Chapter 15

A novel sound source localization method using a global-best guided cuckoo search algorithm for drone-based search and rescue operations

Annesya Banerjee[a], **Achal Nilhani**[a], **Supriya Dhabal**[b], **and Palaniandavar Venkateswaran**[a]

[a]*Dept. of Electronics and Tele-Communication Engineering, Jadavpur University, Kolkata, India,*
[b]*Dept. of Electronics and Communication Engineering, Netaji Subhash Engineering College, Kolkata, India*

15.1 Introduction

In recent years, unmanned aerial vehicles (UAVs) or drones have gained extreme popularity in military operations and for different commercial and civilian use (Okutani et al., 2012). Due to their easy accessibility and affordability, drones have found ample applications in all business areas and fields including photography, filmmaking, agriculture, journalism, construction, entertainment, and several innovative scientific works during the 2010s. With the development and technological advancement, unmanned, autonomous, and semiautonomous systems have accelerated the emergency tasks including security, mapping, search and rescue (SAR) operations, and so on (Restas, 2015). Being cost-effective and quickly accessible, drones have become an inseparable part of SAR operations after disasters like floods, tsunamis, earthquakes, landslides, and avalanches, searching for missing people and immigrants in dense forests, mountains, etc. These unmanned systems are the only option to continue the search in situations which are almost inaccessible to the first response teams. Particularly, the faster mobility offered by aerial systems makes them more advantageous than ground-based vehicles. Further, for effective rescue operation, it is necessary to quicken the process by covering a large region of operation at a time which would require the involvement of huge human resources unless drones were

Unmanned Aerial Systems: Theoretical Foundation and Applications
https://doi.org/10.1016/B978-0-12-820276-0.00022-4

employed. Time efficiency achieved using drone-based search operations has made this technology the most beneficial one for emergency service personnel with police officers, firefighters, and volunteers from rescue teams. Most importantly, the application of drones for searching has reduced the threats offered to the lives of rescuers in disaster-affected areas, thus leading to a reduction in the probability of loss of human resources of a society. Especially in emergency situations like a nuclear explosion, it is harmful to the rescuers to visit the area at that moment but at the same time, it is necessary to look for victims and rescue them as early as possible. Such accelerated operations with minimum capital are achieved by the unmanned aerial systems.

In SAR, the unmanned systems are generally equipped with different sensors for real-time data acquisition to search for victims as well as to map the surrounding environment simultaneously. The traditional sensors used are optical sensors or cameras with high resolution and image quality for vision, thermal cameras for heat sensing, wireless sensor, laser, and sonar (Gala et al., 2018; Kwok et al., 2005). However, the vision-based sensors have gained more importance over the other sensors due to their capability to cover a wider range along with low energy consumption and less ambiguity in the acquired data. These vision-based drones are mainly useful in real-time imaging of the disaster-struck area and identification of distressed people after the natural calamity. Though these UAV-embedded tools with camera system have found numerous applications in critical search operations over several years and have gained the attention of a large community of researchers all over the world, they have some inherent disadvantages. First of all, these embedded optical sensors cannot perform very well in low light areas, in foggy weather, and during the night. In a few situations, such as where the victims are trapped under debris, it is indeed difficult to locate them at the earliest. Thermal sensors have proven to degrade in performance when the environmental temperature rises, thus limiting the application of these sensor-equipped drones to continue SAR in wildfires and assisting firefighters. For accurate identification, through the acquisition of high quality images (or videos) the embedded sensors need to be highly sensitive and sophisticated which in turn increases the net cost of the entire system.

Acoustic sensor-equipped drones have not yet been used for real-time SAR operations, though these systems find potential applications in situations where the traditional sensors cannot provide accurate and satisfactory results. Applying the philosophy of acoustic scene analysis, particularly focusing on sound source localization (SSL) by exploiting the features of sensor data, it is possible to use the drone-embedded system to identify victims in disaster-struck regions. This sensor system provides the drones with audition capability, functioning similarly as human ears to sense audio signals. The system will require an array of microphones embedded in the drone to capture audio data. As the distressed or missing people seek for help by shouting loudly, the drone embedded auditory sensors receive the speech signal. In literature, research to localize sound sources by analyzing the data captured by an array of microphones is ongoing.

Depending on the array structures and features of the recorded data, different studies have been carried out. A similar approach can be implemented for SAR operations. For this application, the drone-embedded systems look for the victim as this is the origin of the acoustic signal. For this mobile system, it is not practically possible to identify the absolute location of the source. Localization needs to be performed with respect to the current position of the drone system, which is useful for the source tracking applications. However, this approach has some inherent difficulties that need to be addressed before implementation of the system. The primary challenge in practical realization of the drone-based audition system lies in the fact that the quality of the audio data captured by the acoustic sensors is heavily reduced in the presence of stationary background and nonstationary ego and wind noises (Löllmann et al., 2014). The effect of such noises is not unique for aerial systems; rather, all robotic systems equipped with rotating components like motor, joints, and fans deal with this type of noise. The rest of the chapter is organized as follows. In Section 15.2, a literature survey related to the SSL and drones is presented. The problem formulation of the proposed denoising approach and localization technique is discussed in Section 15.3. The proposed algorithms are presented in Section 15.4. In Section 15.5, the verification of proposed algorithms is carried out. Finally, a discussion of the proposed approach and our conclusion are presented in Sections 15.6 and 15.7, respectively.

15.2 Literature survey

SSL algorithms have gained profound application in the field of auditory scene analysis – SSL along with source separation techniques are being used intensively for speaker localization, and identification-based technologies. SSL has applications in interactive robotic systems. Several researchers have focused on the study of SSL algorithms. Besides, researchers are working on techniques to deal with the issues related to noise, including ego noise and wind noise, in robotic systems. Kwok et al. (2005) studied the different geometric structures of the microphone array to facilitate SSL using time difference of arrival (TDOA) for human–robot interaction systems. The authors in Kwok et al. (2005) applied an evolutionary genetic algorithm to deal with the noisy signals acquired. The results exhibit angle estimation error in the range of 4.5 degree azimuth and 1.6 degree elevation for one type of array structure and 2.2 degree azimuth and 0.7 degree elevation for two other types of array structures. The effect of wind noise is very prominent for robotics systems and thus needs to be reduced effectively. A study of wind noise reduction using nonnegative sparse coding and wind noise dictionary has been presented in Schmidt et al. (2007). The performance of the algorithm, applied to single-channel noisy speech, has been compared with the Spectral subtraction and Qualcomm-ICSI-OGI noise reduction method in varying SNR conditions. Ince et al. (2009) presented two different approaches on ego motion noise suppression based on template estimation and subtraction,

namely, block-wise template subtraction and parameterized template subtraction. The authors in Ince et al. (2009) mentioned that the template subtraction outperforms existing SSL methods but yet to apply the technique for speech signals it is necessary to maintain the intelligibility and apply the method from single-channel to multichannel as future work. In their other work, the authors used a combination of histogram-based recursive level estimation for stationary noise estimation and template-based nonstationary noise estimation for single-channel speech enhancement (Ince et al., 2011). Fan et al. (2010) presented the localization approach using a special structure of a planar microphone array. The quasi-L1 autocorrelation and interpolation algorithms used in this paper for delay measurement increase the estimation accuracy. The result shows an improved positioning accuracy for the four-microphone systems over the three-microphone systems. For the purpose of automatic speech recognition (ASR) in robots, it is necessary to eliminate the ego noise present in the microphone recordings. Ince et al. (2010) developed an approach using the MUltiple SIgnal Classification (MUSIC) algorithm for the direction of arrival (DOA) estimation and applied it to the localized sound for source separation. For the evaluation of the algorithm, Ince et al. (2010) classified the noise in three different types, i.e., arm motion noise, leg motion noise, and head motion noise, and the results are discussed for the three separate cases. Results given in terms of word correct rates in ASR systems exhibit 50% correctness for arm and leg motion noise and 25% for strong head motion noise. The authors concluded the necessity of both single- and multichannel noise template subtraction for dealing with the head motion noise. An intensive study of the generalized cross-correlation (GCC) algorithm for SSL has been presented in Chen et al. (2011). Different signal weighting techniques like PHAT, ROTH, SCOT, and CC are compared in depth. The algorithm tested here, for positioning of mechanical failure source, produces 93% and 86% accuracy in terms of error less than 0.2 meter and 0.1 meter (variation in distance), respectively. Li et al. (2012) proposed a new approach using GCC-PHAT-$\rho\gamma$ and the guided spectral-temporal position method to reduce noise and reverberation based on GCC in mobile robots. The proposed methods were evaluated and compared with the existing GCC algorithm for different signal-to-noise ratio (SNR) (i.e., 10, 25, 40 dB) environments. The method, presented in Li et al. (2012), achieves a localization accuracy of 99.55% at an SNR of 40 dB. However, no test results have been provided for the negative SNR (very strong noise) scenario. As the authors have mentioned, this approach is useful for real-time processing. Velasco et al. (2012) implemented an approach based on beam forming for indoor acoustic source localization. The algorithm uses a modified version of Steered Response Power (SRP), called the SRP-PHAT, for the acoustic power mapping of the environment and prediction of a generative linear model. In addition, an optimization-based approach is used for model fitting and consequent source localization. The experimental result of a speech database demonstrated an error reduction of up to 30% compared to the traditional SRP-PHAT algorithm. Another work proposed in

Nakadai et al. (2012) takes into account the dynamically varying acoustic environment for human–robot interactions. The study includes MUSIC and its variation based on the generalized eigenvalue decomposition for noise-robust SSL and template-based ego noise suppression. The proposed approach has been evaluated based on the word correct rate in the ASR system and the accuracy in different SNR systems. The template subtraction approach for ego noise suppression has outperformed other methods even with negative SNR. The work presented in Blandin et al. (2012) discussed an in-depth theory on the different TDOA estimation methods including the popular angular spectrum methods and comparatively less explored clustering methods. The authors discussed five new methods for multiple TDOA estimation and source separation based on SNR weighting.

Recently, Löllmann et al. (2014) explored an overview of the challenges related to the development of audition systems and acoustic signal analysis in humanoid robots. The authors in Löllmann et al. (2014) discussed the ego noise and its characteristics along with the acoustic echo control schemes. Another approach for ego motion noise suppression, presented by Tezuka et al. (2014), extracts the noise feature using semiblind infinite nonnegative matrix factorization (SBINMF). SBINMF does not require any means for system motion analysis in noise estimation. The results indicated that SBINFM performs well compared to the extensively used template-based method. Huang and Wang (2014) presented a novel algorithm to develop a spherical estimating signal parameter via rotational invariance technique (ESPRIT), an approach using an array signal model, for spherical microphone array-based source localization. ESPRIT performs well as a 3D localization algorithm with low computational cost. Time domain beam forming methods, developed by Wang and Choy (2015), implement an approach for side-lobe suppression and an increased spatial resolution scanning strategy near the sound field for accurate SSL. The blind source separation (BSS) algorithm plays an important role in multiple SSL as studied in Nogueira and Petraglia (2015). But, in the BSS algorithm the accuracy of SSL largely varies with the distance between the microphone pair in the array structure and needs to be improved further. Schmidt et al. (2016) presented an approach for ego noise suppression in robots based on multichannel dictionary learning where the system learns a dictionary containing the spatial and spectral characteristics of the varying noise and a nonlinear classifier is designed for noise reduction. This motor data-guided method has been tested to perform well for the microphone array structure that was not learned by the system previously. Dorfan et al. (2016) proposed two approaches, namely, batch algorithm based on the maximum likelihood criterion optimized via expectation-maximization iterations and particle filter for sequential Bayesian estimation in SSL with moving microphone system. Jung et al. (2017) developed a method based on the golden selection searching for SSL. The algorithm proposed in Jung et al. (2017) aims at reducing the computational cost by narrowing the search region but maintaining high accuracy of localization at the same time. The method

presented in Löllmann et al. (2017) explored an insight into microphone array signal processing including microphone array arrangement, ego noise removal, echo cancelation, and audio source tracking. Later on, Haubner et al. (2018) implemented a multichannel nonnegative matrix factorization (MNMF) algorithm to suppress ego noise. The results obtained using MNMF reveal that for ego noise suppression multichannel NMF outperforms single-channel NMF, i.e., joint ego noise suppression is advantageous over considering each channel separately. Gala et al. (2018) proposed a novel method for orientation and distance localization of sound sources in 3D using the interaural time difference cue using a self-rotational bimicrophone array. The results exhibit an error of 4 degrees in angle localization and 0.6 m in distance localization in a very low SNR environment. Xenaki and Boldt (2018) estimated the DOA by sparse signal reconstruction using sparse Bayesian learning (SBL). The utility of this high resolution SBL beam forming technique resulted in speech separation along with DOA calculation. Ma et al. (2018) adapted the deep neural network (DNN) tool for binaural source localization. The model-based information of the target source and background source are estimated using the spectral characteristics extracted and these models are used for explaining the mixed observation in the DNN-based source localization system. This approach performs very well in the presence of many interfering sound sources.

