in the far field of the drone, and ignore the scattering of waves from the source signals by the drone structure [3]. Then we can model the drone-related transfer function between the ℓ^{th} sound source and q^{th} microphone as

$$D_{\ell,q}(k) = \exp(-i2\pi f_k \tau_{\ell,q}), \tag{4.3}$$

where f_k denotes the frequency corresponding to the k^{th} frequency bin, and $\tau_{\ell,q}$ is the time difference between the origin and the q^{th} microphone for a signal arriving from the ℓ^{th} source. Note that we make no assumptions about the drone noise as drone motors are in the near-field from the microphones. We use experimental data in a real-life environment to demonstrate the localisation performance.

4.3 Cross-correlation based DOA Estimation

In this section, we propose a general method to estimate the DOA of the sound sources using the time-difference-of-arrivals (TDOAs) based approach considering the cross-correlation between pairs of microphones. Firstly, we introduce the instantaneous angular spectrum of the multichannel signals. We then subtract the drone noise using this instantaneous angular spectrum of the current-specific drone noise to suppress the drone noise from the recordings.

[3] In Chapter 6, we model the scattering and diffraction of the drone body beyond free-field propagation.

4.3.1 The instantaneous angular spectrum

Here, we intend to estimate the unknown TDOAs of the sound sources $\tau(\theta_\ell, \phi_\ell)$, and the directions describing the DOAs of the sound sources i.e., (θ_ℓ, ϕ_ℓ), where $\ell = 1, \ldots, L$, by exploiting the known locations of the microphones on the drone. Let

$$P_q(n,k) = |P_q| \exp(i\Phi_q(n,k)), \tag{4.4}$$

where $|P_q|$ and $\Phi_q(n,k)$ are the amplitude and phase of the q^{th} microphone signal. We denote the instantaneous normalised cross-power spectrum[4] between the microphone channels q and q',

$$\begin{aligned} R_{q,q'}(n,k) &\triangleq \frac{P_q(n,k)P^*_{q'}(n,k)}{|P_q(n,k)P^*_{q'}(n,k)|}, \\ &= \exp(i\Phi_q(n,k) - i\Phi_{q'}(n,k)), \end{aligned} \tag{4.5}$$

where $(\cdot)^*$ denotes the complex conjugate. Note that $R_{q,q'}(n,k)$ only preserves the phase difference between the q^{th} and $q'^{\,th}$ signals. We refer this term as the inter-channel phase difference measured from the q^{th} and $q'^{\,th}$ channels.

Consider a uniform grid of potential angles in three-dimensional space within the search boundaries of $\theta \in [-\pi/2, \pi/2]$ and $\phi \in (-\pi, \pi]$. We calculate the TDOAs between two microphone channels q and q' from an arbitrary angle (θ, ϕ) for sound propagation of a source in a free-field as

$$\tau(q, q', \theta, \phi) = \frac{(\mathbf{x_q} - \mathbf{x_{q'}}) \cdot \widehat{\mathbf{x}}}{c}, \tag{4.6}$$

where ' $\cdot$ ' represents the dot product between two vectors, $\widehat{\mathbf{x}}$ denotes the unit vector in the direction of sound propagation, and c denotes the speed of sound propagation. Let the microphone pair index associated with channel q and q' be denoted as j with $j = 1, \ldots, J$, where $J = Q!/2!(Q-2)!$ is the total number of distinct pairs in the microphone array.

Let $\tau_j(\theta, \phi)$ be the TDOAs of the j^{th} microphone pair. We obtain the instan-

[4]In time domain, this is referred to as generalised cross-correlation with phase transform (GCC-PHAT) in literature [171].

taneous angular spectra for the j^{th} microphone pair at the $(n,k)^{th}$ time-frequency bin as

$$\Psi(n,k,\tau_j) = \exp(i\Phi_q(n,k) - i\Phi_{q'}(n,k)) \exp(-i2\pi f_k \tau_j(\theta,\phi)), \tag{4.7}$$

where $\exp(-i2\pi f_k \tau_j(\theta,\phi))$ is the inter-channel phase differences of the j^{th} microphone pair computed theoretically with respect to the delay $\tau_j(\theta,\phi)$ [169]. From here on, we represent $\tau_j(\theta,\phi)$ as τ_j and ignore the θ and ϕ terms in Ψ for simplicity. Here, $\Psi(n,k,\tau_j)$ is computed for the range of τ_j at the $(n,k)^{th}$ time-frequency bin for the j^{th} microphone pair and reached its maximum when two inter-channel phase differences are consistent with each other.

We aggregate the spectral, temporal, and spatial information of the microphone signals by averaging across the frequency bins, time frames, and microphone pairs, respectively. In particular, we average Ψ over frequency k, time n and microphone pairs j. In result, the local maximum peaks shall yield the TDOAs of the sound sources. Thus, we search for ℓ^{th} local maximum peak over (θ,ϕ) to find the TDOA of the ℓ^{th} sound source in terms of the ℓ^{th} highest peak as

$$\tau(\theta_\ell,\phi_\ell) = \underset{\tau}{\operatorname{argmax}}\big(\sum_{j=1}^{J}\sum_{n=1}^{N}\sum_{k=1}^{K}\Psi(n,k,\tau_j)\big). \tag{4.8}$$

With the known number of sound sources L, we extract the highest L local maximum peaks that corresponding to TDOAs of L sound sources $\tau(\theta_\ell,\phi_\ell)$, $\ell = 1,\ldots,L$. We map $\tau(\theta_\ell,\phi_\ell)$ into (θ_ℓ,ϕ_ℓ) to estimate the direction of L sound sources.

Figure 4.2 shows a block diagram of the computation of $\Psi(n,k,\tau_j)$. Given $P_q(n,k)$, $q = 1,\ldots,Q$ (see Fig. 4.2a), we obtain $R_{q,q'}(n,k)$ between the microphone channel q and q' which are highlighted with green shadows shown in Fig. 4.2b using (4.5). In (4.7), the estimated $R_{q,q'}(n,k)$ of the j^{th} microphone pair, is multiplied by theoretical inter-channel phase difference $\exp(-i2\pi f_k \tau_j(\theta,\phi))$ to obtain $\Psi(n,k,\tau_j)$ for the j^{th} microphone pair at the $(n,k)^{th}$ time-frequency bin as shown in Fig. 4.2c.

Intuitively, we calculate the Relative Transfer Function (ReTF) between the sources and the j^{th} microphone pair in (4.5) which account only for the phase difference while neglecting the amplitude ratio. Hence, we also can use the phase content of ReTF between q and q' microphones for the TDOA estimation.

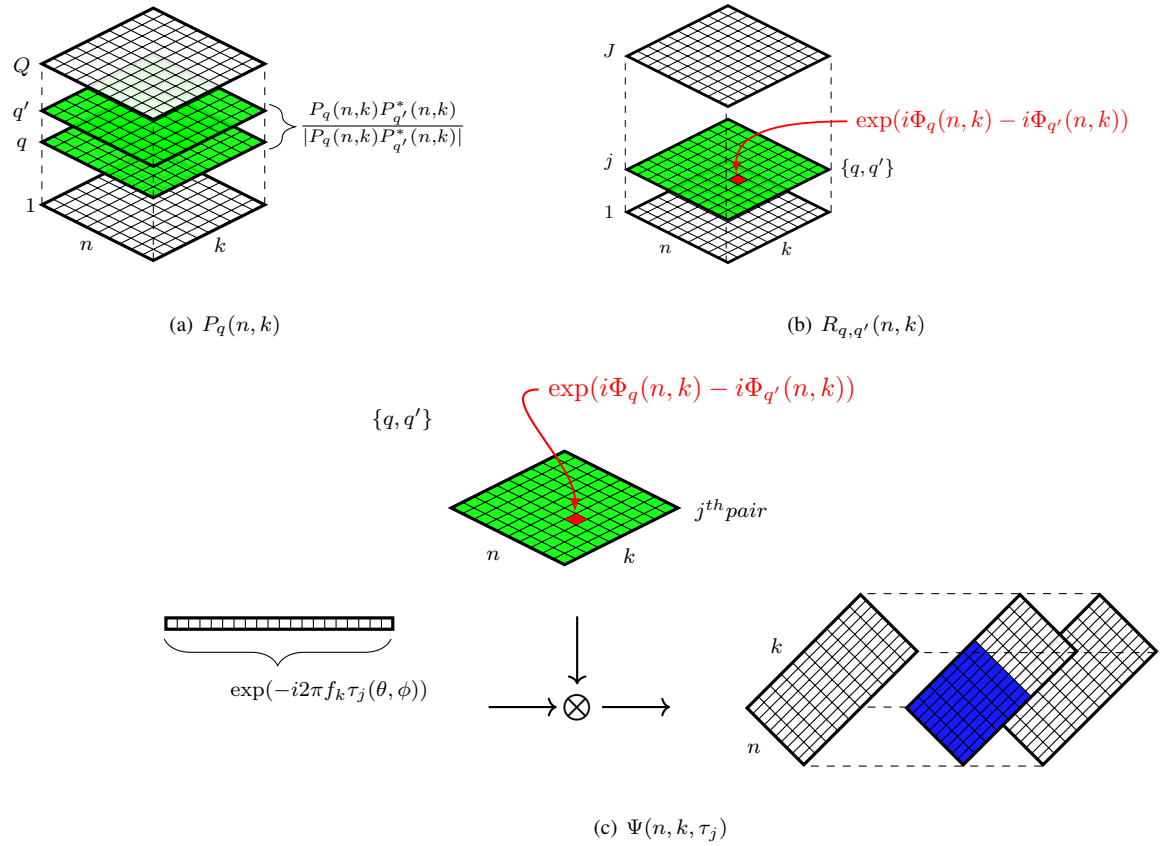


Figure 4.2: Illustration of the computation of the instantaneous angular spectrum, $\Psi(n,k,\tau_j)$ from the recordings.

Note that it is typically difficult to estimate the true TDOAs of the sound sources from (4.8) in practice without suppressing the drone noise. This is because these TDOAs capture information from both sound sources, and noise sources, and introduce spurious peaks in the angular spectrum due to the drone noise, and hinder the true TDOAs of the sound sources. Alternatively, in Sec. 4.3.2 we propose subtracting the drone noise angular spectrum from (4.8) with the use of motor speed or current-specific drone noise, to avoid the false detection of the sound sources from the recordings. For the remainder of this section, we assume that the speed or current-specific drone noise is already known (or measured).

4.3.2 Noise angular spectrum subtraction to remove drone noise

Here, we propose a process for suppressing the drone noise from the multichannel recordings through the use of the drone noise characteristics assuming the fact that the locations of drone motors and propellers are fixed compared to the microphone array. We learn these drone noise characteristics by using either drone-motor current or the speed-specific drone-noise-only recordings. This allows us to spatially remove the drone motors and propellers (noise sources) from the angular spectrum for accurately estimating the sound source position $\mathbf{x}_\ell$. Therefore, we propose subtracting the angular spectrum of drone noise from the angular spectrum of the mixture recordings before averaging over microphone pairs. The reason is to perform the average noise subtraction before doing the pair averaging to preserve the spatial acoustic features. We express the angular spectrum of the sound source as

$$\Psi(\tau)_S = \sum_{j=1}^{J} \left(\sum_{n=1}^{N} \sum_{k=k_1}^{k_2} \Psi(n,k,\tau_j)_{\text{mixture}} - \sum_{n'=1}^{N'} \sum_{k=k_1}^{k_2} \Psi(n,k,\tau_j)_{\text{drone}} \right), \tag{4.9}$$

where $\Psi(\cdot)_{\text{mixture}}$ and $\Psi(\cdot)_{\text{drone}}$ denote the instantaneous angular spectrum of the mixture and drone noise, respectively, and k_1, k_2 represent the lower and upper range of the target sound source signal's frequency band, respectively. For example, if the target sound source is a speech source, we perform frequency summing across the speech frequency bandwidth. This allows us to avoid the low frequencies where drone noise is primarily interfering. Note that the second term in (4.9), is slightly different in time to the mixture, which does not require the same length of recordings. Thus, the TDOA estimation of ℓ^{th} sound source is given by the ℓ^{th} local maximum peak response,

$$\tau(\theta_\ell, \phi_\ell) = \underset{\tau}{\text{argmax}}\ \Psi(\tau)_S. \tag{4.10}$$

4.4 Experimental Verification

To prove the above theory and to evaluate the performance of the proposed sound source localisation method, we designed an experiment with real drone measurements. This section details the experimental methodology we have performed and presents the single and multiple sound source localisation results over significantly noisy (below -10 dB S$\overline{\text{D}}$NR) recordings at the end.

4.4.1 Methodology

Experimental setup

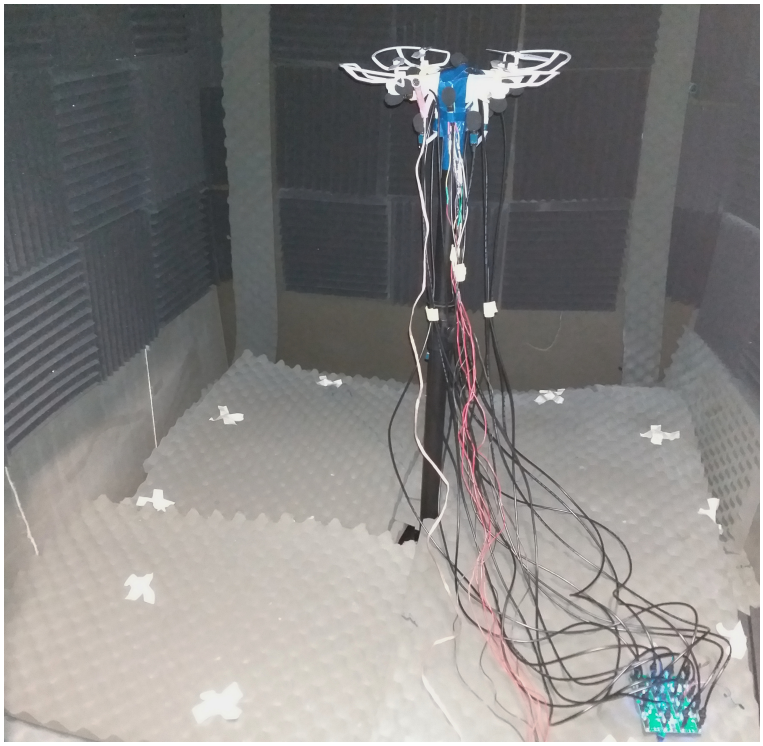
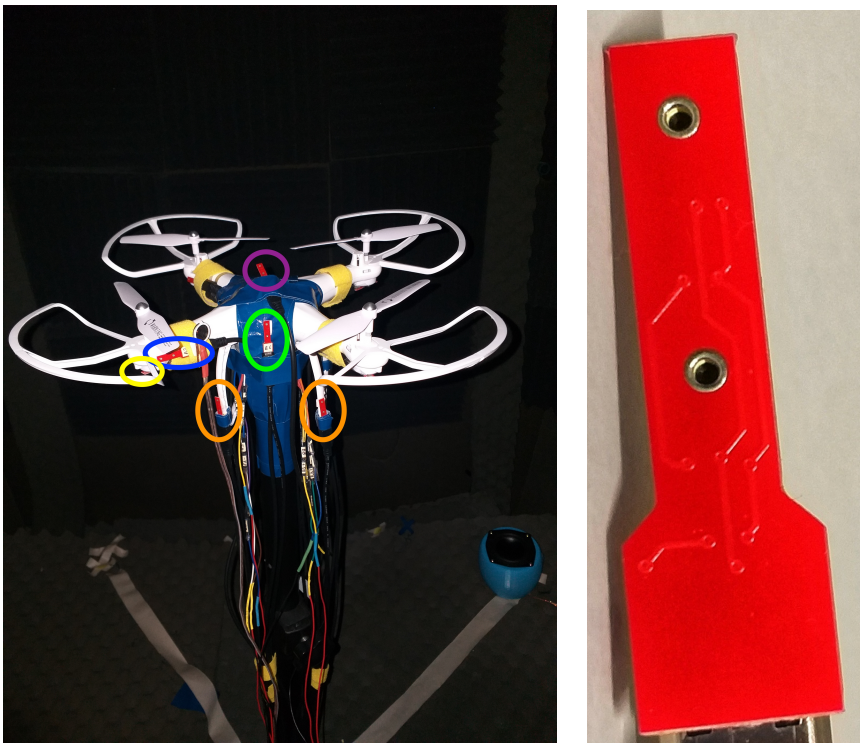


Figure 4.3: Experimental setup inside a semi-anechoic chamber, where the loudspeaker placements are shown by the white markers on the floor.

The experimental setup consists of a drone with a set of microphones attached to its body and a single source on the ground. We conducted the experiment in a semi-anechoic chamber at the Australian National University with dimensions of

[1.45; 1.40; 1.60] m and a room reverberation time of T_{20} of 20 ms. We attached 15 Micro-Electro-Mechanical System (MEMS) microphone modules with two microphones (ICS-43432), to the body of the drone, and fixed the drone on a 1 m tall rigid stand inside the chamber (as shown in Fig. 4.3). We changed the position of the sound source at 12 points, regularly spaced along a circle with radius of 0.6 m on the ground concentric with the base of the stand as shown in Fig. 4.3.



(a) drone-embedded microphone array.

(b) Dual microphone module.

Figure 4.4: The drone-embedded with 15 dual microphone modules.

Figure 4.4 shows the dual microphone module placement on the drone. Here, we attached three microphone modules close to each motor: underneath the motor (as shown in a yellow circle), side of the motor (as shown in a blue circle), and on the landing gear (as shown in orange circles). Also, one module is attached to the top of the drone (as shown in the purple circle), and two modules are connected in between the front and back of the drone arms (as shown in a green circle). These

microphone modules are very close to the drone motors and propellers than the loudspeaker on the ground. Hence, the sound source signal is obscured by the drone noise and ended up with an extremely low $\overline{\text{SDNR}}$ level. We define the $\overline{\text{SDNR}}$ as the power ratio between the source signal and drone noise, i.e.,

$$\overline{\text{SDNR}} = 10 \log_{10} \left(\frac{s_\ell(t)^2}{\sum_{m=1}^{M} v_{m,q}^2(t)} \right), \tag{4.11}$$

where $s_\ell(t)$, and $v_{m,q}(t)$ are given in (4.1). The following experiment is conducted with varying motor current and loudness levels of the sound source to obtain different $\overline{\text{SDNR}}$ levels.

Motor current-specific drone noise recordings

We learn the drone noise characteristics from the motor current-specific pre-recorded noise only recordings. As the noise characteristics of a given drone vary with the flight conditions such as take-off, hover, and forward flight, in this study, we consider the hovering maneuver, where a similar current drives all motors of the drone. We operate the drone at hovering maneuver using different current ratings within its operational range and measure the drone noise against different driving currents. The drone operates using four direct current brushed permanent magnet motors. We use 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.

The loudness of the measured drone noise at the microphone array is between 62 dB to 89 dB with different current ratings from 100 mA to 1000 mA, and the mean loudness of the target sound source signal at the microphone array (with no motor or propeller operating) is 60 dB on average for normal talk. Note that in a distress situation, a person can scream loud so the loudness level reaches up to 90 dB. We simulate our application scenario within the $\overline{\text{SDNR}}$ level range from 0 dB to -30 dB with different motor current ratings.

Experiment scenario

We obtained impulse response measurements from a source at each of the 12 source positions to each microphone channel. We note that the impulse response recordings incorporate incident soundfield as well as the sound scattering, and diffraction by rigid boundaries of the drone body. We convolved the impulse response recordings with the speech signal (approximately five seconds long) from the WSJCAM0 database [297] to simulate the sound source signal received at the microphones.

We added the sound source signal together with drone noise recordings to obtain the mixture recordings. Here, we adjust the S$\overline{\text{D}}$NR level to simulate the real-world scenario which is equivalent to having the sound sources further away from the drone. Scenarios with simultaneously active two sound sources are analysed by adding multiple speech recordings with the drone noise. Note that for the drone noise at the q^{th} microphone in (4.1), we use different recordings of the same motor current rating to obtain the multichannel recordings $p_q(t)$ and the drone noise $v_{m,q}(t)$.

Algorithm parameters

The sampling rate of the recordings is 48 kHz. We compute STFT on 50% overlap sine window with a window length of 4096 samples. We use 15 microphone modules, each have two microphones and hence 30 microphone channels, and there are totally $J = 30!/2!(30-2)! = 435$ pairs of microphones in the array, and from each of them the instantaneous angular spectra are computed for each time-frequency bin using the proposed method. The DOAs of the sound sources are obtained by the angular spectrum subtraction described in Sec. 4.3.2 with the cutoff frequencies k_1 and k_2 being 1 and 3 kHz, respectively.

Evaluation metric

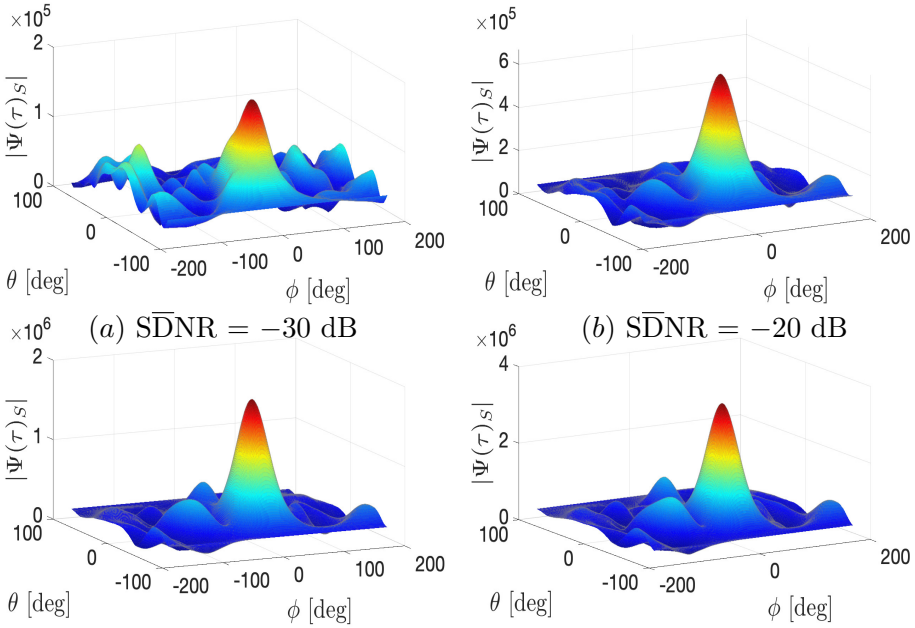
To evaluate the performance of our proposed method, we define the prediction error for the ℓ^{th} source by the angular difference between the true source DOA (θ_ℓ, ϕ_ℓ) and the estimated DOA $(\overline{\theta}_\ell, \overline{\phi}_\ell)$ as

$$\Delta_\ell = \cos^{-1}(\cos(\overline{\theta}_\ell)\cos(\theta_\ell) + \sin(\overline{\theta}_\ell)\sin(\theta_\ell)\cos(\overline{\phi}_\ell - \phi_\ell)). \tag{4.12}$$

Compared methods

We compare the DOA estimation performance of the proposed cross-correlation based drone noise angular spectrum subtraction method with the traditional MUSIC algorithm [14, 123], GEVD-MUSIC [24] and GCC-PHAT [171] methods over very low $\overline{\text{SDNR}}$ levels for both single and multi-source scenarios. Here we include GCC-PHAT which we use to derive the proposed method for a fair comparison.

4.4.2 Results for single source localisation



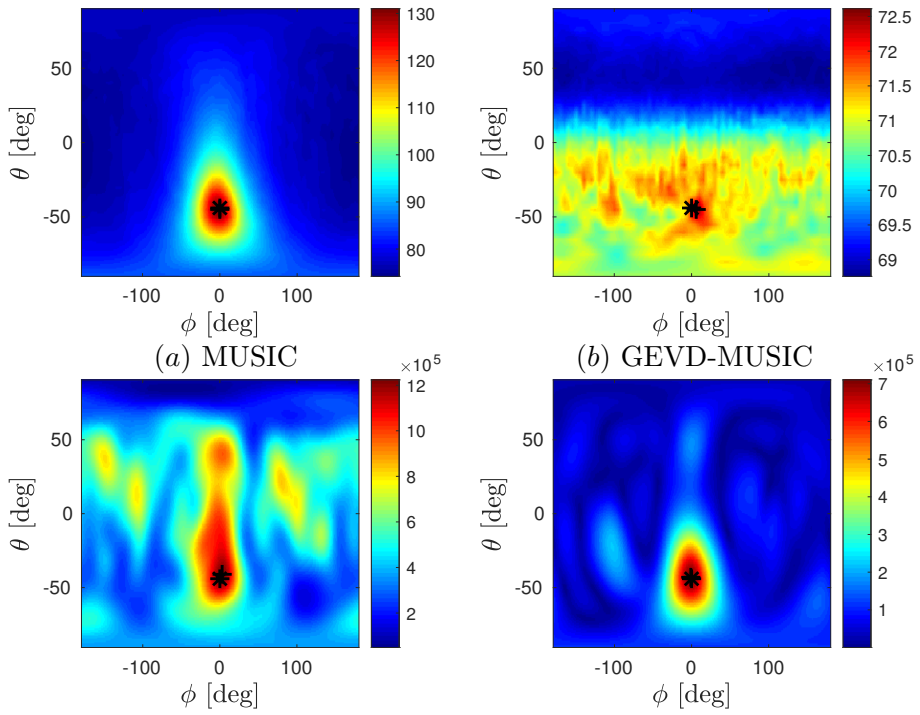
(a) $\overline{\text{SDNR}} = -30$ dB (b) $\overline{\text{SDNR}} = -20$ dB

(c) $\overline{\text{SDNR}} = -10$ dB (d) $\overline{\text{SDNR}} = 0$ dB

Figure 4.5: Instantaneous angular spectra over (θ, ϕ) for the proposed method at low $\overline{\text{SDNR}}$ levels for a sound source position at $(\theta_\ell = -44°, \phi_\ell = 0°)$.

We place a sound source at elevation, $\theta_\ell = -44°$ and azimuth, $\phi_\ell = 0°$ (for $\ell = 1$) on the ground with respect to the drone's center. Here, we vary the $\overline{\text{SDNR}}$ level ranges from -30 dB to 0 dB with an increment of 10 dB. Figure 4.5 illustrates the instantaneous angular spectra of the proposed method. We observe that the proposed method's angular spectrum is able to sharply distinguish with one prominent peak

corresponding to the true DOA of the sound source in all four cases. The peaks are sharper and not influenced by the drone noise[5]. We note that the prediction error is less than 2° for the given S$\overline{\text{D}}$NR levels in all source positions.



(a) MUSIC (b) GEVD-MUSIC

(c) GCC-PHAT (d) Our method

Figure 4.6: Performance comparison over (θ, ϕ) at S$\overline{\text{D}}$NR level of -15 dB for a sound source at $(\theta_\ell = -44°, \phi_\ell = 0°)$.

Let us now compare the source localisation performance with three state-of-the-art DOA estimation methods at low S$\overline{\text{D}}$NR conditions in Fig. 4.6 to Fig. 4.9. The instantaneous angular spectra with the true and estimated sound source position at S$\overline{\text{D}}$NR level of -15 dB and -20 dB, for the four methods are presented in Fig. 4.6 and Fig. 4.7, respectively. Note that we use the marker '$*$' as the true source location, whereas '$+$' as the estimated source position. It is observed that MUSIC, GEVD MUSIC algorithms, and our method perform with an obvious peak closes to

[5]It is claimed that drone propellers push the airflow downwards and microphones capture most of the self-generated wind noise by these propellers [29, 298]

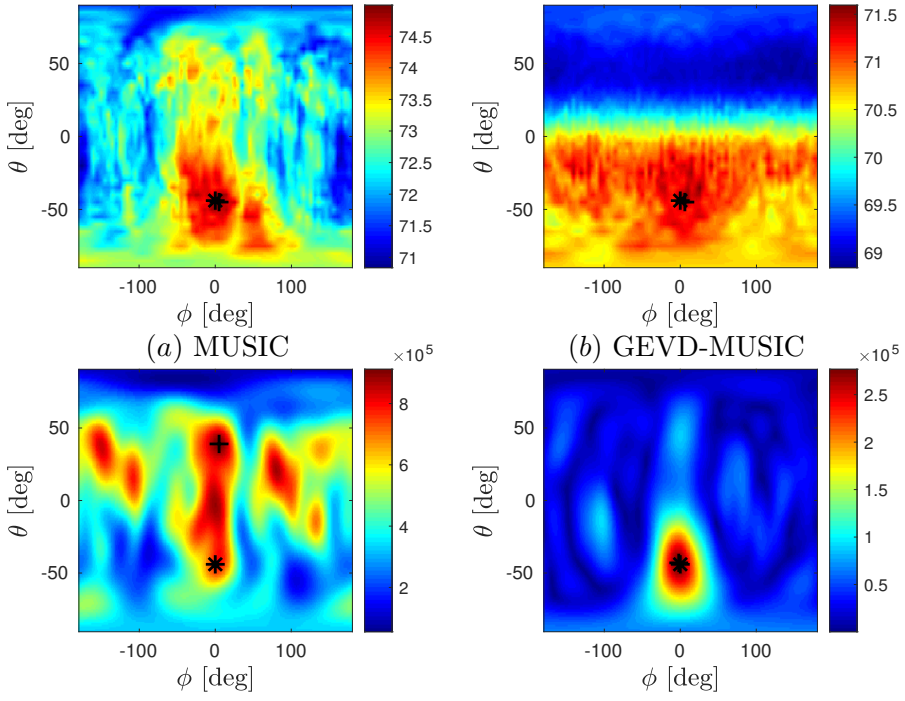


Figure 4.7: Performance comparison over (θ, ϕ) at S$\overline{\text{D}}$NR level is -20 dB with a sound source position at $(\theta_\ell = -44°, \phi_\ell = 0°)$.

the true sound source location. However, with lower S$\overline{\text{D}}$NR condition, i.e., S$\overline{\text{D}}$NR is -20 dB (see Fig. 4.7(c)), the GCC-PHAT fails to localise the sound source due to the drone noise generated by the motors and propellers in the upper hemisphere.

Figures 4.8 and 4.9 plot the localisation performance for all four methods at S$\overline{\text{D}}$NR level of -25 dB, and -30 dB, respectively. It is clear that the MUSIC, GEVD-MUSIC and GCC-PHAT algorithms fail to identify the location of the sound source due to the drone noise. Figure 4.8(c) shows that the GCC-PHAT even worse than the GEVD-MUSIC algorithm when the S$\overline{\text{D}}$NR level is -25 dB. Also, note that even with a restricted peak search in the lower hemisphere to localise the sound source using GCC-PHAT may not perform well based on the instantaneous angular spectrum since the peaks are less obvious. By contrast, the proposed method is still able to detect source position.

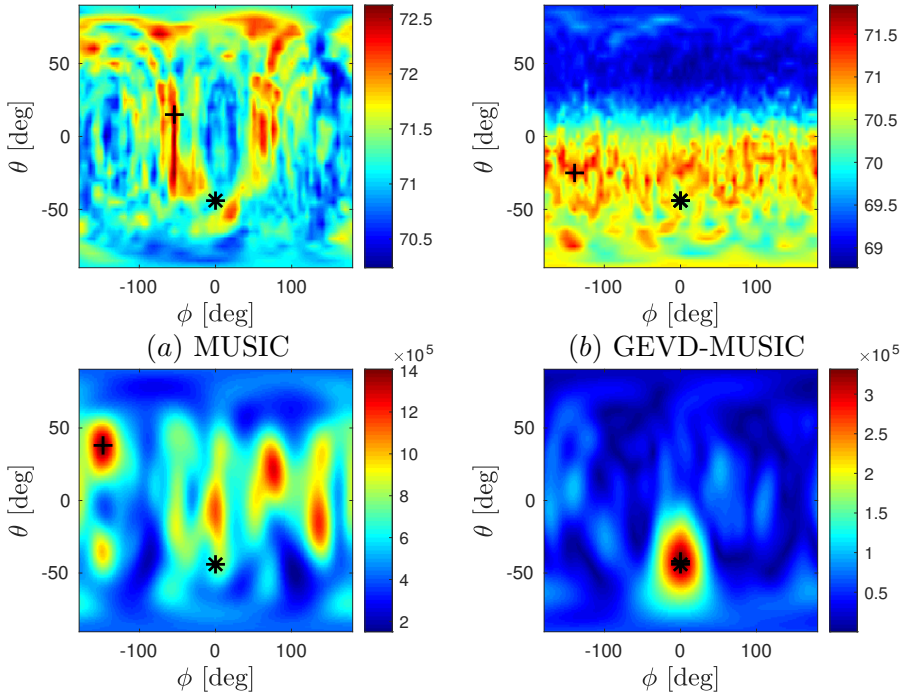


Figure 4.8: Performance comparison over (θ, ϕ) at S$\overline{\text{D}}$NR level is -25 dB with a sound source position at $(\theta_\ell = -44^\circ, \phi_\ell = 0^\circ)$.

The localisation accuracy for the single sound source under very low S$\overline{\text{D}}$NR levels varying from -30 dB to -15 dB for different source positions on the ground for the four methods is given in Table 4.1. We observe that the localisation performance degrades when the S$\overline{\text{D}}$NR levels gradually decrease in all four methods. We notice that when S$\overline{\text{D}}$NR > -20 dB, all three methods except GCC-PHAT are able to localise the sound source with acceptable accuracy. Comparing the proposed method to all three methods illustrates low prediction error in low S$\overline{\text{D}}$NR condition (e.g. -25 dB). The reason behind this is the drone noise characteristics changed with the varying motor current and resulted in faulty peaks in the instantaneous angular spectrum. In general, the conventional cross-correlation based methods, e.g. GCC-PHAT, and subspace-based methods, e.g MUSIC, and GEVD-MUSIC algorithms, are more vulnerable to additive drone noise. However, the subtrac-

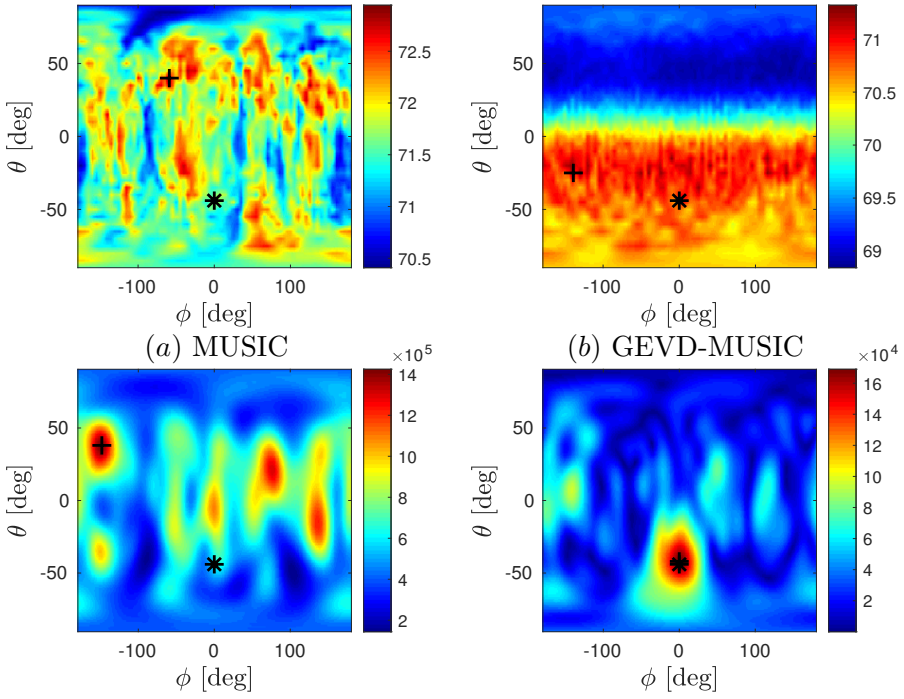


Figure 4.9: Performance comparison over (θ, ϕ) at S$\overline{\text{D}}$NR level is -30 dB with a sound source position at $(\theta_\ell = -44^\circ, \phi_\ell = 0^\circ)$.

tion of drone noise angular spectrum from the mixture in the proposed method will obtain more accurate results at the extreme S$\overline{\text{D}}$NR conditions. For the set of placement of the sound source shown in Table 4.1, the sound source position at $\theta = -44^\circ, \phi = -90^\circ$ has a higher prediction error for the proposed method compared to the other placements. It is clear that the localisation performance is affected by the source position on the ground.

It can be seen that the proposed method achieved and outperformed the MUSIC, GEVD-MUSIC, and GCC-PHAT methods for the single sound source localisation in the extremely low S$\overline{\text{D}}$NR levels.

Table 4.1: The prediction error for single source localisation. Note that the lowest prediction error for a given position is highlighted.

$\overline{\text{SDNR}}$ level (dB)	Position (θ_ℓ, ϕ_ℓ)	MUSIC	GEVD-MUSIC	GCC-PHAT	Our method
−15	$(-44°, 0°)$	**0.05°**	5.95°	2.95°	1.05°
	$(-44°, 90°)$	**1.00°**	5.09°	63.36°	2.83°
	$(-42°, -150°)$	4.11°	1.42°	**2.66°**	3.15°
	$(-43°, -90°)$	9.22°	**5.38°**	90.73°	6.32°
	$(-43°, -30°)$	4.30°	15.61°	7.44°	**1.00°**
−20	$(-44°, 0°)$	4.95°	9.95°	3.99°	**2.05°**
	$(-44°, 90°)$	35.01°	13.42°	64.34°	**2.24°**
	$(-42°, -150°)$	36.03°	14.69°	112.83°	**1.78°**
	$(-43°, -90°)$	24.17°	19.02°	90.83°	**7.61°**
	$(-43°, -30°)$	4.30°	10.08°	31.35°	**0.5°**
−25	$(-44°, 0°)$	55.05°	139.05°	148.01°	**0.05°**
	$(-44°, 90°)$	50.68°	128.34°	121.80°	**3.61°**
	$(-42°, -150°)$	98.55°	14.69°	38.69°	**1.78°**
	$(-43°, -90°)$	44.41°	51.39°	85.24°	**7.80°**
	$(-43°, -30°)$	29.98°	109.97°	133.87°	**2.25°**
−30	$(-44°, 0°)$	59.01°	139.05°	148.01°	**0.05°**
	$(-44°, 90°)$	148.77°	128.34°	94.23°	**1.41°**
	$(-42°, -150°)$	91.02°	14.69°	38.69°	**1.99°**
	$(-43°, -90°)$	31.13°	51.39°	85.25°	**9.43°**
	$(-43°, -30°)$	29.07°	109.98°	133.87°	**3.26°**

4.4.3 Results for multi-source localisation

The DOA estimation of two simultaneously active sound sources is evaluated considering two scenarios.

Scenario A - with adjacent sound sources under same $\overline{\text{SDNR}}$ conditions

We place two sound sources under same $\overline{\text{SDNR}}$ conditions, the level of −10 dB and −15 dB, whose elevation and azimuth angles are $(-44°, 30°)$, $(-45°, 60°)$, and $(-44°, 90°)$ with $(-44°, 0°)$ by varying the azimuth angular distance between them by 30°, 60°, and 90° as shown in Fig. 4.10. Note that the sources' elevation is −45° with respect to the drone. However, due to uneven floor, the elevation angle of

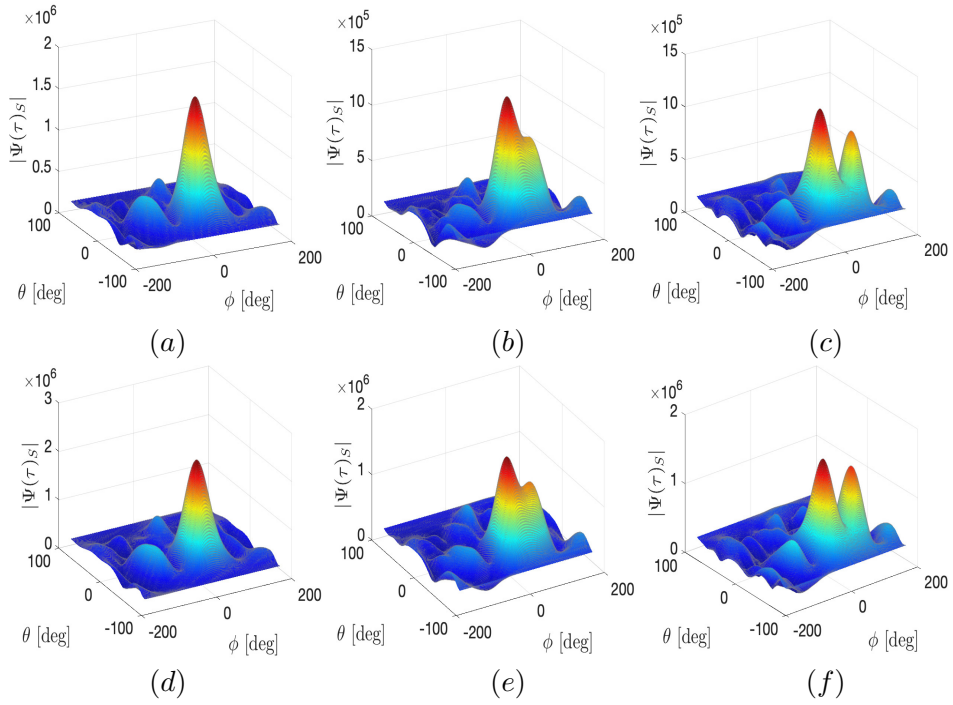


Figure 4.10: Instantaneous angular spectrum over (θ, ϕ) for the proposed method where the azimuth angular difference between the two adjacent sound sources in (a,d) 30°, (b,e) 60°, and (c,f) 90°, and the S$\overline{\text{D}}$NR level of in (a-c) -15 dB, and (d-f) -10 dB.

each sound source differs by $\pm 2.5°$.