15.2.1 Related work with drones

Several works have been carried out to analyze the noises present in the audition system embedded in humanoid and moving robots. A similar analysis for UAVs is a comparatively new area of research and gained focus after 2012. Okutani et al. (2012) proposed a modified algorithm based on MUSIC that particularly deals with the dynamically varying noise present in microphone array recordings of a quadcopter used for auditory scene analysis. The incremental generalized eigenvalue decomposition (iGEVD-MUSIC) algorithm presented by the authors performs well even in high noise, i.e., negative SNR environment. The approach considers an adaptive estimation of the noise correlation matrix with results in suppression of noise of varying nature and accurate source localization is possible, as is shown in their results. Another work published in Basiri et al. (2012) presented a system for narrowband SSL using a microaerial vehicle (MAV). The authors used a method based on particle filter to extract information from cross-correlation of the signals of the spatially separated microphones. However, the algorithm is implemented for whistle and alarm sound localization. The performance of the algorithm for human voice has not been evaluated. Furukawa et al. (2013) presented a method for SSL by multirotor UAV using the adaptive noise correlation matrix. Gaussian process regression has been used to estimate the noise correlation matrix. The results show that the proposed algorithm produces more prominent peaks in the MUSIC spectrum than that produced by other existing algorithms. Ohata et al. (2014) proposed a

version of iGEVD-MUSIC with correlation matrix scaling to improve the SSL performance by soft whitening of noise and it facilitates low computational cost. The algorithm has been evaluated using both circular and spherical microphone array prototypes, and the results show that the method performs well for SSL in the outdoor environment and is efficient for real-time processing. The study by Wang and Cavallaro (2016) suggested a model for the ego noise as the sum of N directional noises and one diffuse noise, where N represents the number of rotors present in the UAV. Besides, considering the microphone array and motor orientation to be stationary, the mixing model of the noises to the clean audio is to be stationary and deterministic as well. The results of the study show that it is possible to achieve SNR improvement using the proposed model given the number of microphones used is less than six, i.e., the number of directional ego noises is five. Beyond this limit, the performance of the noise suppression algorithm decreases significantly. Morito et al. (2016) discussed a modified version of DNN, called partially shared DNN (PS-DNN), for human speech separation and identification using UAV-embedded microphone arrays. The proposed algorithm is particularly useful because, in contrast to traditional DNNs that require huge data for training, the PS-DNN requires fewer amounts of annotated data and can learn multiple tasks simultaneously. Though the experimental results show good performance of this algorithm compared to the DNN-based approach, the performance for multiple source localization and separation using this method is yet to be evaluated. In another approach, Wang and Cavallaro (2018) suggested time frequency processing for the localization and enhancement of target sound by analyzing the spectral and spatial features of the ego noise. The algorithm first estimates the local DOA in each time frequency bin separately and using the statistical analysis implements a spatial filter to localize the target source in a particular direction. A DOA weighting scheme is implemented to achieve accurate SSL even in a very low SNR condition. The recent work in Misra et al. (2018) developed a binaural acoustic sensing system using a pair of microphones for drone-based SSL. This study is important as it deals with a lower number of sensors, thus reducing the payload of the mobile system and still achieving significant SNR improvement in high noise levels.

15.3 Problem formulation

Drones equipped with microphones have promising applications in SAR operations but the studies so far indicated that only SSL algorithms applied to the captured audio cannot yield the desired results. Due to the presence of high amplitude noises, it is necessary to first suppress the noise components from the data. Thus it has become a more challenging task as traditional noise reduction algorithms that are applied to low background noises completely fail in this application. In this work, we have presented a novel approach for the elimination of ego noise and wind noise from drone-captured audio data and hence localization of the sound source is performed for application in SAR operations and

disaster management. An analysis of the major noise components appeared in a drone-embedded systems has been addressed.

15.3.1 Ego noise

In robotic systems, the data captured using the acoustic sensors are heavily affected by the noise of the internal mechanical system of the robot. The airflow noise of the rotating fans and propellers in aerial robots adds to the system noise – this high amplitude noise is termed ego noise (Ince et al., 2009; Tezuka et al., 2014; Wang and Cavallaro, 2016). The special type of ego noise generated during the motion of the robot is called ego motion noise. The ego noise has two components – stationary noise produced by static sources and nonstationary noise from the moving components of the system. For example, the fans inside the robot structure will produce ego noise stationary in nature whereas the motors produce nonstationary noise. The analysis of nonstationary noise is difficult as its pattern varies with the movement and positioning of the robotic system and thus is unpredictable. In particular, for the drone system, the ego noise is generated by its motors and rotating propellers. For the need of position variation and stabilization in air, it is necessary to change the speed of the drone motors very quickly which generates the ego noise. Figs. 15.1a–15.1e show the time domain plots of the ego noise, captured at Microphone 1, generated by Motor 2 of the drone at five different speeds (50, 60, 70, 80, and 90 rpm). From these figures it is obvious that with the variation in motor speed the ego noise signal waveform varies in amplitude. The initial silent portion of the noise waveform represents the time span when the motor just starts to rotate. Once it starts rotating with the predefined speed, the noise amplitude reaches a high value, as illustrated in Figs. 15.1a–15.1e.

Figs. 15.2a–15.2d highlight the variation in the ego noise waveform at 80 rpm captured by Microphone 2 for the four motors of the drone. From these figures, it can be concluded that Motor 2 is closer to Microphone 2 than the other motors because the mean amplitude of noise generated at Motor 2 is higher than that of other motor noises. The time domain nature of the same is different from the other three. In addition, from the analysis of the frequency components, depicted in Figs. 15.3a and 15.3b, it can be observed that for the noise generated by a motor rotating at X rpm, the noise has a fundamental frequency of X Hz and harmonic components are present at multiples of X Hz. Hence the ego motor noise is related to the rotational speed of the motor.

Figs. 15.3a and 15.3b present the power spectral density (PSD) of the noise generated by Motor 1 at a speed of 50 rpm and Motor 4 at 80 rpm, respectively. It can be observed that the PSD in Fig. 15.3a has its first peak at normalized frequency equal to 0.002224 rad/sample, i.e., 50 Hz. The next peaks appear at the multiples of 50 Hz. Similarly, for Fig. 15.3b the first peak is at frequency 80 Hz and consecutive peaks occur at the harmonics of 80 Hz. From this observation, we can conclude that with respect to the speech signal recorded by the

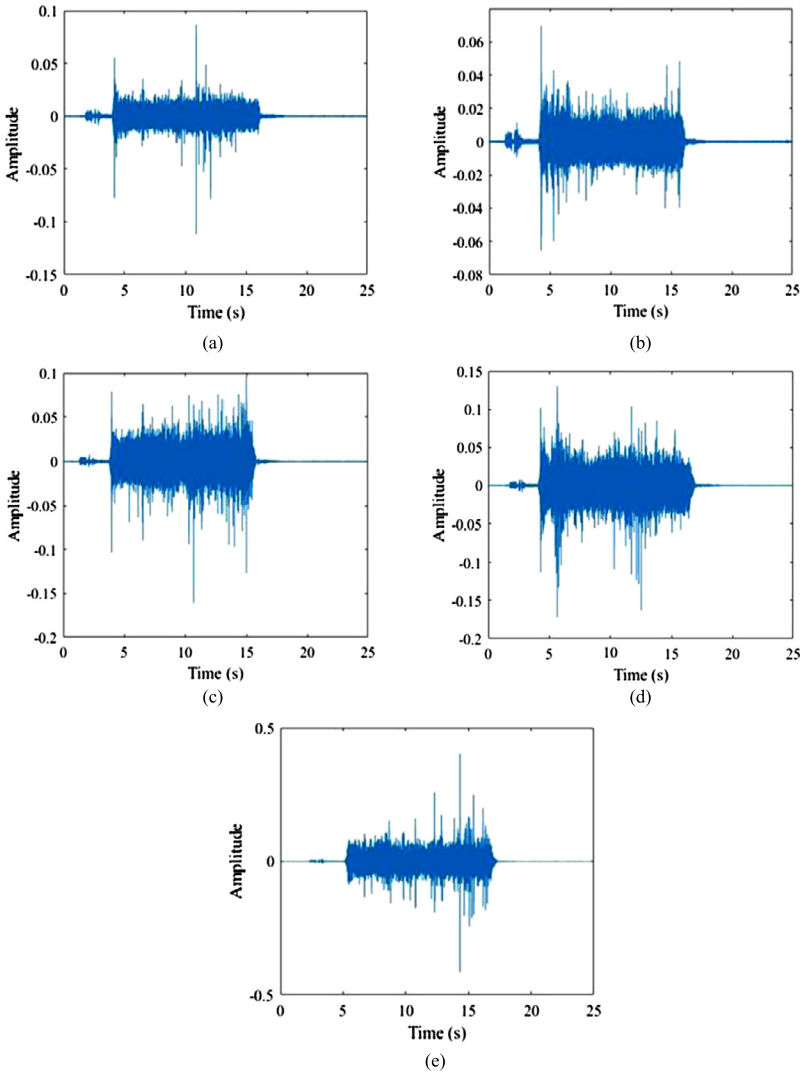

FIGURE 15.1 Time domain plot of ego noise generated by Motor 2 at different speeds captured by microphone 1. (a) 50 rpm. (b) 60 rpm. (c) 70 rpm. (d) 80 rpm. (e) 90 rpm.

microphone array system, the amplitude of the noise components is so high that the original signal gets completely buried in noise, as presented in Figs. 15.4a and 15.4b.

15.3.2 Wind noise

Another noise component that heavily affects the recordings by the drone-embedded microphone array is wind noise caused by the rotating propellers

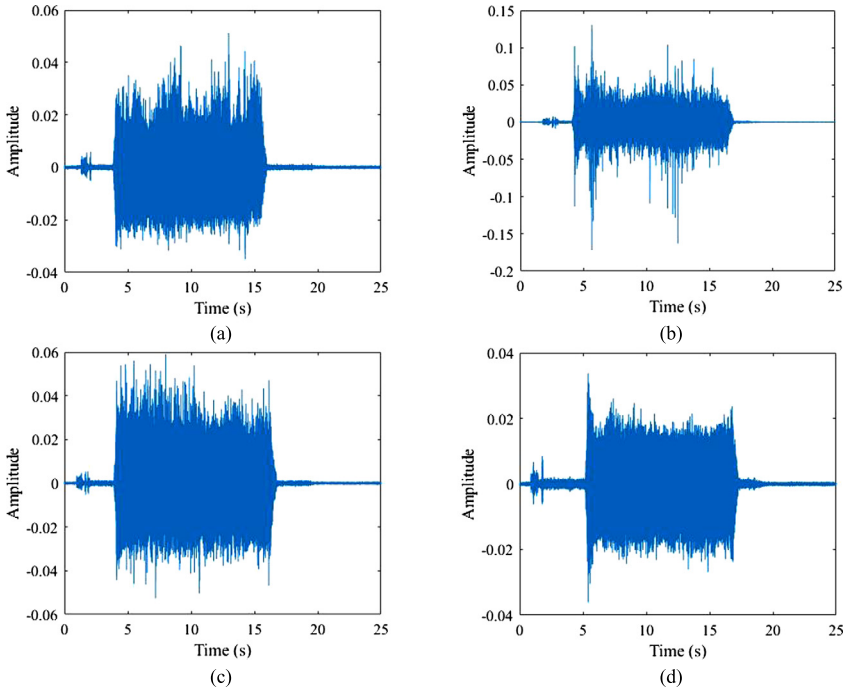

FIGURE 15.2 Time domain plot of ego noise generated by (a) Motor 1, (b) Motor 2, (c) Motor 3, and (d) Motor 4 at 80 rpm and captured by microphone 2.

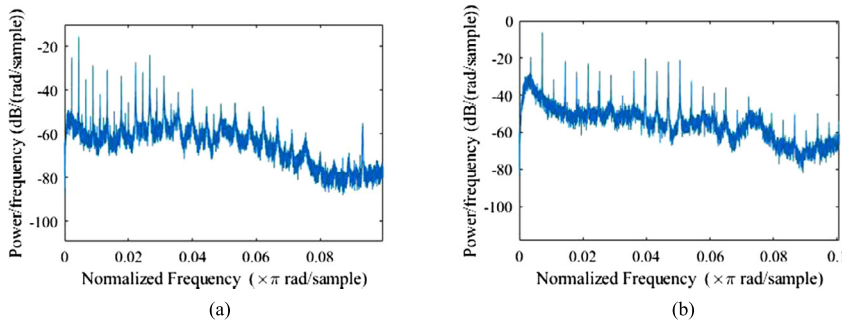

FIGURE 15.3 Relation of ego noise with motor speed for (a) Motor 1 and 50 rpm and (b) Motor 4 and 80 rpm.

(Schmidt et al., 2007). As the four propellers rotate simultaneously, they cut the air and produce heavy noise. In addition, the drone movement and the outdoor environment are responsible for the generation of wind noise. Wind noise is of high power and has low frequency – lying in the frequency range of speech signals. This nature makes it particularly difficult to separate the wind noise from

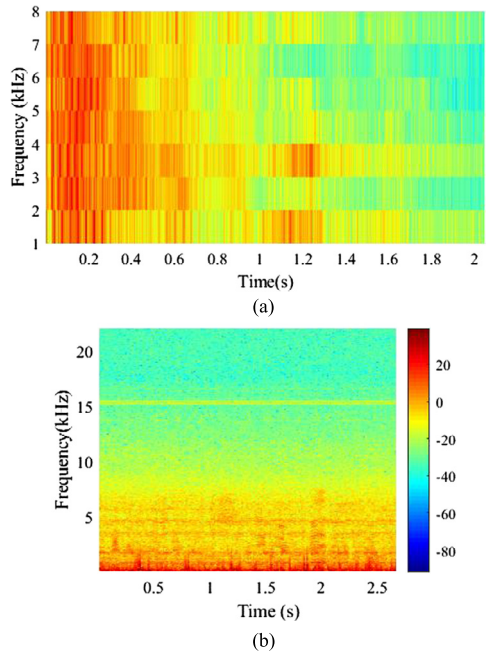

FIGURE 15.4 Time frequency spectrum of (a) clean speech signal and (b) speech signal corrupted by ego noise.

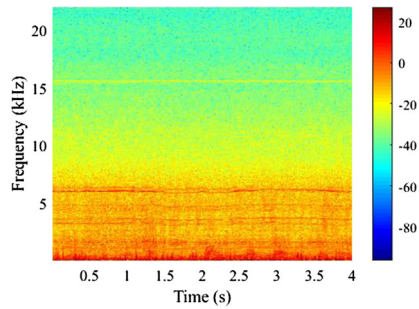

FIGURE 15.5 Spectrum of ego noise- and wind noise-affected clean speech.

clean speech. Fig. 15.5 demonstrates how the spectrum of a clean speech gets affected by heavy wind and ego noise of a drone. It is obvious that any speech signal can hardly be identified from the spectrum as it appears entirely noisy.

15.4 Proposed algorithm

The algorithms proposed in this chapter aim at the suppression of the ego noise and wind noise present in the recordings by the drone microphones. It uses

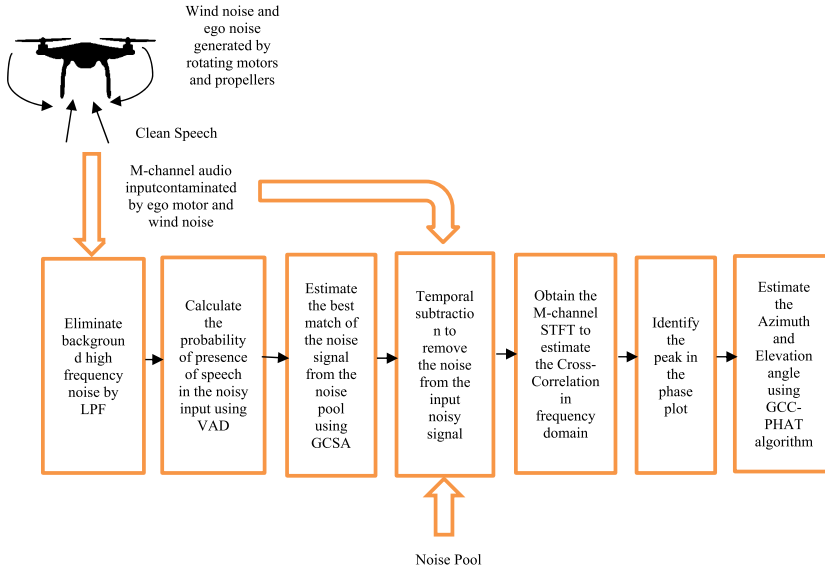

FIGURE 15.6 Block diagram of the localization technique used in proposed Algorithm 1.

an optimization-based approach for noise estimation where we consider a set of known noise signals and their features to estimate an unknown noise. We have presented here two algorithms, named proposed Algorithm 1 and proposed Algorithm 2. These two algorithms adopt two different noise subtraction methods discussed later. After effective noise elimination, it is possible to apply the source localization algorithm for accurate position identification of the speech source. There are three main approaches for the SSL: (a) calculating TDOA, (b) estimating the steered response power (SRP), and (c) the MUSIC approach. We adopt the TDOA estimation-based approach for the source localization. As Blandin et al. (2012) discussed, TDOA for a pair of sensors can be identified using clustering and angular spectrum-based approaches. But the latter being applicable to any microphone spacing and independent of any initialization, we have used this approach in our study. Fig. 15.6 shows the block schematic of proposed Algorithm 1. Each block represents a step in the localization process. The first two blocks are mainly the preprocessing steps. The filtering block removes the background stationary noise that is characterized by high frequencies. Thus, the output of this block contains only the original signal mixed with ego motor noise and the wind noise only. The next block then detects whether any speech signal is actually present in the input. This is determined by the probability measure depending on the SNR value. For this purpose, a particular threshold is chosen as the decision boundary. If the probability is more than the threshold, we consider that it contains a speech signal and further processing is carried out. Otherwise, we decide that it is a noise-only signal. The preprocessed sig-

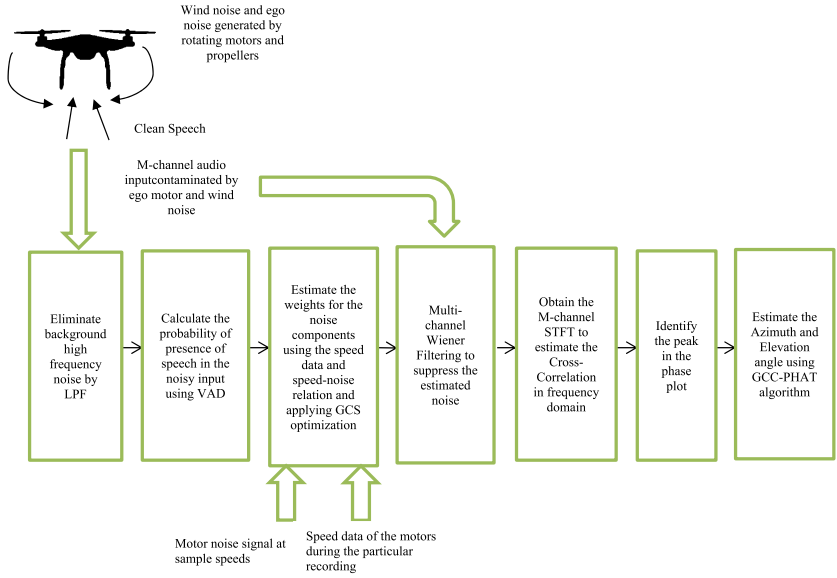

FIGURE 15.7 Block diagram of the localization technique used in proposed Algorithm 2.

nal is now used to estimate the noise by matching the unknown noise with the noise signals present in the noise pool. Once the best fitting of the noise signal is identified we move to the next block for noise elimination from the original noisy signal. After noise suppression, the GCC-PHAT algorithm is used for the localization of the source signal by estimating the cross-correlation peaks. The steps are represented in the last three blocks of the schematic. The details of each block are discussed in Section 15.4.3.3.

Fig. 15.7 presents the schematic block of proposed Algorithm 2. The preprocessing technique applied here is the same as that applied for proposed Algorithm 1. The second block that identifies the speech presence probability detects the portion of the signal that has an active voice signal for a significant time. The speech containing frames of the signal are utilized for noise estimation in the next step. In this algorithm, the noise estimation block takes as input the current speed information of the motors and a set of noise signals at sample speeds. This is a supervised approach and the Gbest-guided cuckoo search (GCS) estimates the optimum weights for the motor noises that yield the best estimate of the overall noise. The linear relationship between motor speed and motor noise has been exploited here. This estimation of noise is the input to the multichannel filtering block. The Wiener filter suppresses the input noise from the original signal. The denoised output of the filter is used for the localization of the source using GCC-PHAT as has been shown in Fig. 15.6. The details are provided in Section 15.4.3.4.

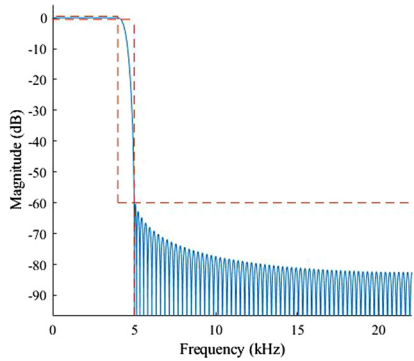

FIGURE 15.8 Response of the low pass filter with cut-off frequency 4 kHz.

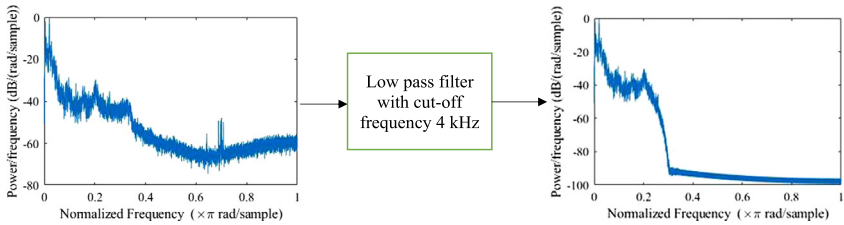

FIGURE 15.9 Block diagram representation of the filter with its input and output.

15.4.1 Preprocessing filter

Except for the nonstationary ego and wind noise, the signals captured by the microphone array contain noise generated from the structural vibration and other electronic components of the system. As these are high frequency stationary noise and do not overlap with the speech signal band (<10 kHz), a low pass filter with proper cut-off frequency can remove these noise components without distorting the original speech component. Fig. 15.8 shows the frequency response of the low pass filter with cut-off frequency 4 kHz and stop-band attenuation of 60 dB. Before applying the ego noise suppression algorithm, we implement this preprocessing filter to eliminate the background noises. However, it should be noted that this filtering technique does not affect the low frequency nonstationary noises. As represented in Fig. 15.9, the input to the filtering block is a noisy signal having a wide range of frequency components. It is passed through a low pass filter with cut-off frequency fixed to 4 kHz. The output signal PSD reveals that the frequency components beyond the cut-off frequency are attenuated to low amplitude where the low frequency components remain unaffected.

15.4.2 Voice activity detector

As a part of the preprocessing block, we first try to identify whether an unknown signal contains any speech signal and this is determined by using a voice activity detector block. As discussed in Sohn et al. (1999), the detector estimates the probability of speech (P_{speech}) in the current frame of analysis by evaluating a likelihood ratio (Λ_k) given by

$$\Lambda_k = \frac{1}{1+\alpha_k} \exp\left(\frac{\beta_k \alpha_k}{1+\alpha_k}\right), \tag{15.1}$$

where α_k denotes the ratio of the variance of clean speech and noise for the k^{th} frame, called the a priori SNR, and β_k represents the ratio of the variance of noise-contaminated speech and noise for the k^{th} frame, called the a posteriori SNR. The decision of speech or no speech for the particular frame is identified based on the value of the quantity Λ_k. As the authors have discussed, having knowledge of the statistical properties of the noise it is possible to evaluate β_k for the unknown signal. From the β_k value, α_k can be estimated using a maximum likelihood estimator as

$$\hat{\alpha}_k = \beta_k - 1. \tag{15.2}$$

To detect if the drone microphone recordings contain any speech content it is beneficial to carry out the time frame-wise analysis approach. As an effect of ego noise on the clean speech, the SNR of the entire signal decreases to a very low value. In addition, the overall P_{speech} in the signal becomes very low, indicating no speech. However, the analysis shows that this inference is erroneous in most of the cases. As an alternative, for each small frame of the signal, we have calculated P_{speech} using the algorithm discussed above. Then we find the frame having maximum and consistent probability of speech. We define consistency by the fact that if for a test signal, the i^{th} frame has a maximum P_{speech} value equal to p_i, then at least n frames before and after the i^{th} frame will have a P_{speech} value in the range $[p_i - \delta, p_i + \delta]$, where δ is a very small quantity. If the consistency is not maintained, then we check for the next highest value of P_{speech} after p_i. This approach ensures that the speech is active in the signal for a significant time. The block diagram in Fig. 15.10 represents the proposed approach to identify the part of the signal containing speech.

Fig. 15.12a shows the probability of speech in different time frames of the noisy speech signal with its time domain representation given in Fig. 15.11. As can be observed, due to the presence of high amplitude noise the variation of speech present in the signal cannot be detected visually. However, from Fig. 15.12b, we can infer which portion of the signal contains speech and detect the start and end time points of the speech frames. For the M-channel input

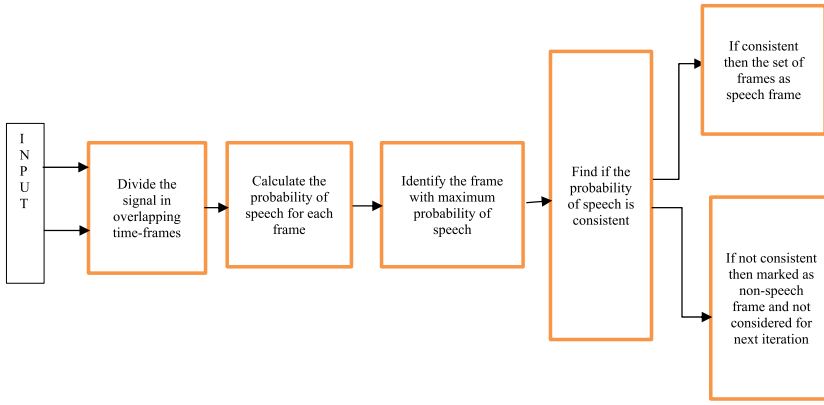

FIGURE 15.10 Block schematic of the voice activity detector.

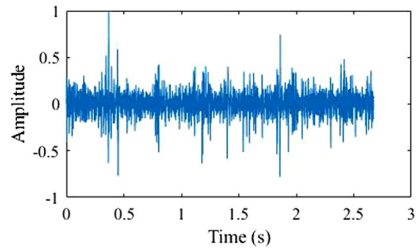

FIGURE 15.11 Time domain representation of noisy signal containing speech signal buried in heavy ego noise.

signal, the probability of speech is given by the average of the probabilities in each of the channels. For consistency identification, we have set the parameter values $n = 7$ and $\delta = 0.1$. The result of the test is shown in Fig. 15.12b. The speech frame (indicated by the green (light gray in print version) window) has a maximum $P_{speech} = 0.949$. The frame (indicated by the red (gray in print version) window) contains a time frame with the highest P_{speech} value but does not follow the consistency criteria, and therefore it cannot be considered as a speech frame.