Table 4.2: The prediction errors of the proposed method for multi-source localisation under various S$\overline{\text{D}}$NR conditions. Note that the prediction error convention is given as (first source/second source).

Angular difference	S$\overline{\text{D}}$NR level (dB)	
	-10	-15
30°	**8.95°**/125.81°	**8.95°**/131.88°
60°	**1.95°**/6.03°	**1.95°**/156.84°
90°	**0.05°**/3.16°	**0.05°**/3.16°

Table 4.2 displays the prediction error under the above situation. We recognise that there are two obvious peaks detected for azimuth angular difference of 90°, and 60°, in both S$\overline{\text{D}}$NR environments, and -10 dB, respectively, and all the other

cases fail to distinguish the two sound sources.

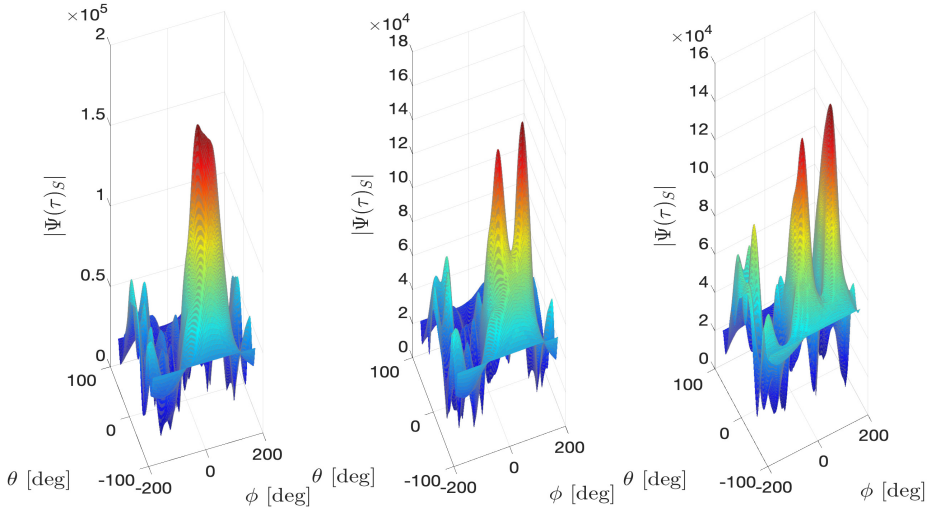


Figure 4.11: Instantaneous angular spectra over (θ, ϕ) for the proposed method at $\overline{\text{SDNR}}$ level of -30 dB and azimuth angular difference between the two adjacent sound sources in (a) 60°, (b) 90°, and (c) 120°.

Figure 4.11 illustrates the two sound sources whose separation distance is 60°, 90° and 120° at the $\overline{\text{SDNR}}$ level of -30 dB. We observe that two sharper peaks in the angular spectrum corresponding to the active sound sources for the azimuth angular difference of 90° and 120° with the prediction error of 0.95°/2.24°, and 0.05°/6.44°, respectively. For the sound source separation of 60°, two peaks are merged and not able to distinguish them separately. The first sound source is detected with the prediction error of 9.95° whereas the second sound source has a higher prediction error of 121.75°.

Let us now confirm the localisation performance for two active sources with the MUSIC algorithm. Similar to Fig. 4.10, we examine two adjacent sound sources to detect the minimum azimuth angular difference between them, for the MUSIC algorithm, under the $\overline{\text{SDNR}}$ level of -10 dB in Fig. 4.12. We can see two obvious peaks when the angular distance is 90°. It is claimed that the limitation for the azimuth angular distance between two adjacent sources for both MUSIC and the proposed method deduce from the microphone array geometry on the drone.

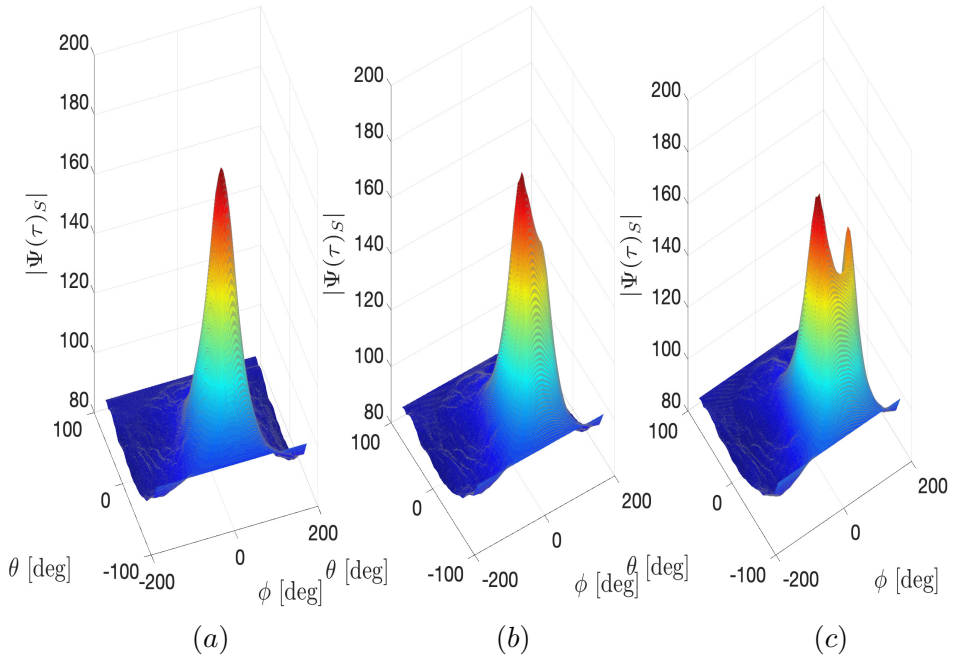
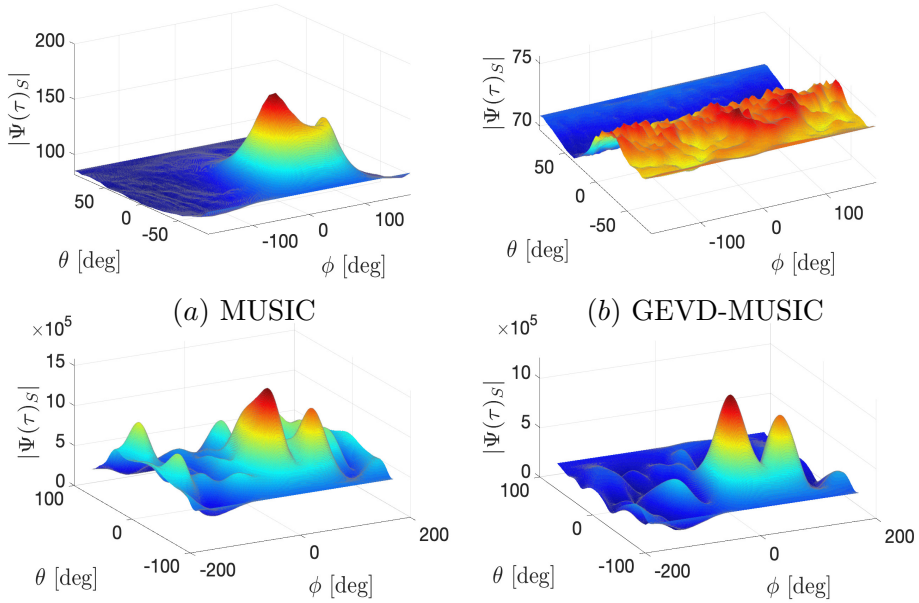


Figure 4.12: MUSIC spectrum over (θ, ϕ) for azimuth angular difference between the two adjacent sound sources in (a) 30°, (b) 60°, and (c) 90°. The S$\overline{\text{D}}$NR level is -10 dB.

Based on the above assumption that the sound source separation angle should be at least 90° to accurately detect the adjacent sound sources at low S$\overline{\text{D}}$NR conditions, we evaluate the localisation performance for the two sound sources position at $(\theta_1 = -44°, \phi_1 = 0°)$ and $(\theta_2 = -44°, \phi_2 = 90°)$ at S$\overline{\text{D}}$NR level of -15 dB for all four methods in Fig. 4.13. It is shown that two local maximum peaks corresponding to the target sound sources are identified for all four methods. However, for the GEVD-MUSIC, we observe that peaks are not distinguishable in the lower hemisphere.

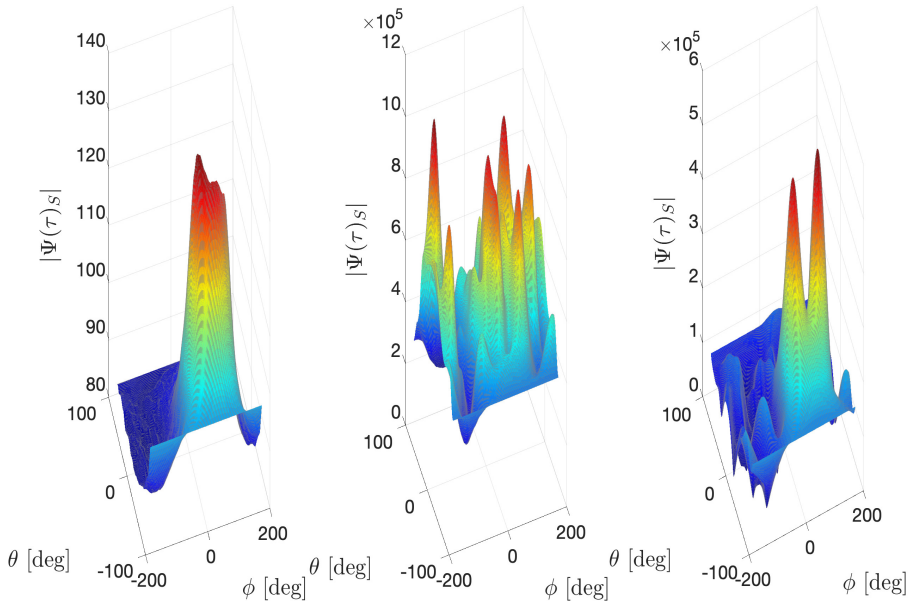


(a) MUSIC (b) GEVD-MUSIC

(c) GCC-PHAT (d) Our method

Figure 4.13: Performance comparison over (θ, ϕ) at the S$\overline{\text{D}}$NR level of -15 dB with the sound sources positions at $(\theta_1 = -44°, \phi_1 = 0°)$ and $(\theta_2 = -44°, \phi_2 = 90°)$.

Next, we look into the effect of the adjacent sound sources at -20 dB in Fig. 4.14. Note that we eliminated the GEVD-MUSIC algorithm from the analysis. It is shown that the GCC-PHAT method fails to localise the true sound source positions at the S$\overline{\text{D}}$NR level of -20 dB. We can observe that two sharper and barely recognisable peaks for the proposed method and MUSIC algorithm, respectively. The MUSIC algorithm detects only one sound source at $(-44°, 0°)$ whereas fails to distinguish the sound source at $(-44°, 90°)$ accurately as the prediction error is 50.88°. In contrast, the proposed method has two obvious peaks close to the true sound source positions. Table 4.3 summarises the prediction errors for the above scenarios. Under these cases, the proposed method has the lowest prediction error compared to the state-of-the-art methods.



(*a*) MUSIC (*b*) GCC-PHAT (*c*) Our method

Figure 4.14: Performance comparison over (θ, ϕ) at the $\overline{\text{SDNR}}$ level of -20 dB with the sound sources positions at $(\theta_1 = -44°, \phi_1 = 0°)$ and $(\theta_2 = -44°, \phi_2 = 90°)$.

Table 4.3: The prediction errors for the sound sources position at $(\theta_1 = -44°, \phi_1 = 0°)$ and $(\theta_2 = -44°, \phi_2 = 90°)$. Note that the prediction error convention is given as (first source/second source).

Method	$\overline{\text{SDNR}}$ level (dB)	
	-15	-20
MUSIC	**0.05°**/5.09°	**0.95°**/50.88°
GEVD-MUSIC	4.95°/6.39°	
GCC-PHAT	2.95°/13.34°	149.01°/65.58°
Proposed method	**0.05°/3.16°**	**0.95°/3.16°**

According to the above discussion, it is clear that all four methods suffer from a limitation which emerges with the given microphone array that the minimum azimuth angular difference between two sound sources needs to be at least 90° to accurately localise the speech source at the $\overline{\text{SDNR}}$ level ≤ -10 dB. The reason behind this constraint unfolds with the microphone placement which only covers four axes on the azimuth plane that leads to the minimum angular distance between

the adjacent sound source for localisation. We can use this approach to deal with more sound sources whose SD̄NR levels are approximately equal, although in many situations performance may drop as soon as azimuth angular distance is less than 90° or the number of sources is larger than four.

Scenario B - SD̄NR level difference between the active sound sources

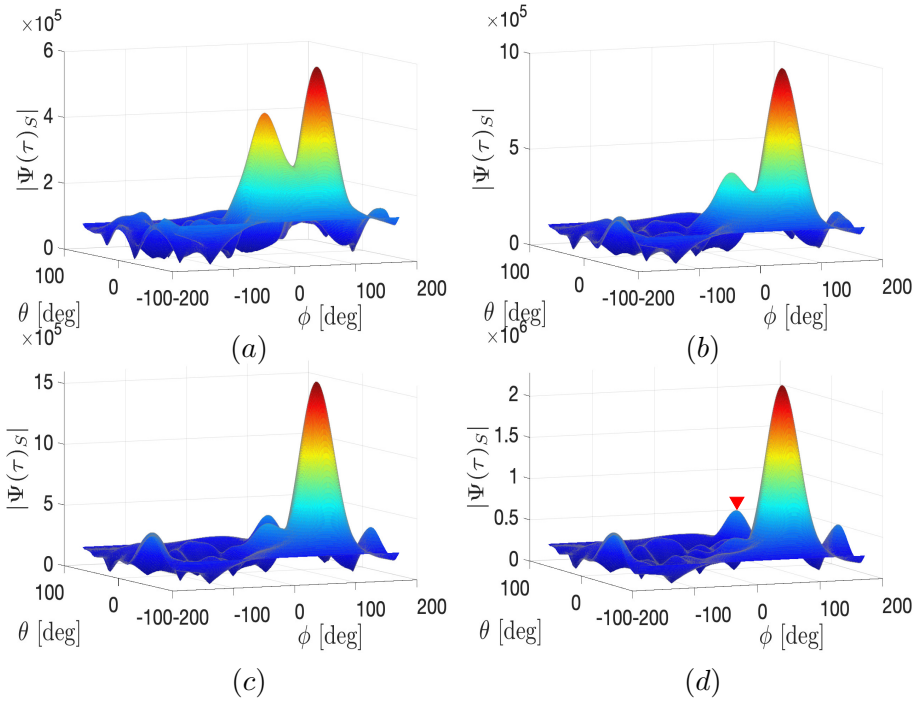


Figure 4.15: Instantaneous angular spectra over (θ, ϕ) for the proposed method where the azimuth angular difference between the two adjacent sound sources is 90°, and the SD̄NR level difference w.r.t to −25 dB in (a) 5 dB, (b) 10 dB, (c) 15 dB, and (d) 20 dB. Note that the red triangle in (d) indicates a false peak detection.

Here, we consider two sound sources similar to the previous scenario with azimuth angular distance between the two sound sources equals to 90° and evaluate the proposed algorithm under various SD̄NR levels. We fix one sound source whose elevation and azimuth is $(-44°, 0°)$ with SD̄NR level of $\{-30, -25, -20, -15, -10\}$ dB, and vary the SD̄NR level of the second sound source whose elevation and azimuth is $(-44°, 90°)$ by incrementally increasing the SD̄NR level by 5 dB with

respect to the first sound source to localise both sound sources. Figure 4.15 shows the instantaneous angular spectrum for S$\overline{\text{D}}$NR level difference with respect to −25 dB. We can observe that for up to 15 dB S$\overline{\text{D}}$NR level difference, the proposed method is still capable of detecting two peaks. When the S$\overline{\text{D}}$NR level difference is greater than 15 dB, the proposed method can only localise the sound source with the highest S$\overline{\text{D}}$NR level (a false peak is detected as shown by the red triangle in Fig. 4.15(d)). Table 4.4 displays the prediction error for this scenario. It is clear that if the sound sources have more than 10 dB S$\overline{\text{D}}$NR level difference, we cannot ensure an accurate DOA estimation for both sound sources.

Table 4.4: The prediction errors of the proposed method for adjacent sound sources in multi-source localisation scenario with varying S$\overline{\text{D}}$NR level difference. Note that the prediction error convention is given as (first source/second source).

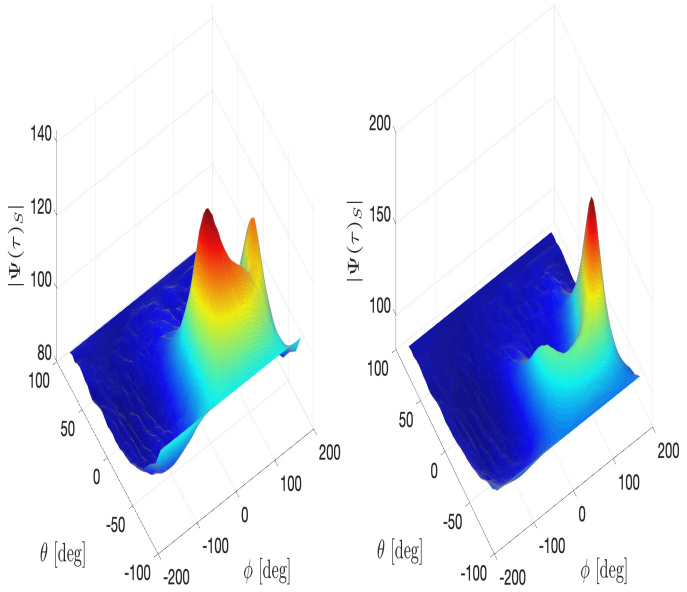
S$\overline{\text{D}}$NR level difference (dB)	S$\overline{\text{D}}$NR level (dB)				
	−10	−15	−20	−25	−30
5	0.95°/3.61°	0.95°/3.61°	1.95°/3.61°	0.95°/3.16°	2.95°/3.61°
10	1.95°/3.61°	1.95°/3.61°	4.95°/3.61°	3.95°/3.61°	6.95°/3.61°
15	3.95°/2.24°	4.95°/3.61°	87.00°/3.61°	6.95°/3.61°	159.05°/3.61°
20	88.00°/2.24°	88.00°/2.24°		87.00°/3.61°	

According to the table 4.3 and the performance comparison of the single sound source localisation (as shown in Fig. 4.8 and Fig. 4.9), we notice that at the lower S$\overline{\text{D}}$NR level, the local maximum peaks are not very prominent to detect using the GEVD-MUSIC algorithm whereas GCC-PHAT indicates false peaks in the upper hemisphere with the high prediction errors due to the presence of the drone noise. From now on, we only compare the performance of the proposed method with the MUSIC algorithm.

Since the MUSIC algorithm is capable of localising the sound sources at −20 dB, we consider two adjacent sound sources with the azimuth angular distance of 90° and analyse the effect of the S$\overline{\text{D}}$NR level difference in Table 4.5. We can observe that the MUSIC algorithm is capable of identifying the two sound sources whose source separation is at 90° over 20 dB S$\overline{\text{D}}$NR level difference for S$\overline{\text{D}}$NR level ≥ -15 dB, unlike the proposed method. It is shown that the MUSIC algorithm performs well in higher S$\overline{\text{D}}$NR level difference, e.g. 5 dB and 10 dB, for the azimuth angular difference between the adjacent sound source of 90°.

Table 4.5: The prediction errors of the MUSIC algorithm for adjacent sound sources in multi-source localisation scenario with varying $\overline{\text{SDNR}}$ level difference. Note that the prediction error convention is given as (first source/second source).

$\overline{\text{SDNR}}$ level difference (dB)	$\overline{\text{SDNR}}$ level (dB)		
	-10	-15	-20
5	0.05°/4.00°	0.05°/4.12°	4.95°/5.09°
10	2.05°/4.00°	2.05°/4.00°	5.95°/5.09°
15	1.05°/4.00°	1.05°/4.00°	179.03°/4.12°
20	0.05°/4.00°	0.05°/4.00°	179.03°/4.12°



(a) (b)

Figure 4.16: MUSIC spectra over (θ, ϕ) for azimuth angular difference between the two adjacent sound sources is 120°, and the $\overline{\text{SDNR}}$ level difference w.r.t to -20 dB in (a) 0 dB, and (b) 15 dB.

Finally, we place two sound sources whose elevation and azimuth angles are $(-44°, 0°)$, and $(-43°, 120°)$ with respect to the drone under different $\overline{\text{SDNR}}$ conditions. Figure 4.16(a) shows that both sound sources are at -20 dB and notice two peaks with a lower prediction error of 5.05°/3.26°. However, MUSIC fails to localise the multiple sound sources with the azimuth angular difference of 90° in the

S$\overline{\text{D}}$NR level of -20 dB (as shown in Fig. 4.14(a)). It is clear that in lower S$\overline{\text{D}}$NR conditions, the MUSIC algorithm is able to identify both sound sources when the azimuth angular difference equals 120°. In Fig. 4.16(b) plots MUSIC spectra for S$\overline{\text{D}}$NR level difference of 15 dB. We can observe that MUSIC is able to detect two peaks with a prediction error of 0.05°/2.59°.

The multi-source localisation performance of the proposed method outperforms the MUSIC based algorithms (MUSIC and GEVD-MUSIC), and the GCC-PHAT method in terms of accuracy under extreme S$\overline{\text{D}}$NR levels.

Overall, as verified by the above experimental results, MUSIC, GEVD-MUSIC algorithms, and the GCC-PHAT method work well over the higher S$\overline{\text{D}}$NR levels (≥ -15 dB). However, the proposed method works well in lower S$\overline{\text{D}}$NR conditions (e.g. -30 dB), for both single and multiple sound source localisation, with a minimum azimuth angular distance (90°) between the sound sources for the given microphone array setup on the drone.

4.5 Summary and Contributions

We have proposed a robust sound source localisation method for both single and multi-source scenarios using an irregular microphone array embedded in a drone. This work has highlighted the practical advantage of embedding the microphones on the drone which allows an arbitrary array geometry as opposed to a standard microphone array on the drone limited to linear, circular, cubic, or spherical arrays. The chapter contributions can be itemised as follows:

- A cross-correlation based DOA estimation method using the TDOA at different microphone pairs, with noise angular spectrum subtraction was introduced to accurately overcome angular spectrum distortions at extreme S$\overline{\text{D}}$NR levels due to drone noise.

- The use of drone noise characteristics such as motor current-specific drone noise was shown to achieve better noise suppression and minimise the prediction error.

- Successful localisation of a single and simultaneously active two sound sources was proven through a practical experiment at low SdNR conditions (up to -30 dB). The localisation of two sound sources located at a certain azimuth angular separation was also illustrated with acceptable accuracy.

In summary, we have proposed an algorithm for sound source localisation using on-board microphones on the drone. The algorithm uses known (estimated or measured) drone-noise-only recordings to cancel the unwanted noise generated by motors and propellers. Due to its simplicity, applicability to any array geometry, and better robustness against drone noise, the proposed method increases the feasibility of localisation under extreme SdNR levels. In the next chapter, we propose an audio signal enhancement framework to extract the target sound source signal, which is robust to adverse noisy environments.

4.6 Related Publications

- **W. N. Manamperi**, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone Audition: Sound Source Localisation using On-Board Microphones," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 30, pp. 508 – 519, January 2022.

Chapter 5

Drone Noise Suppression with Wiener Filtering for Multichannel Audio Signal Enhancement

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 $S\overline{D}NR$ level. In this chapter, we propose a solution using Wiener filtering techniques to remove drone noise and enhance the target sound source signal from the microphone recordings. The framework exploits known statistics of motor current-specific/speed-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.

5.1 Introduction

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 [2]. 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 ($\overline{\text{SDNR}}$) (defined as the power ratio between the source signal and the drone noise) level[1]. In this chapter, we address the problem of audio signal enhancement at very low $\overline{\text{SDNR}}$ conditions (up to -30 dB) for drone mounted or embedded microphone array platforms. Unlike in the conventional speech enhancement in a room scenario [299], it is challenging (depends on the application [300]) 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 aim to enhance the audio signal recorded from drone on-board microphones when the drone hovers above a constant height from a sound source on the ground. To this end, we propose a solution for any arbitrary microphone array embedded/mounted on the drones that differentiates from a specific array geometry attached to the drones as in [9, 15, 19, 20, 74, 75, 78, 266–268, 270] and demonstrate the audio signal enhancement for both (i) speech, and (ii) bird calls.

We present a multichannel signal enhancement framework using Wiener filtering approaches which combines the audio and motor modalities. In particular, we propose a framework of (i) the widely known multichannel Wiener filter (MWF) [186, 301–303] to remove drone noise from microphone recordings, and (ii) an unsupervised learning method; Gaussian mixture model (GMM) based dual-stage parametric Wiener filter (WF) to further reduce of residual noise. The main contributions of this chapter are (i) the single channel GMM WF, and (ii) the

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

combination of these two Wiener filtering techniques in the application of signal enhancement for drones. We use the MWF (similar to [186, 301, 302] but not restricting to a preferred microphone channel) for a significant drone noise suppression and the GMM WF for single channel speech enhancement as a post-filter to obtain the enhanced sound source signal. With an accurate estimate of the drone noise PSD, MWF derived from [186, 301, 302] suppresses the drone noise more effectively at very low S$\overline{\text{D}}$NR conditions below -10 dB.

In Chapter 3, we showed that drone harmonic noise is proportional to the motor current. To this end, we combine the audio and motor modalities to better suppress the drone noise from the microphone recordings. We use non-acoustical information (as in [9, 267, 268, 270]) 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 by updating it recursively with a few seconds of multichannel recording. In this chapter, 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 using the proposed GMM WF. A clean source signal dataset is used to model the source PSD in the *training* phase of the 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 are used to estimate the source and noise-only PSDs from the output of the MWF using dual-stage Wiener filtering.

We provide the evaluation of the proposed framework on (i) our motor current-specific drone noise dataset (in Section 4.4.1), and (ii) motor speed-specific drone noise dataset (the DREGON dataset in [22]) for speech enhancement. We compare the performance in terms of S$\overline{\text{D}}$NR improvement, 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 [208] and MWF in [22] over our drone acoustic dataset.

We then reformulate the audio signal enhancement problem using the noisy drone in the presence of interfering noise sources such as wind, and traffic noise. We generalise the drone noise removal process in the GMM WF for incorporating the interfering sources. We evaluate the modified framework through experimental recordings.

The remainder of this chapter is structured as follows. In Section 5.3, we present the proposed framework using MWF with a set of GMM WF for a drone-embedded microphone array at very low $\overline{\text{SDNR}}$ conditions (up to -30 dB). In Section 5.4, we evaluate and analyse the proposed method using an extensive experimental study including the experimental recordings from two different drone acoustics datasets, and the outdoor measurements from a hovering drone for a bioacoustic application. In Section 5.5, we generalise the GMM WF to remove the interfering sources as well as the drone noise and validate the performance. In Section 5.6, we summarise and provide conclusions.

5.2 Problem Formulation

We aim to extract a desired sound source signal from drone on-board multiple microphone recordings.

Consider a drone with Q microphones embedded to its structure, 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

$$\mathbf{p}(f,t) = \mathbf{x}(f,t) + \mathbf{v}(f,t), \tag{5.1}$$

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 transpose operation. The source component as $\mathbf{x}(f,t)$ can be written

$$\mathbf{x}(f,t) = \mathbf{d}(f)s(f,t), \tag{5.2}$$

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 the noisy source correlation matrix as

$$\mathbf{\Phi}_{pp}(f,t) \triangleq E\{\mathbf{p}(f,t)\mathbf{p}^H(f,t)\}, \tag{5.3}$$

the clean sound source correlation matrix as

$$\mathbf{\Phi}_{xx}(f,t) \triangleq E\{\mathbf{x}(f,t)\mathbf{x}^H(f,t)\}, \tag{5.4}$$

and the drone noise correlation matrix as

$$\mathbf{\Phi}_{vv}(f,t) \triangleq E\{\mathbf{v}(f,t)\mathbf{v}^H(f,t)\}, \tag{5.5}$$

where $E\{\cdot\}$ is the statistical expectation operator. We assume drone noise signals to be uncorrelated with the sound source signal. Therefore, using (5.1), and substituting both (5.4) and (5.5) into (5.3) gives

$$\mathbf{\Phi}_{pp}(f,t) = \mathbf{\Phi}_{xx}(f,t) + \mathbf{\Phi}_{vv}(f,t). \tag{5.6}$$

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 maneuver with constant current flow through the motors allowing us to estimate $\mathbf{\Phi}_{vv}(f,t)$ when there is no sound source present.

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

5.3 Multi-microphone Signal Enhancement Framework using Wiener Filtering

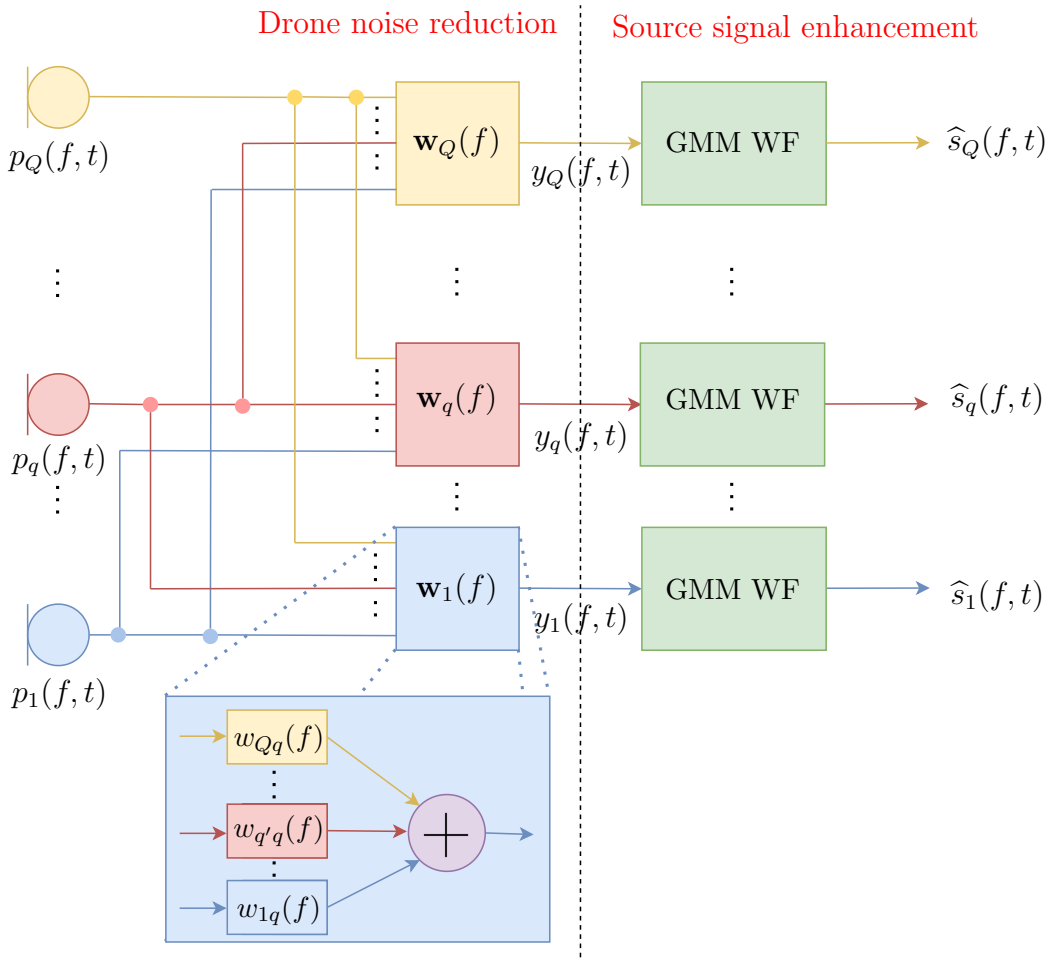


Figure 5.1: A block diagram of the proposed multichannel audio signal enhancement framework.

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

The proposed solution in Fig. 5.1 is comprised of two components:

(i) MWF for drone noise reduction, and

(ii) GMM WF for source signal enhancement.

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 5.3.1. Next, we filter the outputs of MWF $y_q(f,t)$, for $q = 1,\ldots,Q$ through a set of GMM WF for source signal enhancement to obtain Q estimation of original sound source signal $s(f,t)$, as $\widehat{s}_q(f,t)$, for $q = 1,\ldots,Q$. The detailed method is given in subsection 5.3.2.

5.3.1 Drone noise suppression with MWF

The output of the MWF is calculated as

$$\mathbf{y}(f,t) = \mathbf{W}^H(f)\mathbf{p}(f,t), \tag{5.7}$$

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 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,

$$y_q(f,t) = \mathbf{w}_q^H(f)\mathbf{p}(f,t), \tag{5.8}$$

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 minimise the following mean squared error criterion

$$J_q(f,t) = E\{|x_q(f,t) - y_q(f,t)|^2\}. \tag{5.9}$$

The solution to (5.9) derives the conventional form of MWF [303–305] as

$$\mathbf{w}_q(f) = \mathbf{\Phi}_{pp}^{-1}(f,t)E\{\mathbf{p}(f,t)x_q^*(f,t)\}. \tag{5.10}$$

By combining all weights vectors $\mathbf{w}_q(f)$ for $q = 1,\ldots,Q$ in (5.10) together, we obtain

$$\mathbf{W}(f) = \mathbf{\Phi}_{pp}^{-1}(f,t)\mathbf{\Phi}_{xx}(f,t), \tag{5.11}$$

where $\mathbf{W}(f) = [\mathbf{w}_1(f),\ldots,\mathbf{w}_Q(f)]$. Using (5.6) and (5.11), this can be rewritten as

$$\mathbf{W}(f) = \mathbf{I}_Q - \mathbf{\Phi}_{pp}^{-1}(f,t)\mathbf{\Phi}_{vv}(f,t), \tag{5.12}$$

where $\mathbf{I}_Q$ denotes the identity matrix of size $Q \times Q$. Given the very low S$\overline{\text{D}}$NR conditions, we reformulate (5.12) to obtain the optimal weight matrix using the inverse of $\mathbf{\Phi}_{vv}(f,t)$ similar to the work in [186, 301, 302].[2] For that, we expand the inverse of $\mathbf{\Phi}_{pp}(f,t)$ in (5.6). As $\mathbf{\Phi}_{xx}(f,t) = \Phi_{ss}(f,t)\mathbf{d}(f)\mathbf{d}^H(f)$ given that the power spectral density (PSD) of source signal $\Phi_{ss}(f,t) = E\{|s(f,t)|^2\}$, from (5.6), we write

$$\mathbf{\Phi}_{pp}(f,t) = \Phi_{ss}(f,t)\mathbf{d}(f)\mathbf{d}^H(f) + \mathbf{\Phi}_{vv}(f,t). \tag{5.13}$$

Using Sherman–Morrison formula [306] and further simplifications, we obtain

$$\mathbf{\Phi}_{pp}^{-1}(f,t) = \mathbf{\Phi}_{vv}^{-1}(f,t) - \frac{\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{\Phi}_{xx}(f,t)\mathbf{\Phi}_{vv}^{-1}(f,t)}{1 + \Phi_{ss}(f,t)\mathbf{d}^H(f)\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{d}(f)}. \tag{5.14}$$

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

$$tr[\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{\Phi}_{xx}(f,t)] = \Phi_{ss}(f,t)\mathbf{d}^H(f)\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{d}(f), \tag{5.15}$$

where $tr[\cdot]$ denotes the trace operator. Finally, using (5.15), (5.14) in (5.12) together with (5.6), we express the optimum filter weight that needs to be applied

[2]Note that when considering the inverse $\mathbf{\Phi}_{vv}(f,t)$ drone noise correlation matrix, as been seen as means to reduce the drone noise from the mixture correlation matrix $\mathbf{\Phi}_{pp}(f,t)$. We discuss this difference of MWF output results in Section 5.4.4.

to each microphone outputs of a drone to remove drone noise as

$$\mathbf{W}(f) = \frac{\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{\Phi}_{pp}(f,t) - \mathbf{I}_Q}{1 + tr[\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{\Phi}_{pp}(f,t)] - Q}. \tag{5.16}$$

A similar derivation can also be found in [301] with respect to a reference microphone channel. Note that the optimal weight applied to the q^{th} microphone channel $\mathbf{w}_q(f)$ for $q = 1, \ldots, Q$ to obtain the drone noise suppressed signal as in (5.8) 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, i.e., correlation matrices $\mathbf{\Phi}_{pp}(f,t)$ and $\mathbf{\Phi}_{vv}(f,t)$: 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 (5.16), there is no requirement of prior knowledge of the drone-related transfer functions or free-field sound propagation assumption, but only estimates of the statistics of the drone noise, received signal, and the number of microphones in the array.

5.3.2 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 $\widehat{v}_q(f,t)$, for $q = 1, \ldots, Q$. To further enhance the output signals of the MWF, we propose a single channel GMM based dual-stage parametric WF (also known as the GMM WF) to each output of MWF $\mathbf{y}(f,t)$, for $q = 1, \ldots, Q$ (see Fig. 5.1).

The GMM WF has a *training phase* and an *enhancement phase* as shown in Fig. 5.2. 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*.

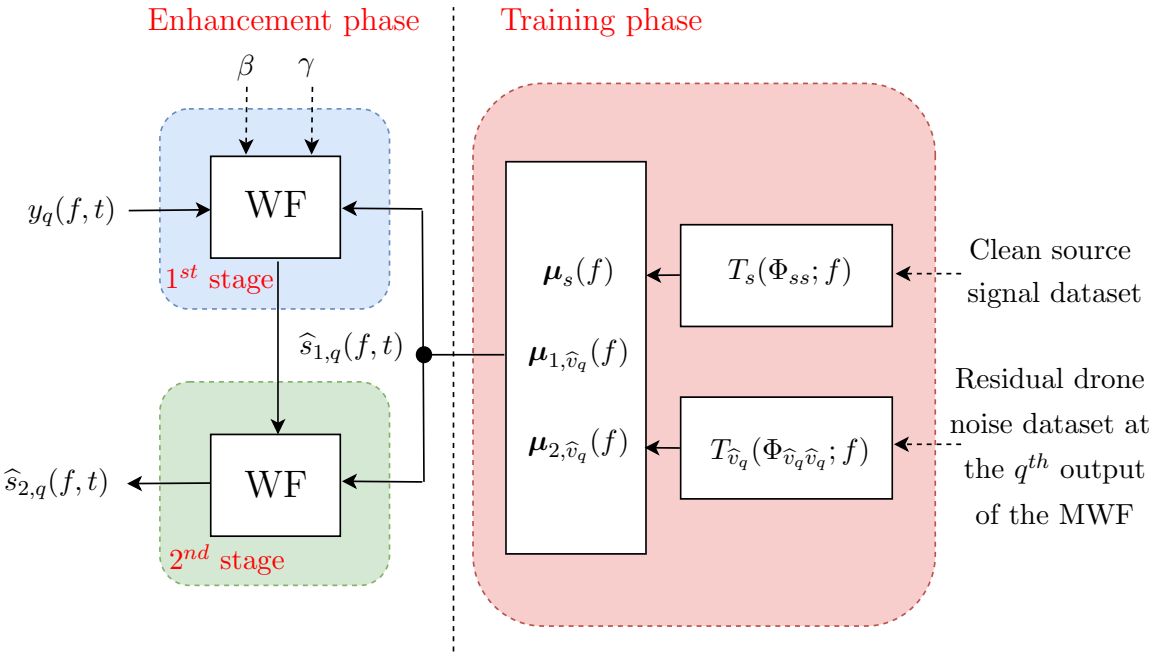


Figure 5.2: Illustration of the single channel GMM WF at the q^{th} output of the MWF. The source and residual drone noise PSDs are obtained using (5.19) and (5.20).