15.4.3 Proposed denoising algorithms based on cuckoo search algorithm

As discussed previously, it is required to suppress the noise components from the captured signals before applying the source localization algorithm. In this work, we propose a cuckoo search optimization-based algorithm for effective noise elimination.

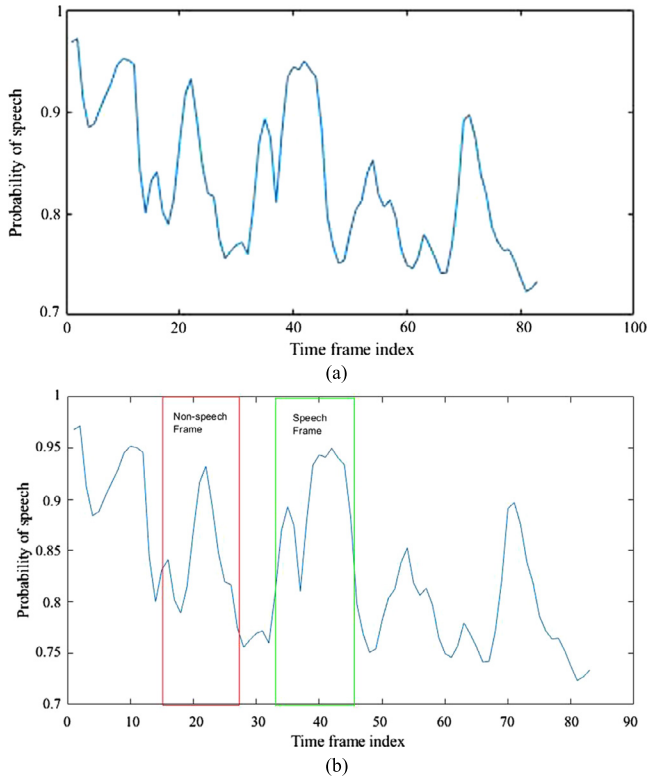

FIGURE 15.12 (a) Variation in the probability of speech evaluated for noisy signal. (b) Result of the probability consistency checking test.

15.4.3.1 Cuckoo search algorithm

Cuckoo search algorithm (CSA) is a newly proposed optimization algorithm (Yang and Deb, 2010). The algorithm is based on the natural reproduction of cuckoo birds and the hypothesis of the Levy flight process. In Levy flight, the random walk is obtained by Levy's distribution as

$$\text{Levy}(\lambda) = \left| \frac{\Gamma(1+\lambda) \times \sin(\pi\lambda/2)}{\Gamma((1+\lambda)/2) \times \lambda \times 2^{(\lambda-1)/2}} \right|^{1/\lambda}. \tag{15.3}$$

Thus, with the help of Levy's distribution, the cuckoo birds' egg laying behavior is described by

$$x_i^{t+1} = x_i^t + \alpha \oplus \text{Levy}(\lambda), \tag{15.4}$$

where $\alpha > 0$ represents the step size, $1 < \lambda \leq 3$, and "$\oplus$" signifies entry-wise multiplication. The traditional CSA follows three assumptions: (a) each cuckoo lays only one egg and dumps it in a randomly selected nest, (b) the best nests

with high quality eggs will survive for the next generation, and (c) the number of available host nests is fixed and a host can discover an alien egg with probability $p_a \in [0, 1]$. The last condition is maintained by replacing a fraction (p_a) of n host nests with new nests.

15.4.3.2 Improved cuckoo search algorithm

In standard CSA, the search equation used by the cuckoos is entirely based on the random walks, which may not guarantee a fast convergence. Therefore, here we have incorporated four modifications in the proposed CSA algorithm (Dhabal and Venkateswaran, 2017), to enhance its convergence rate and simultaneously to make it auto-tuned. The changes are as follows.

(a) Modification in replacement strategy: Normally, in standard CSA, the replacement of old nests is performed at random, which reduces the convergence speed. Thus, instead of searching in random, replacement of old nests is performed based on global-best solution so that better control of the step size is achieved. The modified equation is as follows:

$$nest_{new} = nest_{old} + rand * (nest_{best} - nest_{old}) \oplus K \quad \text{if } K > p_a, \tag{15.5}$$

where $nest_{old}$ and $nest_{best}$ represent the permutation matrix obtained from the old and best nests, respectively, and $nest_{new}$ is the new nest generated in the current iteration. As the generations of new nests depend on the best nest obtained so far, it is named *Gbest-guided cuckoo search* (GCS) algorithm.

(b) Modification in λ: For better exploration in searching, instead of assuming a fixed value of $\lambda = 1.5$ in Levy's distribution, here we vary λ as follows:

$$\lambda = (\lambda_{\max} - \lambda_{\min}) \times \frac{(\text{iter}_{\max} - \text{iter})}{\text{iter}_{\max}} + \lambda_{\min}, \tag{15.6}$$

where $\lambda_{\max} \to \lambda_{\min} = 1.5 \to 1$ and $\text{iter}_{\max}$ and iter indicate the maximum and current iteration, respectively.

(c) Modification in p_a: Yang and Deb (2010) suggested that the CSA outperforms Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) in terms of the number of parameters to be tuned – the probability of abandoned nests (p_a). They found that the convergence rate of CSA is insensitive to p_a and the setting of $p_a = 0.25$ is sufficient for all problems. But experimental results reveal that for complex and multimodal problems, the convergence rate can be improved by suitably adjusting the parameter p_a. Therefore, to make the algorithm self-tuned p_a is varied based on the following equation:

$$p_a = {}^{rand}/_{D}, \tag{15.7}$$

where D denotes the dimension of the problem and $rand \in [0, 1]$ is a random number. In our work, we have used the GCS algorithm to optimize the ego noise

model for best fit with the original noise present in the signal. Here we present two approaches for modeling the noise using GCS.

15.4.3.3 Proposed Algorithm 1

From the knowledge of the ego noise nature generated by the motors of a drone at different speeds, the algorithm takes as input a noise pool containing the motor noise signals. The proposed algorithm selects a noise signal randomly from the noise pool and computes the value of the fitness function. This is performed to identify the noise signal which best matches the noise present in the unknown noisy input. For this particular problem, we have defined the fitness function as the MSE between the noisy audio signal and the noise as follows:

$$\min F = MSE = \frac{1}{N}\sum_{k=1}^{N}\left(xn_i\left(k\right) - n_i\left(k\right)\right)^2, \tag{15.8}$$

where "N" denotes the number of samples of the signal, xn_i is the i^{th} sample of the input noisy signal, and n_i is the i^{th} sample of a randomly selected noise from the noise pool. The task of the GCS is algorithm is to choose the best fit n_i to minimize the value of F. Using temporal subtraction the best fitted n_i is removed from the original signal, resulting in a denoised signal applicable for source localization. In Figs. 15.13a–15.13d, the results obtained by proposed Algorithm 1 are presented. After obtaining the denoised audio signal, the GCC-PHAT technique is applied to localize the sound source. The PSD of an unknown noisy signal is presented in Fig. 15.13a. Considering this unknown input, the GCS algorithm searches for a known noise signal present in the noise pool. The result of the search is a noise signal which best matches the noise present in the input signal. The PSD of this estimated noise is given in Fig. 15.13b. From the two plots, it can be concluded that they have a similar variation of power at different frequencies. The two spectra in Figs. 15.13c and 15.13d correspond to the noisy signal and the signal obtained after the temporal subtraction of the noise. The two spectra are indeed distinguishable. In Fig. 15.13d the speech components are more prominent than in Fig. 15.13c. The flowchart of proposed Algorithm 1 is presented in Fig. 15.14.

15.4.3.4 Proposed Algorithm 2

This algorithm considers a hybrid approach of noise estimation by comparing with the original noisy input signal as stated for the previous approach and exploits the harmonic components present in the ego noise components as has been discussed in Section 15.3.1. As a first step, the algorithm estimates the ego noise generated by each propeller/motor i, $i = 1, \ldots, 4$. As a motor rotating at speed (v) generates noise containing harmonics at frequency $f = v$ Hz, it is possible to estimate the ego noise at any unknown speed (v) from the knowledge of the ego noise at two other speeds of the motor as follows. Let us consider the i^{th}

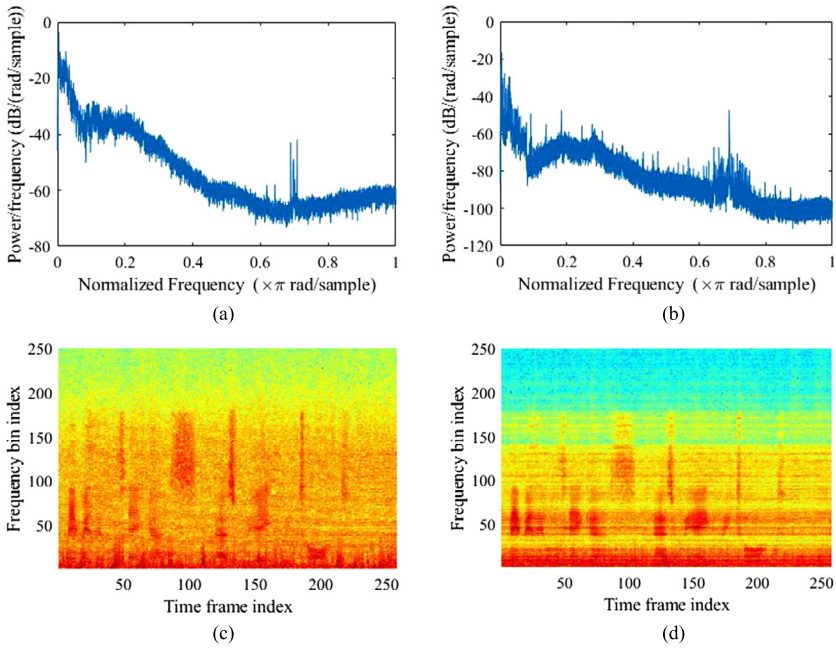

FIGURE 15.13 (a) PSD of unknown signal containing speech buried in high amplitude noise. (b) PSD of the noise estimated for noisy input given in (a) using proposed Algorithm 1. (c) Spectrum of the noisy input signal. (d) Spectrum of the signal after attenuating the estimated noise using temporal subtraction.

motor rotating with speed v_i and the ego noise produced by the same motor at speeds v_i^{upper} and v_i^{lower} are known to be equal to n_i^{upper} and n_i^{lower}, respectively. Then considering the upper and lower weights as w_{ui} and w_{li} we can estimate the ego motor noise at current speed (v_i) as

$$n_i = w_{li} n_i^{lower} + w_{ui} n_i^{upper}, \tag{15.9}$$

where w_{ui}, $w_{li} \in (0, 1)\, \forall i$. Therefore, for a particular input signal xn, the estimated noise component ($n_{estimated}$) can be represented as follows:

$$n_{estimated} = \langle w.n \rangle,$$

or

$$n_{estimated} = w^T n = \sum_{i=1}^{4} w_i n_i, \tag{15.10}$$

where $w = [w_1 w_2 w_3 w_4]^T$ is called the weight vector, $n = [n_1 n_2 n_3 n_4]^T$ is the noise component vector, and n_i is given as in equation (15.9) $\forall i$. The GCS algorithm is used to determine the optimum value of the vector w and the weights

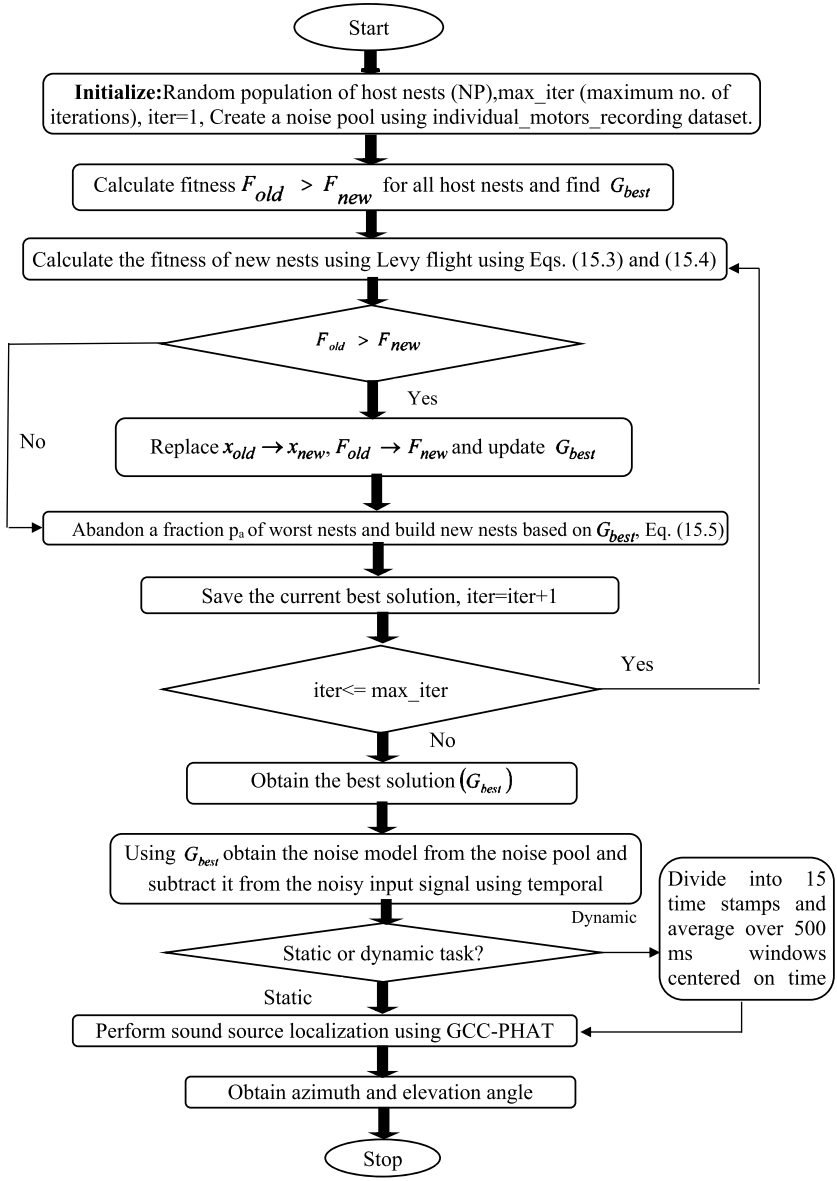

FIGURE 15.14 Flowchart of proposed Algorithm 1.

$\{w_{ui}\, w_{li}\}\forall i$ so that it minimizes the MSE given in equation (15.8) subject to the following constraints:

1. $w_{li} + w_{ui} = 1\forall i$,
2. $\sum_i w_i = 1$.

For better estimation of noise, a two-level approach using GCS is implemented here. In the first level, the unknown noisy signal is divided into several overlapping windows. For each window, a noise signal is estimated using GCS following equations (15.9) and (15.10) and the value of the cost function is calculated. For the next level of estimation, the algorithm considers only that window for which the value of the cost function evaluated in level 1 is minimum because the minimum cost function implies maximum similarity with the unknown noise present in the signal. It has been verified that the signal portion identified by this method has minimum speech probability, calculated using the voice activity detector (VAD). Thus this portion has maximum noise and for estimating the unknown noise this portion performs the best. Then, GCS is applied to the selected window to obtain the updated set of weights and weight vectors from which the estimated noise $n_{estimated}$ can be derived. This two-level approach enhances the performance of the optimization algorithm and results in a better estimation of the unknown noise. The estimated noise is taken as the reference signal for the multichannel Wiener filter as discussed in Strauss et al. (2018). Using the multichannel Wiener filter, the original signal $\hat{S}(t, \omega)$ can be estimated as

$$\hat{S}(t, \omega) = W_{MWF}(f) X(t, \omega), \tag{15.11}$$

$$W_{MWF}(f) = R_{XX}(f)^{-1} (R_{XX}(f) - R_{NN}(f)), \tag{15.12}$$

where $W_{MWF}(f)$ represents the multichannel Wiener filter and $R_{XX}(f)$ and $R_{NN}(f)$ represent the covariance matrix of the unknown noisy signal and the estimated noise:

$$R_{XX}(f) = \text{cov}[X, X] = E\left[(X - \mu_X)(X - \mu_X)^T\right] = E\left[XX^T\right] - \mu_X \mu_X^T, \tag{15.13}$$

$$R_{NN}(f) = \text{cov}[N, N] = E\left[(N - \mu_N)(N - \mu_N)^T\right] = E\left[NN^T\right] - \mu_N \mu_N^T, \tag{15.14}$$

where $E(\cdot)$, $(\cdot)^T$, and μ denote expectation, matrix transpose, and the mean of the variables, respectively.

Fig. 15.15 illustrates the results obtained using proposed Algorithm 2. The PSD of the unknown noisy signal has been presented in Fig. 15.15a and the noise estimated by proposed Algorithm 2 for this unknown signal has a PSD shown in Fig. 15.15b. The spectrum of the noisy signals presented in Fig. 15.15c shows that due to very heavy noise the speech signal present in between the time frame index 100 to 200 is completely attenuated. A comparatively enhanced result is obtained using the multichannel Wiener filter and the spectrum of the speech signal present in the noisy signal is more identifiable than the previous case. This result has been shown in Fig. 15.15d. The GCC-PHAT algorithm is then applied to the Wiener-filtered signal to extract the coordinate (azimuth and elevation)

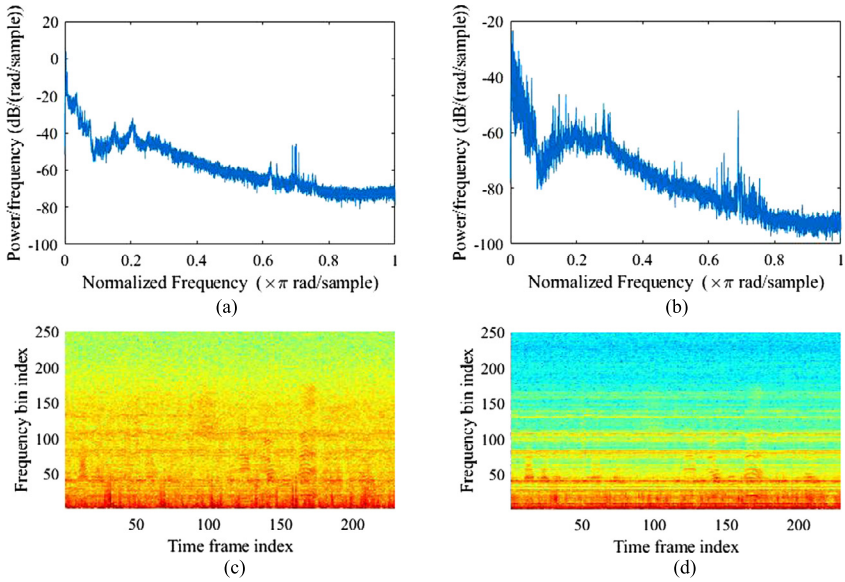

FIGURE 15.15 (a) PSD of unknown signal containing speech buried in high amplitude noise. (b) PSD of the noise estimated for noisy input given in (a) using proposed Algorithm 2. (c) Spectrum of the noisy input signal. (d) Spectrum of the signal after attenuating the estimated noise using multichannel Wiener filter.

information of the sound source. Fig. 15.16 shows the flowchart of proposed Algorithm 2.

15.4.4 Time difference of arrival

For a pair of sensors and a source it is possible to measure the difference in time taken by an acoustic signal emitted by the source to reach the two sensors, generally microphones for SSL. If the propagation speed of the signal is known, then from the knowledge of the difference in arrival time we can estimate the distance difference (Blandin et al., 2012). Suppose there are two microphones, Mic 1 and Mic 2, with coordinates (x_1, y_1) and (x_2, y_2), respectively. We have a source S with a coordinate (x_s, y_s). For our purpose we can safely assume a far-field model, i.e., $d_1, d_2 >> d_m$, for the sound source, where the sound waves contain planar wavefronts when they reach the microphones. Let t_1 and t_2 be the arrival times of the acoustic signals to Mic 1 and Mic 2, respectively, given by $t_1 = d_1/c$ and $t_2 = d_2/c$. If Δt is the difference in arrival time for the two sensors and Δd is the difference in distance, then we can write

$$\begin{aligned} &\Delta t = t_1 - t_2, \\ &\text{or } \Delta t = d_1/c - d_2/c \\ &\therefore \Delta d = \mathrm{c}\Delta t, \end{aligned} \tag{15.15}$$

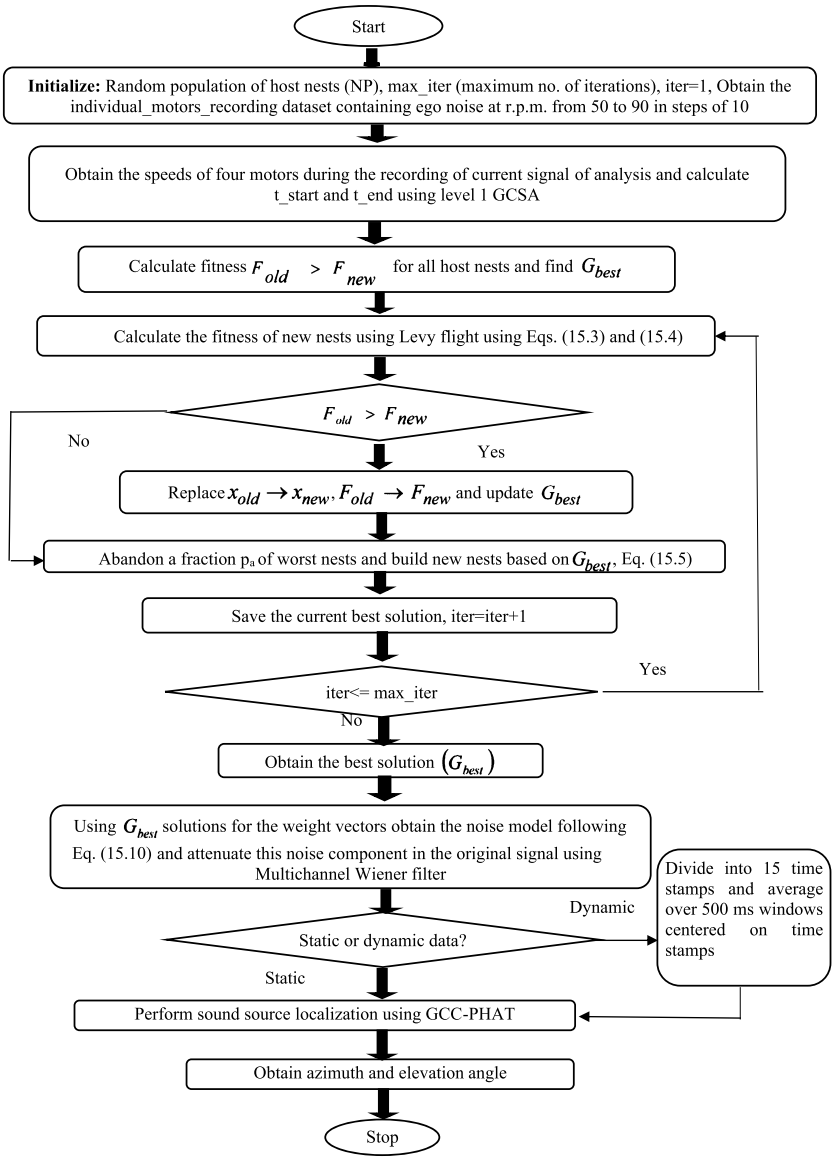

FIGURE 15.16 Flowchart of proposed Algorithm 2.

where $\Delta d = d_1 - d_2$ and c denotes the propagation speed of the signal. Further, from Fig. 15.17, we obtain

$$\Delta d = \sqrt{(x_1 - x_s)^2 + (y_1 - y_s)^2} - \sqrt{(x_2 - x_s)^2 + (y_2 - y_s)^2}. \qquad (15.16)$$

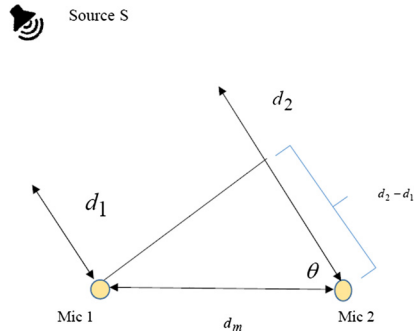

FIGURE 15.17 Orientation of source and microphones.

Similarly, if we have another pair of sensors we can get another difference in distance in terms of (x_s, y_s). Assuming that the microphone sensor array geometry is precisely known, their coordinates are known to us. Hence using numerical methods we find the solution for (x_s, y_s). It should be noted that for the 2D position estimation of the source it is necessary to have two such pairs of receivers, i.e., at least three receivers are required. Again, from the hyperbolic navigation system, we can localize the source using TDOA. As defined, the hyperbola is a set of points having a constant range difference from two points called its foci. So, the point of intersection of all possible curves in the form of equation (15.16) or the point of intersection of all possible hyperbolas considering all the pairs of microphones gives the position of the acoustic emitter. Generalizing this concept for 3D space and n microphones $M_i, i = 1, \ldots, n$, and using vector notation we can write

$$\Delta d_{ij} = \left|\vec{M}_i - \vec{S}\right| - \left|\vec{M}_j - \vec{S}\right|, \qquad i, j = 1, \ldots, n \quad \text{and} i \neq j, \tag{15.17}$$

where Δd_{ij} is the difference in distance for the microphone pair $\left(M_i, M_j\right)$ and is given by

$$\Delta d_{ij} = c\tau_{ij}, \tag{15.18}$$

τ_{ij} being the time difference of arrival for the microphone pair $\left(M_i, M_j\right)$, and $\vec{M}_i$ and $\vec{S}$ denote the position vector of the i^{th} microphone and the source with respect to an inertial frame of reference. Particularly for a UAV-embedded system, we may consider the body-fixed coordinate system of the UAV as the reference frame. Before we calculate the position of the source it is necessary to find the quantity Δt (the difference in arrival time). Cross-correlation is a very powerful tool for estimating the time delays; it has been adopted in this work for calculating Δt. Now considering the received signals are $r_i(t), i = 1, 2$, for a microphone pair we can write

$$r_i(t) = s(t - t_i) * h(t) + n_i(t), \tag{15.19}$$

where $s(t)$ denotes the source signal, $h(t)$ is the impulse response of the medium between the source and the microphones, $n_i(t)$ represents the noise component independent of the source signal, and the symbol '$*$' denotes the convolution operation. The cross-correlation between the two signals is given by

$$R(\tau) = \int_{-\infty}^{\infty} r_1(t)\, r_2(t+\tau)\, dt. \tag{15.20}$$

The function $R(\tau)$ attains the maximum value when τ equals the time delay, i.e., when $\tau = \Delta t$. τ can be obtained by observing the position of occurrence of the peak in the plot of $R(\tau)$ as follows:

$$\tau_{desired} = \arg\max\left(R(\tau)\right), \tag{15.21}$$

$$\Delta t = \tau_{desired}. \tag{15.22}$$

This method of calculating TDOA is called GCC. From this we can find the angular position of the source as follows. If θ is the angle of the source, then

$$\begin{aligned} &\theta = \cos^{-1}\left(c\Delta t / d_m\right) \\ &\Rightarrow \theta = \cos^{-1}\left(\left(d_1 - d_2\right)/d_m\right). \end{aligned} \tag{15.23}$$

A more efficient approach to determine Δt is using frequency domain cross-correlation. A delay in the time domain corresponds to a phase difference in the frequency domain. Therefore, to estimate the time difference we first transform the received signal from the time domain to the frequency domain using short time Fourier transform (STFT). Let $X_i(t,\omega)$ be the STFT of the received signal $r_i(t)$. Then $X_i(t,\omega)$ is given by

$$X_i(t,\omega) = \int_{-\infty}^{\infty} r_i(t)\, w(t-\tau) \exp\left(-j\omega t\right) dt, \tag{15.24}$$

where $w(t)$ represents a window function. From equation (15.19) we obtain

$$X_i(t,\omega) = S(t - t_i, \omega)\, H(t,\omega) + N_i(t,\omega), \tag{15.25}$$

where S, H, and N_i are the signals in a short time Fourier domain and (t,ω) is the corresponding time frequency index. We define

$$\hat{X}(t,\omega) = {}^{X(t,\omega)}/_{|X(t,\omega)|}, \tag{15.26}$$

where $\hat{X}(t,\omega)$ preserves the phase information of $X(t,\omega)$ but it is normalized to obtain unity gain for all frequencies and achieve robustness; $\hat{X}(t,\omega)$, when inverse transformed to time domain, theoretically produces a peak at the corresponding time delay value. Normalization helps in reducing the probability of secondary peaks by reducing the effects of echo, reverberation, and noise. This weighting scheme by taking the phase transform (PHAT) is known as GCC-PHAT.