GMM model representation

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

$$T_s(\Phi_{ss};f) = \sum_{k=1}^{K_s} C_{sk}(f)\mathcal{N}(\mu_{sk}(f),\Sigma_{sk}(f)), \tag{5.17}$$

and

$$T_{\widehat{v}_q}(\Phi_{\widehat{v}_q\widehat{v}_q};f) = \sum_{k=1}^{K_{\widehat{v}_q}} C_{\widehat{v}_qk}(f)\mathcal{N}(\mu_{\widehat{v}_qk}(f),\Sigma_{\widehat{v}_qk}(f)), \tag{5.18}$$

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, its contribution, the mean value, and the variance of k^{th} GMM component at the f^{th} frequency for $i \in \{s,\widehat{v}_q\}$, for source and the

residual drone noise at the q^{th} output of the MWF, respectively.[3]

Average power spectra

As shown in Fig. 5.2, 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, the 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 [307] by a weighted sum of GMM mean values over a number of GMM components. Then, PSDs can be expressed as

$$\Phi_{ss}(f,t) = \boldsymbol{\mu}_s^T(f)\,\boldsymbol{\alpha}_s(t), \tag{5.19}$$

$$\Phi_{n,\widehat{v}_q\widehat{v}_q}(f,t) = \boldsymbol{\mu}_{n,\widehat{v}_q}^T(f)\,\boldsymbol{\alpha}_{n,\widehat{v}_q}(t), \tag{5.20}$$

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

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

$$\widehat{\Phi}_{y_q y_q}(f,t) = \boldsymbol{\mu}_s^T(f)\,\boldsymbol{\alpha}_s(t) + \boldsymbol{\mu}_{n,\widehat{v}_q}^T(f)\,\boldsymbol{\alpha}_{n,\widehat{v}_q}(t). \tag{5.21}$$

By arranging (5.21) for each frequency bin f, we obtain a set of linear equations in a matrix form for a given time frame t as

[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 [307]. Note that we omit $\mathbf{w}_q^H(f)\mathbf{d}(f)$ (in (5.2)) term for simplicity.

$$\widehat{\boldsymbol{\Phi}}_{y_q y_q}(t) = \left[\boldsymbol{U}_s \, \boldsymbol{U}_{n,\widehat{v}_q}\right] \begin{bmatrix} \boldsymbol{\alpha}_s(t) \\ \boldsymbol{\alpha}_{n,\widehat{v}_q}(t) \end{bmatrix}, \tag{5.22}$$

where $\boldsymbol{U}_s = [\boldsymbol{\mu}_s^T(f_1), \ldots, \boldsymbol{\mu}_s^T(f_F)]$, $\boldsymbol{U}_{n,\widehat{v}_q} = [\boldsymbol{\mu}_{n,\widehat{v}_q}^T(f_1), \ldots, \boldsymbol{\mu}_{n,\widehat{v}_q}^T(f_F)]$, and $\widehat{\boldsymbol{\Phi}}_{y_q y_q}(t) = [\widehat{\Phi}_{y_q y_q}(f_1, t), \ldots, \widehat{\Phi}_{y_q y_q}(f_F, t)]^T$. It can be solved to find the power coefficients of the current time frame t at stage n as

$$\begin{bmatrix} \boldsymbol{\alpha}_s(t) \\ \boldsymbol{\alpha}_{n,\widehat{v}_q}(t) \end{bmatrix} = \left[\boldsymbol{U}_s \, \boldsymbol{U}_{n,\widehat{v}_q}\right]^{\ddagger} \widehat{\boldsymbol{\Phi}}_{y_q y_q}(t), \tag{5.23}$$

where $(\cdot)^{\ddagger}$ is the Moore-Penrose inverse.

This estimation may cause some power coefficients to be negative due to the pseudo-inverse of the mean matrix $\left[\boldsymbol{U}_s \, \boldsymbol{U}_{n,\widehat{v}_q}\right]$. The negative power coefficients make PSDs negative which is not possible. To avoid that, we normalise the matrix to set all the values to be positive such that the negative power coefficients are set to zero and the total power of the matrix is the same as the original matrix. An alternative solution in the future may be to estimate the power coefficients directly using non-negative least squares.

Reconstruction of the source

We reconstruct source-only PSD and the residual drone noise-only PSD of the current time frame using (5.19) and (5.20), respectively. Both source and residual drone noise spectrums are smoothed to reduce their temporal fluctuations. Therefore, we estimate the above PSDs recursively as

$$\Phi_{ss}(f, t) = \eta \Phi_{ss}(f, t) + (1 - \eta)\Phi_{ss}(f, t - 1), \tag{5.24}$$

and

$$\Phi_{n,\widehat{v}_q\widehat{v}_q}(f, t) = \eta \Phi_{n,\widehat{v}_q\widehat{v}_q}(f, t) + (1 - \eta)\Phi_{n,\widehat{v}_q\widehat{v}_q}(f, t - 1), \tag{5.25}$$

where η denotes the forgetting factor.

GMM dual-stage source enhancement

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

$$\widehat{s}_{1,q}(f,t) = \Big(\frac{\Phi_{ss}(f,t)}{\Phi_{ss}(f,t) + \beta\Phi_{1,\widehat{v}_q\widehat{v}_q}(f,t)}\Big)^{\gamma} y_q(f,t), \tag{5.26}$$

where β denotes the over- or underestimation factor and γ denotes the power exponent for trading off source signal distortion for noise suppression [308]. 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 [309]. Following the WF procedure in the first stage (5.26), we can filter out the $\widehat{s}_{1,q}(f,t)$ at the second stage to obtain the total framework/stage 2 output, derived as

$$\widehat{s}_{q}(f,t) = \Big(\frac{\Phi_{ss}(f,t)}{\Phi_{ss}(f,t) + \beta\Phi_{2,\widehat{v}_q\widehat{v}_q}(f,t)}\Big)^{\gamma} \widehat{s}_{1,q}(f,t). \tag{5.27}$$

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. The combination/selection of these channels for an optimal output is not considered in this thesis but is suggested for future research directions in Chapter 7.

5.4 Experimental Study

In this section, we evaluate the proposed signal enhancement framework using experimental data for the case of a single target sound source. We first review the algorithm properties and the performance measures. Secondly, we examine the performance using the experimental recordings from outdoor measurements for a bioacoustic application using birds, and the measured impulse responses from Section 4.4 for a search-and-rescue mission using human speakers. We then examine the performance on the DREGON dataset [22] and compare the performance with the baseline. Finally, we provide discussion on how our algorithm is performed in different configurations.

5.4.1 Algorithm properties

Noise statistics at the MWF

We explore drone noise correlation matrix $\mathbf{\Phi}_{vv}(f,t)$ estimation using (i) the multichannel recording on-the-fly (in Section 5.4.2), and (ii) the pre-recorded drone noise-only measurements based on non-acoustical information, especially, use of pre-recorded motor current-specific drone noise (in Section 5.4.3) and pre-recorded motor speed-specific drone noise (in Section 5.4.4).

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 drone noise from the multichannel recordings, effectively. Instead of the motor current-specific drone noise data [13, 31], we can use motor speed-specific drone noise data as in the DREGON dataset [22].

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 [94], 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 [310, 311].

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

$$\mathbf{\Phi}_{vv}(f,t) \triangleq \frac{1}{T}\sum_{t=1}^{T} E\{\mathbf{v}(f,t)\mathbf{v}^{H}(f,t)\}. \tag{5.28}$$

Similarly to (5.28), given T' frames of the multichannel recordings when both source and drone are active, we can directly estimate noisy source correlation matrix

as

$$\mathbf{\Phi}_{pp}(f,t) \triangleq \frac{1}{T'}\sum_{t=1}^{T'} E\{\mathbf{p}(f,t)\mathbf{p}^H(f,t)\}. \tag{5.29}$$

Algorithm parameters of GMM WF

In the GMM WF, the number of GMM components for the desired sound source, e.g., bird, human speaker, and the drone noise are chosen to be $K_s = 6$ and $K_{\widehat{v}_q} = 13$ for $q = 1, \ldots, Q$, respectively following a trial and error approach proposed in [312]. This selection was made as a trade-off with signal distortion. During the *training phase*, we use a clean sound source signal to model the GMMs of the target sound source PSDs. In particular, we use a clean bird signal for bird calls in Section 5.4.2, and a clean speech training dataset from the TIMIT database [313] for speech in both Section 5.4.3 and Section 5.4.4. While the drone noise is reduced significantly in the MWF output $y_q(f,t)$, for $q = 1, \ldots, Q$ (as illustrated in Fig. 5.1), there are some residual drone noise remains as well. To model the GMMs of residual drone noise PSDs, we use a signal 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 silence of the desired sound source. Here, we note that the residual drone noise for the *training phase* of the 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 recorded at 48 kHz in the MWF and down sampled to 16 kHz in the 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 $\overline{\text{SDNR}}$ levels in (5.26), 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 ($n = 2$) in (5.27) to make the less significant residual drone noise GMM mean vectors to be considerable and minimise the speech distortion introduced by the parametric WF [308]. We set η in (5.24), and (5.25) to 0.3. For the experimental setup in Section 5.4.3, we note

that two different speech datasets have been used to construct the multichannel inputs (noisy speech signals) at the MWF and to model the speech GMMs in the training phase. This means that the experimental setup in Section 5.4.3 is both speaker-independent and speech-content independent.

Performance measures

We omit the performance evaluation of the outdoor measurements for the bird calls in Section 5.4.2 due to its unavailability of the ground truth. For a comprehensive evaluation of the proposed algorithm for the human speakers in both Section 5.4.3 and Section 5.4.4, 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 [314] and the speech intelligibility by using the Short-Time Objective Intelligibility (STOI) score [315] (available at [316]). The range of PESQ score is from -0.5 to 4.5, whereas STOI score is from 0% to 100%. Note that the high values of PESQ and STOI imply that the performance is better.

The level of noise reduction achieved through the MWF and the GMM WF is evaluated through the $\overline{\text{SDNR}}$ improvement. The $\overline{\text{SDNR}}$ improvement is measured as the difference between output $\overline{\text{SDNR}}$ and the input $\overline{\text{SDNR}}$ at each step as

$$\overline{\text{SDNR}}_{\text{imp}} = \overline{\text{SDNR}}_{\text{out}} - \overline{\text{SDNR}}_{\text{in}}. \tag{5.30}$$

The definition of the average input $\overline{\text{SDNR}}$ level is found by adding the input $\overline{\text{SDNR}}$ level at each channel and averaging it by all channels.

5.4.2 Evaluations on outdoor measurements

Experimental setup and recordings

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 source signal using the proposed algorithm. We used a Phantom 3 DJI drone [317] to obtain the measurements outdoors at the Australian National University (ANU).

We used a ReSpeaker six-channel circular microphone array [318] 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 source.

The nine seconds of time domain microphone signals is transformed into the short-time Fourier domain by a 512 point STFT using a 20 ms Hanning window with 50% of overlap to calculate $\mathbf{\Phi}_{pp}(f,t)$. Whereas, $\mathbf{\Phi}_{vv}(f,t)$ is estimated using four to five seconds of multichannel recordings prior to the bird calls when the drone is stable at the hovering manoeuvre. We note that the wind noise was also included in all the recordings.

Spectrogram analysis

Figure 5.3(a) provides the spectrogram of the clean bird call that is used for PSD modelling 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. Figures 5.3(b) and (c) show the spectrograms of the input and output signals at the MWF for the channel 3. We observe a great reduction in drone noise, while the wind noise stayed in the MWF output (around the $2-4$ seconds). We note that the clean bird call in Fig. 5.3(a) used in the *training phase* at the GMM WF is not time aligned with the input at the MWF in Fig. 5.3(b). The enhanced output signals at the first and the second stage of the GMM WF are illustrated in Figs. 5.3(d) and (e), respectively. Once more, we observe a considerable reduction in drone noise. This suggests that the enhanced signal obtained by (5.27) when $n = 2$, is expected to offer good intelligibility for manually identifying the bird species for expertise on bioacoustic[4]. We share a link to the audio files in Fig 5.3.[5]

[4]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.

[5]https://github.com/wnilmini/DroneAudition/tree/main/Outdoor%20Measurements

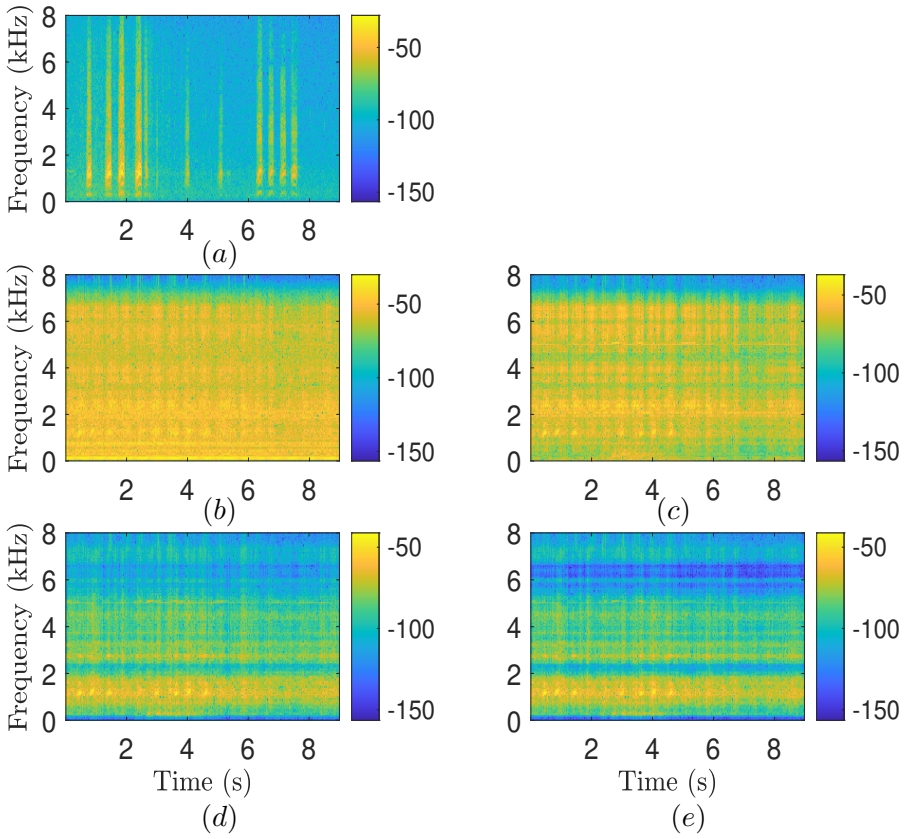


Figure 5.3: Spectrogram of (a) the clean bird call, (b) the mixture at channel 3, (c) the output signal of the MWF at channel 3, (d) stage I, and (e) stage II output signal of the GMM WF at channel 3. We note that the clean bird call in (a) used in the *training phase* at the GMM WF is not time aligned with the input at the MWF as shown in (b).

5.4.3 Evaluations on measured impulse responses

Experimental setup and recordings

The performance of the proposed method is evaluated using the experimental setup described in the previous chapter (Section 4.4). We generate the received noisy speech signal with the impulse response measurements convolved with the speech signal from the WSJCAM0 database [297] to simulate the sound source signal received at the microphones, and add it together with the current-specific drone

noise measurements. We adjust the $\overline{\text{SDNR}}$ level of the noisy speech recordings by rescaling the volume of the speech to simulate the real-world scenario which is equivalent to having the speech sources further away from the drone. We simulate our application scenario within the $\overline{\text{SDNR}}$ level range from -30 dB to -10 dB, with an increment of 5 dB over different motor current ratings. All signals are transformed to the frequency domain by a 512 point STFT using a 20 ms Hanning window with zero padding and 50% of overlap.

Noise reduction performance with MWF

We present the performance of the MWF for multichannel noise reduction. Figures 5.4(a),(b),(c) and (d) show the time varying spectrograms of the clean speech, drone noise, the mixture signal, and the output signal at the MWF of $q = 11$ microphone for average input $\overline{\text{SDNR}}$ level of -30 dB, respectively. We observe more prominent low and higher order harmonics in the drone noise spectrogram for the motor current rating of 1020 mA (as shown in Fig. 5.4 (b)). Specifically, three segments of frequency range with high power for the above motor current rating. The mixture signal of $q = 11$ microphone in Fig. 5.4 (c) has a $\overline{\text{SDNR}}$ level of -24 dB and still contains the harmonics structure of the drone noise. According to the Fig. 5.4 (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 (we confirm this by providing an analysis of the preservation of spatial cues at the MWF in Appendix A), while a fewer amount of drone noise still remains in the three frequency segments that were observed for the given motor current. The output signal at the MWF has a $\overline{\text{SDNR}}$ improvement of around 21 dB and improves the output $\overline{\text{SDNR}}$ level 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 the PESQ score of 0.72 and the STOI score of 20.8%. This proves that the output spectrum is denoised at the MWF at average input $\overline{\text{SDNR}}$ level of -30 dB and suggests the MWF achieves better performance at the very low $\overline{\text{SDNR}}$ conditions.

To further understand the drone noise reduction performance at the MWF, we plot the average $\overline{\text{SDNR}}$ improvement, PESQ, and STOI scores over average

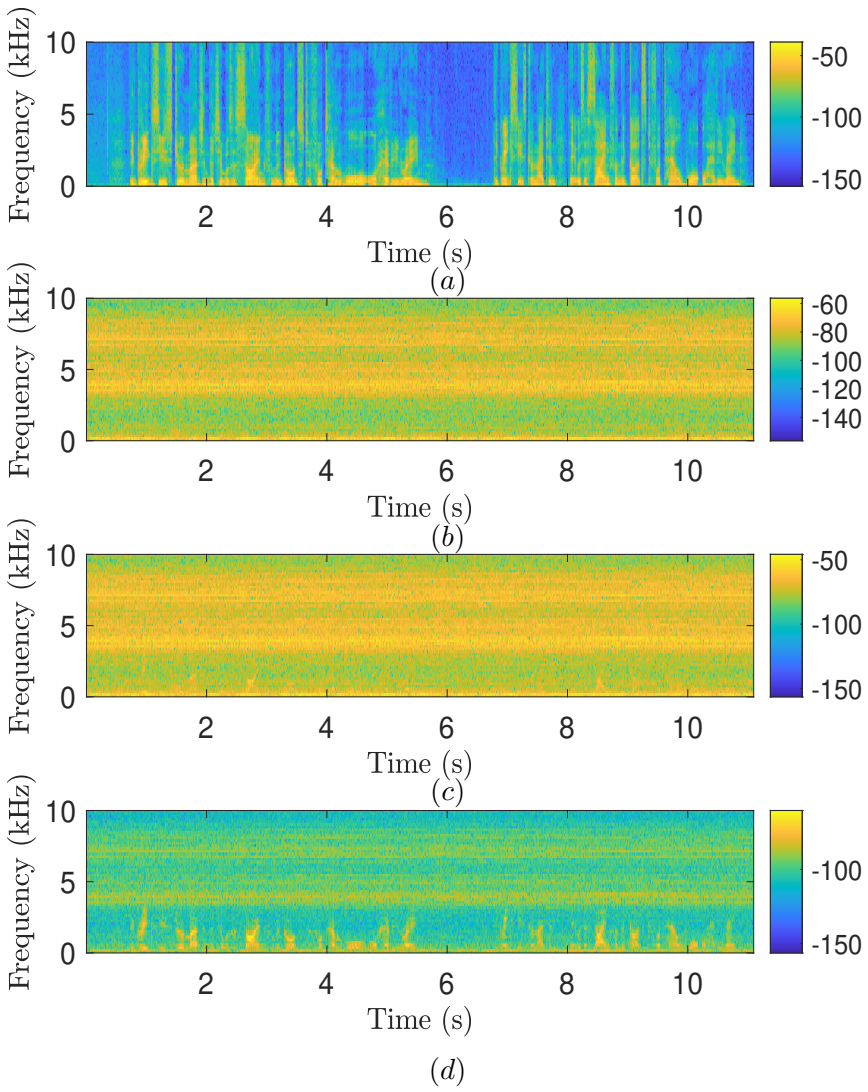


Figure 5.4: Spectrogram of (a) the clean speech, (b) the drone noise at channel 11 for motor current of 1020 mA, (c) the mixture at channel 11, and (d) the output signal of the MWF at channel 11. Here the average input $\overline{\text{SDNR}}$ level is at -30 dB whereas the input $\overline{\text{SDNR}}$ level at channel 11 is -24 dB.

input $\overline{\text{SDNR}}$ levels of -30 dB to -10 dB in Fig. 5.5. The average maximum noise reduction of approximately 25 dB is reached at the average input $\overline{\text{SDNR}}$ level of -30 dB. A continuous decrease in the noise reduction performance is observed when

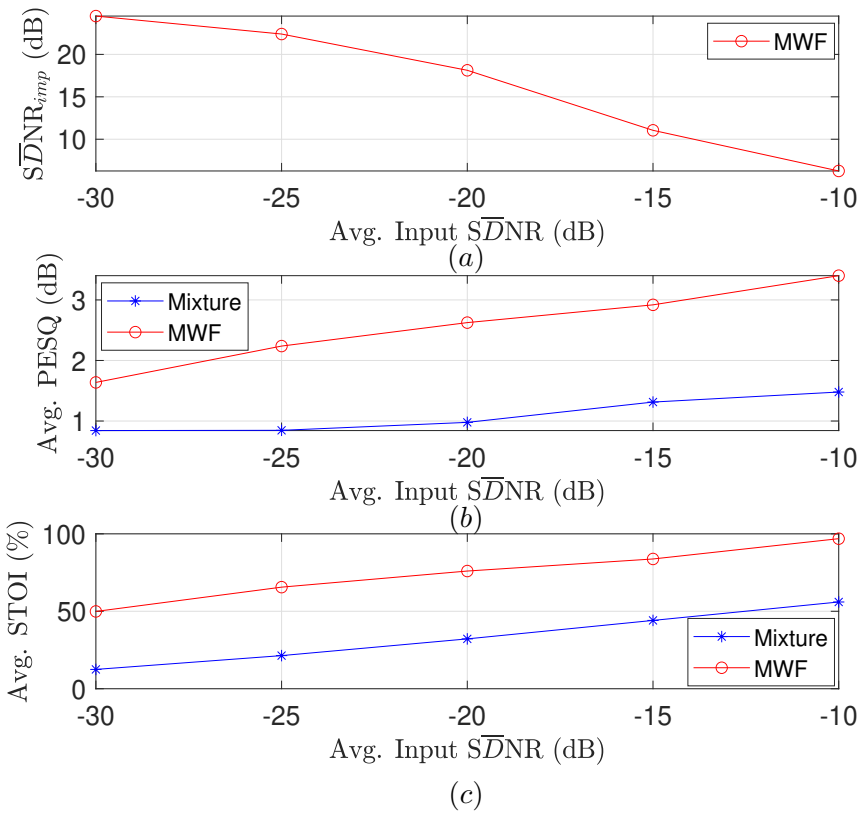


Figure 5.5: Performance evaluation of the proposed framework at the MWF. (a) Average $S\overline{D}NR$ improvement, (b) average PESQ scores, and (c) average STOI scores (in percent).

average input $S\overline{D}NR$ level 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 $S\overline{D}NR$ and speech quality without degrading intelligibility. MWF was still observed to perform well in high $S\overline{D}NR$ conditions (> -10 dB). However, in this chapter, we only focus on $S\overline{D}NR$ levels below -10 dB.

Signal enhancement performance with GMM WF

Figure 5.6 shows the time varying spectrogram of the input and outputs at the GMM WF of $q = 11$ microphone. We 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 through the staging process. Figure 5.7 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 at the two stages. Later, in the *enhancement phase*, this helps to obtain a more accurate estimate of the residual drone noise PSD.

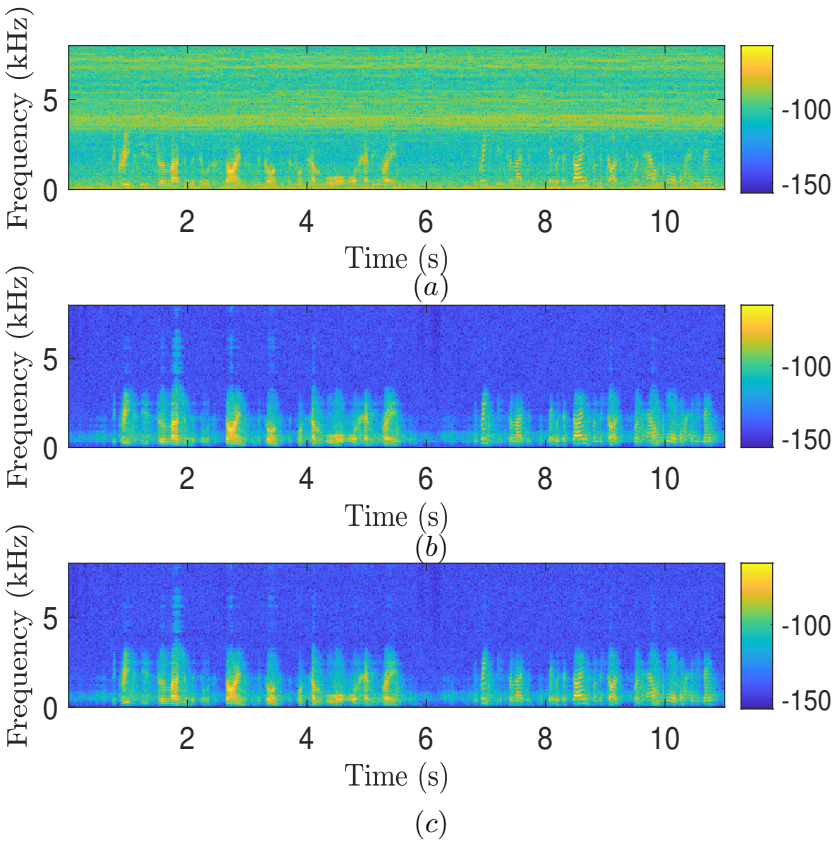


Figure 5.6: Spectrogram of (a) the MWF output, (b) stage I, and (c) stage II of the GMM WF at channel 11.

Based on both Fig. 5.6 and 5.7, 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 4kHz. In the first stage, by selecting the high energy GMM mean values of the residual drone noise, we mainly cancel out the residual drone noise in the frequency

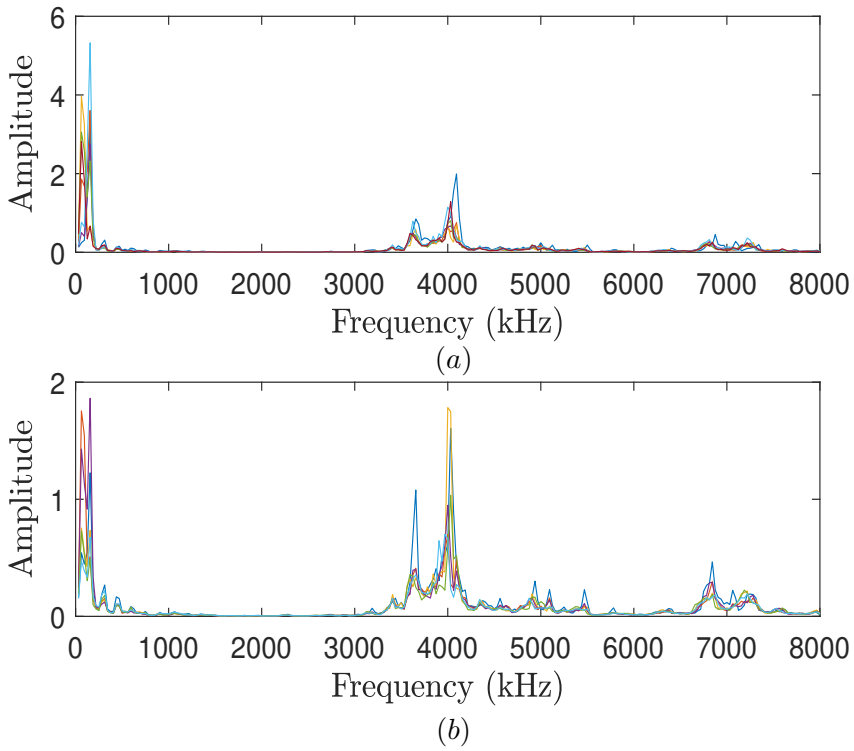


Figure 5.7: GMM mean vector plots of residual drone noise for $K_{\widehat{v}_q} = 13$ for $q = 11$. The selection of the set of mean vectors $K_{n,\widehat{v}_q}$ sets to (a) 7 for $n = 1$, and (b) 6 for $n = 2$.

range of 0 to 300 Hz as shown in Fig. 5.6 (b). Whereas in the second stage, the residual drone noise in both frequency regions is suppressed by selecting the low energy residual drone noise GMM mean values (as shown in Fig. 5.6 (c)). Overall, the enhanced signal has PESQ of 2.2 and STOI of 64.7% scores. We share a link to the audio files in Fig 5.6.[6] We note that the speech enhancement using drone-embedded 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 the heavy drone noise should result in a lower intelligibility.

Our drone-embedded microphone array is arbitrarily distributed around the drone body. We place three dual-microphone modules closer to each motor. This

[6] https://github.com/wnilmini/DroneAudition/tree/main/Measured%20IRs

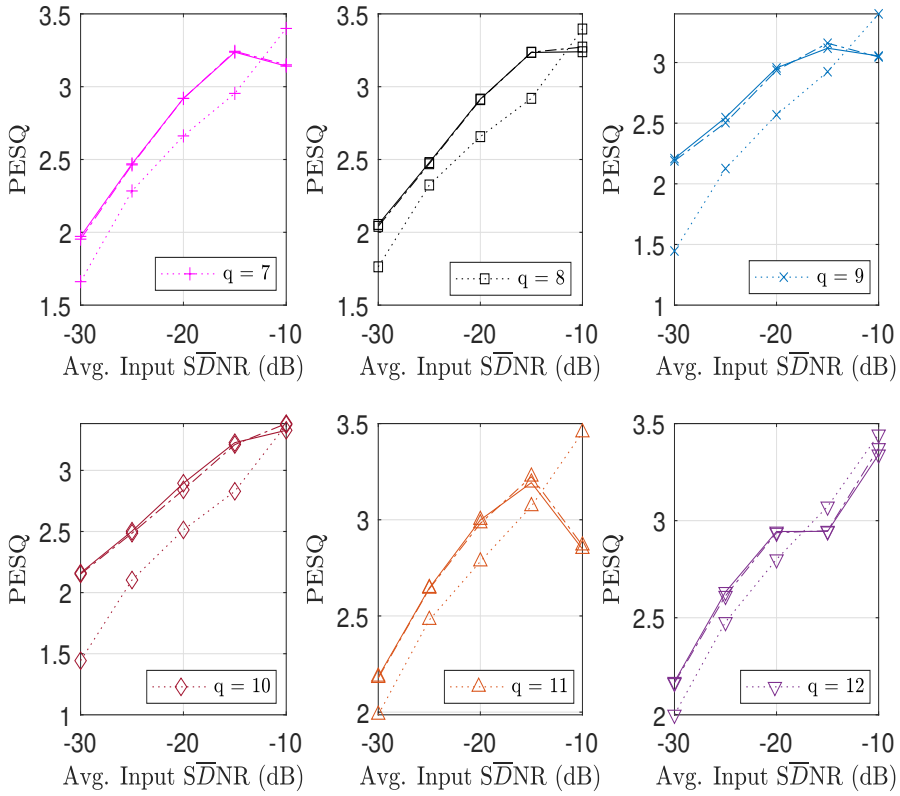


Figure 5.8: PESQ scores at channel q where $q \in [7, 12]$ with very low average input S$\overline{\text{D}}$NR levels. Dotted line represents the input at the GMM WF (output from the MWF) whereas the dash-dotted and solid lines represent the output at the GMM WF at stage 1 and 2, respectively.

indicates that we can take six microphone channels around a single motor to analyse the drone noise reduction and signal enhancement performance. We use the microphone channels of $q \in [7, 12]$ to analyse the speech enhancement performance at the GMM WF in terms of perceptual quality and intelligibility scores. Figures 5.8 and 5.9 illustrate the PESQ and STOI results, respectively. The PESQ scores, shown in Fig. 5.8, show a better improvement to the MWF processed speech over the average input S$\overline{\text{D}}$NR levels ranging from -30 dB to -15 dB except channel 12. However, the PESQ scores decrease compared to the MWF output at the S$\overline{\text{D}}$NR level of -10 dB except for the channel 10. Figure 5.9 indicates that all six channels ($q \in [7, 12]$) at extremely low average input S$\overline{\text{D}}$NR levels, e.g., < -15

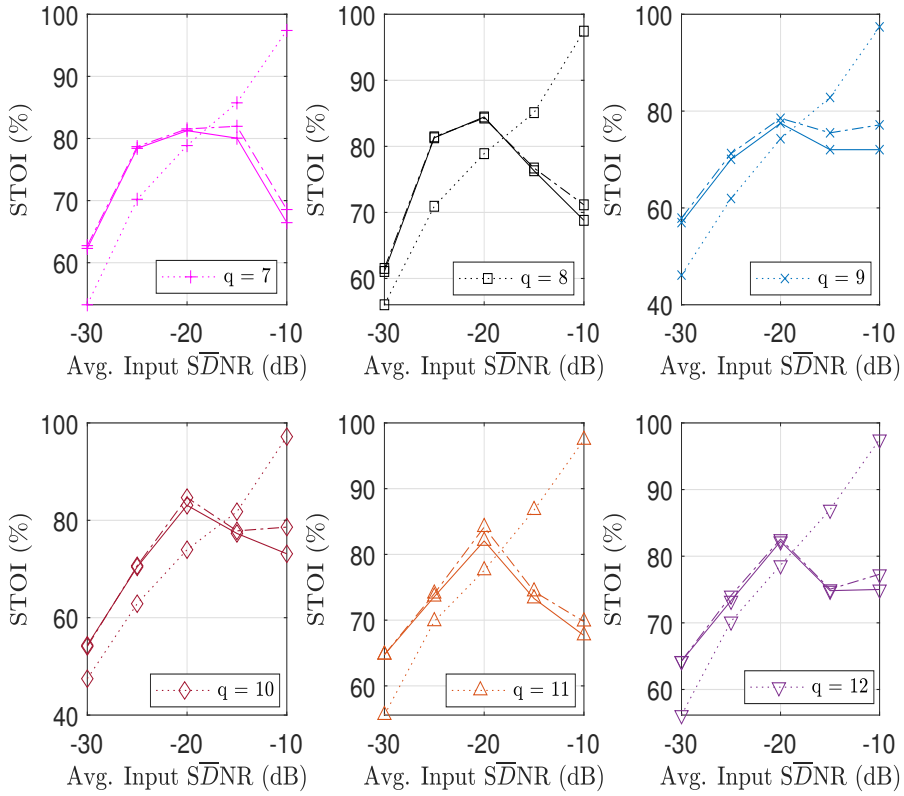


Figure 5.9: STOI scores (in percent) at channel q where $q \in [7, 12]$ with very low average input S$\overline{\text{D}}$NR levels. Dotted line represents the input at the GMM WF (output from the MWF) whereas the dash-dotted and solid lines represent the output at the GMM WF in stage 1 and 2, respectively.

dB, the total framework output achieves higher STOI scores with respect to the processed speech at the MWF. However, STOI scores begin to decrease from -15 dB for all the channels compared to the output at the MWF.

Figure 5.10 shows the S$\overline{\text{D}}$NR improvement at the GMM WF of $q =\in [7, 12]$ microphones. We observe that the S$\overline{\text{D}}$NR improvement is not very significant until -20 dB. All six channels have a similar S$\overline{\text{D}}$NR improvement with the increase of the average input S$\overline{\text{D}}$NR from -15 dB. We noticed earlier that the enhanced speech at each channel has lower perceptual quality and intelligibility scores compared to the MWF output at the low S$\overline{\text{D}}$NR levels (> -15 dB). In general, the parametric WF introduces speech distortion at the cost of noise reduction. However, for those

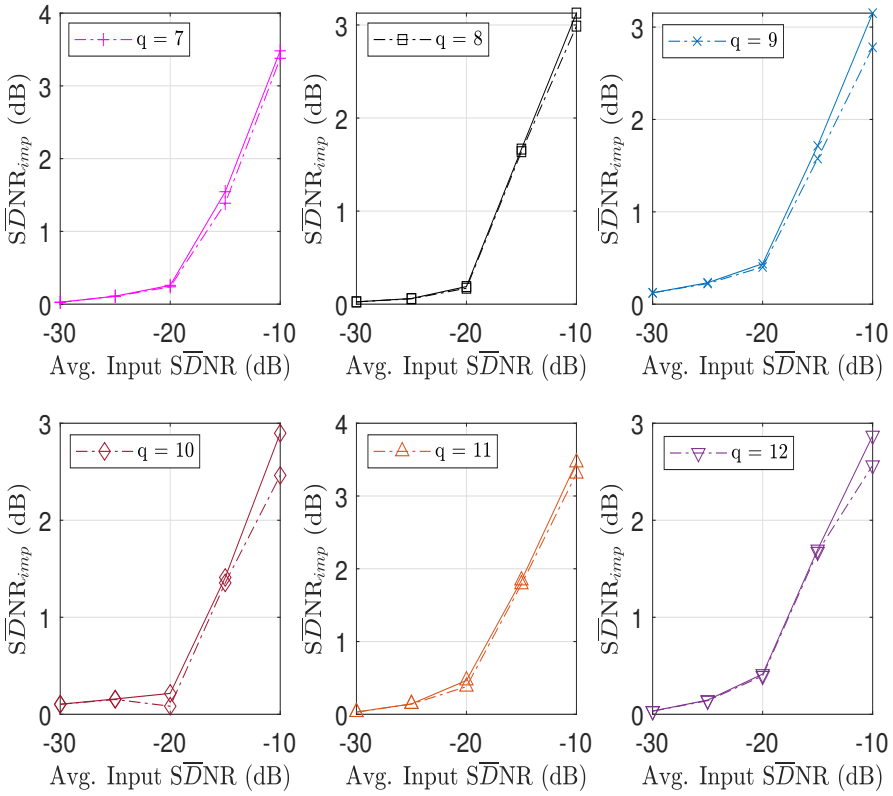


Figure 5.10: $S\overline{D}NR$ improvement in the GMM WF at channel q where $q \in [7, 12]$ with very low average input $S\overline{D}NR$ levels. Dash-dotted and solid lines represent the output at the GMM WF in stage 1 and 2, respectively.

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

5.4.4 Evaluations on DREGON dataset

Experimental setup and recordings

We analyse the performance of the proposed framework on the DREGON dataset [22]. The DREGON dataset has a constellation of eight 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 use 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 $\overline{\text{SDNR}}$ 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').

Signal enhancement performance

Figure 5.11 illustrates the results at the MWF. The DREGON dataset configuration achieves an average maximum noise reduction of nearly 13 dB at average input $\overline{\text{SDNR}}$ level of -30 dB. We can observe that both datasets display the same trend in noise reduction with increasing $\overline{\text{SDNR}}$ level. However, the $\overline{\text{SDNR}}$ improvement over very low $\overline{\text{SDNR}}$ conditions is less compared to our dataset [13]. Moreover, the PESQ and STOI score improvement over the extreme $\overline{\text{SDNR}}$ 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 thirty embedded microphones arbitrarily distributed on the drone structure. It is clear that at the MWF, using a higher number of microphones on the drone and placing them around the drone body will capture more propagation paths and add more diversity to the framework. This suggests that a higher number of drone-embedded microphones will result in more noise suppression at the MWF.

Figure 5.12 shows the speech enhancement performance of the DREGON dataset over the microphone q where $q \in (1, 2)$. Both stages of the 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 $\overline{\text{SDNR}}$ level of -10 dB, we observe

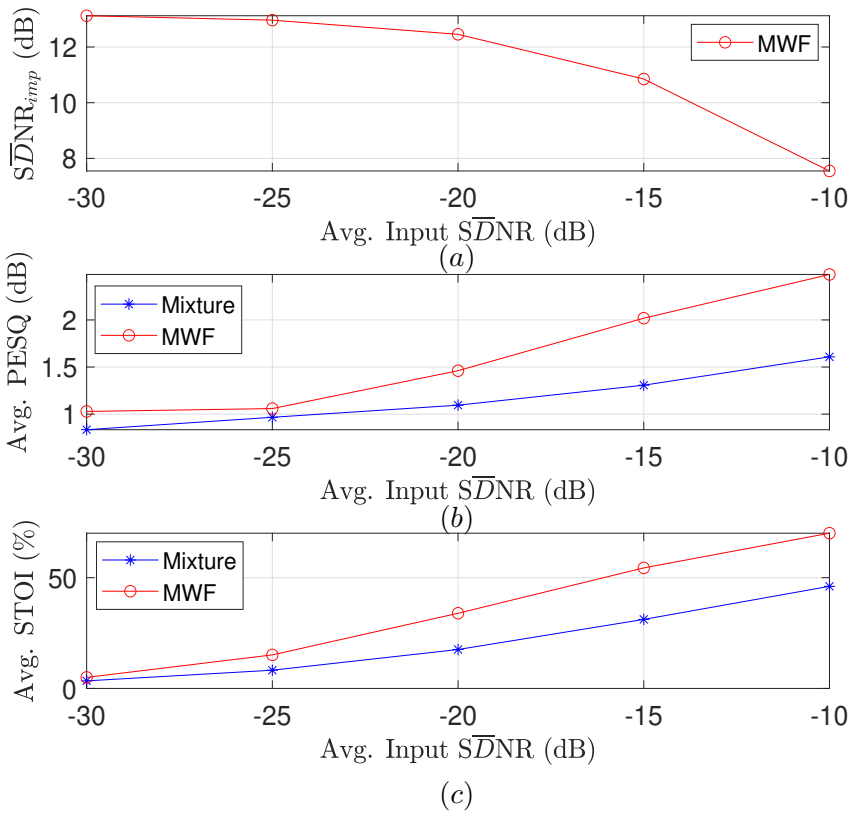


Figure 5.11: Performance of the MWF outputs for the DREGON dataset. (a) Average $\overline{\text{SDNR}}$ improvement, (b) average PESQ scores, and (c) average STOI scores (in percent).

that the enhanced signal has slightly lower PESQ and STOI scores compared to the output signal at the MWF. According to the Fig. 5.12, we observe that the $\overline{\text{SDNR}}$ improvement has reached a peak at -20 dB and decreased with the increase of the $\overline{\text{SDNR}}$ levels for both channels. It is noticed that our experimental setup in Section 5.4.3 achieves greater performance in terms of speech quality and intelligibility compared to the DREGON dataset configurations. 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

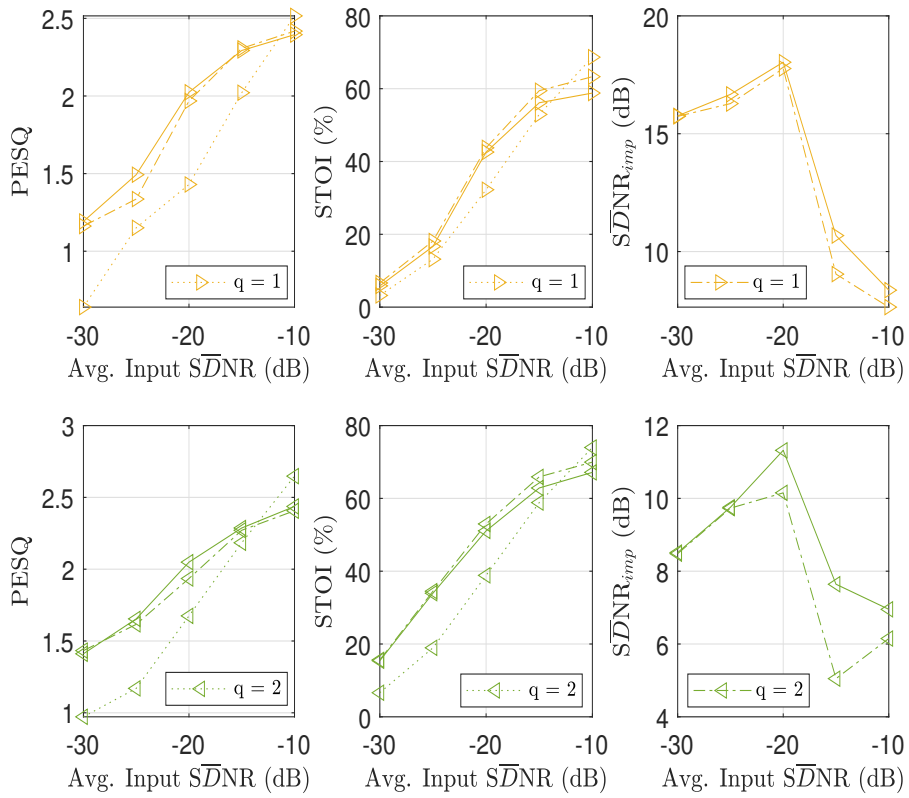


Figure 5.12: PESQ, STOI (in percent) scores and $S\overline{D}NR$ improvement at channel q where $q \in [1, 2]$ with very low average input $S\overline{D}NR$ levels. Dotted line represents the input at the GMM WF whereas the dash-dotted and solid lines represent the output at the GMM WF in stages 1 and 2, respectively.

both MWF and the 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.

5.4.5 Comparison with baseline

We now compare two baseline methods: MVDR [208] and MWF found in [22] for the average input $S\overline{D}NR$ of -30 dB at different microphone channels of the array for our drone acoustic dataset. Figures 5.13(a-c) show the spectrograms of

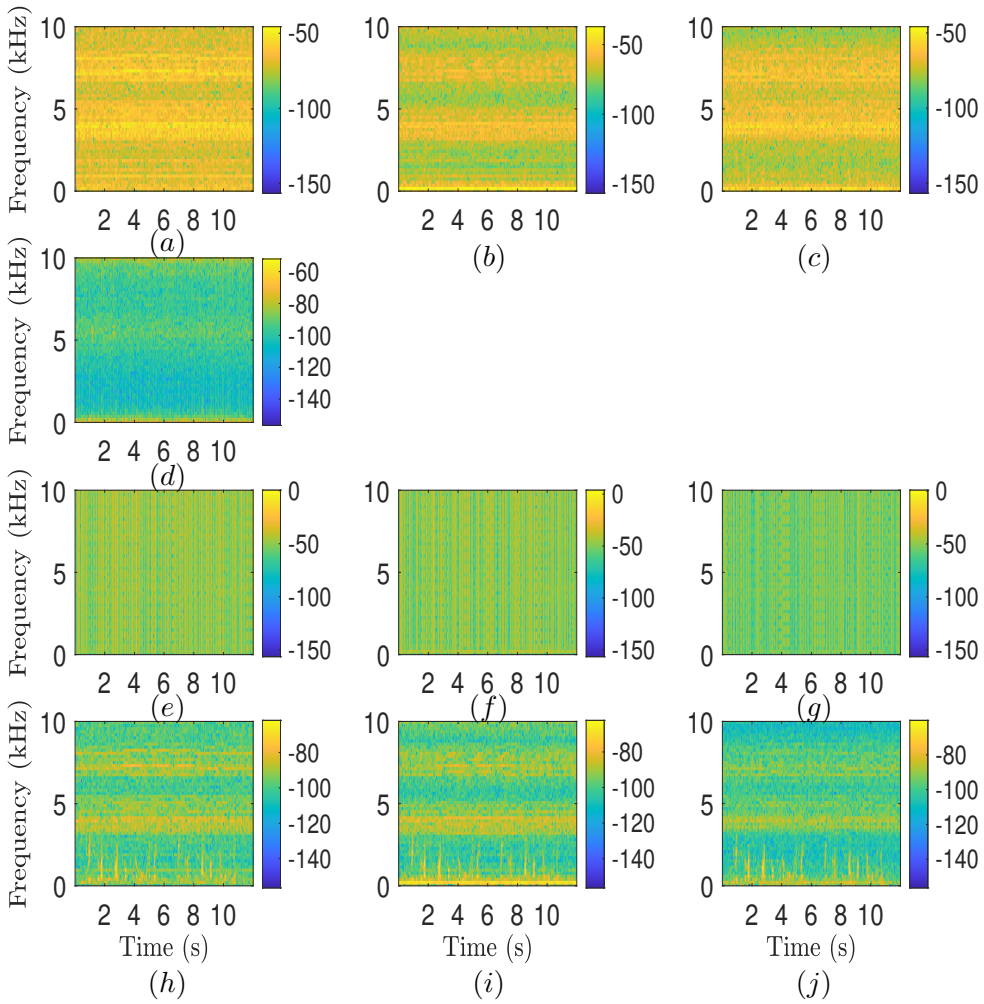


Figure 5.13: Comparison with baseline methods at average input $\overline{\text{SDNR}}$ of -30 dB: Spectrograms (a-c) the mixture for channel $q \in \{7, 9, 11\}$, (d) the MVDR output, (e-g) the output at MWF method from typical MWF for channel $q \in \{7, 9, 11\}$, and (h-j) the output at MWF derived in the proposed method for channel $q \in \{7, 9, 11\}$.

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. 5.13(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. 5.13(c)).

Figure 5.13(d) represents 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 (5.6) and (5.11) as in [22]

$$\mathbf{W}^{(1)}(f) = \mathbf{\Phi}_{pp}^{-1}(f,t)(\mathbf{\Phi}_{pp}(f,t) - \mathbf{\Phi}_{vv}(f,t)). \tag{5.31}$$

Figures 5.13(e-g) illustrate the output spectrograms at the MWF in [22] 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 [22] are not satisfactory. Figures 5.13(h-j) present the MWF in (5.16) 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 MWF implementation 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 (5.16) and (5.31), respectively. This suggests that the optimal filter weights in (5.16) suppress more drone noise from the multichannel recordings.

5.4.6 Discussion

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

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 arbitrarily distributed microphone array on the drone with a large number of microphones ($q = 30$) than a standard microphone array. This may be due to utilising 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.

Comparison with baselines

The output of the baseline methods such as MVDR [208] and MWF found in [22] 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 the noise correlation matrix. We showed audio signal enhancement performance for a hovering drone. The use of recursively updated drone noise, in contrast, to using pre-recorded drone noise data for estimating $\mathbf{\Phi}_{vv}(f,t)$ has shown much less satisfactory results as it is 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 $\overline{\text{SDNR}}$ 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 without the impulse response measurements.

5.5 Interfering Source Removal Process using Improved GMM WF

We extend our proposed algorithm in Section 5.3 to address more practical scenarios in the presence of interfering sound sources such as (i) traffic, and (ii) wind noise. Here, our objective is to suppress both the drone noise and interfering sources. In this section, we first reformulate the problem of the target source enhancement, then modify both MWF and GMM WF to suppress the introduced interfering noise, and finally conduct the experimental evaluation of the proposed framework.

5.5.1 Problem reformulation

Our goal is to enhance the sound source components of the target source from the multichannel recordings. Consider an acoustic scenario comprising L sound sources on the ground and a drone with Q microphones embedded to its structure, hovering at a constant height above the ground. The received noisy source signal in the STFT domain, at the q^{th} microphone, is given by

$$p_q(f,t) = \sum_{\ell=1}^{L} d_{\ell,q}(f) s_\ell(f,t) + v_q(f,t), \tag{5.32}$$

where $d_{\ell,q}(f)$ is the acoustic transfer function from the ℓ^{th} source to the q^{th} microphone, and $s_\ell(f,t)$ is the source signal. Assume a single sound source and multiple interfering sources. We define the STFT of the (i) sound source content as $x_q(f,t) = d_{1,q}s_1(f,t)$, and (ii) interfering source content as $u_q(f,t) = d_{\ell,q}s_\ell(f,t)$, where $\ell = \{2,\ldots,L\}$, at the q^{th} channel. The received signal Q dimensional stacked vector can be written in the same form as in (5.1) with

$$\mathbf{p}(f,t) = \mathbf{x}(f,t) + \mathbf{u}(f,t) + \mathbf{v}(f,t), \tag{5.33}$$

where $\mathbf{u}(f,t) \triangleq [u_1(f,t), \cdots, u_Q(f,t)]^T$.

Additionally, we define the interfering noise correlation matrix as

$$\mathbf{\Phi}_{uu}(f,t) \triangleq E\{\mathbf{u}(f,t)\mathbf{u}^H(f,t)\}. \tag{5.34}$$

In general, we assume the desired sound source, interfering noise, and drone noise signals are mutually uncorrelated. Therefore, using (5.33) similar to (5.6), we write

$$\mathbf{\Phi}_{pp}(f,t) = \mathbf{\Phi}_{xx}(f,t) + \mathbf{\Phi}_{uu}(f,t) + \mathbf{\Phi}_{vv}(f,t). \tag{5.35}$$

5.5.2 Source enhancement using the improved GMM WF

In this subsection, we first modify the MWF in Section 5.3.1 to remove interfering noise from each microphone channel. We then further reduce the residual interfering noise using an improved version of the GMM WF by incorporating the interfering source model into the GMM WF in Section 5.3.2.

Drone noise reduction using MWF

We rewrite (5.16) with interfering source content in (5.33) as

$$\mathbf{W}(f) = \frac{(\mathbf{\Phi}_{uu}(f,t) + \mathbf{\Phi}_{vv}(f,t))^{-1}\mathbf{\Phi}_{pp}(f,t) - \mathbf{I}_Q}{1 + tr[(\mathbf{\Phi}_{uu}(f,t) + \mathbf{\Phi}_{vv}(f,t))^{-1}\mathbf{\Phi}_{pp}(f,t)] - Q}. \tag{5.36}$$

The derivation of the optimal weight matrix of the MWF considering the interfering noise sources is similar to Section 5.3.1. The q^{th} column of $\mathbf{W}(f)$ matrix equals to the q^{th} output of the MWF for $q = 1, \ldots, Q$. Then, we can write

$$\mathbf{w}_q(f) = \frac{(\mathbf{\Phi}_{uu}(f,t) + \mathbf{\Phi}_{vv}(f,t))^{-1}\mathbf{\Phi}_{pp}(f,t) - \mathbf{I}_Q}{1 + tr[(\mathbf{\Phi}_{uu}(f,t) + \mathbf{\Phi}_{vv}(f,t))^{-1}\mathbf{\Phi}_{pp}(f,t)] - Q}\mathbf{e}_q, \tag{5.37}$$

where $\mathbf{e}_q = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ is an Q dimensional vector, in which the q^{th} index equals to one.

The second order statistics of the interfering noise sources $\mathbf{\Phi}_{uu}(f,t)$, and drone noise $\mathbf{\Phi}_{vv}$ can be calculated together from a few seconds of multichannel recording when the drone hovers in a stable flight and the interfering source is present prior to the target sound source.

Signal enhancement using improved GMM WF

We extend the GMM WF in Section 5.3.2 to model the residual interfering noise from the q^{th} output of MWF using an improved GMM WF approach. We sum-

marise the additional steps of the improved GMM WF below.

(i) Let $T_{\widehat{u}_q}(\Phi_{\widehat{u}_q\widehat{u}_q}; f)$ be the PDF of $\Phi_{\widehat{u}_q\widehat{u}_q}(f,t)$, we then express

$$T_{\widehat{u}_q}(\Phi_{\widehat{u}_q\widehat{u}_q}; f) = \sum_{k=1}^{K_{\widehat{u}_q}} C_{\widehat{u}_q k}(f)\mathcal{N}(\mu_{\widehat{u}_q k}(f), \Sigma_{\widehat{u}_q k}(f)), \tag{5.38}$$

where $K_{\widehat{u}_q}$, $C_{\widehat{u}_q k}(f)$, $\mu_{\widehat{u}_q k}(f)$ and $\Sigma_{\widehat{u}_q k}(f)$ denote the number of GMM components, the contribution, the mean value, and the variance of k^{th} GMM component at the f^{th} frequency bin for the residual interfering noise at the q^{th} output of the MWF.

(ii) Similar to (5.19) and (5.20), the interfering source PSD can be written as

$$\Phi_{n,\widehat{u}_q\widehat{u}_q}(f,t) = \boldsymbol{\mu}^T_{n,\widehat{u}_q}(f)\,\boldsymbol{\alpha}_{n,\widehat{u}_q}(t), \tag{5.39}$$

where $\boldsymbol{\mu}_{n,\widehat{u}_q}(f) = [\mu_{n,\widehat{u}_q 1}(f), \ldots, \mu_{n,\widehat{u}_q K_{n,\widehat{u}_q}}(f)]^T$ and $\boldsymbol{\alpha}_{n,\widehat{u}_q}(t) = [\alpha_{n,\widehat{u}_q 1}(t), \ldots, \alpha_{n,\widehat{u}_q K_{n,\widehat{u}_q}}(t)]^T$.

(iii) Assuming the source signal, residual interfering source, and residual drone noise are uncorrelated, we obtain the PSD of $y_q(f,t)$ similar to (5.21) as

$$\widehat{\Phi}_{y_q y_q}(f,t) = \boldsymbol{\mu}^T_s(f)\,\boldsymbol{\alpha}_s(t) + \boldsymbol{\mu}^T_{n,\widehat{v}_q}(f)\,\boldsymbol{\alpha}_{n,\widehat{v}_q}(t) + \boldsymbol{\mu}^T_{n,\widehat{u}_q}(f)\,\boldsymbol{\alpha}_{n,\widehat{u}_q}(t). \tag{5.40}$$

(iv) By writing (5.40) 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 similar to (5.23) as

$$\begin{bmatrix} \boldsymbol{\alpha}_s(t) \\ \boldsymbol{\alpha}_{n,\widehat{v}_q}(t) \\ \boldsymbol{\alpha}_{n,\widehat{u}_q}(t) \end{bmatrix} = \left[\boldsymbol{\Lambda}_s\, \boldsymbol{\Lambda}_{n,\widehat{v}_q}\, \boldsymbol{\Lambda}_{n,\widehat{u}_q}\right]^{\ddagger} \widehat{\boldsymbol{\Phi}}_{y_q y_q}(t), \tag{5.41}$$

where $\boldsymbol{\Lambda}_{n,\widehat{u}_q} = [\boldsymbol{\mu}^T_{n,\widehat{u}_q 1}(f_1), \ldots, \boldsymbol{\mu}^T_{n,\widehat{u}_q K^{(n)}_{\widehat{u}}}(f_F)]$.

(v) Reconstruct the residual interfering noise-only PSDs of the current time frame using (5.39) by updating it recursively.

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

$$\widehat{s}_{1,q}(f,t) = \Big(\frac{\Phi_{ss}(f,t)}{\Phi_{ss}(f,t) + \beta(\Phi_{1,\widehat{v}_q\widehat{v}_q}(f,t) + \Phi_{1,\widehat{u}_q\widehat{u}_q}(f,t))}\Big)^{\gamma} y_q(f,t), \tag{5.42}$$

and the total framework output as

$$\widehat{s}_{2,q}(f,t) = \Big(\frac{\Phi_{ss}(f,t)}{\Phi_{ss}(f,t) + \beta(\Phi_{2,\widehat{v}_q\widehat{v}_q}(f,t) + \Phi_{2,\widehat{u}_q\widehat{u}_q}(f,t))}\Big)^{\gamma} \widehat{s}_{1,q}(f,t). \tag{5.43}$$

5.5.3 Experimental verification

We evaluate the performance of the interfering source removal process using both outdoor measurements and measured impulse responses.

Practical analysis

We examine the performance using the experimental setup in Section 5.4.2. Outdoor recordings of two sound source signals using drone on-board microphones were considered: these are target bird calls and wind noise. Here, we model the wind noise as the interfering source in (5.32) using the wind noise recordings from the Zenodo dataset [319] with $K_{\widehat{u}_q} = 9$.

Figures 5.14(a) and (b) show the spectrograms of the input and output signals of the MWF at $q = 3$ microphone. We observed a great reduction in the drone noise, while the wind noise stayed in the MWF output with spectral energy below 1 kHz between $0 - 10$ seconds as shown in Fig. 5.14(b). The enhanced output signals at the first and second stages of the improved GMM WF are illustrated in Figs. 5.14(c) and (d), respectively. We observe a considerable reduction in drone noise and additionally, a suppression of wind noise at both stages. This suggests that the improved GMM WF removes the interfering source effectively compared to the modified MWF in (5.36). The audio files in Fig 5.14 can be found.[7]

[7] https://github.com/wnilmini/iGMM-WF/tree/main/Outdoor%20Measurements

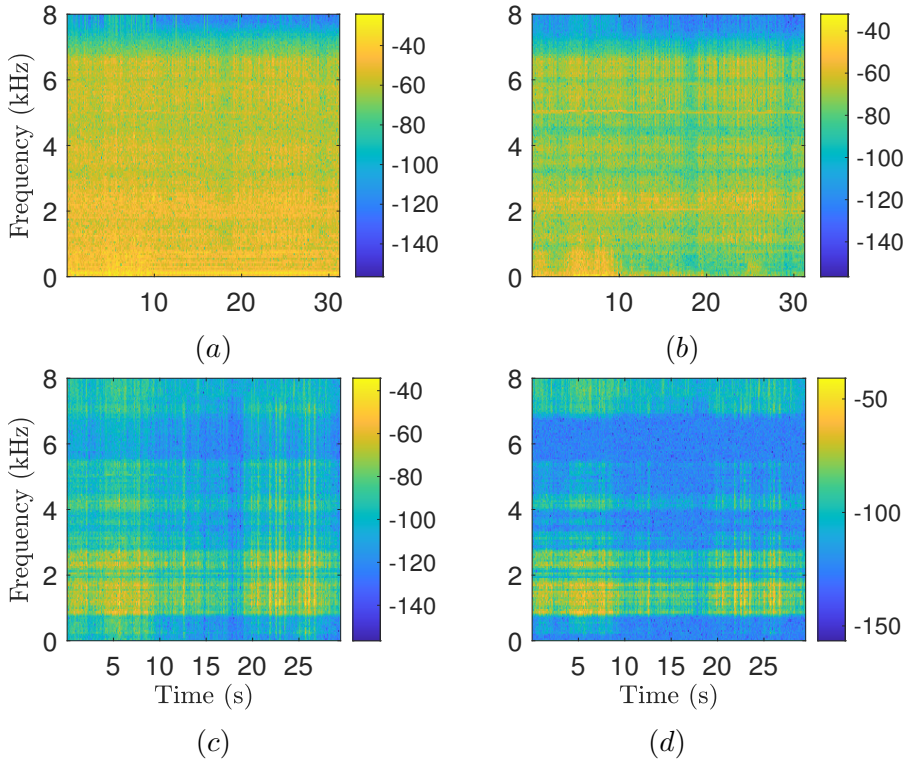


Figure 5.14: Spectrogram plots at $q = 3$ microphone for the outdoor measurements: (a) input at the MWF, (b) output at the MWF, (c) stage one, and (d) stage two output at the improved GMM WF.

Simulation analysis

We simulate a scenario of a desired speech source and two interfering noise sources ($L = 3$) located on the ground at an azimuth angle of 0°, 90°, and −90°, respectively, with a hovering drone above using the experimental recordings in Section 5.4.3. We use traffic noise as the interfering source from the MAVD-traffic dataset [320] with $K_{\widehat{u}_q} = 9$. We generate multichannel recordings with signal-to-interfering noise ratio (SINR) level of −10 dB, −5 dB, and 0 dB, and the input $\overline{\text{SDNR}}$ level range from −30 dB to −10 dB, with an increment of 5 dB.

Figure 5.15 represents the time-varying spectrograms of the input, and the output signals at different stages of the modified framework for the average input $\overline{\text{SDNR}}$ level of −30 dB and input SINR of −10 dB at $q = 11$ microphone. From

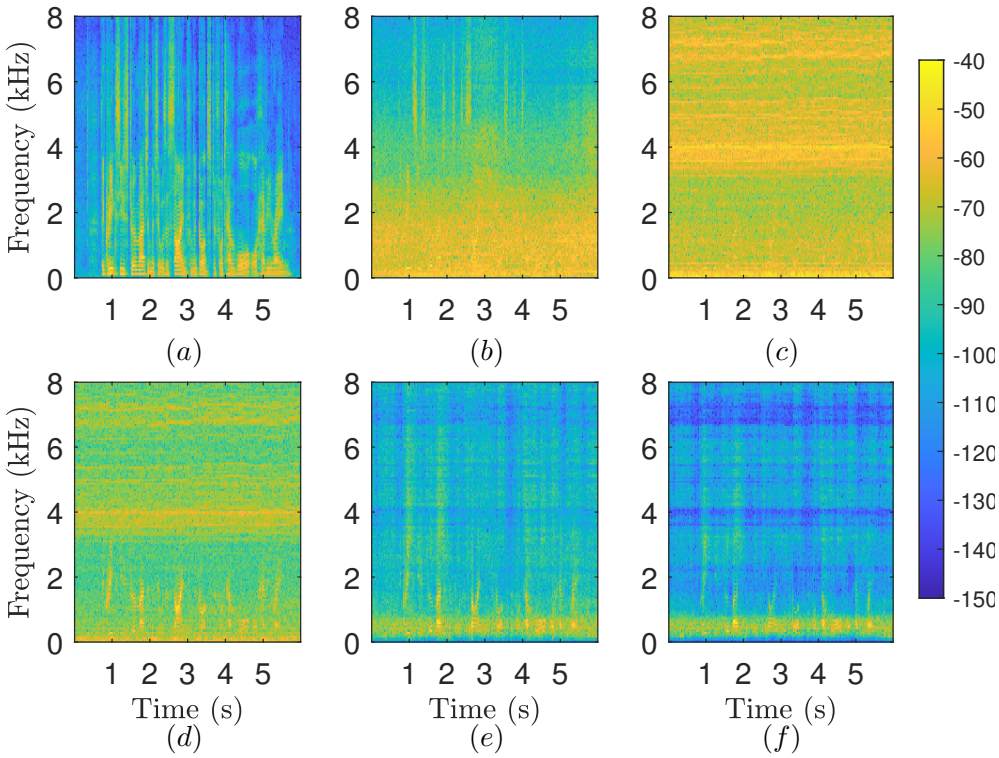


Figure 5.15: Spectrogram plots at $q = 11$ microphone: (a) the clean speech, (b) speech and traffic noise only, (c) input at the MWF, (d) output at the MWF, (e) stage one, and (f) stage two output at the improved GMM WF.

Fig. 5.15(b), we observe that traffic noise contains high energy in frequencies below 2 kHz. It is clear that more prominent low and higher order drone noise harmonics are presented in the mixture spectrogram as shown in Fig. 5.15(c). Figure 5.15(d) shows that the large residual noise remains with respect to the frequency range of the drone noise and traffic noise in MWF output. In the post-processing stage, we further suppress this residual noise using dual-stage improved GMM WF as shown in Figs. 5.15(e) and (f). We share a link to the audio files in Fig 5.15.[8]

We now examine the performance over different average input $\overline{\text{SDNR}}$ conditions and SINR levels in Fig. 5.16. We observe a significant improvement in the performance in terms of $\overline{\text{SDNR}}$ improvement, PESQ, and STOI over -10 dB to

[8] https://github.com/wnilmini/iGMM-WF/tree/main/Measured%20IRs

−30 dB $\overline{\text{SDNR}}$ range for SINR of 0 dB. The maximum noise reduction of approximately 24 dB is reached at the average input $\overline{\text{SDNR}}$ level of −30 dB. We find the modified algorithm has the same trend in performance degradation with decreasing the SINR and confirm the extraction of understandable speech content from the informal listening test [321].

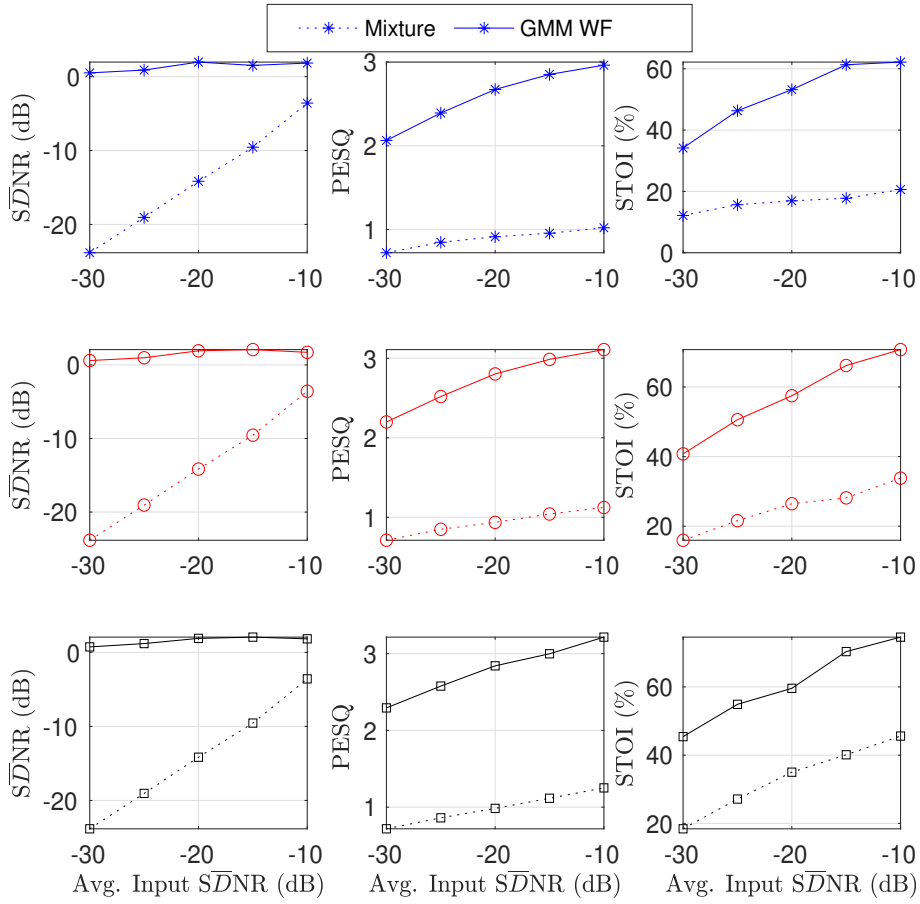


Figure 5.16: Performance evaluation of the proposed framework with the improved GMM WF at different average input $\overline{\text{SDNR}}$ levels at $q = 11$ microphone. Row 1 (in blue): Input SINR of −10 dB, Row 2 (in red): Input SINR of −5 dB, Row 3 (in black): Input SINR of 0 dB.

5.6 Summary and Contributions

This chapter presents a framework for audio signal enhancement using a drone-embedded arbitrary microphone array. We have proposed a solution using Wiener filtering techniques and conducted an extensive experimental study using both outdoor measurements and measured impulse response recordings. The algorithm is low in complexity and has a highly applicable outcome for drone applications with any microphone array geometry. The following outcomes and insights were obtained:

- We showed that the widely known MWF can be used to suppress a significant amount of drone noise from the multichannel recordings.
- We proposed a single channel dual-stage GMM WF which removes the residual drone noise and enhances the desired sound source signal.
- We achieve this drone noise suppression process through either motor current-specific, or motor speed-specific data to inform the Wiener filtering approaches.
- Through extensive experimental data, we found that the proposed framework is robust to extremely low $\overline{\text{SDNR}}$ scenarios up to -30 dB.
- Furthermore, we show that the drone noise reduction and signal enhancement performance is improved with arbitrarily distributed drone-embedded microphones compared to the drone-mounted microphone arrays.
- We have also generalised the drone noise removal process in GMM WF to incorporate the interfering sources, and showed the modified framework provides robustness to both drone noise as well as interfering sources in very low $\overline{\text{SDNR}}$ scenarios.

The multi-microphone audio signal enhancement using the framework of the MWF and the Gaussian-Mixture based post-filtering has been shown to be an effective solution for drone embedded/mounted microphone arrays in adverse noisy environments. For example, using the proposed framework, the drone on-board microphones will assist bioacoustic applications in identifying and monitoring wildlife

[322]. In the next chapter, we propose an efficient measuring and modelling of the drone acoustic transfer function approach for drone audition functions such as sound source localisation and source enhancement focusing on incorporating the scattering and diffraction response of a drone to an incident soundfield.

5.7 Related Publications

Much of this chapter's work has been published or submitted for publication in the following journal papers and conference proceedings.

- **W. N. Manamperi**, P. N. Samarasinghe, T. D. Abhayapala, and J. Zhang, "GMM based Multi-stage Wiener Filtering for Low SNR Speech Enhancement," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1-5, Bamberg, Germany, September 2022.

- **W. N. Manamperi**, T. D. Abhayapala, P. N. Samarasinghe, and J. Zhang, "Drone Audition: Improved GMM Wiener Filtering Approach for Signal Enhancement", in *INTERNOISE and NOISE-CON Congress and Conference Proceedings*, Greater Tokyo, Japan, August 2023.

- **W. N. Manamperi**, T. D. Abhayapala, P. N. Samarasinghe, and J. Zhang, "Drone Audition: Audio Signal Enhancement from drone-embedded Microphones using Multichannel Wiener Filtering and Gaussian-Mixture based Post-Filtering," in *Applied Acoustics*, vol. 216, pp. 109818, January 2024.

Chapter 6

Efficient Measurements and Parameterisation of the Drone Related Transfer Function Model

Following the recent advancements to commercial drones in industrial applications, considerable attention has been received to enabling drone audition. Microphone arrays are beginning to be embedded into the body of drones for sound source localisation and signal enhancement. In practical acoustical environments, the performance of drone audition is degraded due to the free-field sound propagation assumption around the drone body and its structure. In this chapter, we propose modelling the scattering and diffraction of a drone body beyond free-field propagation, as the Drone Related Transfer Functions (DRTFs) that are estimated with a discrete set of measurements by multiple loudspeaker placements on circular arrays. We re-define the efficient parameterisation and practically feasible measurement procedure of a continuous DRTF model based on microphone alignment using spherical harmonic expansions. The method is a practically viable solution for typical impulse response measurement setups and models the scattering and diffraction response of a drone to an incident soundfield that can be used to enhance drone audition functions. We use a real drone-embedded with microphones to show the DRTF measurement procedure from a circular loudspeaker array.

6.1 Introduction

Recent emergence of drones across the world enables applications and services in search and rescue missions, wildlife monitoring, and video capturing for media and filming industries [2, 18]. Processing of audio recordings from microphone arrays embedded on drones on real-world scenes has been referred to as drone audition, whose function includes sound source localisation [11–14], source separation [15], source tracking [16, 17], and signal enhancement [18, 19]. There exists audio source localisation and speech enhancement algorithms that hold free-field assumptions to derive the microphone array steering matrix or array transfer functions [12–14, 67, 74, 175, 267, 271]. However, in practical acoustical environments, the performance of those functions may suffer due to such free-field assumption [13] as the drone acts as a secondary sound source due to the scattering of the drone body with the sound propagation from any arbitrary incident soundfield (primary source). Alternatively, the performance of those functions may be enhanced by carefully accounting for the drone's acoustic properties such as scattering and diffraction to an incident soundfield. This motivates us to require and make use of an accurate representation of the drone's acoustic properties.

In this chapter, we offer a method to model the drone scattering into a transfer function from a source to a microphone on the drone and to estimate its parameters using a practically viable method. Earlier works on measuring the scattering and diffraction of the human head are known as the head-related transfer functions (HRTF) which are modelled using the dummy head or human subjects and have been widely studied in the past few decades [273–277]. There also exist direction-dependent responses or device-related transfer functions from microphone arrays mounted on complex-shaped devices incorporating the device's scattering [104, 283–286, 290]. In another approach similar to the HRTF, *Donley et al.* [291] presented a device related transfer function dataset by employing the Knowles Electronics Mannequin for Acoustics Research (KEMAR) manikin with an augmented reality glasses microphone array on a head and torso simulator in an anechoic environment for a discrete set of positions on a sphere. This HRTF and the device-related transfer function are used in the application of binaural reproduction [292] using binaural headphones or virtual reality (VR) and augmented reality (AR) headsets

for an immersive experience. We note that the robot audition system utilises the HRTF jointly with the interaural phase difference (IPD) and interaural intensity difference (IID) extracted from the binaural microphone recordings to localise the sound sources [104]. However, the proposed DRTF is intended to be used in drone audition applications, e.g., sound source localisation, source signal enhancement, and source tracking. To cope with the outdoor environments, external factors such as wind noise, and reflections from the ground should be considered in conjunction with the anechoic DRTF as it only captures the sound wave propagation effect around the drone body and its structure due to a sound source.

As such, the existing methods for modeling acoustic scattering from complex-shaped rigid objects (e.g., the human body, and head-worn devices), utilise measured IRs, despite being conducted in an anechoic or non-anechoic environment. To this end, our aims are (i) to use a discrete set of experimental measurements that are practically viable; and (ii) to model a continuous drone-related transfer function (DRTF): the spatial acoustic response of a drone due to a sound source to an on-board microphone of the drone, contains all of the relevant spatial cues incorporating scattering and diffraction, by decomposing impulse response measurements in the spherical harmonic domain. To the best of our knowledge, there has not been prior work for measuring and modelling embedded microphone responses for drones.

The method emphasises the DRTF measurement and modelling procedure using spherical harmonics. The major problem with the existing HRTF-type measurements is that a large number of discrete measurements on a sphere have to be used for modelling with a high spherical harmonic decomposition order. However, these traditional measurements result in a high computational cost for practical applications and increase the acquisition time. In contrast, we focus on the DRTF representation of each microphone on a drone-embedded microphone array using a discrete set of measurements from a set of sample positions that are practically viable. In particular, we present (i) a parametric DRTF model for each microphone attached to the drone in spatial and spectral domains; and (ii) its parameter estimation using a set of measurement points on multiple horizontal circles.

There exists a parametric transfer function estimation method using a set of measurements for robot audition [104]. The method incorporates scattering from

the robot to the direct path by modelling the transfer function relies on the assumption that the robot head is a sphere. However, due to a drone being a complex-shaped object, modelling the transfer function by scattering theory is difficult. Another parametric approach proposes on-the-spot calibration of microphone array transfer functions incorporating the robot body and room acoustics for robot audition applications [289]. This method estimates the relative transfer function over a given plane using a set of microphones. By contrast, our parametric model is independent of the number of microphones and obtains the continuous representation of the transfer function for each microphone on the drone over the elevations. Furthermore, the transfer function approximation by performing a Fourier series expansion with terminating the approximation at an appropriate order has been introduced in [287, 288]. In contrast, this chapter proposes an alternative form of continuous, orthogonal representation by expressing the transfer function as a weighted sum of surface spherical harmonics utilising a set of orthogonal functions. We emphasise that the continuous transfer function model for drone audition application would be a benefit to utilise moving sound sources.

The acoustic reciprocity principle [323] states that the transfer function stays the same when the source and the receiver are interchanged. We follow this concept to propose a parametric DRTF by reversing the directly measured DRTF with the loudspeaker attached to the drone body and microphone around the drone.

The proposed parametric model is built upon the hybrid array structure that uses a set of parallel circular arrays to decompose a soundfield into spherical harmonic components in [324, 325]. This array structure provides practical sensor locations compared to typical spherical arrays. We note that both works in [324,325] are on soundfield reproduction, but neither of these works provides a novel measurement process for the transfer function modeling. The advantage of using such an array structure is that IR measurements can be performed using a single height-adjustable loudspeaker without either a spherical or circular loudspeaker array setup.

With the proposed measurement method, we use a discrete set of loudspeaker positions in circular arrays which are at different elevations and parallel to the x-y plane. We use these circular arrays to derive the 3D spherical harmonic DRTF coefficients from the 2D circular harmonic coefficients that are obtained from eigen-

functions of 2D soundfields [293]. We validate the parameter estimation using a known soundfield on a simple rigid sphere acting analogous to a drone through both theoretical simulation and experimental measurements using em32 Eigenmike spherical microphone array [326]. Furthermore, we also conduct an error analysis over the number of measurement points through simulations.

We use a real drone-embedded with microphones to show the DRTF measurement procedure from a circular loudspeaker array. After deriving DRTF coefficients from the proposed parameter estimation method, we propose to model the DRTF as a continuous function of angles over frequency considering three inner shell radii based on the drone's structure. Due to the higher order nature of the DRTF and the drone being a complex-shaped rigid body, this consideration of setting separate radii over frequency is desirable to select a suitable truncation order. We illustrate the continuous DRTF model and the reconstruction error in terms of a Normalised Mean Square Error (NMSE) for each embedded microphone on a drone as a function of direction for frequencies up to 1125 Hz. We note that utilising this DRTF model and validating its improvement in drone audition functions are left for future work.

The pre-processing of the HRTF measurements based on the ear alignment was suggested to reduce the required number of measurements and its effective spherical harmonic order using either time-alignment or phase-correction [327–330]. Alternatively, in this chapter, we re-define microphone aligned DRTF model as a function of both frequency and incident source direction by using a free-field translation operator, which derives through the spherical wave model [329], thus performs both magnitude and phase correction accordingly to address the mismatch between the default origin and the microphone position, while achieving improved reconstruction performance. We utilise this microphone aligned DRTF model in a beamforming example using real-life recordings.

The rest of this chapter is organised as follows. In Section 6.3, we propose an efficient parameterisation of the DRTF in spatial and spectral domains using finite-order spherical harmonic expansion and their drone dependent weightings defined by a set of DRTF coefficients. In Section 6.4, we estimate the DRTF coefficients through a practically viable method. In Section 6.5, we validate the proposed parameterisation approach using a simple rigid sphere acting as a drone.

In Section 6.6, we conduct the DRTF measurement procedure using a circular loudspeaker array and illustrate the continuous DRTF model for each embedded microphone on a drone as a function of direction for frequency up to 1125 Hz. In Section 6.7, we re-define the continuous DRTF model based on microphone alignment and evaluate and compare its reconstruction performance over different measurement setups. Finally, in Section 6.8 we summarise and provide concluding remarks.

6.2 Problem Formulation

In this chapter, we introduce a parametric DRTF model in the spherical harmonic domain.