15.5 Evaluation

In this section, we discuss the results obtained by applying the denoising algorithm on the DREGON dataset and a simulated dataset.

15.5.1 Dataset and design of data acquisition system

Our proposed algorithm has been evaluated using the dataset provided by the IEEE Signal Processing (SP) Cup, 2019 along with the DREGON dataset. This dataset can be divided into two broad sections as follows.

Static data: The Static dataset contains 300 audio data each of eight channels, i.e., each signal is recorded using an array of eight microphones. During the recording of this type of signal, the drone was kept hovering at a particular position, so for the Static data, the coordinate of the receiver is fixed. In addition, the source is assumed to be fixed at a particular point throughout the study. Hence it requires estimating only a single direction (azimuth and elevation) for the sound source. The sampling frequency of the signals was fixed to the value 44.1 kHz and the length of each captured signal is 2 s. For each recording, the rotor speeds of all four motors were varied and knowledge about the rpm value of the rotors is available to the user. As the rotor speeds for each noisy signal were different from the other, it is obvious that the nature of the ego noise varies in relation to the speeds. Thus, in spite of the source and receiver system being static, the ego noise varies for this situation. It is necessary to consider an ego noise model in accordance with the variation of the rotor speeds.

Dynamic data: The Dynamic dataset contains 36 audio data each of eight channels, similar to the Static data. The sampling frequency of the signals was fixed to the value 44.1 kHz and the length of each captured signal is 4 s. During the recording of these signals, the drone was in the flight mode, covering a certain area. The flights were performed in a large room with a moderate reverberation level. Thus the coordinate of the drone system was varying continuously. Hence, with respect to the frame of reference of the drone, the coordinate of the sound source is time-varying. For this reason, it was required to estimate the position (azimuth and elevation) of the source after every small interval of time; for our evaluation we considered the interval to be 0.25 s, i.e., the entire signal was divided into 15 windows. In the dynamic case, the speed of all the four rotors was varied and knowledge about the rpm value of the rotors was available to the user. The primary goal of this task was to identify the average azimuth and elevation of the source when the drone was flying. As the system was in flight mode, the effect of wind noise and the nature of the ego noise were different from that in the static data and it was required to eliminate that noise properly. The dynamic data have two subparts:

1. Broadband data: This dataset was created by emitting white noise from the loudspeaker and the drone was flown over the speaker. The task was to iden-

FIGURE 15.18 The drone-embedded microphone array system used for creating the DREGON dataset (Strauss et al., 2018; Deleforge et al., 2019; IEEE Signal Processing Cup Syllabus, 2019).

tify the white noise source accurately. There were 20 such eight-channel recordings.

2. Speech data: This dataset was created by emitting speech from the loudspeaker and the drone was flown over the speaker. The task was to identify the speech source coordinate after small time intervals. There were 16 such eight-channel recordings.

Development data: This dataset contains the ego noise recordings of the drone. These ego noises are generated mainly by the drone motors rotating at different rpm (varying from 50 to 90 with step size 10). The recordings were captured while rotating a single motor of the drone at a time. From these data, an understanding of the nature of the drone ego noise can be obtained.

Specification of the data acquisition system: As discussed in Strauss et al. (2018), the DREGON dataset has been collected using a customizable quadrotor UAV manufactured by MikroKopter, Germany. The UAV was equipped with four MK2832-35 motors and the sound recording system containing eight microphones in a cubic array and a sound card. The UAV control has been set up using ODROID-XU4 Linux Computer which runs the Robot Operating System. The TeleKyb-genom3 framework for implementing the low level flight control receiving the body-frame velocity commands and Wi-Fi communication has been built for the purpose of drone data transfer with the ground station. The propellers' speeds range from 15 turns/s when they start to around 95 turns/s at maximum power. The total weight of the UAV system is 1.68 kg. The UAV setup along with the microphone array is shown in Fig. 15.18. Figs. 15.19a and 15.19b show the positions of the eight microphones on the surface of the cube. The azimuth and elevation angles, represented by θ and φ, have been calculated with reference to the center of the cubic structure. This is evident from the orientation shown in Fig. 15.19a. The alignment of the microphone array with respect to the UAV frame and its four arms is shown in Fig. 15.19b.

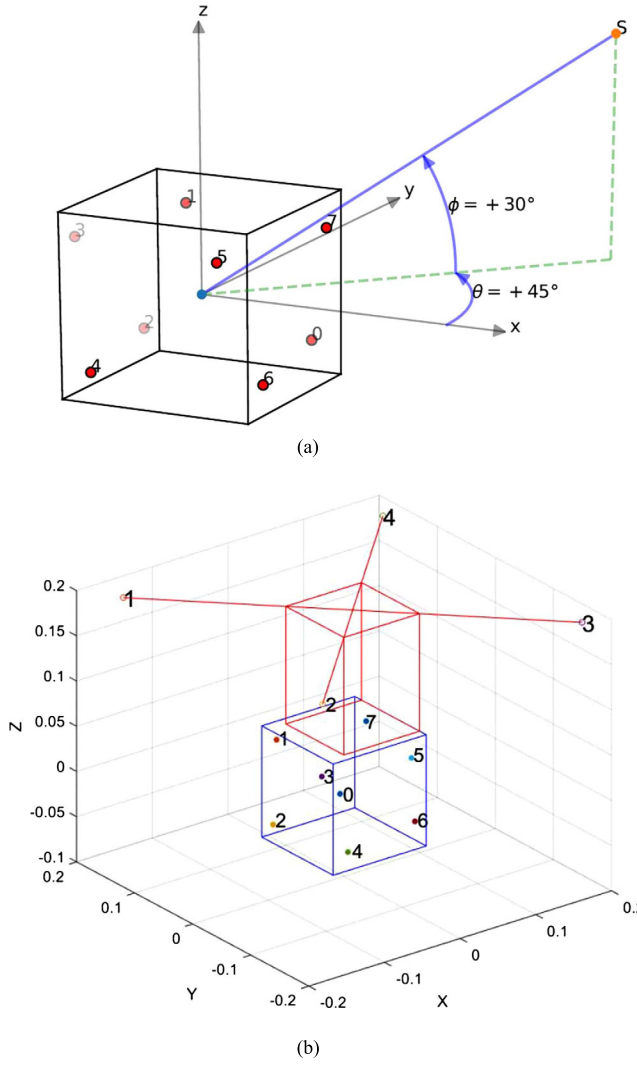

FIGURE 15.19 Eight-MEMS microphone array, (a) placed as a cubical structure and (b) with drone and the rotors in 3D space (Strauss et al., 2018; Deleforge et al., 2019; IEEE Signal Processing Cup Syllabus, 2019).

15.5.2 Evaluation measures

For the evaluation of the results, we used the error estimates. The error denotes the difference in the estimated and actual location of the source. The aim of the algorithm is to minimize the distance between the two points; hence the permissible value of the angle difference has been fixed to less than 10 degrees. The evaluation is mainly based on two types of errors. Absolute error is the absolute

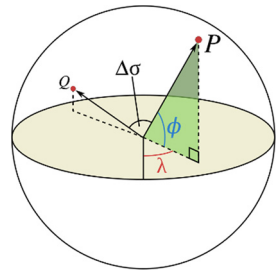

FIGURE 15.20 The central angle ($\Delta\sigma$) between two points p and q, where λ and φ are the azimuth and elevation angles, respectively.

value of the difference between the estimated angle value and the ground truth value. If $(\theta_{estimated}, \varphi_{estimated})$ and $\left(\theta_{ground-truth}, \varphi_{ground-truth}\right)$ are the estimated and ground truth values of the azimuth and elevation angles, respectively, then the absolute error is given by

$$error_{\theta} = \left|\theta_{estimated} - \theta_{ground-truth}\right|, \tag{15.27}$$

$$error_{\varphi} = \left|\varphi_{estimated} - \varphi_{ground-truth}\right|. \tag{15.28}$$

The second type of error is determined from the value of the great circle distance. This is the least distance between two points on a sphere measured along the spherical surface (Fig. 15.20).[1]

The error value ($\Delta\sigma$) called the central angle is given by the following equation:

$$\Delta\sigma = \arctan \frac{\sqrt{\left(\cos\phi_2 \sin\left(\Delta\lambda\right)\right)^2 + \left(\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2 \cos\left(\Delta\lambda\right)\right)^2}}{\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos\left(\Delta\lambda\right)}, \tag{15.29}$$

where (λ_1, ϕ_1) and (λ_2, ϕ_2) are the geographical longitude and latitude in radians of two points 1 and 2 and $\Delta\lambda$, $\Delta\phi$ are their absolute differences. The localization is considered to be correct if $\Delta\sigma$ is less than 10 degrees. The SNR of the unknown input signal and the output signal are estimated as a measure of evaluation. The SNR is calculated as follows:

$$SNR = 10\log_{10} \frac{\sum_{k=0}^{N-1} s^2(k)}{\sum_{k=0}^{N-1} n^2(k)}. \tag{15.30}$$

The improvement in SNR (ΔSNR_{dB}) is given by

$$\Delta SNR_{dB} = SNR_{dB}^{out} - SNR_{dB}^{in}. \tag{15.31}$$

[1] https://en.wikipedia.org/wiki/Great-circle_distance.

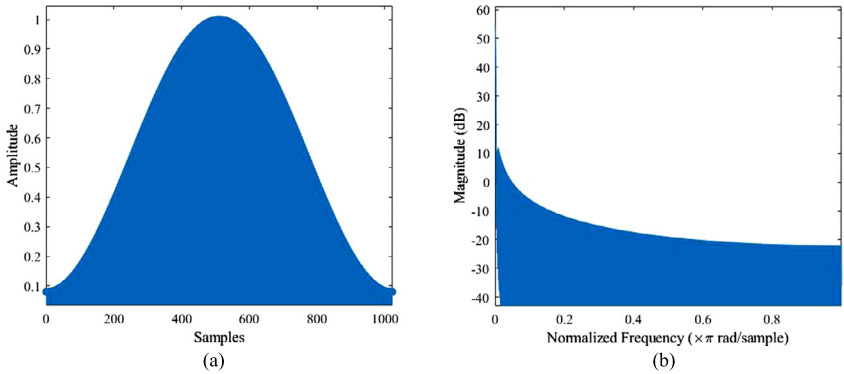

FIGURE 15.21 Response of Hamming window with 1024 points. (a) Time domain. (b) Frequency domain.

A comparative study of the different algorithms has been presented based on the value of the absolute error as well as the root mean square deviation (RMSD),

$$RMSD = \sqrt{\frac{\sum_{k=0}^{N-1}(\hat{\alpha}_k - \alpha_k)^2}{N}}, \tag{15.32}$$

where α and $\hat{\alpha}$ represent the actual and estimated values of the variable of interest, respectively, and N is the total number of samples.

15.5.3 Parameter initialization

From the geometry, it is clear that the value of azimuth and elevation angles should be in the range of $[-179°, 180°]$ and $[-90°, 90°]$, respectively. For the evaluation, we have tried with different values of population size (NP) and maximum number of iterations (max_iter). NP and max_iter are finally fixed at 10 and 10, respectively, in order to obtain the optimum results. In addition, the speed of sound has been considered to be equal to 343 m/s and the cut-off frequency of the preprocessing filter has been fixed to 4 kHz as it produces a good extent of elimination of background noise. For computing, the STFT of the multichannel signals window length was fixed to be 1024 samples and the following Hamming window (Proakis and Manolakis, 1996) has been used:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1. \tag{15.33}$$

This is a type of cosine sum window whose time and frequency domain responses are given in Figs. 15.21a and 15.21b, respectively.

TABLE 15.1 Azimuth and elevation angles (in degrees) for the 50 signals from the DREGON dataset.