Consider a source located at (r, θ, ϕ) from the origin of the spherical coordinate system $\mathbf{o} = (0, 0, 0)$ where r denotes the Euclidean distance between the origin and the source position, (θ, ϕ) denote the angular directions with elevation angle $\theta \in [0, \pi]$ downwards from the z-axis and azimuth angle $\phi \in [0, 2\pi)$ counterclockwise from the x-axis. Let there be Q microphones attached/embedded to the drone body. We define the DRTF with respect to a coordinate system centered at $\mathbf{o}$, $D^{(q)}(k, r, \theta, \phi; \mathbf{o})$ as the frequency response between a source located at (r, θ, ϕ) and the q^{th} $(q \in \{1, \cdots, Q\})$ microphone where $k = 2\pi f/c$ is the wave number, f is the frequency, c is the speed of sound propagation. We consider both Cartesian and Spherical coordinate systems with respect to the centre $\mathbf{o}$ of the drone as shown in Fig. 6.1.

The main objectives of this chapter are to develop

(i) a parametric DRTF model for each microphone as a continuous function of direction (θ, ϕ) and frequency k;

(ii) a practical method to estimate the parameters of the DRTF model.

We wish to measure a discrete set of measurements to each microphone on the drone with a set of loudspeaker locations that are practically viable.

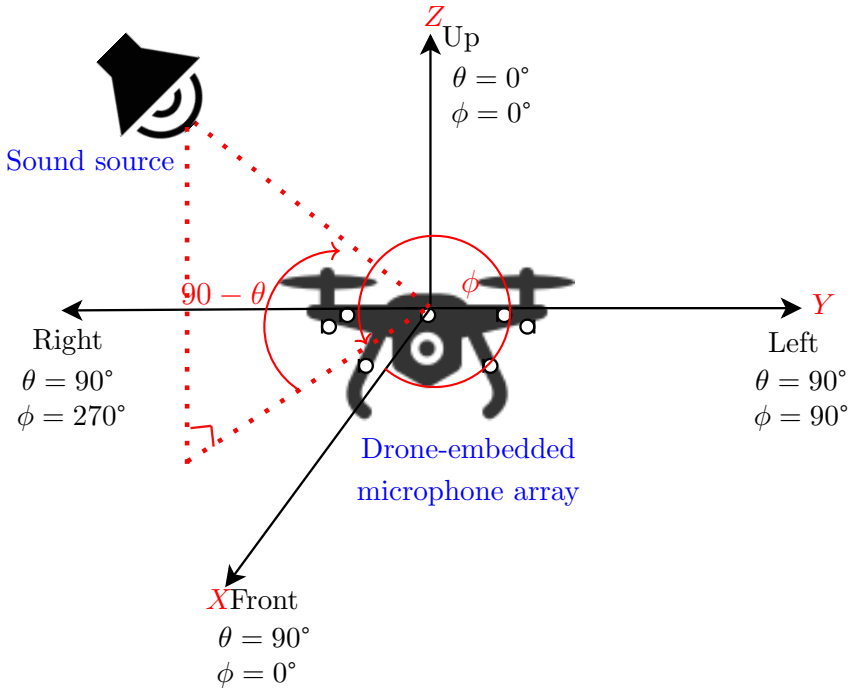


Figure 6.1: The spherical coordinates system in DRTF measurement for drone-embedded microphones.

6.3 Parameterisation of the DRTF

In this section, we present the DRTF parameterisation in the range, elevation, and azimuth three-dimensions using a spherical loudspeaker measurement method which we consider as the baseline method for estimating DRTF coefficients.

6.3.1 Spherical harmonic representation of the DRTF

The DRTF for the q^{th} microphone, $D^{(q)}(k, r, \theta, \phi; \mathbf{o})$ represents the acoustic response of the q^{th} microphone embedded on a drone subject to stimulus from a sound source at (r, θ, ϕ). Following the work in [277] on HRTF which uses the principle of acoustic reciprocity [279, 323], we model the $D^{(q)}(k, r, \theta, \phi; \mathbf{o})$ to a sound source with arbitrary position (r, θ, ϕ), shown in Fig. 6.1, as a function of frequency and

source location

$$D^{(q)}(k, r, \theta, \phi; \mathbf{o}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_{nm}^{(q)}(k) h_n^{(1)}(kr) Y_{nm}(\theta, \phi), \tag{6.1}$$

where n is the order, and m is the degree, $d_{nm}^{(q)}(k)$ are the 3D spherical harmonic *DRTF coefficients* of the q^{th} microphone that describes the acoustic scattering and diffraction by the drone which are dependent on frequency k but independent of spatial position (r, θ, ϕ). The DRTF coefficients are considered to be estimated prior from the measured DRTF $D^{(q)}(k, r, \theta, \phi; \mathbf{o})$ which can be continuously characterised over space (r, θ, ϕ) with only a set of coefficients $d_{nm}^{(q)}(k)$. The middle term, $h_n^{(1)}(\cdot)$ is the first kind spherical Hankel function, and the final term, $Y_{nm}(\theta, \phi)$ are the spherical harmonic basis functions, defined as

$$Y_{nm}(\theta, \phi) = \mathcal{P}_{n|m|}(\cos\theta) E_m(\phi), \tag{6.2}$$

where $\mathcal{P}_{n|m|}(\cdot)$ are the normalised associated Legendre functions given by

$$\mathcal{P}_{n|m|}(\cos\theta) = \sqrt{\frac{2n+1}{2}} \sqrt{\frac{(n-|m|)!}{(n+|m|)!}} P_{n|m|}(\cos\theta), \tag{6.3}$$

and $E_m(\phi) = (1\sqrt{2\pi}) e^{im\phi}$ are the normalised exponential functions, $e \approx 2.71828$ is the Euler's number and $i = \sqrt{-1}$. The functions $P_{n|m|}(\cos\theta)$ and $E_m(\phi)$ form orthonormal basis sets in elevation $\theta \in [0, \pi]$ and azimuth $\phi \in [0, 2\pi)$, respectively.

The DRTF coefficients of the spatially continuous DRTFs for the q^{th} microphone by the inverse spherical Fourier transform [331] in (6.1) as

$$d_{nm}^{(q)}(k) = \frac{1}{h_n^{(1)}(kr)} \int_{\theta} \int_{\phi} D^{(q)}(k, r, \theta, \phi; \mathbf{o}) Y_{nm}^{*}(\theta, \phi) \sin(\theta)\, d\theta\, d\phi, \tag{6.4}$$

where $(\cdot)^*$ denotes the complex conjugation operation.

In practice, the required number of loudspeakers over a unit sphere is obtained

by approximating the integral in (6.5) using a set of samples (θ_ℓ, ϕ_ℓ) as

$$d_{nm}^{(q)}(k) \approx \frac{1}{h_n^{(1)}(kr)} \sum_{\ell=1}^{\mathcal{L}} D^{(q)}(k, r, \theta_\ell, \phi_\ell) Y_{nm}^*(\theta_\ell, \phi_\ell), \tag{6.5}$$

where $\mathcal{L}$ is the total number of samples.

With the equivalent mode-limited functionality of the HRTF [277], the number of modes to represent the DRTF spatial variation is given by taking a truncated expansion

$$N = \lceil ekR/2 \rceil, \tag{6.6}$$

R is the radius of the spherical region encompassing the drone, and $\lceil \cdot \rceil$ denotes ceiling operation. Hence, the summation of the infinite series in (6.1) is truncated to finite order N, derive the minimum number of discrete loudspeakers required over the unit sphere as $\mathcal{L} \geq (N+1)^2$ for the spherical sampling geometry [332–334]. Thus,

$$\mathcal{L} \geq (\lceil ekR/2 \rceil + 1)^2. \tag{6.7}$$

Once $d_{nm}^{(q)}(k)$, $n = 0, \cdots, N$ and $m = -n, \cdots, n$, are determined, the continuous DRTF model over space (r, θ, ϕ) and frequency k in (6.1) can be reconstructed with the closed-form expansion

$$D^{(q)}(k, r, \theta, \phi; \mathbf{o}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} d_{nm}^{(q)}(k) h_n^{(1)}(kr) \mathcal{P}_{n|m|}(\cos\theta) E_m(\phi). \tag{6.8}$$

using a set of coefficients by choosing a sufficiently large truncation order N.

Note that the DRTF coefficients are naturally extracted from the measured IRs over all directions by placing the drone inside a spherical loudspeaker array, that requires large specialised equipment and consumes significant time [276].

In the next section, we propose a practically viable method to estimate these 3D spherical harmonic DRTF coefficients from a set of measured DRTFs on parallel circular arrays.

6.4 Estimation of the DRTF Coefficients using Multiple Circular Arrays

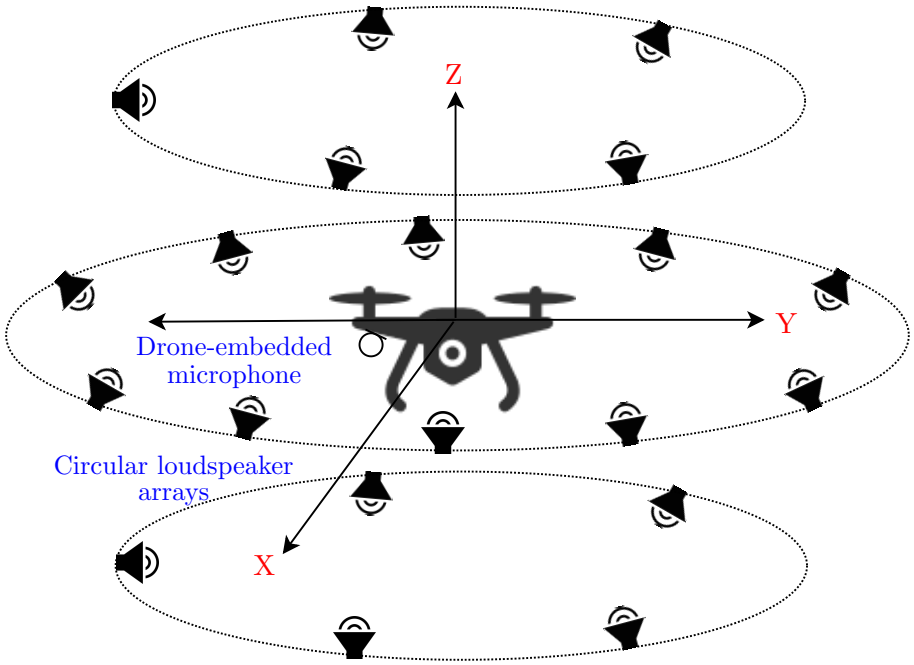


Figure 6.2: Example of multiple circular array structure for DRTF measurement.

In this section, we propose an alternative design for DRTF measurements which enables a viable solution compared to the traditional DRTF measurement setup (as in (6.5)). We refer to the array setup in [324] and [325] to provide a more feasible design for DRTF measurements. While [324,325] use a set of parallel circular arrays instead of a spherical loudspeaker array for soundfield reproduction, neither of these works yields a novel measurement process for the device-related transfer functions. We propose a set of measurements by the multiple circles (as shown in Fig. 6.2) to obtain the DRTF coefficients $d_{nm}^{(q)}(k)$. We also provide the performance measures to evaluate the reconstruction accuracy of the proposed parameterisation method. In the following, we estimate the DRTF coefficients for the q^{th} microphone. Note that all other microphones on a drone-embedded microphone array can be similarly estimated.

The first step of the parameterisation of the DRTF is to decompose the measured DRTFs into circular harmonics [293]. This involves obtaining the 2D circular harmonic coefficients using multiple circular arrays (Section 6.4.1). We then estimate the 3D spherical harmonic DRTF coefficients $d_{nm}^{(q)}(k)$ in Section 6.4.2. We derive each of these steps in the following.

6.4.1 DRTF decomposition by multiple circles

Re-writing $D^{(q)}(k, r, \theta, \phi; \mathbf{o})$ by substituting (6.2) into (6.8) with the expansion of the spherical harmonic basis function as

$$D^{(q)}(k, r, \theta, \phi; \mathbf{o}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} d_{nm}^{(q)}(k) h_n^{(1)}(kr) \mathcal{P}_{n|m|}(\cos\theta) E_m(\phi). \tag{6.9}$$

Consider a circular loudspeaker aperture (r_ν, θ_ν) for $r = r_\nu$ and $\theta = \theta_\nu$ ($\nu = 1, \ldots, \mathcal{V}$) that is parallel to the x-y plane (see Fig. 6.2 for 3 circular apertures of loudspeakers). Hence, the DRTF $D^{(q)}(k, r_\nu, \theta_\nu, \phi; \mathbf{o})$ for any arbitrary circular aperture (r_ν, θ_ν) is spatially dependent on the azimuth angle ϕ and is expressed as

$$D^{(q)}(k, r_\nu, \theta_\nu, \phi; \mathbf{o}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} d_{nm}^{(q)}(k) h_n^{(1)}(kr_\nu) \mathcal{P}_{n|m|}(\cos\theta_\nu) E_m(\phi). \tag{6.10}$$

We can extract the 3D spherical harmonic DRTF coefficients $d_{nm}^{(q)}(k)$ by exploiting (6.10) from multiple circle apertures (r_ν, θ_ν) using the properties of the associated Legendre functions. We multiply (6.10) by $E_{-m}(\phi)$ and integrate it with respect to ϕ over the circular aperture $[0, 2\pi)$ to obtain the circular harmonics coefficients, $b_m^{(q)}(k, r_\nu, \theta_\nu)$ as

$$b_m^{(q)}(k, r_\nu, \theta_\nu) = \int_0^{2\pi} D^{(q)}(k, r_\nu, \theta_\nu, \phi; \mathbf{o}) E_{-m}(\phi)\, d\phi, \tag{6.11}$$

$$= \sum_{n=|m|}^{N} d_{nm}^{(q)}(k) h_n^{(1)}(kr_\nu) \mathcal{P}_{n|m|}(\cos\theta_\nu). \tag{6.12}$$

This circular coefficient for a given mode m is equivalent to the weighted sum of the 3D spherical harmonic DRTF coefficients for a specific m and $n = |m|, |m| +$

$1, \ldots, N$. Evaluating (6.11) over a finite number of sampling points L_ν[1] on a given circle (r_ν, θ_ν), we approximate the integral in (6.11) by a summation

$$b_m^{(q)}(k, r_\nu, \theta_\nu) \approx \frac{2\pi}{L_\nu} \sum_{v=1}^{L_\nu} D^{(q)}(k, r_\nu, \theta_\nu, \phi_v; \mathbf{o}) E_{-m}(\phi_v). \tag{6.13}$$

Since $b_m^{(q)}(k, r_\nu, \theta_\nu)$ is estimated from (6.13) for $m = -N, \ldots, N$, (6.12) can be solved to find the 3D spherical harmonic DRTF coefficients in the least squares sense.

6.4.2 Extracting DRTF coefficients

Let there be $\mathcal{V} \geq (N+1)^2$ multiple circles of microphones located parallel to x-y plane given as $(r_\nu, \theta_\nu), \nu = 1, \ldots, \mathcal{V}$. By writing (6.12) for $\mathcal{V}$, we obtain a set of simultaneous equations given by

$$\mathbf{b}_m^{(q)} = \mathbf{H}_m \mathbf{d}_m^{(q)}, \text{for } m = -N, \ldots, N \tag{6.14}$$

where

$$\mathbf{H}_m = \begin{bmatrix} h_{|m|}^{(1)}(kr_1)\mathcal{P}_{|m||m|}(\cos\theta_1) & \ldots & h_N^{(1)}(kr_1)\mathcal{P}_{N|m|}(\cos\theta_1) \\ \vdots & \ddots & \vdots \\ h_{|m|}^{(1)}(kr_\mathcal{V})\mathcal{P}_{|m||m|}(\cos\theta_\mathcal{V}) & \ldots & h_N^{(1)}(kr_\mathcal{V})\mathcal{P}_{N|m|}(\cos\theta_\mathcal{V}) \end{bmatrix},$$

$\mathbf{d}_m^{(q)} = [d_{|m|m}^{(q)}, d_{(|m|+1)m}^{(q)}, \ldots, d_{Nm}^{(q)}]^T$, and $\mathbf{b}_m^{(q)} = [b_{1m}^{(q)}, \ldots, b_{\mathcal{V}m}^{(q)}]^T$. For a circular aperture (r_ν, θ_ν), the spherical harmonic coefficients $\mathbf{d}_m^{(q)}$ can be calculated by solving the system of linear equations described by (6.14) for each m using least squares

[1]Here, we can use nonuniform samples on each circle as long as the separation between any two samples is less than the maximum separation required for the Shannon sampling theorem.

[2]Note that we estimate $(N+1)^2$ DRTF coefficients for the N^{th} order continuous DRTF model.

method provided that $\mathbf{H}_m$ is nonsingular,[3] as in, [324]

$$\mathbf{d}_m^{(q)} = \mathbf{H}_m^{\ddagger}\mathbf{b}_m^{(q)}, \text{for } m = -N, \ldots, N \tag{6.15}$$

where $(\cdot)^{\ddagger}$ denotes the Moore-Penrose inverse of a matrix.

Therefore, the DRTF of a drone-embedded microphone array can be estimated via a set of circular loudspeaker arrays. This provides a novel parameterisation and measurement approach of the DRTF. Note that, unlike the circular loudspeaker array, a single height-adjustable loudspeaker can be used for the DRTF measurements.

6.4.3 Reconstruction error

We evaluate the reconstruction performance in terms of the normalised Mean Square Error (MSE) by defining the pressure error between the true and reconstructed DRTF over all $\mathcal{L}$ angles ($\ell = 1, \cdots, \mathcal{L}$) at each frequency as

$$\epsilon^{(q)}(k) = \frac{\sum_{\ell}^{\mathcal{L}} |D^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o}) - \widetilde{D}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})|^2}{\sum_{\ell}^{\mathcal{L}} |D^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})|^2}, \tag{6.16}$$

where $\widetilde{(\cdot)}$ denotes the reconstructed soundfield by substituting the estimated $d_{nm}^{(q)}(k)$ from (6.8). Similarly, we define the magnitude and phase error [330] as

$$\epsilon_{\text{mag}}^{(q)}(k) = \frac{\sum_{\ell}^{\mathcal{L}} |D_{\text{mag}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o}) - \widetilde{D}_{\text{mag}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})|^2}{\sum_{\ell}^{\mathcal{L}} |D_{\text{mag}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})|^2}, \tag{6.17}$$

$$\epsilon_{\text{phase}}^{(q)}(k) = \frac{1}{\mathcal{L}} \sum_{\ell}^{\mathcal{L}} (D_{\text{phase}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o}) - \widetilde{D}_{\text{phase}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})), \tag{6.18}$$

where $D_{\text{mag}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})$, $\widetilde{D}_{\text{mag}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})$ are magnitude values of the true and reconstructed DRTFs, respectively, and $D_{\text{phase}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})$, $\widetilde{D}_{\text{phase}}^{(q)}(k, r_\ell, \theta_\ell, \phi_\ell; \mathbf{o})$ are phase values of the DRTFs.

[3]Note that we set up an overdetermined system such that the number of active circles is greater than $N - |m|$. We can use nonuniform samples on each circle as long as the separation between any two samples is less than the maximum separation required for the Shannon sampling theorem.

6.5 Validation of the Proposed Parameterisation and Measurement Approach using a Rigid Sphere

In this section, we validate the proposed method's reconstruction accuracy of the DRTF using a simple rigid sphere that acts analogous to a drone. We first theoretically simulate the soundfield on the rigid sphere (Section 6.5.1). We then estimate the DRTF coefficients through both theoretical simulation and experimental measurements using em32 Eigenmike spherical microphone array (Section 6.5.2). Finally, we evaluate the performance to verify our proposed parameterisation approach.

6.5.1 Simulating the soundfield on a rigid sphere

The total soundfield at the q^{th} microphone can be written as the summation of the incident soundfield and scattered field as

$$D^{(q)}(k, r^{(q)}, \theta^{(q)}, \phi^{(q)}; \mathbf{o}) = D^{(q)}_{\text{dir}}(k, r^{(q)}, \theta^{(q)}, \phi^{(q)}; \mathbf{o}) + D^{(q)}_{\text{ref}}(k, r^{(q)}, \theta^{(q)}, \phi^{(q)}; \mathbf{o}). \quad (6.19)$$

The incident sound pressure at $(r^{(q)}, \theta^{(q)}, \phi^{(q)})$ due to a point source at (r, θ, ϕ) is given by the free-field Green's function can be written in the form of spherical harmonic expansions as

$$D^{(q)}_{\text{dir}}(k, r^{(q)}, \theta^{(q)}, \phi^{(q)}; \mathbf{o}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} ikh_n^{(1)}(kr) Y_{nm}(\theta, \phi) j_n(kr^{(q)}) Y^*_{nm}(\theta^{(q)}, \phi^{(q)}), \quad (6.20)$$

where $j_n(\cdot)$ is spherical Bessel function of order n. The scattered field can be obtained by

$$D^{(q)}_{\text{ref}}(k, r^{(q)}, \theta^{(q)}, \phi^{(q)}; \mathbf{o}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \gamma_{nm} h_n^{(1)}(kr^{(q)}) Y^*_{nm}(\theta^{(q)}, \phi^{(q)}), \quad (6.21)$$

where $\gamma_{nm} = -i(j_n'(kr_0)/h_n'^{(1)}(kr_0))kh_n^{(1)}(kr)Y_{nm}(\theta,\phi)$ is the scattering coefficients, r_0 denotes the radius of the rigid sphere, and $(\cdot)'$ denotes the first derivative with respect to the argument. Substituting (6.20), and (6.21) into (6.19), we obtain

$$D^{(q)}(k, r^{(q)}, \theta^{(q)}, \phi^{(q)}; \mathbf{o}) = \sum_{n=0}^{N}\sum_{m=-n}^{n} \underbrace{ik\left(j_n(kr^{(q)}) - \frac{j_n'(kr_0)}{h_n'^{(1)}(kr_0)}h_n^{(1)}(kr^{(q)})\right)Y_{nm}^*(\theta^{(q)},\phi^{(q)})}_{\triangleq d_{nm}^{(q)}(k)} h_n^{(1)}(kr)Y_{nm}(\theta,\phi). \tag{6.22}$$

Therefore, the DRTF at microphone position $(r^{(q)}, \theta^{(q)}, \phi^{(q)})$ can be numerically solved in the same form as in (6.8) with [293]

$$d_{nm}^{(q)}(k) = ik\left(j_n(kr^{(q)}) - \frac{j_n'(kr_0)}{h_n'^{(1)}(kr_0)}h_n^{(1)}(kr^{(q)})\right)Y_{nm}^*(\theta^{(q)},\phi^{(q)}). \tag{6.23}$$

6.5.2 Reconstruction performance of the proposed parameterisation

Implementation of the coefficient estimation process

The implementation of the proposed algorithm comprises four steps as summarised below.

(i) Measure (or simulate) the impulse responses $D^{(q)}(k, r_\nu, \theta_\nu, \phi; \mathbf{o})$ for $\nu = 1, \ldots, \mathcal{V}$ (over the setting as listed in Table 6.1 for the rigid sphere).

(ii) Extract the circular harmonic coefficients $b_m^{(q)}(k, r_\nu, \theta_\nu)$ for each circle ν from (6.13).

(iii) Use (6.15) to estimate the DRTF coefficients $d_{nm}^{(q)}(k)$.

(iv) Reconstruct the DRTF model in space domain with (6.8). Note that the frequency k determines the truncation order $N(\approx \lceil kR \rceil)$ of the reconstructed DRTF.

Table 6.1: Design formulation of the DRTF measurements for a rigid sphere using multiple circles.

Parameter	Value
$\mathcal{V}$	5
r_ν	1m
θ_ν	$31.72^\circ, 58.28^\circ, 90^\circ, 121.72^\circ, 148.28^\circ$
$\Delta\phi_\nu$	12°

Table 6.2: Microphone placement on the rigid sphere.

q	$\theta^{(q)}$	$\phi^{(q)}$
1	69°	0°
2	159°	89°
3	69°	180°
4	21°	269°

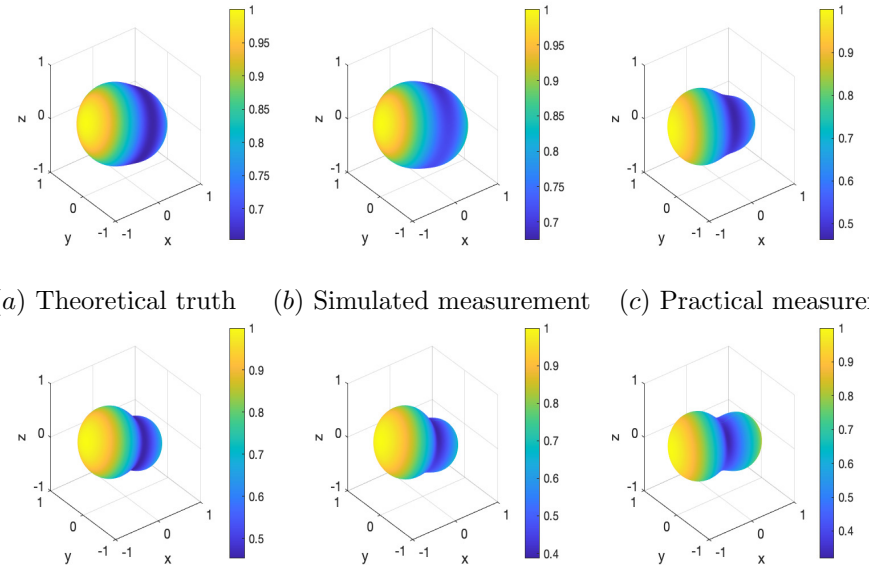
We consider an ideal rigid sphere of $r_0 = 4.2$ cm with 4 microphones placed at locations given by Table 6.2. We simulate the DRTF of the rigid sphere using the theoretical DRTF coefficients in (6.23). The simulated DRTF at each microphone is added with 30 dB signal-to-noise ratio (SNR) of white Gaussian noise and is referred to as *simulated measurement*.

We also experimentally measure the DRTF on the rigid sphere for the same settings depicted in Table 6.1. An em32 Eigenmike rigid spherical microphone array was used for the experimental setup in a non-anechoic room.[4] We truncate the IRs to remove the room reverberation.

Performance at a single frequency

In Fig. 6.3(a), (b), and (c), we show the magnitude of the DRTF for microphone $q = 3$ at the frequency of 750 Hz for the truncation order of $N = 1$ in theory, simulated measurements, and practical measurements, respectively. We observe that all three implementations have the same pattern, and similar orientation at the same locations. We have also displayed the magnitude of the DRTF at the frequency of 1125 Hz for truncation order of $N = 1$ in Fig. 6.3(d), (e) and (f) and observe similar results for theory and the simulated measurement. Contrary,

[4]Note that the room condition is detailed in Section 6.6.1.

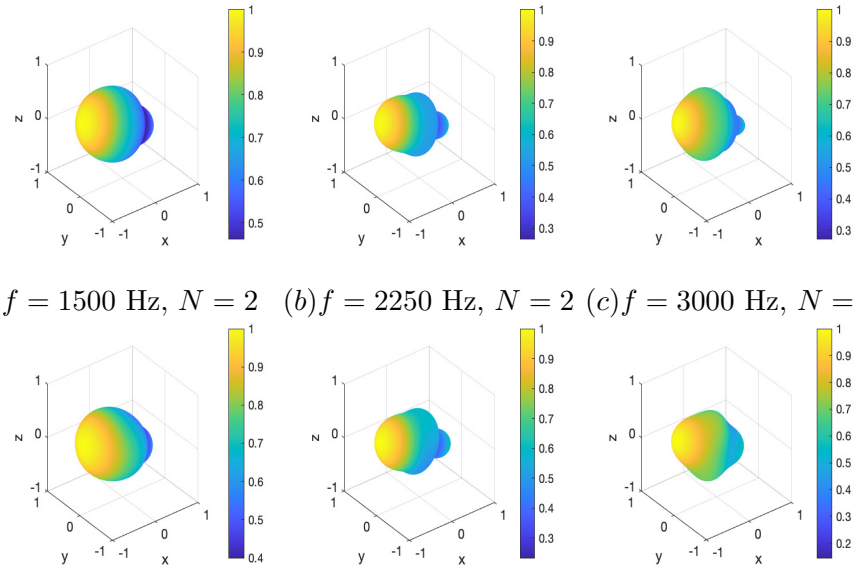


(*a*) Theoretical truth (*b*) Simulated measurement (*c*) Practical measurement

(*d*) Theoretical truth (*e*) Simulated measurement (*f*) Practical measurement

Figure 6.3: Magnitude of the DRTF of $q = 3$ microphone on the rigid sphere at $f = 750$ Hz with truncation order $N = 1$ ((a), (b) and (c)) and $f = 1125$ Hz with $N = 1$ ((d), (e) and (f)).

a different orientation and similar directivity were seen in Fig. 6.3(f) for the real measurements. This may be due to the material properties of the Eigenmike, e.g., rough sphere surface and the attached stand, failing to accurately model real-world phenomena such as scattering and diffraction through the simulations. In such a case, we expect using non-anechoic impulse response measurements would have minor impact on the higher frequencies. Therefore, we omit the real measurements from the DRTF model analysis at the higher frequency bands due to limitations in our experimental facilities at the Australian National University (ANU).



$(a) f = 1500$ Hz, $N = 2$ $(b) f = 2250$ Hz, $N = 2$ $(c) f = 3000$ Hz, $N = 3$

$(d) f = 1500$ Hz, $N = 2$ $(e) f = 2250$ Hz, $N = 2$ $(f) f = 3000$ Hz, $N = 3$

Figure 6.4: Magnitude of the DRTF of $q = 3$ microphone on the rigid sphere at high frequency $f \in \{1500, 2250, 3000\}$ Hz for theory ((a), (b) and (c)) and simulated measurement ((d), (e) and (f)).

Figure 6.4 gives the magnitude of the DRTF of $q = 3$ microphone for both theory and simulated measurements at $f \in \{1500, 2250, 3000\}$ Hz. The simulated measurement closely matches the theory at high frequencies corresponding to the speech signal frequency band (up to 3000 Hz). We observe that more directive beam patterns are seen to have modelled with higher order soundfield estimation at higher frequencies.

Broadband performance

We next investigate the DRTF measurement errors from the simulated measurement shown in Fig. 6.5 for different microphones given in Table 6.2 for frequency range up to 5 kHz. As expected, as a consequence of the truncation order and the required dimensionality for the reconstruction, the measurement errors are shown to be minimised in low frequencies. Observe that both the magnitude DRTF reconstruction errors $\epsilon^{(q)}$, and $\epsilon^{(q)}_{\text{mag}}$ are not significant at the higher frequencies (less

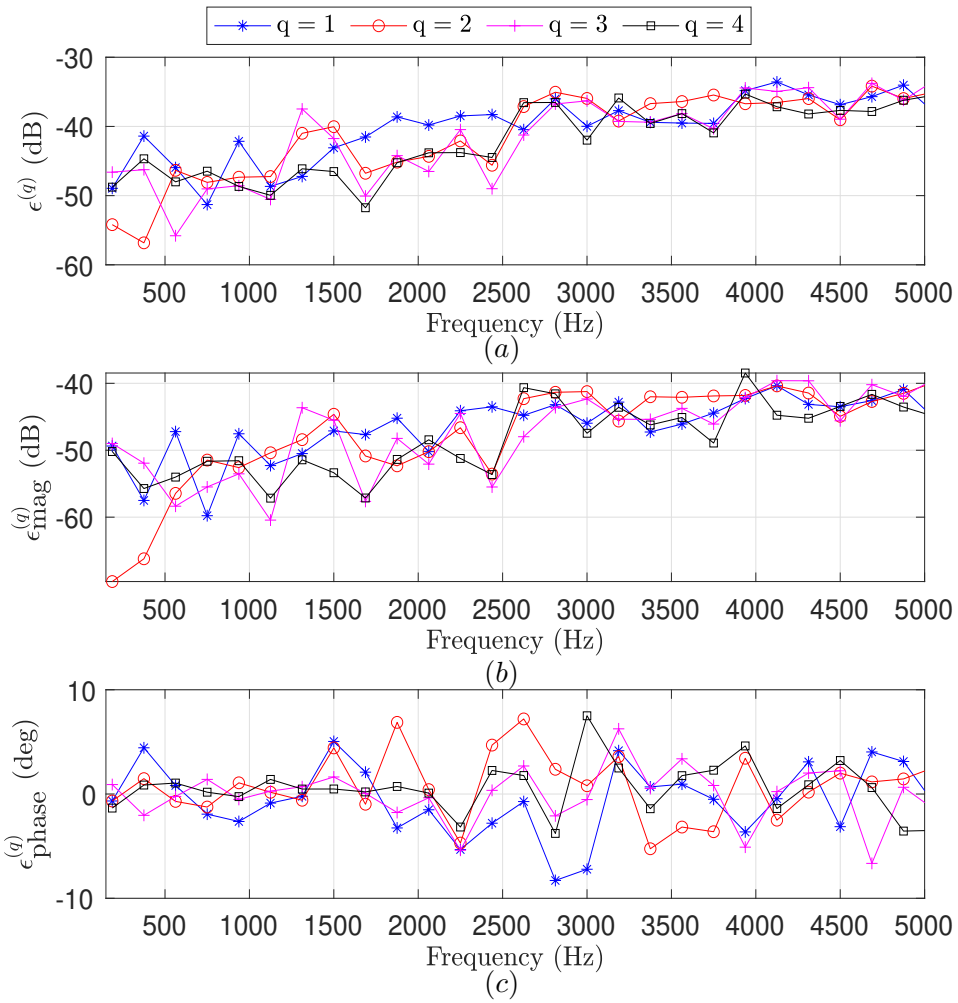


Figure 6.5: The simulated measurement: DRTF reconstruction error performance the proposed parameter estimation method (a) $\epsilon^{(q)}$, (b) $\epsilon^{(q)}_{\text{mag}}$, and (c) $\epsilon^{(q)}_{\text{phase}}$, computed as in (6.16), (6.17), and (6.18), respectively, at the frequency range of [0.2, 5] kHz for the q^{th} microphone where $q \in \{1, 2, 3, 4\}$

than -30 dB), the phase error $\epsilon^{(q)}_{\text{phase}}$ is less than 10°, and thus the accuracy of the proposed parameter estimation approach confirms. In the next subsection, we show that the error performance can be further reduced by increasing the number of multiple circles to include more elevation angles.

Investigation of reconstruction error vs number of measurement points

We also study the DRTF parameter estimation procedure using a discrete set of measurements and evaluate the normalised MSE, magnitude, and phase errors. We introduce six setups by arranging a set of DRTF measurements in multiple circular arrays as depicted in Table 6.3. This way we can examine the proposed method over a fewer number of circles that cover different elevations and azimuth angles.

Table 6.3: Loudspeaker locations of the DRTF measurement setups using multiple circular arrays.

Setups	Directions ($\mathcal{L}$)	$\mathcal{V}$	Loudspeaker locations (θ_ν, ϕ)
$\mathcal{S}1$	150	5	$\theta_\nu \in \{30^\circ : 30^\circ : 150^\circ\}$, $\phi \in \{0^\circ : 12^\circ : 359^\circ\}$
$\mathcal{S}2$	450	15	$\theta_\nu \in \{30^\circ : 8.57^\circ : 150^\circ\}$, $\phi \in \{0^\circ : 12^\circ : 359^\circ\}$
$\mathcal{S}3$	750	25	$\theta_\nu \in \{30^\circ : 5^\circ : 150^\circ\}$, $\phi \in \{0^\circ : 12^\circ : 359^\circ\}$
$\mathcal{S}4$	90	5	$\theta_\nu \in \{30^\circ : 30^\circ : 150^\circ\}$, $\phi \in \{0^\circ : 20^\circ : 359^\circ\}$
$\mathcal{S}5$	270	15	$\theta_\nu \in \{30^\circ : 8.57^\circ : 150^\circ\}$, $\phi \in \{0^\circ : 20^\circ : 359^\circ\}$
$\mathcal{S}6$	450	25	$\theta_\nu \in \{30^\circ : 5^\circ : 150^\circ\}$, $\phi \in \{0^\circ : 20^\circ : 359^\circ\}$

Figure 6.6 shows the $\epsilon^{(q)}$, $\epsilon^{(q)}_{\text{mag}}$, and $\epsilon^{(q)}_{\text{phase}}$ for all six setups for $q = 3$ microphone. We discuss the results of these errors using (6.7) over the desired lower and upper frequency region. As such, the frequency region we are interested in is [0.2 5] kHz, and the required number of discrete measurement points derives to a minimum of 254. As a result, all four setups excluding the $\{\mathcal{S}1,\mathcal{S}4\}$ are shown to be successfully reduced reconstruction error, verifying the proposed parameterisation method's applicability over a set of discrete measurement points. Below this number of measurement points, both $\{\mathcal{S}1,\mathcal{S}4\}$ setups are considered to be shown higher normalised MSE, and magnitude error as well as phase error.

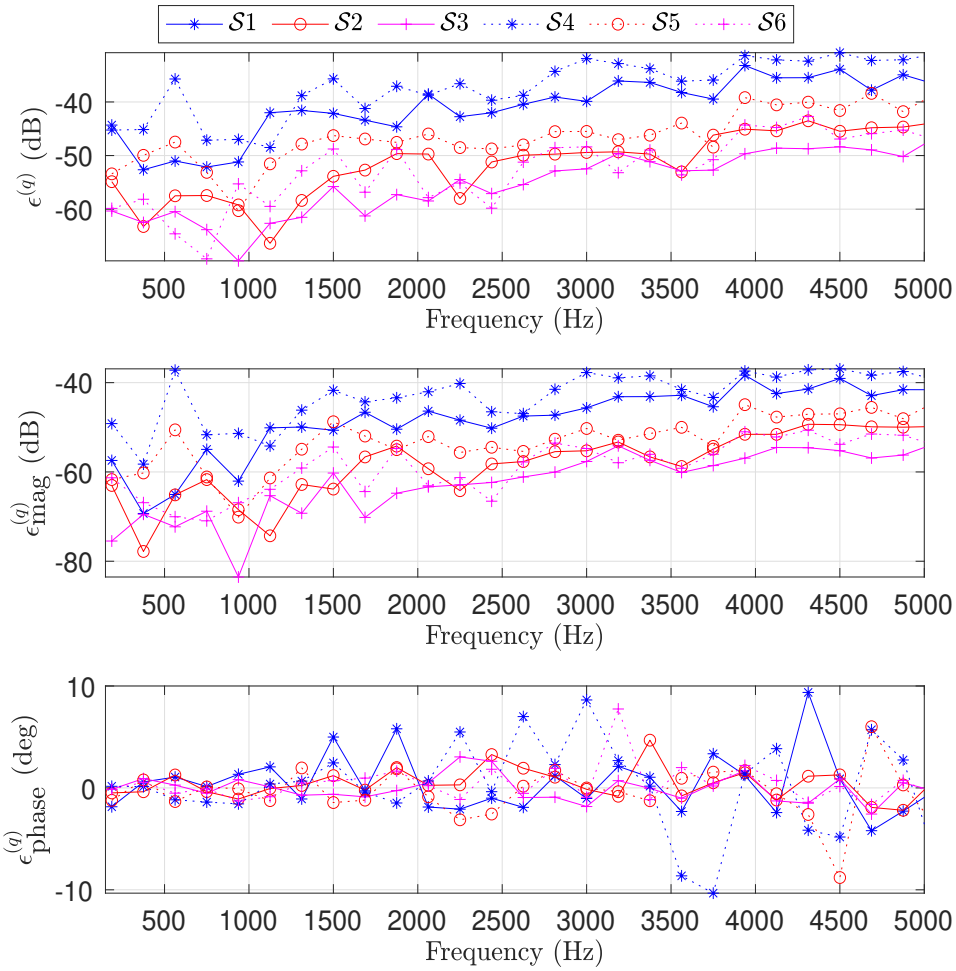


Figure 6.6: The simulated DRTF reconstruction error performance through the proposed parameter estimation method at the frequency range of $[0.2,\ 5]$ kHz for the $q = 3$ microphone over different sets of measurements.