Signal number	Ground truth		Without denoising		Source separation		Wavelet denoising		Proposed Algorithm 1		Proposed Algorithm 2	
	λ	φ	λ	φ	λ	φ	λ	φ	λ	φ	λ	φ
2	60	−30	−95	90	−109	49	47	90	111	52	56	−19
3	45	0	45	0	45	0	−159	90	45	−30	45	0
4	75	0	78	1	75	1	115	90	75	−1	76	−1
5	90	−15	−179	90	−23	55	115	90	−36	66	90	−14
6	75	−15	−101	76	140	49	158	90	−100	76	78	−12
7	90	−15	−34	62	−44	54	115	90	−29	66	88	−13
8	90	0	90	0	89	1	70	90	90	0	90	0
10	45	−15	45	−16	43	−11	115	90	47	−14	45	−15
11	**90**	**−15**	**−136**	**90**	**20**	**54**	**−133**	**90**	**99**	**80**	**88**	**−14**
12	90	0	90	0	−108	49	−114	90	90	0	90	0
13	75	−15	47	90	−42	36	−133	90	−14	51	76	−15
16	90	−30	87	−28	90	−28	−179	−90	90	−28	90	−28
17	45	−15	−108	44	−106	45	−179	90	−108	51	45	−15
18	90	−30	−114	90	90	−23	−174	90	89	−27	89	−27
19	60	0	25	90	162	50	−179	90	−134	89	56	−3
20	45	0	−179	90	−25	59	−179	90	−25	59	45	0
22	75	−15	79	−15	81	−12	−179	90	75	−14	79	−15
25	**75**	**0**	**−179**	**90**	**76**	**−1**	**115**	**90**	**75**	**−1**	**79**	**−1**
31	60	−15	−25	57	−61	74	−179	90	−26	58	64	−15
33	75	−15	79	−11	79	−16	115	90	75	−14	77	−13
35	60	−15	25	90	108	43	−114	90	110	53	62	−12
36	60	−15	−179	90	63	−15	−133	90	59	−14	58	−14
37	75	−15	68	90	77	−15	115	90	75	−14	78	−14
39	45	−30	25	90	109	55	115	90	161	55	43	−29
40	75	−30	79	−28	73	−25	−179	90	76	−28	74	−25
41	75	−30	44	90	74	−22	156	90	75	−25	73	−24
42	75	0	79	0	75	1	115	90	76	0	76	0
43	60	−30	115	90	−124	85	−179	−90	58	−26	57	−21
45	60	−30	56	−26	57	−25	115	90	57	−25	57	−25
47	45	−15	−159	90	43	−11	−159	90	47	−15	45	−17
48	45	0	−31	60	45	0	115	90	45	0	45	0
49	60	−15	58	−16	−20	48	115	90	60	−15	59	−14

continued on next page

15.5.4 Experimental results

The evaluation of the proposed algorithms is performed using the Static data from the DREGON dataset. Fifty test signals from the dataset were used for the performance analysis. The results are presented in Table 15.1 in terms of

TABLE 15.1 (*continued*)

Signal number	Ground truth		Without denoising		Source separation		Wavelet denoising		Proposed Algorithm 1		Proposed Algorithm 2	
	λ	φ	λ	φ	λ	φ	λ	φ	λ	φ	λ	φ
51	60	−15	25	90	−25	59	115	90	−37	67	59	−13
55	90	−15	−136	90	−21	44	−159	90	−135	81	83	−11
57	45	0	−11	45	145	51	47	90	−17	55	45	0
58	60	−30	25	90	−63	53	25	90	161	55	61	−26
59	90	−15	−107	55	88	−15	70	90	89	−15	90	−14
60	90	−30	−36	54	−47	54	−11	11	−44	54	87	−26
62	60	0	56	−1	60	0	115	90	61	0	59	−2
67	45	−30	−106	55	−27	57	−159	90	−110	53	44	−26
69	90	−15	−26	55	−35	83	−159	90	−27	65	89	−13
72	45	0	−11	56	43	3	25	90	45	0	45	0
73	90	−30	85	90	80	76	115	90	114	54	88	−28
75	60	0	115	90	118	48	25	90	69	85	47	90
80	90	0	101	56	85	48	25	90	110	54	91	2
85	45	0	112	59	−134	89	−65	90	112	54	45	0
86	75	0	−133	90	30	87	−133	90	−114	90	79	−1
87	90	0	−174	90	−21	49	115	90	−18	48	90	0
88	45	−30	25	90	44	−31	115	90	45	−28	45	−26
89	90	−30	−101	45	−100	46	115	90	−107	50	89	−27

azimuth (λ) and elevation (φ) angles. The ground truth represents the original azimuth and elevation angle values of the detected source. The estimations of source location using our algorithm have been presented under columns named proposed Algorithm 1 and proposed Algorithm 2. The column named "Without denoising" represents the location estimation when only the GCC-PHAT algorithm was used without any estimation of noise. The results have been compared with that of the preexisting algorithms. The source separation algorithm by Gao et al. (2013) has been extended from single-channel to multichannel and is used to separate the true speech from the noise. This original speech signal was then used for source position estimation. Another approach based on wavelet denoising (Jain and Tiwari, 2017; Ali et al., 2017) for comparison. For example, if we consider signal 11, it contains a speech signal originated from the location with azimuth and elevation angles (in degrees) (90, −15). The location estimated by proposed Algorithm 1 is (99, 80) and that of proposed Algorithm 2 is (88, −14). With GCC-PHAT only the localization result equals (−136, 90). Using source separation we get the result (20, 54) and estimation by wavelet algorithm gives (−133, 90).

Fig. 15.22 shows the position localization results, i.e., azimuth (λ) and elevation (φ) angles of the static speech source represented in $\lambda - \varphi$ plane. For test signal 11, Figs. 15.22a–15.22e depict the results of localization and

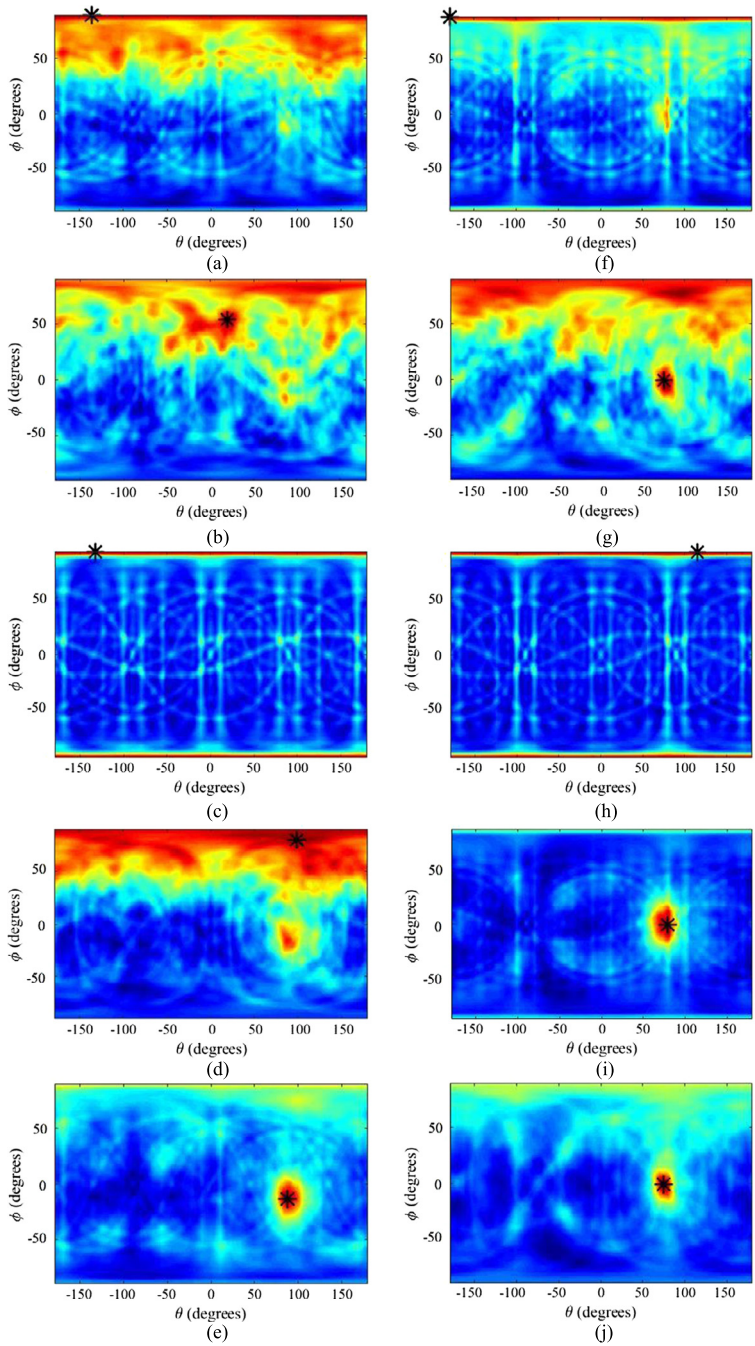

FIGURE 15.22 Azimuth and elevation angles for test signals DS11 and DS25 estimated by (a, f) GCC only, (b, g) source separation, (c, h) wavelet denoising, (d, i) proposed Algorithm 1, and (e, j) proposed Algorithm 2.

Figs. 15.22f–15.22j illustrate the positions of signal 25. The result values are given in Table 15.1. The marker represents the location of the static source. It should be observed that if speech is present, the SNR automatically increases and we can infer that the locations having a higher value of SNR are the probable positions of the signal source. This is evident from the fact that in each of the localization results the markers are pointed at the high SNR regions. Figs. 15.22e and 15.22j show the results for proposed Algorithm 2, and it yields the correct result without any ambiguity. For both cases, it contains a single region with high SNR and that is the correct coordinate of the source. However, proposed Algorithm 1 does not provide the correct result for every case. As can be seen in Figs. 15.22d and 15.22i, the result is correct for Fig. 15.22i, but for Fig. 15.22d there is ambiguity in the result. The algorithm is not able to mark the correct region of high SNR; rather it identifies some other erroneous regions as high SNR region and localizes the source there. Applying only the source localization algorithm GCC, it is not possible to achieve correct coordinates. Both results shown in Figs. 15.22a and 15.22f are incorrect. This is because the high amplitude noise almost completely masks the speech source. Thus it indicates the necessity of a noise suppression approach before applying the position identification methods. It was mentioned previously that due to the nature of the noise signals, traditional denoising algorithms cannot perform well for this case. It is evident from the results in Figs. 15.22b, 15.22c, 15.22g, and 15.22h that the estimation of both the source separation algorithm and the wavelet-based algorithm are absolutely incorrect.

Figs. 15.23a and 15.23b show the variation in the value of the absolute errors for azimuth and elevation angles estimated by GCC-PHAT, source separation algorithm, wavelet denoising, proposed Algorithm 1, and proposed Algorithm 2 with respect to the ground truth values of the coordinate. As can be observed for both plots, the error for proposed Algorithm 2 is lower than that for the other approaches. For some signals, proposed Algorithm 1 performs as good as proposed Algorithm 2 in estimating the angle values, but that is not consistent for all signals. On the other hand, only GCC-PHAT localization cannot estimate the coordinate of the source for noisy signals. The signals for which GCC-PHAT estimation error is close to zero have been tested to contain very low noise. The same is true for the other denoising algorithms. Thus, we can conclude that proposed Algorithm 2 outperforms the other approaches for both azimuth and elevation angle estimation. Fig. 15.24 shows the values of the central angle ($\Delta\sigma$) calculated using equation (15.29) and the coordinate estimates using the three different algorithms. From the plot it is clear that using proposed Algorithm 2, we obtain $\Delta\sigma$ less than 10 degrees for most of the test signals. But it is not true for the other cases. Particularly, when only GCC is applied, the values of $\Delta\sigma$ are much higher than 10 degrees. As stated previously, if $\Delta\sigma$ is less than 10 degrees, then only the localization is assumed to be correct, and we can now conclude that proposed Algorithm 2 outperforms proposed Algorithm 1 in all cases.

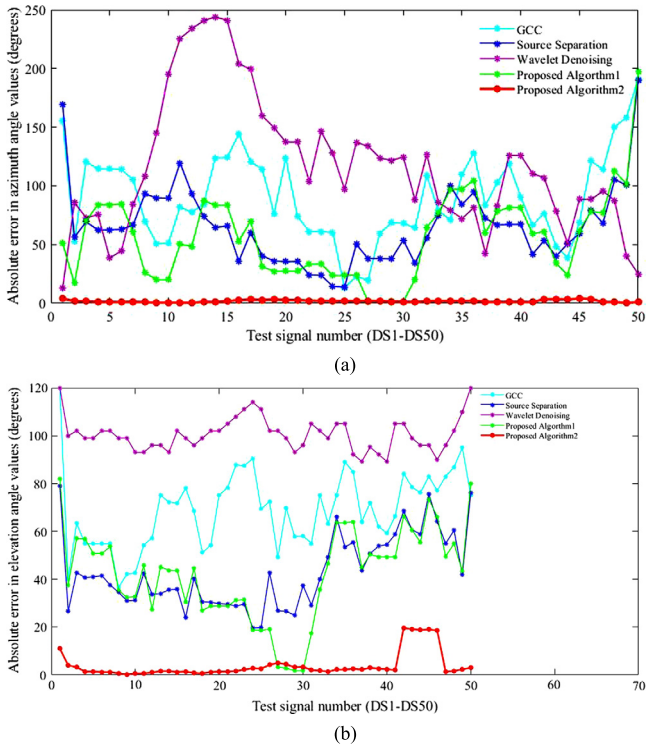

FIGURE 15.23 Absolute errors with respect to ground truth estimated by three different algorithms. (a) Azimuth angle. (b) Elevation angle.

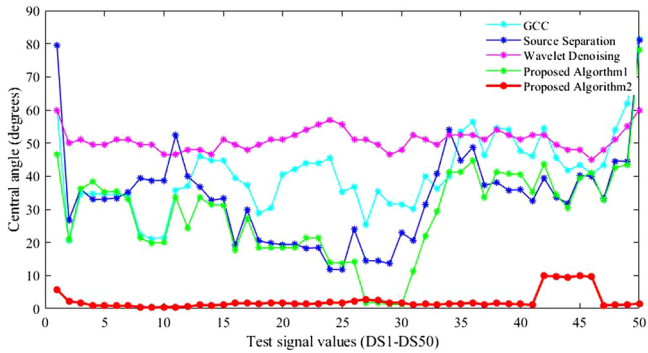

FIGURE 15.24 Values of central angle estimated using three different algorithms.

Fig. 15.25 illustrates the SNR improvement achieved for the test signals DS1-DS50 using proposed Algorithms 1 and 2. It can be observed that for proposed Algorithm 2 SNR improvement values are very high, i.e., SNR of the output signal is much higher than that of the input signal. So, proposed Al-

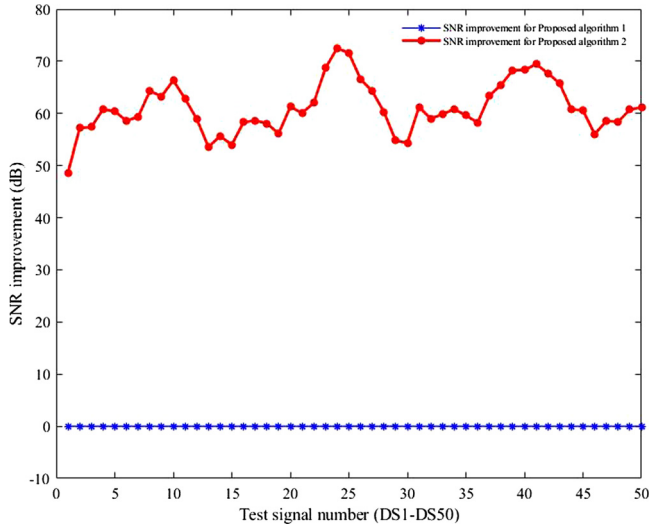

FIGURE 15.25 Variation of output SNR with respect to input SNR for proposed Algorithms 1 and 2.

TABLE 15.2 Mean and RMSD of the error values; best results are indicated in boldface.

Algorithm	Absolute error in azimuth angle		Absolute error in elevation angle		Central angle	
	Mean	RMSD	Mean	RMSD	Mean	RMSD
GCC-PHAT	89.02	90.92	68.4	44.52	40.18	25.13
Source separation algorithm	62.60	78.81	41.94	37.18	32.26	28.22
Wavelet denoising	118.72	92.47	100.12	16.44	50.70	7.03
Proposed Algorithm 1	55.16	68.51	41.12	38.95	27.98	27.32
Proposed Algorithm 2	**1.76**	**2.25**	**3.78**	**12.65**	**2.25**	**6.29**

gorithm 2 not only offers noise suppression but it also enhances the original speech signal. On the other hand, SNR improvement achieved from proposed Algorithm 1 is almost equal to zero for all the signals. Thus, we conclude that this algorithm does not perform any signal enhancement though it can yield correct localization of the signal source. In this context, it should be noted that for the GCC-PHAT algorithm noise suppression is not performed. Hence, for that case, there is no possibility of SNR improvement, and the output signal is the same as the input noisy signal. Table 15.2 presents the mean and RMSD of the

errors obtained using the three algorithms. For proposed Algorithm 2 the mean and RMSD for all the three cases, namely, absolute error in azimuth angle, absolute error in elevation angle, and the central angle, are very small, indicating that it yields very accurate localization. Larger mean and deviation of errors indicate erroneous position estimation. GCC-PHAT and wavelet denoising yields the worst localization.

15.6 Discussion

For the evaluation of our proposed algorithms, 50 test signals were taken from the dataset provided by Strauss et al. (2018). The localization was obtained using GCC-PHAT, source separation algorithm, wavelet denoising, proposed Algorithm 1, and proposed Algorithm 2, and the results have been compared with the ground truth values. It has been shown that for all the signals the GCC-PHAT algorithm fails to yield the correct result. The traditional noise suppression approaches yield completely incorrect estimations. However, when this same algorithm is applied after noise suppression, the error values are reduced drastically. In addition, depending on the effectiveness of the noise suppression block the error varies. We have considered both absolute errors and central angle errors for the evaluation and for both cases proposed Algorithm 2 provides the best results with both types of errors being in the range of less than 10 degrees. Proposed Algorithm 2 achieves accuracy in the range of 1.76 degree for the azimuth angle, 3.78 degree for the elevation angle, and 2.25 degree for the central angle. The mean and standard deviation measures of the results as depicted in Table 15.2 indicate the same. It can be concluded that proposed Algorithm 2 outperforms the other two algorithms. In addition, this algorithm offers signal enhancement along with source localization.

15.7 Conclusion

In this work, we have presented a novel approach for ego noise suppression and SSL for the drone-embedded microphone array-based recordings that can be applied during disaster management and for SAR operations to identify the position of the victim. Two separate algorithms are proposed for the estimation of unknown ego noise using GCS optimization. In proposed Algorithm 1, temporal subtraction of noise from the original noisy signal is applied, whereas proposed Algorithm 2 uses the multichannel Wiener filter for attenuating the estimated noise from the noisy input signal. Further GCC-PHAT localization is used for speech source coordinate estimation. For the experimental validation, the proposed algorithms have been evaluated using noisy signals containing a single speech source, exhibiting significant performance improvements. The proposed work can be further extended for localizing multiple sources simultaneously. A hybrid approach of the GCS algorithm with other existing heuristic search-based optimization algorithms can be implemented and tested for more accurate noise elimination and SSL.

References

Ali, M.N., El-Dahshan, E.S.A., Yahia, A.H., 2017. Denoising of heart sound signals using discrete wavelet transform. Circuits, Systems, and Signal Processing 36 (11), 4482–4497.

Basiri, M., Schill, F., Lima, P.U., Floreano, D., 2012. Robust acoustic source localization of emergency signals from micro air vehicles. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4737–4742.

Blandin, C., Ozerov, A., Vincent, E., 2012. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Signal Processing 92, 1950–1960.

Chen, L., Liu, Y., Kong, F., He, N., 2011. Acoustic source localization based on generalized cross-correlation time-delay estimation. Procedia Engineering 15, 4912–4919.

Deleforge, A., Di Carlo, D., Strauss, M., Serizel, R., Marcenaro, L., 2019. Audio-based search and rescue with a drone: highlights from the IEEE Signal Processing Cup 2019 Student Competition [SP Competitions]. IEEE Signal Processing Magazine 36 (5), 138–144.

Dhabal, S., Venkateswaran, P., 2017. An efficient Gbest-guided Cuckoo Search algorithm for higher order two channel filter bank design. Swarm and Evolutionary Computations 33, 68–84.

Dorfan, C.E.Y., Gannot, S., Naylor, P.A., 2016. Speaker localization with moving microphone arrays. In: Proc. of 24th European Signal Processing Conference (EUSIPCO), pp. 1003–1007.

Fan, J., Luo, Q., Ma, D., 2010. Localization estimation of sound source by microphones array. Procedia Engineering 7, 312–317.

Furukawa, K., et al., 2013. Noise correlation matrix estimation for improving sound source localization by multirotor UAV. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, pp. 3943–3948.

Gala, D., Lindsay, N., Sun, L., 2018. Three-dimensional sound source localization for unmanned ground vehicles with a self-rotational two-microphone array. In: Proc. of the 5th International Conference of Control, Dynamic Systems, and Robotics, Vol. 104, pp. 1–11.

Gao, B., Woo, W.L., Dlay, S.S., 2013. Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and Itakura–Saito nonnegative matrix two-dimensional factorizations. IEEE Transactions on Circuits and Systems I: Regular Papers 60 (3), 662–675.

Haubner, T., Schmidt, A., Kellermann, W., 2018. Multichannel nonnegative matrix factorization for ego-noise suppression, speech communication. In: Proc. of 13th ITG-Symposium, pp. 1–5.

Huang, Q., Wang, T., 2014. Acoustic source localization in mixed field using spherical microphone arrays. EURASIP Journal on Advances in Signal Processing 90, 1–16.

IEEE Signal Processing Cup Syllabus, 2019. http://dregon.inria.fr/SPCup2019/SPCup2019_syllabus.pdf.

Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., Imura, J., 2009. Ego noise suppression of a robot using template subtraction. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 199–204.

Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., Imura, J., 2010. A hybrid framework for ego noise cancellation of a robot. In: Proc. of IEEE International Conference on Robotics and Automation, pp. 3623–3628.

Ince, G., Nakadai, K., Rodemann, T., Imura, J., Nakamura, K., Nakajima, H., 2011. Assessment of single-channel ego noise estimation methods. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 106–111.

Jain, P.K., Tiwari, A.K., 2017. An adaptive thresholding method for the wavelet based denoising of phonocardiogram signal. Biomedical Signal Processing and Control 38, 388–399.

Jung, C., Liu, R., Lian, K., 2017. A fast searching algorithm for real-time sound source localization. In: Proc. of 56th Annual Conference of the Society of Instrument and Control Engineers of Japan, pp. 1413–1416.

Kwok, N.M., Buchholz, J., Fang, G., Gal, J., 2005. Sound source localization: microphone array design and evolutionary estimation. In: Proc. of IEEE International Conference on Industrial Technology, pp. 281–286.

Li, X., Shen, M., Wang, W., Liu, H., 2012. Real-time sound source localization for a mobile robot based on the guided spectral-temporal position method. International Journal of Advanced Robotic Systems 9, 1–8.

Löllmann, H.W., et al., 2017. Microphone array signal processing for robot audition. In: Proc. of Hands-Free Speech Communications and Microphone Arrays (HSCMA), pp. 51–55.

Löllmann, H.W., Hendrik, B., Deleforge, A., Meier, S., Walter, K., 2014. Challenges in acoustic signal enhancement for human-robot communication. In: Proc. of 11th ITG Symposium on Speech Communication, pp. 1–4.

Ma, N., Gonzalez, J.A., Brown, G.J., 2018. Robust binaural localization of a target sound source by combining spectral source models and deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (11), 2122–2131.

Misra, P., Kumar, A.A., Mohapatra, P., Balamuralidhar, P., 2018. DroneEARS: robust acoustic source localization with aerial drones. In: Proc. of IEEE International Conference on Robotics and Automation, pp. 80–85.

Morito, T., Sugiyama, O., Kojima, R., Nakadai, K., 2016. Partially shared deep neural network in sound source separation and identification using a UAV-embedded microphone array. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1299–1304.

Nakadai, K., Ince, G., Nakamura, K., Nakajima, H., 2012. Robot audition for dynamic environments. In: Proc. of IEEE International Conference on Signal Processing, Communication and Computing, pp. 125–130.

Nogueira, L.C.F., Petraglia, M.R., 2015. Robust localization of multiple sound sources based on BSS algorithms. In: Proc. of IEEE 24th International Symposium on Industrial Electronics, pp. 579–583.

Ohata, T., Nakamura, K., Mizumoto, T., Taiki, T., Kazuhiro, N., 2014. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array. In: Proc. of IEEE International Conference on Intelligent Robots and Systems, pp. 1902–1907.

Okutani, K., Yoshida, T., Nakamura, K., Nakadai, K., 2012. Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3288–3293.

Proakis, J.G., Manolakis, D.G., 1996. Digital Signal Processing: Principles, Algorithms, and Applications, 3rd ed. Prentice-Hall, Inc.

Restas, A., 2015. Drone applications for supporting disaster management. World Journal of Engineering and Technology 3, 316–321.

Schmidt, A., Deleforge, A., Kellermann, W., 2016. Ego-noise reduction using a motor data-guided multichannel dictionary. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1281–1286.

Schmidt, M.N., Larsen, J., Hsiao, F., 2007. Wind noise reduction using non-negative sparse coding. In: Proc. of IEEE Workshop on Machine Learning for Signal Processing, pp. 431–436.

Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Processing Letters 6 (1), 1–3.

Strauss, M., Mordel, P., Miguet, V., Deleforge, A., 2018. DREGON: dataset and methods for UAV-embedded sound source localization. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1–8.

Tezuka, T., Yoshida, T., Nakadai, K., 2014. Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization. In: Proc. of IEEE International Conference on Robotics and Automation (ICRA), pp. 6293–6298.

Velasco, J., Pizarro, D., Macias-Guarasa, J., 2012. Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints. Sensors 12, 13781–13812.

Wang, L., Cavallaro, A., 2016. Ear in the sky: ego-noise reduction for auditory micro aerial vehicles. In: Proc. of 13th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 152–158.

Wang, L., Cavallaro, A., 2018. Acoustic sensing from a multi-rotor drone. IEEE Sensors Journal 18, 4570–4582.

Wang, T., Choy, Y., 2015. An approach for sound sources localization and characterization using array of microphones. In: Proc. of International Conference on Noise and Fluctuations (ICNF), pp. 1–4.

Xenaki, A., Boldt, J.B., 2018. Sound source localization and speech enhancement with sparse Bayesian learning beamforming. The Journal of the Acoustical Society of America 143 (6), 3912–3921.

Yang, X.S., Deb, S., 2010. Engineering optimisation by cuckoo search. International Journal of Mathematical Modelling and Numerical Optimisation 1, 330–343.