6.6 DRTF Measurements

In this section, we detail a practical example of estimating the DRTF for each microphone attached to a drone using the proposed parameterisation method. We first present the impulse response measurement using a circular loudspeaker array. We then present the DRTF extraction from the measurements. Finally, we illustrate the results of the continuous DRTF model. Note that for the circular loudspeaker

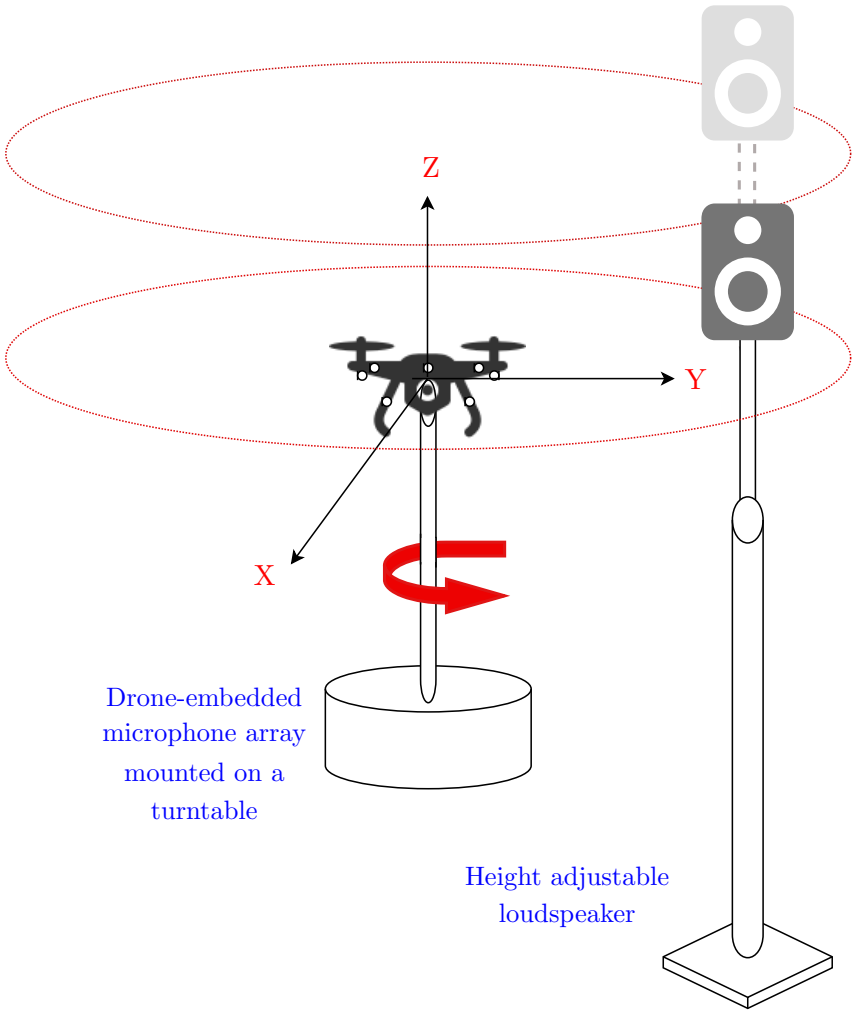


Figure 6.7: A more practical setup for the proposed DRTF measurements using a turntable and a height adjustable loudspeaker.

array, in practice a turntable with a single height adjustable loudspeaker should be sufficient (as shown in Fig. 6.7).

6.6.1 Experimental methodology

The experiment is conducted in the acoustic lab of the ANU with the room dimensions of $[6.7, 3.6, 2.8]$ m for the length, width, and height, respectively, and the

reverberation time of approximately $T_{60} = 400$ ms. Figure 6.8 shows the DRTF measurement setup with a drone-embedded microphone array inside a loudspeaker array. We use a circular loudspeaker array (dodecahedron loudspeaker) with 30 loudspeakers in 5 circular rings. The drone is placed at the centre of the multiple loudspeaker array, and attached with 15 Micro-Electro-Mechanical System (MEMS) dual-microphone modules on the structure of the drone similar to our previous experiment as discussed in Chapter 4 [13] (shown in Fig. 6.9).

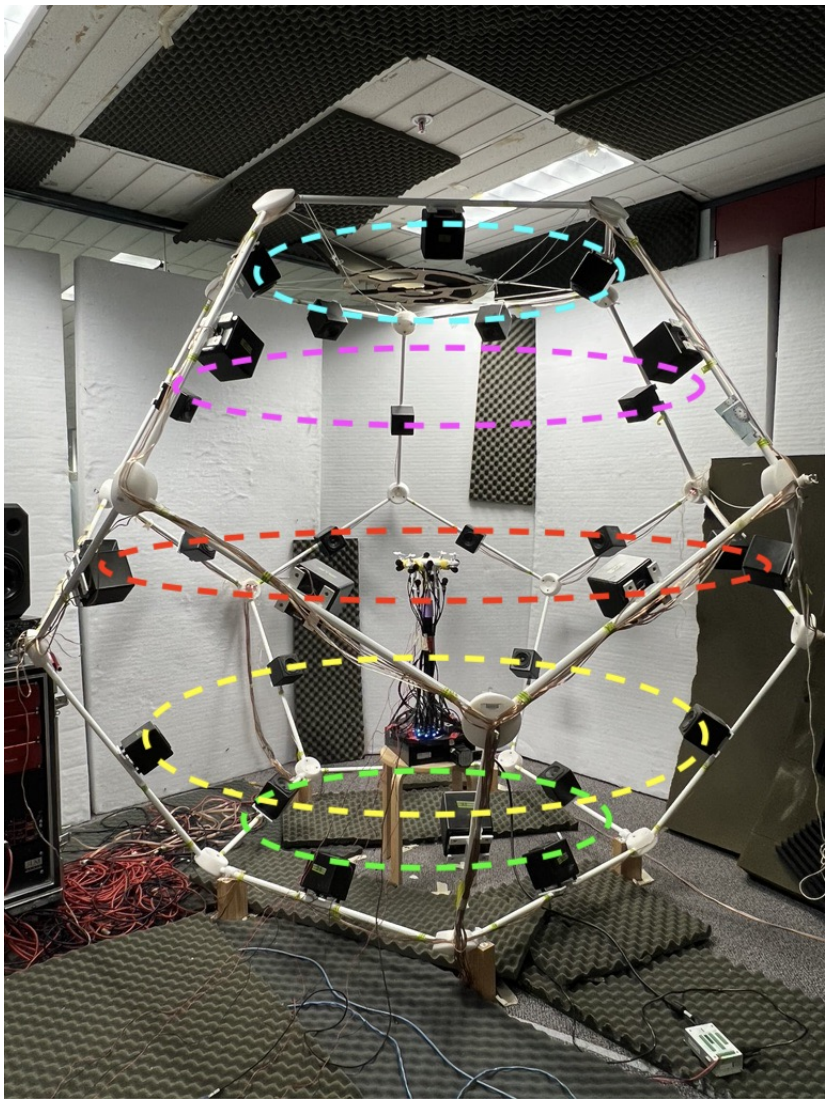


Figure 6.8: Experimental setup using direct DRTF measurements. The dotted circles show the multiple circular loudspeaker arrays.

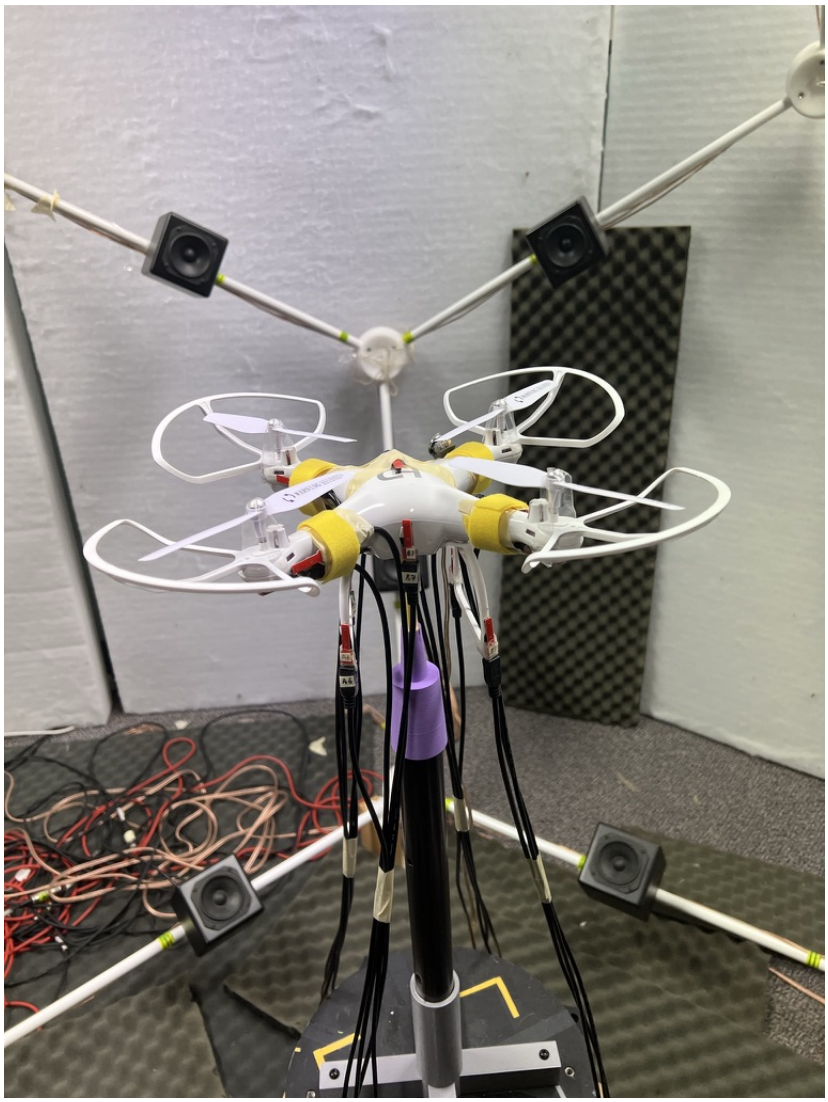


Figure 6.9: The drone-embedded with 15 dual-microphones.

We take measurements by playing a sweep signal of 0.3 s duration. The impulse responses are measured with a sampling rate of 48 kHz. We rotate the setup by mounting the drone on a LinTech 300 series turntable that connects to a QuickSilver Controls Inc. SilverDust D2 IGB servo motor controller, which allows accurate positioning of the drone in the horizontal plane. The measurements are made at a minimum resolution of approximately 12° in the horizontal plane. For the given setup, we obtain impulse responses for 30 different azimuths and 5 different elevation angles for each microphone channel. To obtain more elevation angles,

the setup height is changed by ±20 cm and ±40 cm with respect to the origin of the loudspeaker array by lifting the drone up/down. Therefore, altogether, we obtain impulse responses for 30 different azimuths and 25 different elevation angles for each microphone channel. Table 6.4 summarises the measurement grid of the DRTF measurements.

Table 6.4: Experimental settings of the DRTF measurements using 25 circles to determine DRTF coefficients.

ν	r_ν	θ_ν	$\Delta\phi_\nu$	$\phi_{\nu 0}$	ν	r_ν	θ_ν	$\Delta\phi_\nu$	$\phi_{\nu 0}$
1	1m	31.72°	12°	36°	14	1.26m	137.42°	12°	35.99°
2	1m	58.29°	12°	0°	15	1.36m	157.20°	12°	0°
3	1m	90°	12°	17.99°	16	1.18m	26.58°	12°	36°
4	1m	121.72°	12°	35.99°	17	1.12m	49.53°	12°	0°
5	1m	148.29°	12°	0°	18	1.02m	78.69°	12°	17.99°
6	0.84m	38.93°	12°	36°	19	0.91m	110.95°	12°	35.99°
7	0.91m	69.05°	12°	0°	20	0.83m	141.07°	12°	0°
8	1.02m	101.31°	12°	17.99°	21	1.36m	22.80°	12°	36°
9	1.12m	130.47°	12°	35.99°	22	1.26m	42.58°	12°	0°
10	1.17m	153.42°	12°	0°	23	1.08m	68.19°	12°	17.99°
11	0.69m	49.39°	12°	36°	24	0.86m	98.40°	12°	35.99°
12	0.86m	81.59°	12°	0°	25	0.69m	130.61°	12°	0°
13	1.08m	111.80°	12°	17.99°					

6.6.2 Post-processing to extract the DRTFs

A mini-loudspeaker (manufacturer: VISATON, external Diameter: 16 mm) placed close to one of the microphones on the drone, is used to remove the system delay from the setup by playing an impulse signal. We perform bandpass filtering with the cut-off frequency of $[20, 22\text{k}]$ Hz on the impulse response measurements to avoid the higher frequencies above 22 kHz to remove the unwanted noise from MEMS microphone recordings.

Note that the DRTF measurements ideally performed in an anechoic environment. However, our experiment is conducted in a reverberant room, therefore, in this work we use direct path impulse responses by chopping off late reflections from non-anechoic measurements. The direct path impulse responses can be obtained by

identifying the onset of the first reflections and eliminating that from the impulse response measurements. In order to differentiate the point of the room reflections from the drone scattering, we measure the impulse responses (as described in Section 6.6.1) by removing the drone and leaving one of the microphones in the same position as it was embedded on the drone. We use the microphone pair attached to the top of the drone to identify the drone scattering from the impulse response measurements. This way, we can truncate the impulse responses to obtain the incident soundfield from the acoustic impulse responses measured in a reverberant environment. The impulse responses are truncated to the same length (120 samples) and zero-padded to the length of 256 samples. We perform the fast Fourier transform (FFT) with the FFT length of 256 samples on the impulse responses to obtain the DRTFs. These processed impulse response measurements can be found in the DRTF database.[5]

6.6.3 Experimental representation of the DRTF

As the drone is a complex-shaped rigid body and its ground truth is unknown, therefore, this subsection provides insightful results looking at the DRTF for different microphones. In particular, we illustrate the results of the continuous DRTF representation for a given microphone in (6.8) using the DRTF coefficients derived through the proposed parameter estimation method. We also show an error analysis using a set of measurements to estimate DRTF parameters and using the remaining measurements as ground truth. Let us consider two microphone placements on the drone: $q = 1$ and 2 as shown in Fig. 6.10, attached to one side of the rotor: *rot-microphone*, and on to the top of the drone: *top-microphone*, respectively.

Since the drone is a complex-shaped rigid body, we can encompass it with three radii r_i, $i = 0, 1, 2$ based on its structure: core, arms, and propeller guards/outer radius (as illustrated in Fig. 6.10). We propose to set the truncation order N for frequency f based on r_i. Table. 6.5 provides $N(\approx \lceil kr_i \rceil)$ over three radii. These values are calculated for different frequencies for each radius. A suitable N for a given f may depend on the microphone position on the drone. We simply select the optimal values of N required for DRTF representation as a function of f as

[5] https://zenodo.org/records/10907024

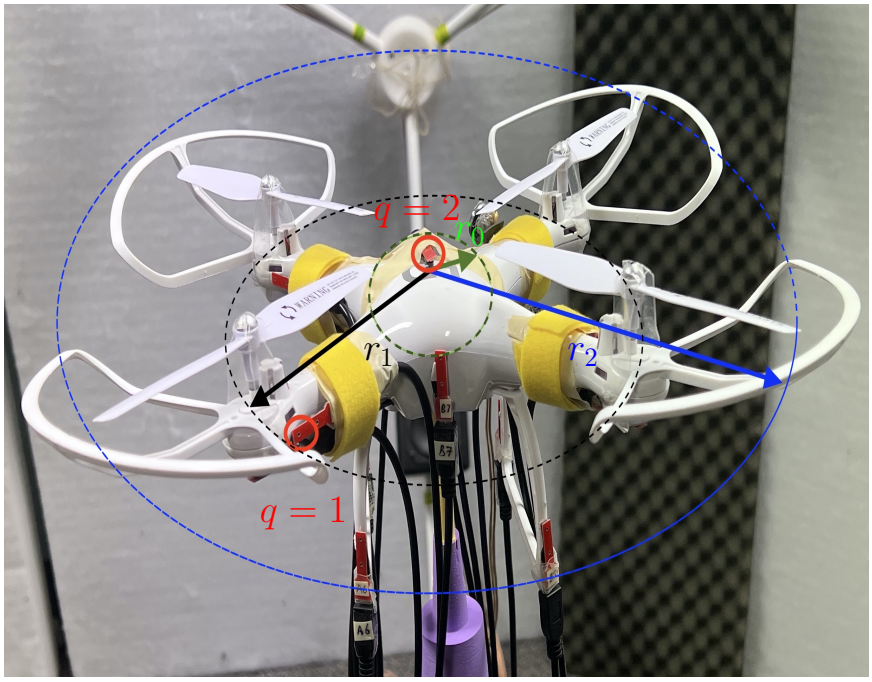
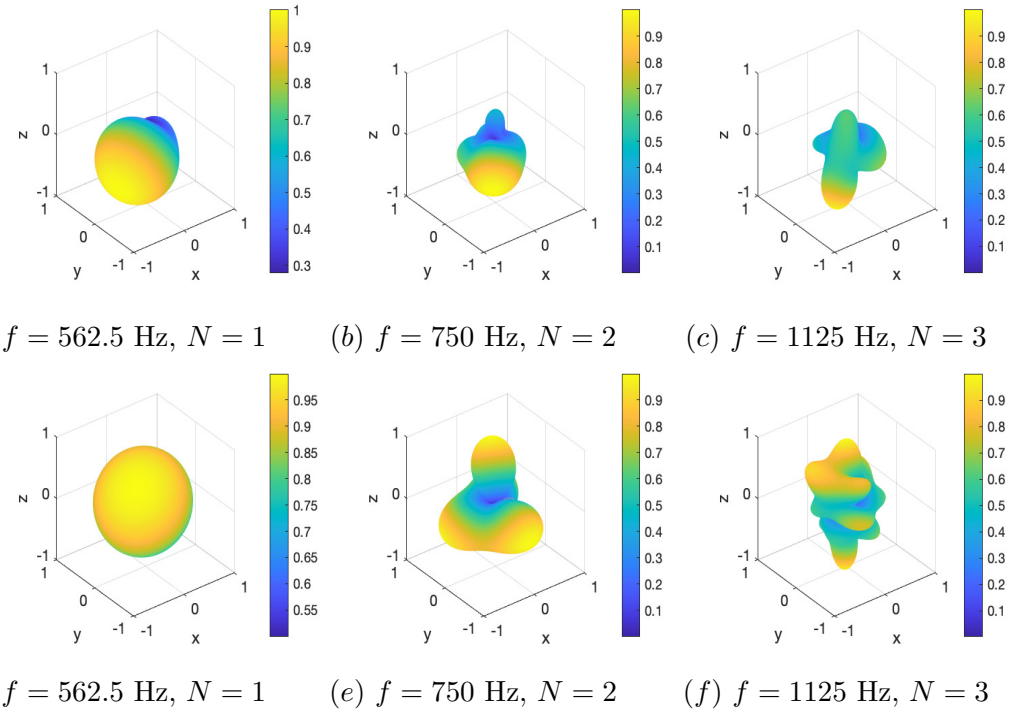


Figure 6.10: Microphone placements of $q \in \{1, 2\}$ on the drone with different radii based on its structure.

highlighted in Table. 6.5 from trial and error. Note that the DRTF model uses the smaller inner shell radius of 3.5 cm for $f < 600$ Hz, and the middle shell with a radius of 12 cm in the frequency range of $[750, 1125]$ Hz.

Table 6.5: List of number of truncation orders for the choice of radii on the drone with $r_0 = 3.5$ cm, $r_1 = 12$ cm, and $r_2 = 22.3$ cm.

f	λ	$\lceil kr_0 \rceil$	$\lceil kr_1 \rceil$	$\lceil kr_2 \rceil$
562.5 Hz	0.61 m	**1**	2	3
750 Hz	0.46 m	1	**2**	4
1125 Hz	0.30 m	1	**3**	5



(a) $f = 562.5$ Hz, $N = 1$ (b) $f = 750$ Hz, $N = 2$ (c) $f = 1125$ Hz, $N = 3$

(d) $f = 562.5$ Hz, $N = 1$ (e) $f = 750$ Hz, $N = 2$ (f) $f = 1125$ Hz, $N = 3$

Figure 6.11: Magnitude of the DRTFs for *rot-microphone* ((a), (b) and (c)) and *top-microphone* ((d), (e) and (f)).

Figure 6.11(a), (b), and (c) presents the magnitude of DRTFs of *rot-microphone* on the drone at $f = 562.5$ Hz, 750 Hz, and 1125 Hz, respectively. We observe similar directivity and orientation across all three DRTF representations. In Fig. 6.11(d), (e), and (f), we show the magnitude of the DRTFs for the *top-microphone* over frequency. We observe that the DRTF has a distribution in an upward direction, that is consistent in all three frequencies. However, in Fig. 6.11(e), more side lobes are seen to have picked up with higher order DRTF coefficients estimation at higher frequencies. The continuous representation of the DRTF inherits maximum truncation order error, however, this mode selection based on the radius is suitable for keeping the artifact to a minimum.

Figures 6.12 and 6.13 illustrate the proposed parameterisation method's accuracy across all measured points and using a set of measurement points as the ground truth while using the rest for parameterisation, respectively. Here, we use the lowest elevation angles θ_ν where $\nu \in [5, 10, 15, 20, 25]$ as the ground truth in Fig. 6.13. All three errors are observed to slightly raise with increasing frequency as shown in both Fig. 6.12 and 6.13. However, it is clear that at low frequencies below 600 Hz, with reconstruction errors being minimised, the proposed parameterisation method obtains accurate modelling of the DRTF.

However, we note that the above DRTF model can be mismatched and cause a reduction in performance in higher frequencies. We identify the following factors would impact the accuracy of the DRTF measurement process such as loudspeaker array placement error: measurements of dodecahedron loudspeaker (containing 30 loudspeakers in 1 m radius sphere), drone attached to a stand, room reverberation/non-anechoic DRTF measurements, and thermal noise in the microphone recordings. We observe that the results above the frequency of 1125 Hz were inconclusive from testing due to the aforementioned reasons, and therefore, we only model and discuss the results of the DRTF representation for the maximum frequency of 1125 Hz.

Here we only illustrate results at low frequencies where drone noise is primarily interfering and drone audition requires significant processing, we see that the drone's rigid body does not match the free-field model, and therefore, measuring and modeling of the DRTF is expected to be a viable and beneficial approach to future drone audition applications.

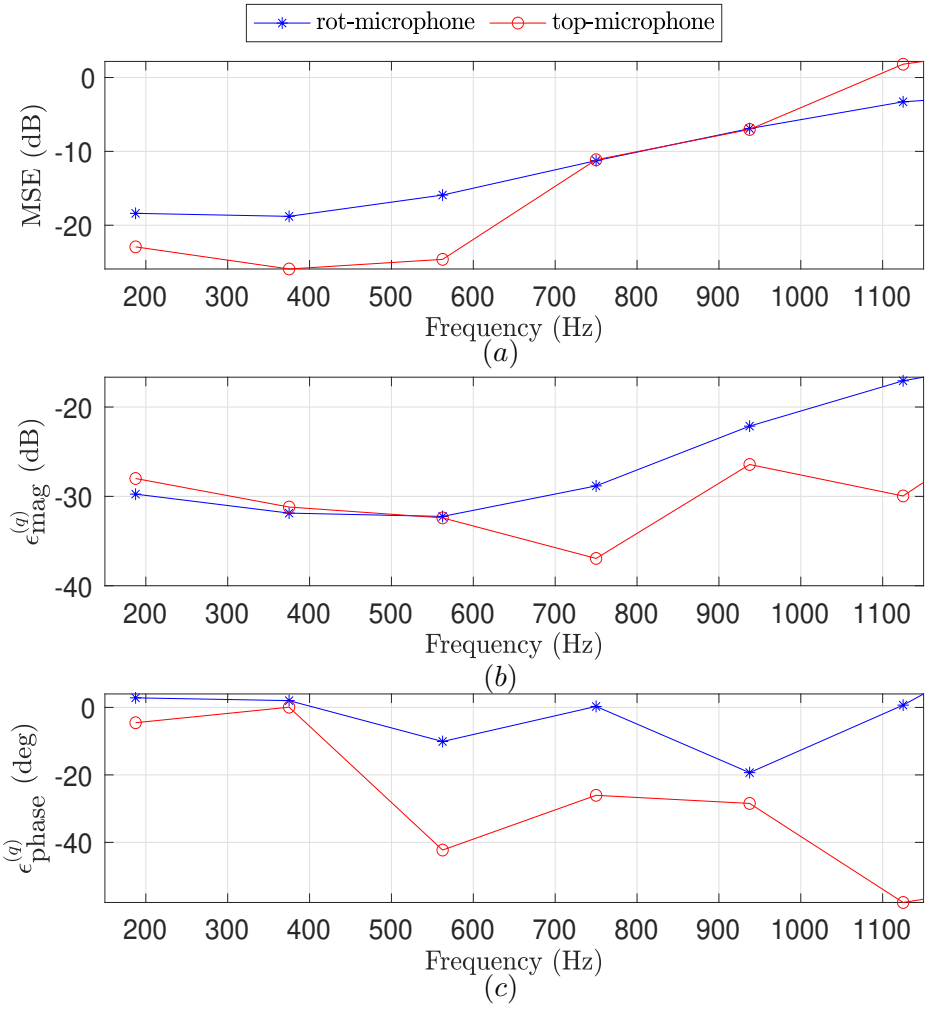


Figure 6.12: The experimental DRTF reconstruction error performance through the proposed parameter estimation method across the measurement points (a) $\epsilon^{(q)}$, (b) $\epsilon^{(q)}_{\text{mag}}$, and (c) $\epsilon^{(q)}_{\text{phase}}$, computed as in (6.16), (6.17), and (6.18), respectively, at the frequency range of $[0.2,\ 2]$ kHz for the q^{th} microphone where $q \in \{1, 2\}$

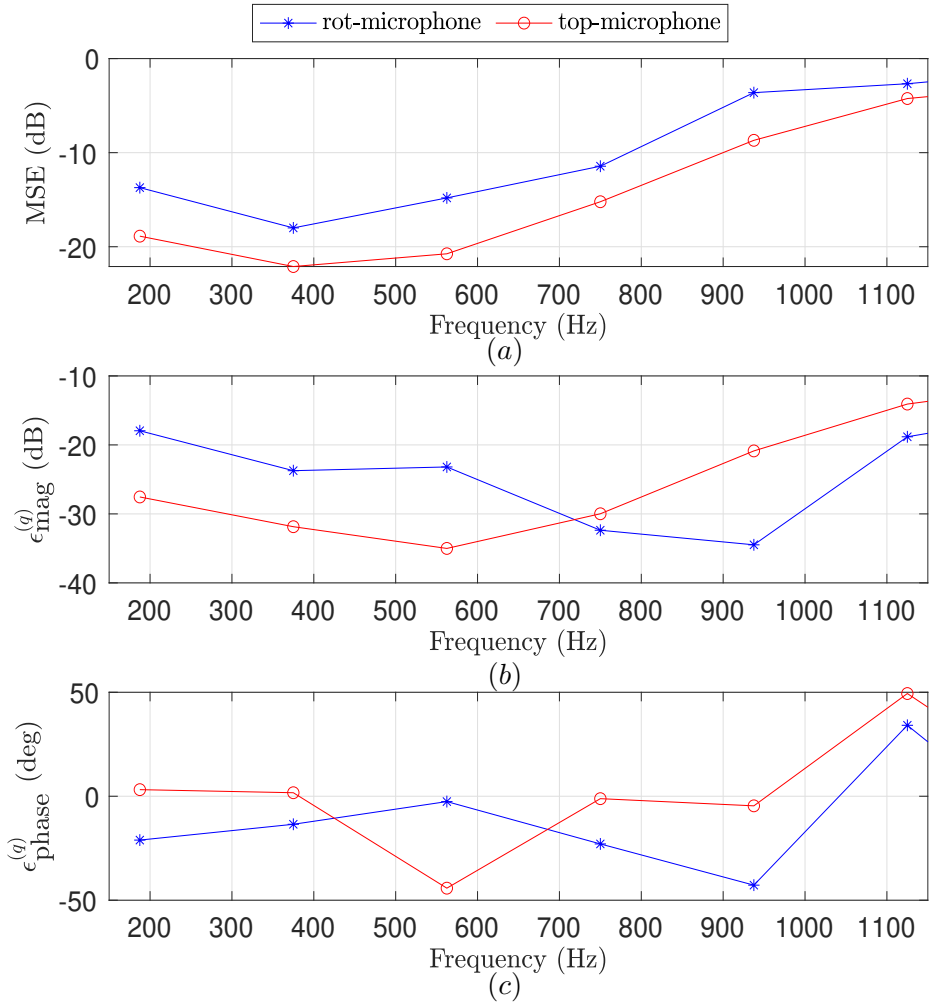


Figure 6.13: The experimental DRTF reconstruction error performance through the proposed parameter estimation method with respect to the ground truth (a) $\epsilon^{(q)}$, (b) $\epsilon^{(q)}_{\text{mag}}$, and (c) $\epsilon^{(q)}_{\text{phase}}$, computed as in (6.16), (6.17), and (6.18), respectively, at the frequency range of $[0.2,\ 2]$ kHz for the q^{th} microphone where $q \in \{1, 2\}$

6.7 Parameterisation of the Microphone Aligned DRTF

The previous sections considered the modelling of DRTF centred at the drone centre. The required number of measurements and the spherical harmonic order can be minimised by addressing the mismatch between the default origin and the microphone position. Therefore, in this section, we pre-process the DRTF measurements prior to modelling the DRTF, based on the microphone alignment with respect to the drone centre. The main contributions of this section when contrasted with Sections 6.3 [335] as follows: (i) a microphone-aligned continuous DRTF model (Section 6.7.1), whereas Section 6.3 proposed the DRTF around the drone centre; (ii) the experimental validation of the proposed parameterisation procedure using anechoic impulse response measurements (Section 6.7.2); and (iii) utilise the DRTF model on a real-life application in beamforming examples (Section 6.7.3).

6.7.1 Extraction of microphone-aligned DRTF coefficients

Here, we first parameterise the microphone-aligned DRTF that is expanded with respect to a given microphone of the drone-embedded microphone array to decompose into spherical harmonic components, and we then estimate its coefficients using a set of measurements by the multiple circular arrays [336].

Translation of the DRTF based on microphone alignment

Consider a source located at $\mathbf{y} = (x, y, z) = (r, \theta, \phi)$ from the origin of the spherical coordinate system $\mathbf{o} = (0, 0, 0)$ where $|\mathbf{y}| = r$ denotes the Euclidean distance between the origin and the source position, $\widehat{\mathbf{y}} = (\theta, \phi)$ denote the angular directions. Let $D^{(q)}(k, \mathbf{y}; \mathbf{x}^{(q)})$ be the microphone-aligned DRTF. With the equivalent ear-aligned HRTF [337], we write

$$D^{(q)}(k, \mathbf{y}; \mathbf{x}^{(q)}) = \mathcal{T}(\mathbf{y}; \mathbf{o} \rightarrow \mathbf{x}^{(q)}) D^{(q)}(k, \mathbf{y}; \mathbf{o}), \tag{6.24}$$

where the spherical wave translation operator for microphone aligned DRTF is given by

$$\mathcal{T}(\mathbf{y};\mathbf{o}\rightarrow\mathbf{x}^{(q)})=\frac{|\mathbf{y}-\mathbf{x}^{(q)}|}{|\mathbf{y}-\mathbf{0}|}e^{-ik(|\mathbf{y}-\mathbf{0}|-|\mathbf{y}-\mathbf{x}^{(q)}|)}, \tag{6.25}$$

and $|\cdot|$ is the ℓ_2-norm. Following the spherical harmonics decomposition in Section 6.3, the microphone-aligned DRTF coefficients are given by

$$d_{nm}^{(q)(\vec{\mathcal{L}})}(k)=\frac{1}{h_n^{(1)}(k|\mathbf{y}|)}\int_{\widehat{\mathbf{y}}}D^{(q)}(k,\mathbf{y};\mathbf{x}^{(q)})Y_{nm}^{*}(\widehat{\mathbf{y}})\,d\widehat{\mathbf{y}}, \tag{6.26}$$

where $(\cdot)_{nm}^{(q)(\vec{\mathcal{L}})}$ denotes harmonics decomposed about the q^{th} microphone position over L directions.

Summary of the coefficient estimation process

The implementation of the proposed parameterisation algorithm (as discussed in Section 6.4) comprises six steps as summarised below.

(i) Translate $D^{(q)}(k,\mathbf{y};\mathbf{o})$ using (6.24) to obtain the microphone-aligned DRTFs.

(ii) Re-write the microphone-aligned DRTFs using the expansion of spherical harmonic basis function:

$$D^{(q)}(k,\mathbf{y};\mathbf{x}^{(q)})=\sum_{n=0}^{N}\sum_{m=-n}^{n}d_{nm}^{(q)(\vec{\mathcal{L}})}(k)h_n^{(1)}(k|\mathbf{y}|)\mathcal{P}_{n|m|}(\cos\theta)E_m(\phi). \tag{6.27}$$

Consider $D^{(q)}(k,\mathbf{y}_\nu;\mathbf{x}^{(q)})$ on a circular aperture (r_ν,θ_ν) for $r=r_\nu$ and $\theta=\theta_\nu$ $(\nu=1,\ldots,\mathcal{V})$ that is parallel to the x-y plane, where $\mathbf{y}_\nu=(r_\nu,\theta_\nu,\phi)$. Multiply (6.27) by $E_{-m}(\phi)$ and integrate it with respect to ϕ over $[0,2\pi)$ to obtain the circular harmonics coefficients [324, 335]:

$$b_m^{(q)}(k,r_\nu,\theta_\nu)=\int_0^{2\pi}D^{(q)}(k,\mathbf{y}_\nu;\mathbf{x}^{(q)})E_{-m}(\phi)\,d\phi, \tag{6.28}$$

$$=\sum_{n=|m|}^{N}d_{nm}^{(q)(\vec{\mathcal{L}})}(k)h_n^{(1)}(k|\mathbf{y}_\nu|)\mathcal{P}_{n|m|}(\cos\theta_\nu), \tag{6.29}$$

for $n = |m|, |m|+1, \ldots, N$.

(iii) Approximate the integral in (6.29) by a summation:

$$b_m^{(q)}(k, r_\nu, \theta_\nu) \approx \frac{2\pi}{L_\nu} \sum_{v=1}^{L_\nu} D^{(q)}(k, \mathbf{y}_\nu; \mathbf{x}^{(q)}) E_{-m}(\phi_v), \tag{6.30}$$

where $L_\nu \geq (2N+1)$ denotes the number of sampling points on a given circle (r_ν, θ_ν).

(iv) Write (6.29) over $\mathcal{V} \geq (N+1)$[6] multiple circles to obtain a set of simultaneous equations given by

$$\mathbf{b}_m^{(q)} = \mathbf{H}_m \mathbf{d}_m^{(q)(\vec{\mathcal{L}})}, \text{for } m = -N, \ldots, N \tag{6.31}$$

where $\mathbf{b}_m^{(q)} = [b_{1m}^{(q)}, \ldots, b_{\mathcal{V}m}^{(q)}]^T$, $\mathbf{d}_m^{(q)(\vec{\mathcal{L}})} = [d_{|m|m}^{(q)(\vec{\mathcal{L}})}, d_{(|m|+1)m}^{(q)(\vec{\mathcal{L}})}, \ldots, d_{Nm}^{(q)(\vec{\mathcal{L}})}]^T$ and $\mathbf{H}_m = [\boldsymbol{h}_{m1}, \cdots, \boldsymbol{h}_{m\nu}, \cdots, \boldsymbol{h}_{m\mathcal{V}}]^T$ where $\boldsymbol{h}_{m\nu} = [h_{|m|}^{(1)}(k|\mathbf{y}_\nu|)\mathcal{P}_{|m||m|}(\cos\theta_\nu), \ldots, h_N^{(1)}(k|\mathbf{y}_\nu|)\mathcal{P}_{N|m|}(\cos\theta_\nu)]^T$ and $(\cdot)^T$ is the transpose operation.

(v) Solve (6.31) for each m using least squares method provided that $\mathbf{H}_m$ is nonsingular, as in [324, 335],

$$\mathbf{d}_m^{(q)(\vec{\mathcal{L}})} = \mathbf{H}_m^{\ddagger} \mathbf{b}_m^{(q)}, \text{for } m = -N, \ldots, N \tag{6.32}$$

where $(\cdot)^{\ddagger}$ denotes the Moore-Penrose inverse of a matrix.

(vi) Use $d_{nm}^{(q)(\vec{\mathcal{L}})}(k)$ from (6.15) to obtain the microphone-aligned continuous DRTF model similar to (6.8) and translate the microphone-aligned DRTF back to the origin as

$$D^{(q)}(k, \mathbf{y}; \mathbf{o}) = \mathcal{T}^{-1}(\mathbf{y}; \mathbf{x}^{(q)} \to \mathbf{o}) D^{(q)}(k, \mathbf{y}; \mathbf{x}^{(q)}), \tag{6.33}$$

[6]Note that we estimate $(N+1)^2$ WDRTF coefficients for the N^{th} order continuous WDRTF model.

where the inverse spherical wave translation operator is given by

$$\mathcal{T}^{-1}(\mathbf{y};\mathbf{x}^{(q)} \rightarrow \mathbf{o}) = \frac{|\mathbf{y}-\mathbf{0}|}{|\mathbf{y}-\mathbf{x}^{(q)}|} e^{-ik(|\mathbf{y}-\mathbf{x}^{(q)}|-|\mathbf{y}-\mathbf{0}|)}. \tag{6.34}$$

6.7.2 Simulation study

Here, we illustrate the proposed microphone-aligned DRTF model's performance evaluation and comparison in various simulation setups using anechoic impulse responses. Due to only having non-anechoic impulse response measurements for the drone (as discussed in Section 6.6), we use anechoic impulse response measurements from a head-worn device from the EasyCom dataset [291]. We model the wearable device related transfer function (WDRTF) similar to DRTF that incorporates scattering, and diffraction from the device and wearer's body to an incident soundfield, and evaluate the performance of the proposed parameterisation of the microphone aligned continuous DRTF model.

WDRTF measurements

Figure 6.14: The approximate microphone positions on the AR glasses are shown with the corresponding channel number. Channels 5 and 6 are binaural microphones.

We use the anechoic impulse responses from the EasyCom dataset that was introduced for the speech enhancement [291]. The impulse responses are measured from the spherical loudspeaker array to the KEMAR manikin with an AR glasses egocentric arbitrary six-channel microphone array (as shown in Fig. 6.14) in an anechoic room over $L = 1020$ directions with 17 different elevation angles and 60 different azimuth angles on a unit sphere. The measurement grid is shown in

Fig. 6.15. We perform the FFT with the FFT length of 256 samples on the impulse responses to obtain the WDRTFs.

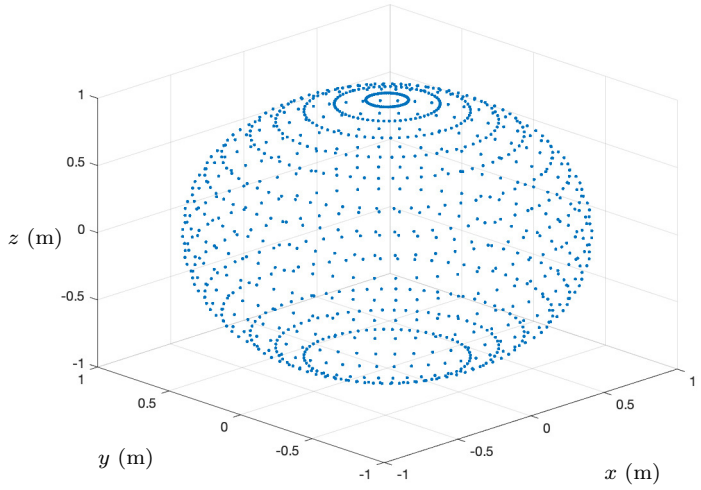


Figure 6.15: Impulse response measurement grid of the EasyCom dataset.

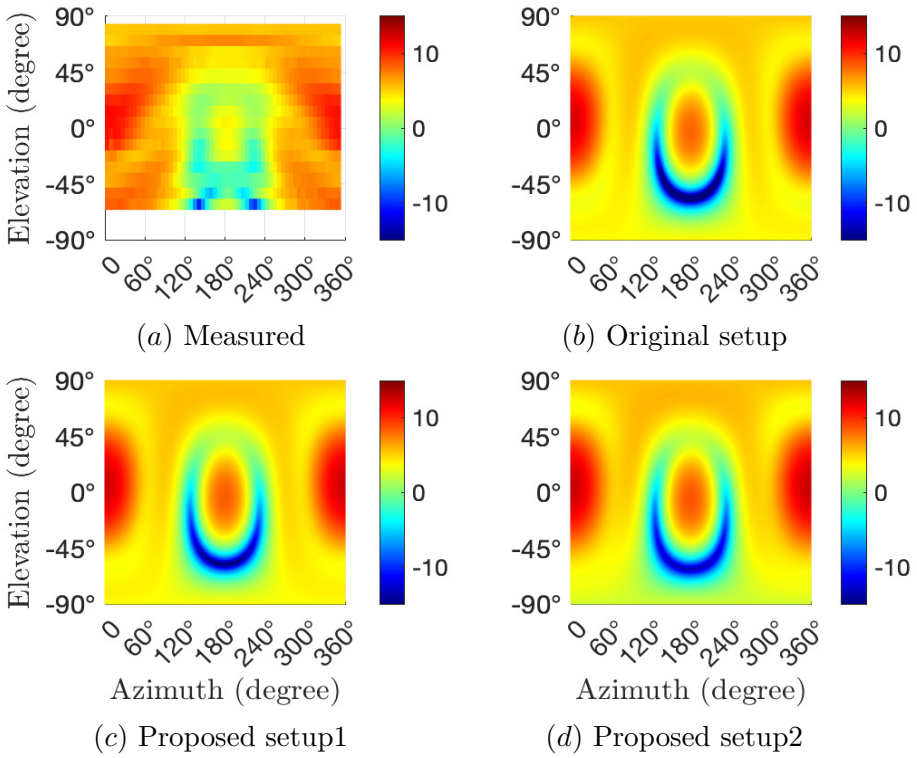
Table 6.6: Loudspeaker locations of the WDRTF measurement setups using multiple circular arrays.

Setup	Directions	Loudspeaker locations (θ_ν, ϕ)
Original setup	1020	$\theta_\nu \in \{6°, 15°, 24°, 33°, 42°, 54°, 63°, 72°, 81°, 90°, 99°, 108°, 117°, 126°, 138°, 147°, 156°\}$, $\phi \in \{0° : 6° : 354°\}$
Proposed setup1	300	$\theta_\nu \in \{15°, 33°, 42°, 63°, 72°, 90°, 108°, 126°, 138°, 156°\}$, $\phi \in \{0° : 12° : 348°\}$
Proposed setup2	210	$\theta_\nu \in \{15°, 42°, 63°, 90°, 126°, 156°\}$, $\phi \in \{0° : 12° : 348°\}$

We introduce two additional setups by arranging a set of WDRTF measurements in multiple circular array positions as depicted in Table 6.6. This way we can examine the proposed method over a fewer number of circles that cover different elevation angles.

Reconstruction performance evaluation and comparison

We illustrate the results of the continuous WDRTF representation of a particular microphone ($q = 2$) in (6.8) using the WDRTF coefficients derived through the proposed parameterisation method. Figure 6.16 shows the magnitude of the WDRTF at the frequency of 1125 Hz for $N = 2$. We observe that all three setups have similar WDRTF reconstruction with respect to the discrete measured WDRTF.



(a) Measured (b) Original setup

(c) Proposed setup1 (d) Proposed setup2

Figure 6.16: The magnitude of the WDRTF of $q = 2$ microphone at 1125 Hz over θ and ϕ using 2^{nd} order $d_{nm}^{(q)}$ using our WDRTF parameter estimation method.

We next investigate the WDRTF reconstruction error in Fig. 6.17 for frequency range up to 5 kHz. As expected, the consequence of the truncation order and the required dimensionality for the reconstruction, the measurement error is shown to be minimised in low frequencies for all three setups. Although the error is decreased in frequencies below 3 kHz for the 'proposed setup2', it is only gradually increasing with increased frequency. This suggests the MSE can be further reduced by increasing the number of multiple circles to include more elevation angles. The results confirm that both 'original setup' and 'proposed setup1' have very similar error patterns, but the 'original setup' needs much more measurements (additional 720 directions). We also evaluate the MSE of WDRTF model derived without microphone centering for the 'proposed setup1' and show errors are larger than that of the microphone-aligned WDRTFs. Observe that the magnitude WDRTF reconstruction error for the 'proposed setup1' is not significant at the higher frequencies (less than -10 dB), and thus the accuracy of the proposed WDRTF modelling approach based on microphone alignment confirms.

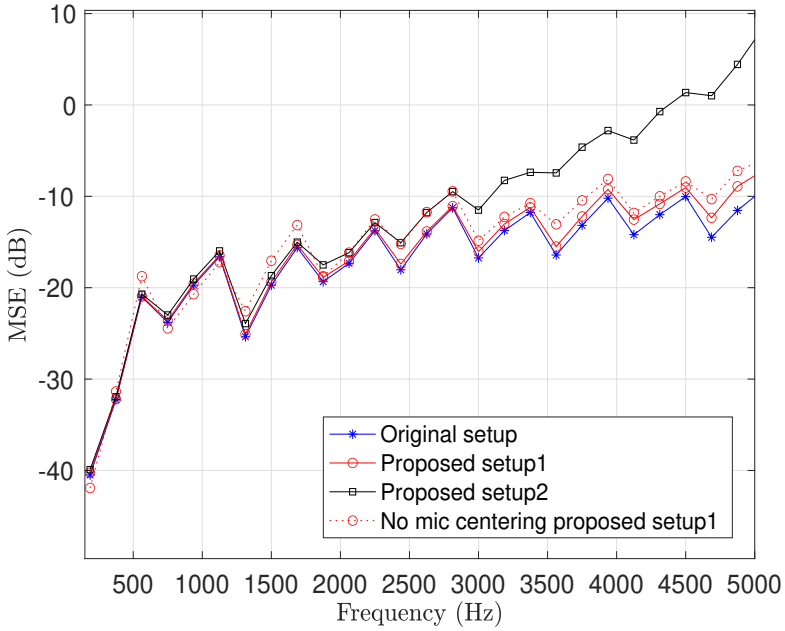


Figure 6.17: The microphone-aligned WDRTF measurement reconstruction error performance through the proposed method at the frequency range of $[0.2,\ 5]$ kHz of $q = 2$ microphone.

6.7.3 Application in speech enhancement

We utilise this WDRTF model and demonstrate its performance on a speech enhancement application. We use the real-life recordings from EasyCom dataset in which the AR glasses wearer (speaker no. 2) was sat down with two talkers (speakers no. 4 and 6) at a table surrounded by ten loudspeakers playing restaurant-like ambient noises as shown in Fig. 6.18. The recordings also contain both target Direction-of-Arrivals (DoAs) and device rotations. We only consider the four microphones mounted on the AR glasses (exclude the two in-ear microphones, e.g., $q = \{5, 6\}$). We downsample the audio signals to 10 kHz. We adopt (i) the MVDR [338] and (ii) the subspace hybrid (SS-Hybrid) as in [339] beamforming methods to extract the target speech active after 2 s by integrating the device rotation data to the target speaker DoAs.

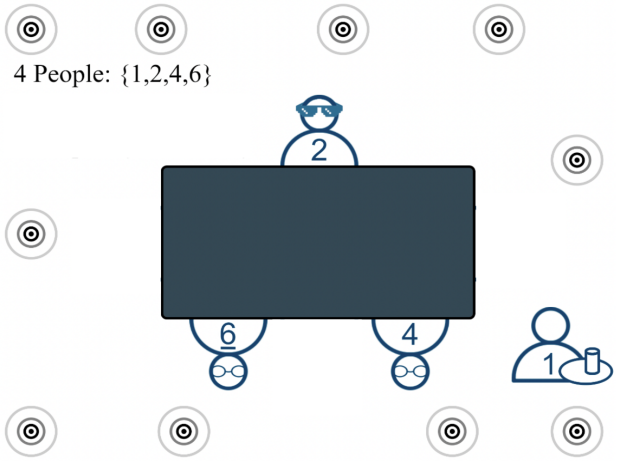


Figure 6.18: The recording setup is shown along with the participant ID numbers, and the concentric rings indicate approximate loudspeaker locations around the room.

Figure 6.19 shows the spectrograms of recorded, MVDR and SS-Hybrid output for two scenarios: (i) one speaker (4) active at a given time, and (ii) two talkers (4 and 6) are active simultaneously. We observe that SS-Hybrid suppresses more noise and enhances the target speech compared to the MVDR. However, the results confirm that both methods accurately extract the target source from the multichannel signals. The audio files can be found in GitHub[7].

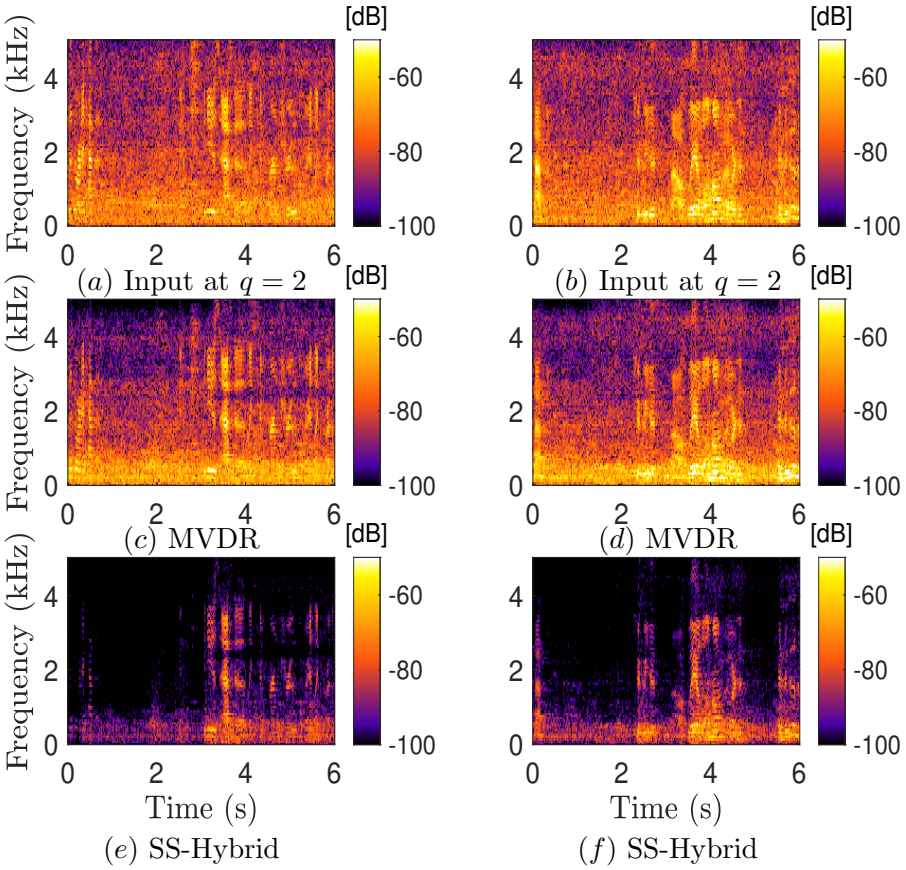


Figure 6.19: Spectrograms of beamforming examples. (a,c,e) Scenario 1: one speaker is active at a given time and (b,d,f) scenario 2: two talkers are active simultaneously.

[7] https://github.com/wnilmini/WDRTF

6.8 Summary and Contributions

In this chapter, we have proposed measuring and modelling a drone's rigid body acoustic response for the intention of extending drone audition processing beyond free-field assumptions as the DRTF. We re-define the microphone aligned DRTF to translate DRTF from the drone centre into microphone centre. We draw the following main conclusions from this chapter:

- A continuous DRTF representation was developed in both spatial and spectral domains using the spherical harmonic expansions.
- A practical method was described to obtain the DRTF coefficients using circular harmonic decomposition with a set of discrete source locations on multiple circular arrays parallel to the x-y plane and different elevations.
- Simulation results using a simple rigid sphere acting analogous to a drone show that the method has also improved reconstruction performance over freuqnecy.
- The DRTF has been parameterised to microphone centering and introduced microphone-aligned DRTF. It was shown to achieve good DRTF frequency and spatial components reconstruction.
- Furthermore, the DRTF model can utilise in beamforming to extract the target source signal.

In summary, we have proposed an efficient parameterisation of a continuous DRTF model using a practically viable method with a discrete set of measurements by multiple circles. The approach is more practical for traditional DRTF measurement setup. The proposed method models the scattering and diffraction response of a drone to an incident soundfield that can be used to enhance drone audition functions such as sound source localisation, and source enhancement.

6.9 Related Publications

Much of this chapter's work has been submitted in the following journal papers and conference proceedings.

- **W. N. Manamperi**, T. D. Abhayapala, L. I. Birnie, P. N. Samarasinghe, and J. Zhang, " Drone Audition: On Measurements and Modelling of Drone-Related Transfer Functions," submitted to *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, February 2023.
- **W. N. Manamperi**, T. D. Abhayapala and P. Holmberg, "Microphone Aligned Continuous Wearable Device-related Transfer Function: Efficient Modelling and Measurements," in *Hands-free Speech Communication and Microphone Arrays (HSCMA) Satellite Workshop* (Accepted), April 2024.

Chapter 7

Conclusions and Future Work

In this final chapter, we state the general conclusions drawn from this thesis. The summary of contributions can be found at the end of each chapter and are not repeated here. We also suggest some future research directions arising from this work.

7.1 Conclusions

This thesis examines spatial soundfield modelling, and spatial audio processing using on-board microphone array on a drone to support the deployment of acoustic scene analysis to facilitate applications of drone audition. The major contributions of this thesis are outlined in the following conclusions:

(i) An analysis model for acoustic characterisation of drone *harmonics* has been proposed in Chapter 3. The proposed current to noise transfer function is derived from the preliminary measurements considering a single motor propeller combination.

(ii) The acoustic source localisation using on-board microphone array on a drone has been answered in Chapter 4. The proposed algorithm exploits known characteristics of drone noise as introduced in Chapter 3 for better noise suppression and localises single and two sound sources in extremely low SdNR conditions with acceptable accuracy compared to the baseline. The method is simple and applicable for any microphone array geometry.

(iii) A multichannel audio signal enhancement framework from drone-embedded microphone array has been developed in Chapter 5. The proposed method has the capability to suppress a significant amount of drone noise from the recordings and works well for both speech and bird calls.

(iv) An efficient parameterisation of the drone related transfer function (DRTF) has been proposed in Chapter 6, which is useful in enhancing the drone audition functions. The DRTF modelling is performed using spherical harmonics. The DRTF measurements are conducted using a discrete sey of measurements by multiple loudspeaker placements on circular arrays. The method is a practically viable solution for typical impulse response measurement setups.

All of the above contributions can be collectively considered as a complete framework that gives the drone's listening capability for the successful deployment of drone audition. We have presented commercially viable and cost effective solutions that enable realistic integration to most drones. *Therefore, we believe that drone-embedded microphone array signal processing techniques will be an integral part of drone audition in delivering drone based services in the near future.*

7.2 Future Work

We list the following improvements and possible problems as potential future deployment to drone audition applications, and the spatial audio research field in general.

Future improvements

- *Harmonic amplitude modelling*: While the preliminary measurements are conducted to model the drone motor-related acoustic transfer function (Chapter 3), the strength behaviour of the drone noise *harmonics* needs to derive for suppressing the drone noise from known *harmonics*. We have shown that the harmonic strength appears to have a negatively underdamped sinusoidal structure from the initial measurements. Thus, we suggest to model these harmonic amplitudes as a sum of complex exponentials in the future.

- *Implementation in real-life scenarios*: We have conducted a few outdoor measurements to confirm our results in practice. Moreover, we have performed an extensive experimental study using indoor measurements inside a semi-anechoic chamber. Nevertheless, we suggest here to verify the performance of our algorithms using on-field measurements in more complex outdoor environments, e.g., for moving target sound sources, and different flight configurations of the drone.

- *Incorporate machine learning techniques*: In Chapter 3, we have shown drone noise can be accurately modelled using on-board microphones. We believe, having a large number of drone noise recordings, this can be efficiently done with machine learning approaches, even to include extrinsic characteristics such as natural wind conditions in the outdoor environment. The results of such a modelling approach would benefit generalising the algorithms proposed throughout this thesis. For example, the work presented in Chapter 4 can be further extended by applying the proposed cross-correlation based DOA estimation method between microphone pairs to train drone specific learning algorithm to predict noise for an accurate subtraction of the drone noise angular spectrum.

- *Accurate estimation of drone noise correlation matrix*: The proposed multichannel audio signal enhancement framework in Chapter 5 is sensitive to noise correlation matrix. We have shown that how one can use informed approaches such as prior-measured motor current-specific drone noise to derive the noise correlation matrix of the drone noise with speech enhancement. However, the framework heavily relies on the estimation accuracy of the second order statics. Therefore, it is required to have a noise correlation matrix estimation method for drones to improve the drone audition performance.

- *Applications of the DRTF*: We have conducted our DRTF measurements in a reverberant room. Hence, we have not performed the DRTF model validation in beamforming applications for drones. However, we have confirmed the proposed parameterisation approach using a simple rigid sphere acting as a drone. Moreover, we have performed model validation using head-worn

device anechoic measurements, and utilised the continuous DRTF model in beamforming application to extract the target speech in a noisy environment. Nevertheless, we suggest here to conduct the DRTF measurement approach in an anechoic environment by multiple circular loudspeaker array, and demonstrate the DRTF model performance in drone audition functions such as sound source localisation, and speech enhancement.

Future additions

- *Device-related transfer function modelling*: Throughout this thesis, we consider signal processing techniques for microphone arrays embedded/attached to the drone. In Chapter 6, we have developed a DRTF measurement procedure and parameterisation approach by circular array positions (desirable a simple setup with a single loudspeaker and height adjustable turn table.). This measurements and modelling approach has the potential usage in deriving a continuous device-related transfer function model in a variety of other acoustic applications. A few examples are robot auditions and wearable devices, e.g., AR glasses microphone arrays, and VR headsets, to improve audio capturing capabilities.

- *Signal enhancement in extremely low SNR scenarios*: Though the MWF has been well established for noise suppression in noisy environments, in Chapter 5, we have shown that the reformulation with the inverse of a noise correlation matrix cancels more noise compared to the traditional implementation by taking all microphones as the input. However, this implementation of the MWF outputs a multichannel signal. As shown in Appendix A, these multiple outputs of the MWF preserves spatial information. Thus, the combination of the enhanced signals for an optimal output is suggested for future work.

Appendix A

Analysis of Preservation of Spatial Cues in MWF

A.1 Introduction

This appendix analyses the preservation of spatial cues at the MWF from on-board microphone recordings on a drone at very low signal-to-drone noise ratio ($\overline{\text{SDNR}}$) levels related to Chapter 5 [340]. We first review the relative transfer function (ReTF). We then explore the spatial information of the desired sound sources such as human speakers and birds through experimental recordings from simulated and real-life environments. Finally, we demonstrate the effectiveness of the preservation of spatial cues at the MWF using an acoustic localisation application.

A.2 The Relative Transfer Function

The ReTF is a spatial function that represents the coupling between microphones in response to a desired source and is given by the ratio of acoustic transfer functions between a source and microphones [341–345] as

$$\mathcal{D}_{qq'}(f) = \frac{d_q(f)}{d_{q'}(f)}, \tag{A.1}$$

where $d_q(f)$ and $d_{q'}(f)$ are the transfer function between the source and q^{th} and q'^{th} microphone, respectively.

The ReTF between the two microphones captures the spatial properties of a sound source and is independent of the emitted sound source signals. Therefore, the ReTF preserves the spatial features of the source [344] and uses in sound source localisation algorithms [344].

A.3 Use of ReTF to Analyse the Preservation of Spatial Cues

Here, we estimate the ReTF using the commonly used covariance whitening method as it outperforms the covariance subtraction method [342–344].

By joint diagonalisation of the matrix pencil $\{\mathbf{\Phi}_{vv}(f,t), \mathbf{\Phi}_{pp}(f,t)\}$, we define the generalised eigenvalue decomposition of $\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{\Phi}_{pp}(f,t)$ [342] as

$$\mathbf{\Phi}_{vv}^{-1}(f,t)\mathbf{\Phi}_{pp}(f,t) = \mathbf{U}(f)\mathbf{\Lambda}(f)\mathbf{U}^H(f), \tag{A.2}$$

where $\mathbf{U}$ represents an orthogonal matrix comprising of the generalised eigenvectors, and $\mathbf{\Lambda}$ denotes a diagonal matrix with the generalised eigenvalues.

Now, considering the principal generalised eigenvector $\mathbf{u}_{\max}$ that corresponds to the maximal generalised eigenvalue in $\mathbf{\Lambda}$, we write the input ReTF at the MWF as in [344]

$$\mathbf{d}_{\text{in}}(f) = \frac{\mathbf{\Phi}_{vv}(f,t)\mathbf{u}_{\max}(f)}{\mathbf{e}_{q_0}^T\mathbf{\Phi}_{vv}(f,t)\mathbf{u}_{\max}(f)}, \tag{A.3}$$

where $\mathbf{e}_{q_0}$ is an all zero Q dimensional vector except for one in the q_0 position for $q_0 \in \{0, \cdots, Q\}$. We select q_0 to be the microphone with the maximum input $\overline{\text{SDNR}}$ level to obtain the ReTF vector.

Similar to (A.3), it can easily be shown that the output ReTF is given by

$$\mathbf{d}_{\text{out}}(f) = \frac{\mathbf{\Phi}_{\widehat{vv}}(f,t)\widehat{\mathbf{u}}_{\max}(f)}{\mathbf{e}_{q_0}^T\mathbf{\Phi}_{\widehat{vv}}(f,t)\widehat{\mathbf{u}}_{\max}(f)}, \tag{A.4}$$

where $\mathbf{\Phi}_{\widehat{v}\widehat{v}}(f,t)$ is the residual drone noise correlation matrix and is given by

$$\mathbf{\Phi}_{\widehat{v}\widehat{v}}(f,t) \triangleq E\{\widehat{\mathbf{v}}(f,t)\widehat{\mathbf{v}}^H(f,t)\}, \tag{A.5}$$

and $\widehat{\mathbf{v}}$ is the residual drone noise signal vector at the MWF output. Whereas $\widehat{\mathbf{u}}_{\max}(f)$ is the principal generalised eigenvector in $\widehat{\mathbf{U}}(f)$ that obtains from the generalised eigenvalue problem of

$$\mathbf{\Phi}_{\widehat{v}\widehat{v}}^{-1}(f,t)\mathbf{\Phi}_{yy}(f,t) = \widehat{\mathbf{U}}(f)\widehat{\mathbf{\Lambda}}(f)\widehat{\mathbf{U}}^H(f), \tag{A.6}$$

where

$$\mathbf{\Phi}_{yy}(f,t) \triangleq E\{\mathbf{y}(f,t)\mathbf{y}^H(f,t)\}, \tag{A.7}$$

and $\widehat{\mathbf{U}}(f)$, and $\widehat{\mathbf{\Lambda}}(f)$ corresponding generalised eigenvectors and eigenvalues matrix, respectively.

Note that the covariance whitening method utilises estimates of the microphone correlation matrices of both noisy source signals and noise-only signals for a better estimate of the ReTF. We discuss this second order statistics estimation more briefly for a practical scenario in Chapter 5.

A.4 Performance Measure

We examine the preservation of spatial cues at the MWF using quantitative evaluation. We evaluate the similarity between the input and output ReTFs using a distance measure based on the Hermitian angle as in [345]

$$\Theta(\mathbf{d}_{\text{in}}(f), \mathbf{d}_{\text{out}}(f)) = \arccos \frac{|\mathbf{d}_{\text{in}}^H(f)\mathbf{d}_{\text{out}}(f)|}{||\mathbf{d}_{\text{in}}(f)||_2||\mathbf{d}_{\text{out}}(f)||_2}, \tag{A.8}$$

where $\arccos(\cdot)$ denotes inverse of the cosine function, and $|\cdot|$ denotes the absolute value.

A.5 Experimental Validation

A.5.1 Measured impulse responses

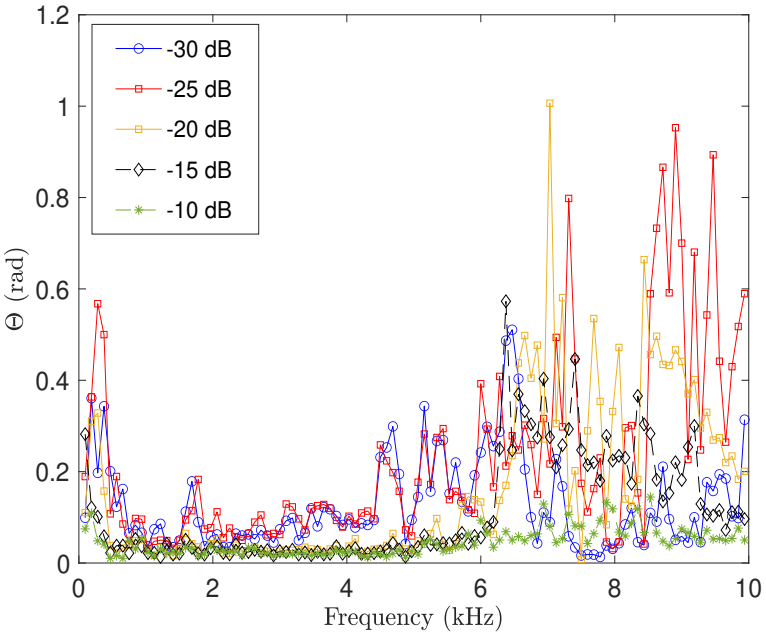


Figure A.1: The distance function: Hermitian angle Θ between the input and output ReTF vectors over frequencies at very low S$\overline{\text{D}}$NR levels.

We analyse the preservation of spatial cues of the desired speech source at the MWF using the experimental setup in Section 5.4.3. Figure A.1 depicts the calculated distance of the input ReTF to the output ReTF over different frequencies. We observe a smaller distance value throughout the wide frequency band under extreme S$\overline{\text{D}}$NR levels (≤ -10 dB). However, a larger distance can be observed above 6 kHz except for S$\overline{\text{D}}$NR level of -10 dB. Results indicate that the input and output ReTFs have a strong similarity and change more significantly above 6 kHz when the S$\overline{\text{D}}$NR levels gradually decrease. However, it also demonstrates a slightly lower distance value at the S$\overline{\text{D}}$NR level of -30 dB. For drone noise suppression and speech enhancement at the MWF, we are only interested in the speech frequency range below 6 kHz in accurately preserving the spatial features of the target sound

source.

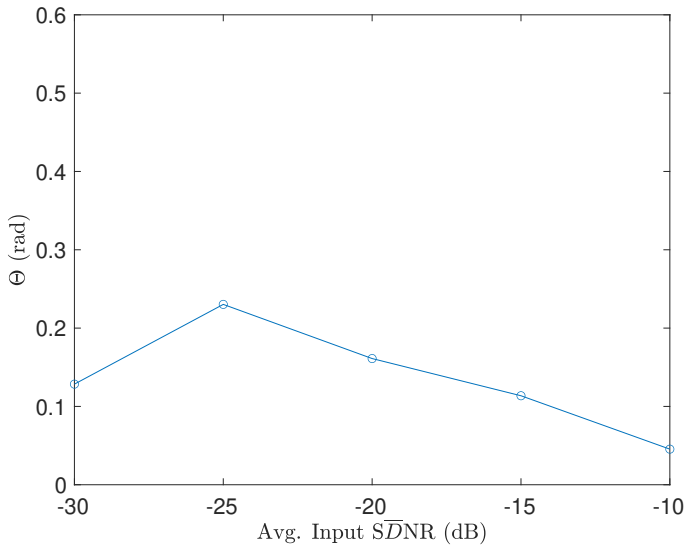


Figure A.2: Hermitian angle Θ between the input and output ReTF vectors averaged over frequency bins at very low SD̄NR levels.

We represent the average Hermitian angles over all frequencies at SD̄NR levels in Fig. A.2. When the distance measure is examined at different SD̄NR conditions, we again found that the performance is better with a smaller Hermitian angle below 0.3. This confirms that the ReTF is accurately estimated at both the input and output of the MWF, thus preserves the spatial cues of the target sound source.

A.5.2 Real-life environments

Figure A.3 exhibits the Hermitian angles over the frequencies for the experiment conducted in the outdoor settings in Section 5.4.2. We observe the distance measure between the input and output ReTFs fluctuates over the frequencies and the Hermitian angle stays below 0.6. Once more, we observe that the distance measure remains low, indicating the preservation of the sound source's spatial property. As expected, the performance of the simulations using the measured impulse responses slightly outperforms the real-world measurements.

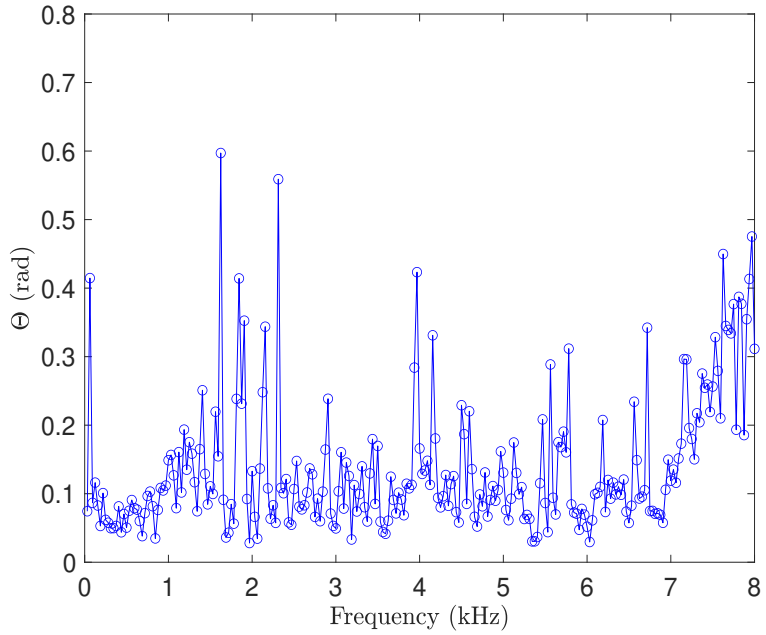


Figure A.3: The distance function: Hermitian angle Θ between the input and output ReTF vectors over frequencies.

A.6 Application in Acoustic Localisation and Conclusion

To conclude, we demonstrate the effectiveness of the preservation of spatial cues at the MWF using an acoustic localisation application. We evaluate the localisation accuracy of the input signals at the MWF using two baseline methods. We adopt (i) the generalised cross-correlation with phase transformation (GCC-PHAT) [171] and (ii) the noise angular spectrum subtraction method as in [13] (Chapter 4). In this context, we use the output signals at the MWF for the scenario in Section 5.4.3 and adopt the GCC-PHAT referred to here as GCC-PHAT$_{\text{MWF}}$ for the comparison.

We evaluate the localisation performance using the prediction error (Δ) in terms of the angular difference between the true source DOA and the estimated DOA from (4.12).

Table A.1 summarizes the localisation accuracy for a single sound source under lower $\overline{\text{SDNR}}$ levels varying from -30 dB to -10 dB over three DOA estimation

Table A.1: The prediction errors Δ over extreme S$\overline{\text{D}}$NR levels.

Method	S$\overline{\text{D}}$NR level (dB)				
	-10	-15	-20	-25	-30
GCC-PHAT	2.95°	2.95°	3.95°	148.01°	148.01°
GCC-PHAT$_{\text{MWF}}$	**0.95°**	**0.05°**	**0.05°**	**0.05°**	1.95°
[13]	1.05°	1.05°	2.05°	**0.05°**	**0.05°**

approaches: GCC-PHAT; GCC-PHAT$_{\text{MWF}}$; and [13]. We observe that all three methods localise the sound source with acceptable accuracy for S$\overline{\text{D}}$NR levels ≥ -20 dB. However, in extreme S$\overline{\text{D}}$NR conditions (below -20 dB), GCC-PHAT fails to localise the sound source as the prediction error is more significant due to the strong drone noise [13]. We observe that both the GCC-PHAT$_{\text{MWF}}$ and [13] obtain more accurate estimates of the target sound source with the prediction error of less than 3° for very low S$\overline{\text{D}}$NR conditions.

Thus, we conclude that the MWF preserves the spatial cues of the target sound source and accurately localises the sound source in drone-based environments using the baseline localisation approach.

Bibliography

[1] S. M. Shavarani, M. G. Nejad, F. Rismanchian, and G. Izbirak, "Application of hierarchical facility location problem for optimization of a drone delivery system: A case study of Amazon prime air in the city of San Francisco," *Int. J. Adv. Manuf. Technol.*, vol. 95, no. 9 - 12, pp. 3141 – 3153, Apr. 2018.

[2] 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. Robots and Syst.*, Oct. 2012, pp. 4737–4742.

[3] C. Corrado and K. Panetta, "Data fusion and unmanned aerial vehicles (UAVs) for first responders," in *Proc. IEEE Int. Symp. Technol. for homeland security*, Apr. 2017, pp. 1–6.

[4] H. E. de Bree, "Acoustic vector sensors increasing UAV's situational awareness," Tech. Rep. 2009-01-3249, SAE Technical Paper, Nov. 2009.

[5] L. Meier, P. Tanskanen, F. Fraundorfer, and M. Pollefeys, "Pixhawk: A system for autonomous flight using onboard computer vision," in *Proc. IEEE Int. Conf. Robot. and Automat.*, May 2011, pp. 2992 – 2997.

[6] D. Floreano and R. J. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, no. 7553, pp. 460 – 466, May 2015.

[7] J. C. Hodgson, S. M. Baylis, R. Mott, A. Herrod, and R. H. Clarke, "Precision wildlife monitoring using unmanned aerial vehicles," *Sci. Rep.*, vol. 6, no. 1, pp. 22574, Mar. 2016.

[8] TDK, "Sensors for drones," https://product.tdk.com/en/techlibrary/applicationnote/sensors-for-drones.html, Accessed: Sept. 2023.

[9] 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 [SP competitions]," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 138–144, Sep. 2019.

[10] A. Deleforge, "Drone audition for search and rescue: Datasets and challenges," in *Proc. QUIET DRONES Int. Symp. on UAV/UAS Noise*, Oct. 2020.

[11] M. Wakabayashi, H. G. Okuno, and M. Kumon, "Drone audition listening from the sky estimates multiple sound source positions by integrating sound source localization and data association," *Adv. Robot.*, vol. 34, no. 11, pp. 744–755, Jun. 2020.

[12] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Sep. 2014, pp. 1902–1907.

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

[14] 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. 2535, Nov. 2017.

[15] T. Morito, O. Sugiyama, R. Kojima, and K. Nakadai, "Partially Shared Deep Neural Network in sound source separation and identification using a UAV-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2016, pp. 1299–1304.

[16] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Sound source tracking by drones with microphone arrays," in *Proc. IEEE/SICE Int. Symp. on Syst. Integr.*, Jan. 2020, pp. 796–801.

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

[18] W. N. Manamperi, T. D. Abhayapala, P. N. Samarasinghe, and J. Zhang, "Drone audition: audio signal enhancement from drone embedded microphones using multichannel Wiener filtering and Gaussian-Mixture based post-filtering," *Appl. Acoust.*, vol. 216, pp. 109818, Jan. 2024, https://doi.org/10.1016/j.apacoust.2023.109818.

[19] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Nov. 2019, pp. 5320–5325.

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

[21] A. J. Torija and C. Clark, "A psychoacoustic approach to building knowledge about human response to noise of unmanned aerial vehicles," *Int. J. Environ. Res. Public Health*, vol. 18, no. 2, pp. 682, Jan. 2021.

[22] 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. Robots and Syst.*, Oct. 2018, pp. 1–8.

[23] 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. Robots and Syst.*, Oct. 2012, pp. 3288–3293.

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

[25] A. W. Christian and R. Cabell, "Initial investigation into the psychoacoustic properties of small unmanned aerial system noise," in *Proc. AIAA/CEAS aeroacoustics Conf.*, 2017, p. 4051.

[26] A. J. Torija, Z. Li, and R. H. Self, "Effects of a hovering unmanned aerial vehicle on urban soundscapes perception," *Transportation Research Part D: Transport and Environment*, vol. 78, pp. 102195, Jan. 2020.

[27] C. J. Hui, M. J. Kingan, Y. Hioka, G. Schmid, G. Dodd, K. N. Dirks, S. Edlin, S. Mascarenhas, and Y. Shim, "Quantification of the psychoacoustic effect of noise from small unmanned aerial vehicles," *Int. J. of Environ. Res. Public Health*, vol. 18, no. 17, pp. 8893, Aug. 2021.

[28] K. S. Brentner and F. Farassat, "Modeling aerodynamically generated sound of helicopter rotors," *Prog. Aerospace Sci.*, vol. 39, no. 2-3, pp. 83–120, Feb. 2003.

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

[30] D. H. Kim, C. H. Park, and Y. J. Moon, "Aerodynamic analyses on the steady and unsteady loading-noise sources of drone propellers," *Int. J. of Aeronautical and Space Sci.*, vol. 20, no. 3, pp. 611–619, Sep. 2019.

[31] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Estimating drone motor related acoustic transfer function: A preliminary investigation," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit and Conf. (APSIPA)*, Dec. 2020, pp. 156–160.

[32] O. Gur and A. Rosen, "Comparison between blade-element models of propellers," *The Aeronautical J.*, vol. 112, no. 1138, pp. 689–704, Dec. 2008.

[33] A W Christian, D D Boyd, N S Zawodny, and S A Rizzi, "Auralization of tonal rotor noise components of a quadcopter flyover," in *Proc. INTER-NOISE and NOISE-CON Congr. and Conf.*, Aug. 2015.

[34] D B Hanson and D J Parzych, "Theory for noise of propellers in angular inflow with parametric studies and experimental verification," Tech. Rep., NASA, Mar. 1993.

[35] T Zhou and R Fattah, "Tonal noise characteristics of two small-scale propellers," in *Proc. AIAA/CEAS aeroacoustics Conf.*, Jun. 2017, vol. 4054, p. 2017.

[36] M. Merchant and L. S. Miller, "Propeller performance measurement for low reynolds number UAV applications," in *AIAA aerospace sci. meeting and exhibit*, Jan. 2006, p. 1127.

[37] R. Cabell, F. Grosveld, and R. McSwain, "Measured noise from small unmanned aerial vehicles," in *Proc. INTER-NOISE and NOISE-CON Congr. and Conf.*, Jun. 2016, vol. 252, pp. 345–354.

[38] T. Zhou, H. Jiang, Y. Sun, R. J. Fattah, X. Zhang, B. Huang, and L. Cheng, "Acoustic characteristics of a quad-copter under realistic flight conditions," in *Proc. AIAA/CEAS aeroacoustics Conf.*, 2019, p. 2587.

[39] W. N. Alexander, J. Whelchel, N. Intaratep, and A. Trani, "Predicting community noise of sUAS," in *Proc. AIAA/CEAS aeroacoustics Conf.*, 2019, p. 2686.

[40] W. N. Alexander and J. Whelchel, "Flyover noise of multi-rotor sUAS," in *Proc. INTER-NOISE and NOISE-CON Congr. and Conf.*, Sep. 2019, vol. 259, pp. 2548–2558.

[41] J. Whelchel, W. N. Alexander, and N. Intaratep, "Propeller noise in confined anechoic and open environments," in *AIAA Scitech 2020 Forum*, 2020, p. 1252.

[42] N. S. Zawodny, A. Christian, and R. Cabell, "A summary of nasa research exploring the acoustics of small unmanned aerial systems," *AHS Technical Meeting on Aeromechanics Design for Transformative Vertical Flight*, , no. NF1676L-27827, Jan. 2018.

[43] D. L. Huff and B. S. Henderson, "Electric motor noise for small quadcopters: Part 1–acoustic measurements," in *Proc. AIAA/CEAS aeroacoustics Conf.*, 2018, p. 2952.

[44] B. S. Henderson and D. Huff, "Electric motor noise for small quadcopters: Part ii-source characteristics and predictions," in *Proc. AIAA/CEAS aeroacoustics Conf.*, 2018, p. 2953.

[45] N. Intaratep, W. N. Alexander, W. J. Devenport, S. M. Grace, and A. Dropkin, "Experimental study of quadcopter acoustics and performance at static thrust conditions," in *Proc. AIAA/CEAS aeroacoustics Conf.*, 2016, p. 2873.

[46] C. E. Tinney and J. Sirohi, "Multirotor drone noise at static thrust," *AIAA J.*, vol. 56, no. 7, pp. 2816–2826, Jul. 2018.

[47] N. S. Zawodny and D. D. Boyd, "Investigation of rotor–airframe interaction noise associated with small-scale rotary-wing unmanned aircraft systems," *J. of the Amer. Helicopter Soc.*, vol. 65, no. 1, pp. 1–17, Jan. 2020.

[48] R. S. McKay, M. J. Kingan, S. T. Go, and R. Jung, "Experimental and analytical investigation of contra-rotating multi-rotor UAV propeller noise," *Appl. Acoust.*, vol. 177, pp. 107850, Jun. 2021.

[49] R. J. Fattah, W. Chen, H. Wu, Y. Wu, and X. Zhang, "Noise measurements of generic small-scale propellers," in *Proc. AIAA/CEAS aeroacoustics Conf.*, 2019, p. 2498.

[50] N. S. Zawodny, D. D. Boyd Jr, and C. L. Burley, "Acoustic characterization and prediction of representative, small-scale rotary-wing unmanned aircraft system components," in *Amer. Helicopter Soc. (AHS) Annu. Forum*, May 2016.

[51] Y. Yang, Y. Liu, Y. Li, E. Arcondoulis, and Y. Wang, "Aerodynamic and aeroacoustic performance of an isolated multicopter rotor during forward flight," *AIAA J.*, vol. 58, no. 3, pp. 1171–1181, Mar. 2020.

[52] 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. 2013.

[53] S. O. Afari, *Prediction of Noise Associated with an Isolated UAV Propeller*, M.S. thesis, Dept. Aerosp. Eng., Embry-Riddle Aeronautical Univ., Daytona Beach, Florida, 2019, [Online]. Available: https://commons.erau.edu/cgi/viewcontent.cgi?article=1463&context=edt.

[54] M. T. Kotwicz Herniczek, D. Feszty, S. A. Meslioui, J. Park, and F. Nitzsche, "Evaluation of acoustic frequency methods for the prediction of propeller noise," *AIAA J.*, vol. 57, no. 6, pp. 2465–2478, Jun. 2019.

[55] W. A. Anemaat, M. Schuurman, W. Liu, and A. A. Karwas, "Aerodynamic design, analysis and testing of propellers for small unmanned aerial vehicles," in *Proc. AIAA Aerospace Sci. Meeting*, 2017, p. 0721.

[56] J. A. Feight, S. Whyte, J. D. Jacob, and R. J. Gaeta, "Acoustic characterization of a multi-rotor UAS as a first step towards noise reduction," in *Proc. AIAA Aerospace Sci. Meeting*, 2017, p. 1174.

[57] H. Bi, F. Ma, T. D. Abhayapala, and P. N. Samarasinghe, "Spherical array based drone noise measurements and modelling for drone noise reduction via propeller phase control," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2021, pp. 286–290.

[58] S. Al-Emadi, A. Al-Ali, A. Mohammad, and A. Al-Ali, "Audio based drone detection and identification using deep learning," in *Proc. Int. Wireless Commun. and Mobile Comput. Conf.*, Jun. 2019, pp. 459 – 464.

[59] A. Bernardini, F. Mangiatordi, E. Pallotti, and L. Capodiferro, "Drone detection by acoustic signature identification," *Electron. Imaging*, , no. 10, pp. 60 – 64, Jan. 2017.

[60] A. A. Cabrera-Ponce, J. Martinez-Carranza, and C. Rascon, "Detection of nearby UAVs using CNN and Spectrograms," in *Proc. Int. Micro Air Vehicle Conf. and Competition*, Oct. 2019.

[61] S. Al-Emadi, A. Al-Ali, A. Mohammad, and A. Al-Ali, "Audio based drone detection and identification using deep learning," in *Proc. Int. Wireless Comm. & Mobile Computing Conf.*, Jun. 2019, pp. 459–464.

[62] Z. Lu, M. Debiasi, and B. C. Khoo, "Acoustic characteristics of a multi-rotor MAV and its noise reduction technology," in *Proc. INTER-NOISE and NOISE-CON Congr. and Conf.*, Aug. 2016, vol. 253, pp. 393 – 403.

[63] Y. Hioka, M. Kingan, G. Schmid, R. McKay, and K. A. Stol, "Design of an unmanned aerial vehicle mounted system for quiet audio recording," *Appl. Acoust.*, vol. 155, no. 7, pp. 423–427, Dec. 2019.

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

[65] M. B. Andra, B. P. A. Rohman, and T. Usagawa, "Feasibility evaluation for keyword spotting system using mini microphone array on uav," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 2264–2267.

[66] 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. Assoc. Annu. Summit and Conf. (APSIPA)*, Nov. 2019, pp. 1885–1892.

[67] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic source localization from multirotor UAVs," *IEEE Trans. Ind. Electron.*, , no. 10, pp. 8618–8628, Oct. 2019.

[68] O. Ruiz-Espitia, J. Martinez-Carranza, and C. Rascon, "Aira-uas: an evaluation corpus for audio processing in unmanned aerial system," in *Proc. Int. Conf. Unmanned Aircr. Syst.*, Jun. 2018, pp. 836–845.

[69] M. Clayton, L. Wang, A. McPherson, and A. Cavallaro, "An embedded multichannel sound acquisition system for drone audition," *IEEE Sensors J.*, May 2023.

[70] Dotterel Technologies, "Enabling Audio in High Noise Environments," https://www.dotterel.com/, Accessed: Sept. 2023.

[71] C. Rascon, O. Ruiz-Espitia, and J. Martinez-Carranza, "On the use of the AIRA-UAS corpus to evaluate audio processing algorithms in unmanned aerial systems," *Sensors*, vol. 19, no. 18, pp. 3902, Sep. 2019.

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

[73] T. Ishiki and M. Kumon, "Design model of microphone arrays for multi-rotor helicopters," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Sep./Oct. 2015, pp. 6143–6148.

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

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

[76] 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.

[77] 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, Feb. 2017.

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

[79] P. Marmaroli, X. Falourd, and H. Lissek, "A UAV motor denoising technique to improve localization of surrounding noisy aircrafts: proof of concept for anti-collision systems," in *Proc. Acoust.*, Apr. 2012, pp. 1–6.

[80] A. Schmidt and W. Kellermann, "Informed ego-noise suppression using motor data-driven dictionaries," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* May 2019, pp. 116–120, IEEE.

[81] A. Schmidt, H. W. Löllmann, and W. Kellermann, "A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* Apr. 2018, pp. 6583–6587, IEEE.

[82] A. Schmidt, A. Deleforge, and W. Kellermann, "Ego-noise reduction using a motor data-guided multichannel dictionary," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2016, pp. 1281–1286.

[83] T. Tezuka, T. Yoshida, and K. Nakadai, "Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2014, pp. 6293–6298.

[84] 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. Consumer Electron.*, Jan. 2015, pp. 26–29.

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

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

[87] T. Morito, O. Sugiyama, R. Kojima, and K. Nakadai, "Partially shared deep neural network in sound source separation and identification using a uav-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2016, pp. 1299–1304.

[88] L. Wang and A. Cavallaro, "Deep learning assisted time-frequency processing for speech enhancement on drones," *IEEE Trans. on Emerging Topics in Computational Intelligence*, vol. 5, no. 6, pp. 871–881, Aug. 2020.

[89] C. Chun, K. M. Jeon, T. Kim, and W. Choi, "Drone noise reduction using deep convolutional autoencoder for UAV acoustic sensor networks," in *Proc.*

IEEE Int. Conf. Mobile Ad Hoc and Sensor Syst. Workshops, Nov. 2019, pp. 168–169.

[90] H. Bi, F. Ma, T. D. Abhayapala, and P. N. Samarasinghe, "Spherical sector harmonics based directional drone noise reduction," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, Sep. 2022, pp. 1–5.

[91] H. Bi, T. D. Abhayapala, F. Ma, and P. N. Samarasinghe, "Spherical sector harmonics based soundfield radial extrapolation and robustness analysis," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Jun. 2023, pp. 1–5.

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

[93] T. Ishiki, K. Washizaki, and M. Kumon, "Evaluation of microphone array for multirotor helicopters," *J. of Robot. and Mechatronics*, vol. 29, no. 1, pp. 168–176, Feb. 2017.

[94] B. Yen, Y. Li, and Y. Hioka, "Rotor noise-aware noise covariance matrix estimation for unmanned aerial vehicle audition," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, Jun. 2023.

[95] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 692–730, Jan. 2017.

[96] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition," in *Proc. Eurospeech*, 2001, p. 1012–1016.

[97] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F. Stöter, M. Hu, and J. M. Martín-Doñas, "Asteroid: the pytorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, May 2020.

[98] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, May 2020.

[99] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 4, pp. 1409–1415, Dec. 2011.

[100] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, Jun. 2021.

[101] H. G. Okuno, K. Nakadai, K. I. Hidai, H. Mizoguchi, and H. Kitano, "Human-robot interaction through real-time auditory and visual multiple-talker tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2001, vol. 3, pp. 1402–1409.

[102] K. Nakadai, H. Nakajima, M. Murase, H. G. Okuno, Y. Hasegawa, and H. Tsujino, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2006, pp. 852–859.

[103] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition hark and its evaluation," in *IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, Dec. 2008, pp. 561–566.

[104] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2003, vol. 2, pp. 1147–1152.

[105] J. M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2003, vol. 2, pp. 1228–1233.

[106] Kai Wu and Andy WH Khong, "Sound source localization and tracking," *Context Aware Human-Robot and Human-Agent Interaction*, pp. 55–78, 2016.

[107] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Process.*, Sept. 2012, pp. 1–6.

[108] J. Sanchez-Riera, X. Alameda-Pineda, J. Wienke, A. Deleforge, S. Arias, J. Čech, S. Wrede, and R. Horaud, "Online multimodal speaker detection for humanoid robots," in *Proc. IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, Nov. 2012, pp. 126–133.

[109] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 2015.

[110] N. Yalta, K. Nakadai, and T Ogata, "Sound source localization using deep learning models," *J. Robot. and Mechatronics*, vol. 29, no. 1, pp. 37–48, Feb. 2017.

[111] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano, "Active audition system and humanoid exterior design," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2000, vol. 2, pp. 1453–1461.

[112] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, pp. 1–19, Dec. 2006.

[113] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

[114] M. R. Bai, S. Lan, and J. Huang, "Time difference of arrival (TDOA)-based acoustic source localization and signal extraction for intelligent audio classification," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, Jul. 2018, pp. 632–636.

[115] K. Wu and A. W. H. Khong, "Acoustic source tracking in reverberant environment using regional steered response power measurement," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit and Conf. (APSIPA)*, Oct. 2013, pp. 1–6.

[116] D. Liu and A. W. Khong, "A wavelet-based GCC prefiltering algorithm for speech DOA estimation," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, 2010.

[117] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, and R. Horaud, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *Proc. IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, Oct. 2013, pp. 203–210.

[118] R. Lebarbenchon, E. Camberlein, D. Di Carlo, C. Gaultier, A. Deleforge, and N. Bertin, "Evaluation of an open-source implementation of the srp-phat algorithm within the 2018 locata challenge," *arXiv preprint arXiv:1812.05901*, 2018.

[119] K. Chen, J. Geiger, W. Jin, M. Taghizadeh, and W. Kellermann, "Robust phase replication method for spatial aliasing problem in multiple sound sources localization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2017, pp. 195–199.

[120] K. Nakadai, H. G. Okuno, and H. Kitano, "Epipolar geometry based sound localization and extraction for humanoid audition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2001, vol. 3, pp. 1395–1401.

[121] U. Kim, K. Nakadai, and H. G. Okuno, "Improved sound source localization in horizontal plane for binaural robot audition," *Appl. Intell.*, vol. 42, pp. 63–74, Jan. 2015.

[122] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition.," in *Proc. INTERSPEECH*, Sept. 2002.

[123] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. antennas and Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[124] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[125] X. Li, S. Yan, X. Ma, and C. Hou, "Spherical harmonics MUSIC versus conventional MUSIC," *Appl. Acoust.*, vol. 72, no. 9, pp. 646–652, Sep. 2011.

[126] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Jul. 2014.

[127] L. Kumar and R. M. Hegde, "Near-field acoustic source localization and beamforming in spherical harmonics domain," *IEEE Signal Process. Mag.*, vol. 64, no. 13, pp. 3351–3361, Mar. 2016.

[128] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*, pp. 157–180. Springer, Jan. 2001.

[129] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2009, pp. 221–224.

[130] V. V. Reddy, B. P. Ng, and A. W. H. Khong, "Insights into MUSIC-like algorithm," *IEEE Trans. on Signal Process.*, vol. 61, no. 10, pp. 2551–2556, Mar. 2013.

[131] Y. Liu, Z. Tan, A. W. H. Khong, and H. Liu, "Joint source localization and association through overcomplete representation under multipath propagation environment," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 5123–5127.

[132] K. Chen, W. Jin, and B. Desikan, "A signal subspace rotation method for localization of multiple wideband sound sources," *arXiv preprint arXiv:1906.08847*, Jun. 2019.

[133] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2009, pp. 664–669.

[134] Y. X. Zou, P. Wang, Y. Q. Wang, C. H. Ritz, and J. Xi, "Speech enhancement with an acoustic vector sensor: an effective adaptive beamforming and post-filtering approach," *EURASIP J. Audio, Speech, and Music Process.*, vol. 2014, no. 1, pp. 1–12, Dec. 2014.

[135] Y. Zou, B. Li, and C. H. Ritz, "Multi-source DOA estimation using an acoustic vector sensor array under a spatial sparse representation framework," *Circuits Syst. Signal Process*, vol. 35, pp. 993–1020, Mar. 2016.

[136] Y. Jin, Y. Zou, and C. H. Ritz, "Robust speaker DOA estimation based on the inter-sensor data ratio model and binary mask estimation in the bispectrum domain," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Mar. 2017.

[137] Y. Zou, Y. Guo, W. Zheng, C. H. Ritz, and J. Xi, "An effective DOA estimation by exploring the spatial sparse representation of the inter-sensor data ratio model," in *Proc. IEEE China Summit & Int. Conf. on Signal and Info. Process. (ChinaSIP)*, Jul. 2014, pp. 42–46.

[138] J. Zhao and C. Ritz, "Semi-adaptive beamforming for co-prime circular microphone arrays," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit and Conf. (APSIPA)*, Dec. 2020, pp. 426–431.

[139] M. Jia, J. Sun, C. Bao, and C. Ritz, "Multiple-to-single sound source localization by applying single-source bins detection," *Appl. Acoust.*, vol. 138, pp. 28–38, Sep. 2018.

[140] V. V. Reddy, A. W. H. Khong, and B. P. Ng, "Unambiguous speech DOA estimation under spatial aliasing conditions," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 2133–2145, Jul. 2014.

[141] K. Wu, V. G. Reju, and A. W. H. Khong, "Multisource DOA estimation in a reverberant environment using a single acoustic vector sensor," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 10, pp. 1848–1859, Jun. 2018.

[142] M. Jia, Y. Wu, C. Bao, and C. Ritz, "Multi-source DOA estimation in reverberant environments by jointing detection and modeling of time-frequency points," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 29, pp. 379–392, Dec. 2020.

[143] J. Zhao and C. Ritz, "Co-prime circular microphone arrays and their application to direction of arrival estimation of speech sources," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2019, pp. 800–804.

[144] Y. X. Zou, W. Shi, B. Li, C. H. Ritz, M. Shujau, and J. Xi, "Multisource DOA estimation based on time-frequency sparsity and joint inter-sensor data ratio with single acoustic vector sensor," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2013, pp. 4011–4015.

[145] M. Shujau, C. H. Ritz, and I. S. Burnett, "Designing acoustic vector sensors for localisation of sound sources in air," in *Proc. Eur. Signal Process. Conf.*, Aug. 2009, pp. 849–853.

[146] K. Wu, V. G. Reju, and A. W. H. Khong, "Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 2015, pp. 444–448.

[147] K. Wu, S. T. Goh, and A. W. H. Khong, "Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2013, pp. 365–369.

[148] K. Wu, V. G. Reju, A. W. H. Khong, and S. T. Goh, "Swarm intelligence based particle filter for alternating talker localization and tracking using microphone arrays," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 6, pp. 1384–1397, Apr. 2017.

[149] V. V. Reddy, B. P. Ng, Y. Zhang, and A. W. H. Khong, "DOA estimation of wideband sources without estimating the number of sources," *Signal Process.*, vol. 92, no. 4, pp. 1032–1043, Apr. 2012.

[150] V. G. Reju, R. S. Rashobh, A. H. T. Nguyen, and A. W. H. Khong, "An efficient multi-source DOA estimation algorithm for underdetermined system," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, Sept. 2018, pp. 86–90.

[151] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 38, no. 7, pp. 1132–1143, Jul. 1990.

[152] H. Ye and D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Signal Process. Mag.*, vol. 43, no. 4, pp. 938–949, Apr. 1995.

[153] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Signal Process. Mag.*, vol. 50, no. 8, pp. 1843–1854, Aug. 2002.

[154] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 10, pp. 1997–2012, Aug. 2017.

[155] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 8, pp. 1913–1928, Apr. 2010.

[156] V. V. Reddy, B. P. Ng, and A. W. H. Khong, "Derivative-constrained frequency-domain wideband DOA estimation," *Multidimensional Syst. and Signal Process.*, vol. 25, pp. 211–233, Jan. 2014.

[157] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *Int. J. Neural Syst.*, vol. 25, no. 01, pp. 1440003, Feb. 2015.

[158] A. Deleforge and R. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction*, Mar. 2012, pp. 431–438.

[159] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Mar. 2016, pp. 405–409.

[160] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Selected Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Dec. 2018.

[161] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Selected Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, Feb. 2019.

[162] R. Lee, M. Kang, B. Kim, K. Park, S. Q. Lee, and H. Park, "Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments," *IEEE Access*, vol. 8, pp. 7373–7382, Jan. 2020.

[163] W. Zhang, Y. Zhou, and Y. Qian, "Robust DOA Estimation Based on Convolutional Neural Network and Time-Frequency Masking," in *Proc. INTERSPEECH*, Sep. 2019, pp. 2703–2707.

[164] V. Tourbabin, J. Donley, B. Rafaely, and R. Mehra, "Direction of arrival estimation in highly reverberant environments using soft time-frequency mask,"

in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2019, pp. 383–387.

[165] J. Zhao and C. Ritz, "Coprime microphone arrays for estimating speech direction of arrival using deep learning," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit and Conf. (APSIPA)*, Dec. 2021, pp. 1–8.

[166] J. Zhao and C. Ritz, "Adapting GCC-PHAT to Co-Prime Circular Microphone Arrays for Speech Direction of Arrival Estimation Using Neural Networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit and Conf. (APSIPA)*, Nov. 2022, pp. 815–819.

[167] W. Q. Zheng, Y. X. Zou, and C. Ritz, "Spectral mask estimation using deep neural networks for inter-sensor data ratio model based robust DOA estimation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 2015, pp. 325–329.

[168] A. Deleforge, F. Forbes, and R. Horaud, "Variational em for binaural sound-source separation and localization," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2013, pp. 76–80.

[169] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950–1960, Aug. 2012.

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

[171] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[172] Thomas Padois, Olivier Doutres, and Franck Sgard, "On the use of modified phase transform weighting functions for acoustic imaging with the generalized

cross correlation," *J. Acoust. Soc. Amer.*, vol. 145, no. 3, pp. 1546–1555, Mar. 2019.

[173] L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Mar. 2017, pp. 496–500.

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

[175] B. Yen and Y. Hioka, "Noise power spectral density scaled SNR response estimation with restricted range search for sound source localisation using unmanned aerial vehicles," *EURASIP J. Audio, Speech, and Music Process.*, , no. 1, pp. 1–26, Dec. 2020.

[176] Jeonghwan Choi and Joon-Hyuk Chang, "Convolutional neural network-based direction-of-arrival estimation using stereo microphones for drone," in *Proc. Int. Conf. on Electron., Info., Commun.*, Jan. 2020, pp. 1–5.

[177] W. Jin, M. J. Taghizadeh, K. Chen, and W. Xiao, "Multi-channel noise reduction for hands-free voice communication on mobile phones," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Mar. 2017.

[178] W. Jin, B. Desikan, A. Kumar, and K. Chen, "Multi-channel noise reduction with interference suppression on mobile phones," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, Sept. 2018, pp. 201–205.

[179] Z. Zhang, J. Geiger, J. Pohjalainen, A. E. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Trans. on Intell. Syst. and Techn. (TIST)*, vol. 9, no. 5, pp. 1–28, Apr. 2018.

[180] J. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. G. Okuno, "Robust recognition of simultaneous speech by a mobile robot," *IEEE Trans. on Robot.*, vol. 23, no. 4, pp. 742–752, Aug. 2007.

[181] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno, "Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory," in *Proc. IEEE Int. Conf. on Robot. Automat.*, Apr. 2004, vol. 2, pp. 1517–1523.

[182] Shipra J Arora and Rishi Pal Singh, "Automatic speech recognition: a review," *Int. J. Comput. Appl.*, vol. 60, no. 9, Jan. 2012.

[183] P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, V. Tourbabin, and T. Lunner, "An introduction to the speech enhancement for augmented reality (spear) challenge," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, Sept. 2022, pp. 1–5.

[184] H. Beit-On, M. Lugasi, L. Madmoni, A. Menon, A. Kumar, J. Donley, V. Tourbabin, and B. Rafaely, "Audio signal processing for telepresence based on wearable array in noisy and dynamic scenes," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 8797–8801.

[185] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound separation on the texai telepresence system using hark," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2011, pp. 2130–2137.

[186] H. W. Löllmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Speech Comm.; 11. ITG Symp. Proc. of VDE*, Sep. 2014, pp. 1–4.

[187] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind source separation with parameter-free adaptive step-size method for robot audition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 6, pp. 1476–1485, Oct. 2009.

[188] S. Yamamoto, J. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. G. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *Proc. IEEE Int. Conf. Robot. Automat.*, Apr. 2005, pp. 1477–1482.

[189] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[190] M. Parchami, W. P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time fourier transform domain," *IEEE Circuits Syst. Mag.*, vol. 16, no. 3, pp. 45–77, Aug. 2016.

[191] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 1997, vol. 2, pp. 1167–1170.

[192] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[193] H. Gustafsson, S. E Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[194] F. Kallel, M. Frikha, M. Ghorbel, A. B. Hamida, and C. Berger-Vachon, "Dual-channel spectral subtraction algorithms based speech enhancement dedicated to a bilateral cochlear implant," *Appl. Acoust.*, vol. 73, no. 1, pp. 12–20, Jan. 2012.

[195] S. Vihari, A. S. Murthy, P. Soni, and D. C. Naik, "Comparison of speech enhancement algorithms," *Procedia computer science*, vol. 89, pp. 666–676, Jan. 2016.

[196] P. Vary and M. Eurasip, "Noise suppression by spectral magnitude estimation—mechanism and theoretical limits—," *Signal Process.*, vol. 8, no. 4, pp. 387–400, Jul. 1985.

[197] Y. Hu, M. Bhatnagar, and P. C. Loizou, "A cross-correlation technique for enhancing speech corrupted with correlated noise," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2001, vol. 1, pp. 673–676.

[198] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2007.

[199] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 1, pp. 163–176, Dec. 2005.

[200] Y. Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 5, pp. 457–465, Aug. 2003.

[201] Y. Hu and P. C. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 270–273, Jan. 2004.

[202] C. H. You, S. N. Koh, and S. Rahardja, "β order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 4, pp. 475–486, Jun. 2005.

[203] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[204] B. Liu, V. G. Reju, A. W. H. Khong, and V. V. Reddy, "A GMM post-filter for residual crosstalk suppression in blind source separation," *IEEE Signal Process. Letters*, vol. 21, no. 8, pp. 942–946, Apr. 2014.

[205] A. H. T. Nguyen, V. G. Reju, and A. W. H. Khong, "Directional sparse filtering for blind estimation of under-determined complex-valued mixing matrices," *IEEE Trans. Signal Process.*, vol. 68, pp. 1990–2003, Mar. 2020.

[206] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[207] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Jun. 2010.

[208] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[209] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jun. 2007.

[210] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Eigenvector-based speech mask estimation for multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 12, pp. 2162–2172, Sep. 2019.

[211] R. Mars, V. G. Reju, A. W. H. Khong, Y. Hioka, and K. Niwa, "Beamforming techniques using microphone arrays," in *Academic Press Library in Signal Process., Volume 7*, pp. 585–612. Elsevier, Jan. 2018.

[212] J. Donley, V. Tourbabin, B. Rafaely, and R. Mehra, "Adaptive multi-channel signal enhancement based on multi-source contribution estimation," in *Proc. Eur. Signal Process. Conf.*, Aug. 2021, pp. 276–280.

[213] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 2, pp. 617–631, Jan. 2007.

[214] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[215] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 12, pp. 2449–2464, Oct. 2015.

[216] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ica-based blocking matrix for improved noise estimation," *EURASIP J. Adv. Signal Process.*, vol. 2014, pp. 1–24, Dec. 2014.

[217] J. M. Martín-Doñas, A. M. Peinado, I. López-Espejo, and A. Gomez, "Dual-channel speech enhancement based on extended kalman filter relative transfer function estimation," *Appl. Sci.*, vol. 9, no. 12, pp. 2520, Jun. 2019.

[218] R. Mars, V. G. Reju, and A. W. H. Khong, "A frequency-invariant fixed beamformer for speech enhancement," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit and Conf. (APSIPA)*, Dec. 2014, pp. 1–6.

[219] S. Hadei and M. Iotfizad, "A family of adaptive filter algorithms in noise cancellation for speech enhancement," *arXiv preprint arXiv:1106.0846*, Jun. 2011.

[220] M. Djendi and R. Bendoumia, "A new efficient two-channel backward algorithm for speech intelligibility enhancement: A subband approach," *Appl. Acoust.*, vol. 76, pp. 209–222, Feb. 2014.

[221] R. Bendoumia and M. Djendi, "Acoustic noise reduction by new two-channel proportionate forward symmetric adaptive decorrelating algorithms in sparse systems," *Appl. Acoust.*, vol. 137, pp. 69–81, Aug. 2018.

[222] P. G. Thaitangam, R. Laishram, K. U. Devi, R. Khwairakpam, M. R. Singh, and C. Oinam, "Speech enhancement using adaptive filter with bat algorithm," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, Dec. 2018, pp. 1–5.

[223] R. Henni, M. Djendi, and M. Djebari, "A new efficient two-channel fast transversal adaptive filtering algorithm for blind speech enhancement and acoustic noise reduction," *Comput. & Elect. Eng.*, vol. 73, pp. 349–368, Jan. 2019.

[224] Osamu Hoshuyama, Akihiko Sugiyama, and Akihiro Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using con-

strained adaptive filters," *IEEE Signal Process. Mag.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[225] M. Shujau, C. H. Ritz, and I. S. Burnett, "Speech enhancement via separation of sources from co-located microphone recordings," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Mar. 2010, pp. 137–140.

[226] S. Choi, A. Cichocki, H. Park, and S. Lee, "Blind source separation and independent component analysis: A review," *Neural Info. Processing-Letters and Reviews*, vol. 6, no. 1, pp. 1–57, Jan. 2005.

[227] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[228] W. Lu and J. C. Rajapakse, "Approach and applications of constrained ica," *IEEE trans. Neural Networks*, vol. 16, no. 1, pp. 203–212, Jan. 2005.

[229] J. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," in *IEE proc. F (radar and signal process.)*, Dec. 1993, vol. 140, pp. 362–370.

[230] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 2, pp. 715–726, Jan. 2007.

[231] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Aug. 2004.

[232] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Int. Conf. on Independent Component Analysis and Blind Signal Separation*, Mar. 2006, pp. 601–608.

[233] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive bss of short mixtures by ica recursively regularized across frequencies," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 3, pp. 624–639, Jun. 2010.

[234] X. Zheng, C. Ritz, and J. Xi, "Collaborative blind source separation using location informed spatial microphones," *IEEE signal process. letters*, vol. 20, no. 1, pp. 83–86, Nov. 2012.

[235] M. Jia, J. Sun, C. Bao, and C. Ritz, "Separation of multiple speech sources by recovering sparse and non-sparse components from b-format microphone recordings," *Speech Commun.*, vol. 96, pp. 184–196, Feb. 2018.

[236] M. Shujau, C. H. Ritz, and I. S. Burnett, "Separation of speech sources using an acoustic vector sensor," in *Proc. IEEE 13th Int. Workshop on Multimedia Signal Process.*, Oct. 2011, pp. 1–6.

[237] B. Liu, V. G. Reju, and A. W. H. Khong, "A linear source recovery method for underdetermined mixtures of uncorrelated AR-model signals without sparseness," *IEEE Trans. on Signal Process.*, vol. 62, no. 19, pp. 4947–4958, Jun. 2014.

[238] A. Deleforge and W. Kellermann, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 2015, pp. 355–359.

[239] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Jun. 2013.

[240] B. Wang and M. D. Plumbley, "Musical audio stream separation by non-negative matrix factorization," in *Proc. DMRN summer conf.*, Jul. 2005, pp. 23–24.

[241] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization.," in *Proc. Interspeech*, Sep. 2006, vol. 2, pp. 2–5.

[242] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low snr conditions via noise estimation using sparse and low-rank nmf with

kullback–leibler divergence," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, Apr. 2015.

[243] S. U.N. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind speech separation and enhancement with gcc-nmf," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 745–755, Jan. 2017.

[244] S. U. N. Wood and J. Rouat, "Unsupervised low latency speech enhancement with rt-gcc-nmf," *IEEE J. of Selected Topics in Signal Process.*, vol. 13, no. 2, pp. 332–346, Apr. 2019.

[245] Y. Feng and C. Ritz, "Single-channel speech separation by including spectral structure information within non-negative matrix factorization," in *Proc. IEEE China Summit and Int. Conf. on Signal and Info. Process. (ChinaSIP)*, Jul. 2015, pp. 620–624.

[246] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Sound source separation of moving speakers for robot audition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 2009.

[247] R. Takeda, K. Nakada, K. Komatani, T. Ogata, and H. G. Okuno, "Exploiting known sound source signals to improve ica-based robot audition in speech separation and recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2007, pp. 1757–1762.

[248] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Aug. 2014.

[249] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[250] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 3, pp. 483–492, Dec. 2015.

[251] H. S. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *Proc. IEEE Int. Conf. on Learn. Representations*, Sept. 2018.

[252] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.

[253] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 2018, pp. 5069–5073.

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

[255] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2019, pp. 6875–6879.

[256] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 29, pp. 2840–2849, Jul. 2021.

[257] N. Saleem, M. Irfan Khattak, M. Y. Ali, and M. Shafi, "Deep neural network for supervised single-channel speech enhancement," *Arch. Acoust.*, vol. 44, 2019.

[258] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Jun. 2021, pp. 3415–3419.

[259] Z. Tan, A. H. T. Nguyen, Y. Liu, and A. W. H. Khong, "Multichannel noise reduction using dilated multichannel U-net and pre-trained single-channel network," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 266–270.

[260] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders," *arXiv preprint arXiv:1905.01209*, May 2019.

[261] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "Time-domain ad-hoc array speech enhancement using a triple-path network," *arXiv preprint arXiv:2110.11844*, Oct. 2021.

[262] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "Tparn: Triple-path attentive recurrent network for time-domain multichannel speech enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 6497–6501.

[263] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "Multichannel speech enhancement without beamforming," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 6502–6506.

[264] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[265] Y. Li, B. Yen, and Y. Hioka, "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.*, Aug. 2021, vol. 263, pp. 3584–3594.

[266] L. Wang and A. Cavallaro, "Deep learning assisted time-frequency processing for speech enhancement on drones," *IEEE Trans. Emerg. Topics in Comput. Intell.*, vol. 5, no. 6, pp. 871–881, Aug. 2020.

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

[268] B. Yen, Y. Hioka, G. Schmid, and B. Mace, "Multi-sensory sound source enhancement for unmanned aerial vehicle recordings," *Appl. Acoust.*, vol. 189, no. 108590, Feb. 2022.

[269] Y. Kim and E. Kim, "CNN based Complex Spectrogram Enhancement in Multi-Rotor UAV Environments," *J. Korea Inst. Info. and Comm. Eng.*, vol. 24, no. 4, pp. 459–466, 2020.

[270] Y. Li, B. Yen, and Y. Hioka, "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*, Jun. 2022.

[271] P. Misra, A. A. Kumar, P. Mohapatra, and P. Balamuralidhar, "Droneears: Robust acoustic source localization with aerial drones," in *Proc. IEEE Int. Conf. Robot. and Automat.*, May 2018, pp. 80–85.

[272] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

[273] W. William and K. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Amer.*, vol. 97, pp. 3907–3908, Jun. 1995.

[274] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *J. Audio Eng. Soc.*, vol. 55, no. 7/8, pp. 623–637, Jul. 2007.

[275] M. Zhang, W. Zhang, R. A. Kennedy, and T. D. Abhayapala, "HRTF measurement on KEMAR manikin," in *Proc. Acoust.*, Nov. 2009, vol. 9.

[276] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 300–321, 1995.

[277] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights into head-related transfer function: Spatial dimensionality and continuous representation," *J. Acoust. Soc. Amer.*, vol. 127, no. 4, pp. 2347–2357, Apr. 2010.

[278] P. M. Morse and K. U. Ingard, *Theoretical acoustics*, Princeton university press, 1986.

[279] R. Duraiswami and N. A. Gumerov, "Method for measurement of head related transfer functions," May 18 2010, U.S. Patent 7,720,229.

[280] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov, "Fast head-related transfer function measurement via reciprocity," *J. Acoust. Soc. Amer.*., vol. 120, no. 4, pp. 2202–2215, Oct. 2006.

[281] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Proc. of 16th Intl. Audio Eng. Soc. Conf. on Spatial Sound Reproduction*, Mar. 1999.

[282] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE J of Selected Topics in Signal Process.*, vol. 9, no. 5, pp. 921–930, Apr. 2015.

[283] D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Broadband DOA estimation using sensor arrays on complex-shaped rigid bodies," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 8, pp. 1573–1585, Mar. 2013.

[284] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 11, pp. 2046–2058, 2015.

[285] D. N. Zotkin, N. A. Gumerov, and R. Duraiswami, "Incident field recovery for an arbitrary-shaped scatterer," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Mar. 2017, pp. 451–455.

[286] A. Bastine, L. Birnie, T. D. Abhayapala, P. Samarasinghe, and V. Tourbabin, "Ambisonics capture using microphones on head-worn device of arbitrary geometry," in *Proc. Eur. Signal Process. Conf.*, Aug. 2022, pp. 309–313.

[287] Y. Asahara, K. Matsuda, H. Nakajima, and K. Nakadai, "A fourier series based data compression model for acoustic transfer function," in *Proc. IEEE/SICE Int. Symp. on Syst. Integr.*, Jan. 2020, pp. 664–668.

[288] Y. Sudo, M. Takigahira, H. Tsuru, K. Nakadai, and H. Nakajima, "Online adaptation of fourier series based acoustic transfer function model to improve sound source localization and separation," https://www.researchgate.net/publication/373633090_Online_Adaptation_of_Fourier_Series_Based_Acoustic_Transfer_Function_Model_to_Improve_Sound_Source_Localization_and_Separation, Accessed: Sept. 2023.

[289] K. Nakamura, S. Ambrose, and K. Nakadai, "On-the-spot calibration of microphone array transfer functions for robot audition," in *Proc. IEEE Int. Conf. Robot. and Automat.*, May 2015, pp. 3354–3359.

[290] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Tsujino, "Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots," *Speech Comm.*, vol. 44, no. 1-4, pp. 97–112, Oct. 2004.

[291] J. Donley, V. Tourbabin, J. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv preprint arXiv:2107.04174*, Jul. 2021.

[292] W. Jin, W. B. Kleijn, Y. Lang, and P. Grosche, "System and method for evaluating an acoustic transfer function," aug 2016, US Patent App. 15/142,063.

[293] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*, Academic Press, London, UK, 1999.

[294] T. K. Sarkar and O. Pereira, "Using the matrix pencil method to estimate the parameters of a sum of complex exponentials," *IEEE Antennas and Propag. Mag.*, vol. 37, no. 1, pp. 48–55, Feb. 1995.

[295] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, Society for Ind. and Appl. Math., Jan. 1995.

[296] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Letters*, vol. 14, no. 5, pp. 337–340, Apr. 2007.

[297] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 1995, pp. 81–84.

[298] J. Klapel, *Acoustic measurements with a quadcopter: Embedded system implementations for recording audio from above*, M.S. thesis, Institutt for elektronikk og telekommunikasjon, 2014.

[299] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Signal Process. Mag.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[300] S. L. Marple Jr and W. M. Carey, *Digital spectral analysis with applications*, Upper Saddle River, NJ: Prentice-Hall, Inc., 1989.

[301] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 2, pp. 260–276, Jun. 2009.

[302] I. Cohen, J. Benesty, and S. Gannot, *Speech processing in modern communication: challenges and perspectives*, Springer Science & Business Media, 2009.

[303] C. Zheng, A. Deleforge, X. Li, and W. Kellermann, "Statistical analysis of the multichannel wiener filter using a bivariate normal distribution for sample covariance matrices," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 5, pp. 951–966, Jan. 2018.

[304] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 5, pp. 1368–1381, Nov. 2010.

[305] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2006, pp. 1–4.

[306] G. H. Golub and C. F. Van Loan, *Matrix computations*, JHU press, Feb. 2013.

[307] S. Chehresa and M. H. Savoji, "MMSE speech enhancement based on GMM and solving an over-determined system of equations," in *Proc. IEEE Int. Symp. Intell. Signal Process.*, Sep. 2011, pp. 1–5.

[308] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[309] C. S. Doire, M. Brookes, P. A. Naylor, C. M. Hicks, D. Betts, M. A. Dmour, and S. H. Jensen, "Single-channel online enhancement of speech corrupted by reverberation and noise," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 3, pp. 572–587, Dec. 2016.

[310] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "Ego noise suppression of a robot using template subtraction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2009, pp. 199–204.

[311] A. Ito, T. Kanayama, M. Suzuki, and S. Makino, "Internal noise suppression for speech recognition by small robots," in *Proc. Eur. Conf. Speech Comm. and Technol.*, 2005.

[312] S. Chehrehsa, *Single-Channel Speech enhancement using statistical modelling*, Ph.D. thesis, Auckland University of Technology, 2016.

[313] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[314] 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," *Rec. ITU-T P.862*, 2001.

[315] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Feb. 2011.

[316] C. Taal, "MATLAB code for algorithms," http://www.ceestaal.nl/matlab-code/, Accessed: Jan. 2022.

[317] DJI, "Phantom 3 standard," https://www.dji.com/au/phantom-3-standard, Accessed: Sept. 2023.

[318] Seeed Studio, "Respeaker 6-mic circular array kit for raspberry pi," https://wiki.seeedstudio.com/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi/, Accessed: Sept. 2023.

[319] Yang, "Wind noise dataset," https://doi.org/10.5281/zenodo.6687982/, Accessed: Sept. 2023.

[320] P. Zinemanas, P. Cancela, and M. Rocamora, "MAVD-traffic dataset," https://zenodo.org/record/3338727/, Accessed: Sept. 2023.

[321] W. N. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Improved GMM wiener filtering approach for signal enhancement," in *Proc. INTER-NOISE and NOISE-CON Congr. and Conf.*, Aug. 2023, pp. 3584–3594.

[322] S. Fischer, A. C. Edwards, S. T. Garnett, T. G. Whiteside, and P. Weber, "Drones and sound recorders increase the number of bird species identified: A combined surveys approach," *Ecological Informatics*, vol. 74, pp. 101988, May 2023.

[323] J. W. S. B. Rayleigh, "Some general theorems relating to vibrations," *Proc. London Math. Soc.*, vol. 1, no. 1, pp. 357–368, Nov. 1871.

[324] T. D. Abhayapala and A. Gupta, "Spherical harmonic analysis of wavefields using multiple circular sensor arrays," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 6, pp. 1655–1666, Dec. 2009.

[325] A. Gupta and T. D. Abhayapala, "Three-dimensional sound field reproduction using multiple circular loudspeaker arrays," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 5, pp. 1149–1159, Oct. 2010.

[326] MH Acoustics, "Em32 eigenmike microphone array release notes (v17. 0)," Tech. Rep., 25 Summit Ave, Summit, NJ 07901, USA, 2013.

[327] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Amer.*, vol. 104, no. 4, pp. 2400–2411, Oct. 1998.

[328] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3616–3627, Jun. 2018.

[329] A. Urviola, S. Sakamoto, and C. D. Salvador, "Ear centering for near-distance head-related transfer functions," in *Proc. IEEE Int. Conf. Immersive and 3D Audio: from Architecture to Automotive*, Sep. 2021, pp. 1–8.

[330] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Efficient representation and sparse sampling of head-related transfer functions using phase-correction based on ear alignment," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 12, pp. 2249–2262, Oct. 2019.

[331] G. B. Arfken, H. J. Weber, and F. E. Harris, *Mathematical Methods for Physicists: A Comprehensive Guide*, Academic Press, Dec. 2012.

[332] B. Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, Oct. 2015.

[333] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, Sep. 2001.

[334] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Dec. 2004.

[335] W. N. Manamperi, T. D. Abhayapala, L. I. Birnie, J. A. Zhang, and P. N. Samarasinghe, "Drone audition: On measurements and modeling of drone-related transfer functions," *Submitted to IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, Feb. 2024.

[336] W. Manamperi, T. D. Abhayapala, and P. Holmberg, "Microphone aligned continuous wearable device-related transfer function: Efficient modeling and measurements," in *Proc. Hands-free Speech Commun. and Microphone Arrays (HSCMA) Satellite Workshop (accepted)*, Apr. 2024.

[337] A. Urviola, S. Sakamoto, and C. D. Salvador, "Ear centering for accurate synthesis of near-field head-related transfer functions," *Appl. Sci.*, vol. 12, no. 16, pp. 8290, Aug. 2022.

[338] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[339] S. Hafezi, A. H. Moore, P. Guiraud, P. A. Naylor, J. Donley, V. Tourbabin, and T. Lunner, "Subspace hybrid beamforming for head-worn microphone arrays," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Jun. 2023, pp. 1–5.

[340] W. Manamperi, P. N. Samarasinghe, J. Zhang, and T. D. Abhayapala, "Drone audition: Analysis of the preservation of spatial cues in multichannel wiener filtering," in *Proc. QUIET DRONES Conf. (accepted)*, Sept. 2024.

[341] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech and Audio Process.*, vol. 12, no. 5, pp. 451–459, Aug. 2004.

[342] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 785–799, Feb. 2014.

[343] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech

signals," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Jun. 2009.

[344] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 2015, pp. 544–548.

[345] R. Varzandeh, M. Taseska, and E. A. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Joint Workshop Hands-free Speech Comm. and Microphone Arrays (HSCMA)*, Mar. 2017, pp. 11–15.