TIME-FREQUENCY PROCESSING FOR SOUND SOURCE LOCALIZATION FROM A MICRO AERIAL VEHICLE

Lin Wang, Andrea Cavallaro

Centre for Intelligent Sensing, Queen Mary University of London
{lin.wang; a.cavallaro}@qmul.ac.uk

ABSTRACT

We address the problem of sound source localization with a microphone array mounted on a micro aerial vehicle (MAV). Due to the noise generated by motors and propellers, this scenario is characterized by extremely low signal-to-noise ratios (SNR). Based on the observation that the energy of MAV sound recordings is usually concentrated at isolated time-frequency bins, we propose a time-frequency processing framework to address this problem. We first estimate the direction of arrival of the sound at individual time-frequency bins. Then we formulate a set of spatially informed filters pointing at candidate directions in the search space. The output of the filtering tends to present high non-Gaussianity when the spatial filter is steered towards the target sound source. Finally, by measuring the non-Gaussianity of the spatial filtering outputs we build a spatial likelihood function from which we estimate the direction of the target sound. Experimental results with real-recorded MAV ego-noise show the superiority of the proposed method over the state of the art in performing source localization robustly.

Index Terms— Ego-noise, micro aerial vehicle, microphone array, source localization, time-frequency processing

1. INTRODUCTION

Multirotor micro aerial vehicles (MAV) are increasingly used as mobile sensing platforms equipped with a variety of sensors in a wide range of applications, such as surveillance, broadcasting, and search and rescue [1–8]. When an MAV is equipped with microphones for multichannel recording, sound source localization is an important task that aims to estimate the location of a target sound, such as a human speech or an emergency whistle [8–11]. However, this task is affected by strong ego-noise, which masks the target sound and degrades the recording quality significantly [12]. The motors and propellers, which contribute to the ego-noise, are much closer to the microphones than the emitter of the target sound, thus resulting in extremely low signal-to-noise ratios (SNR), e.g. -20 dB [13]. The spectrum of the nonstationary ego-noise depends on the rotation speed of each motor and changes with the varying MAV behavior.

Although microphone array-based source localization has been investigated intensively [14–17], most algorithms are developed for indoor environments with a relatively high SNR (e.g., >0 dB). Only a few works have been reported on the challenging MAV-based source localization in extremely low-SNR scenarios (e.g., < -15 dB). These works can be classified as unsupervised and supervised. Unsupervised approaches perform source localization using microphone signals only. Examples include steered response power with phase transform (SRP-PHAT) [8, 18] and multiple signal classification (MUSIC) [9, 19]. SRP-PHAT, which computes a spatial likelihood map by exploiting the correlation between microphone signals, tends to show degraded performance in MAV-based applications when the target sound is masked by strong ego-noise [8, 18]. MUSIC computes a spatial likelihood map by decomposing the observed signal into orthogonal signal and noise subspaces. The algorithm assumes uncorrelated noise components at the microphones and thus the signal and noise subspaces can be discriminated easily. In MAV-based applications the discrimination becomes difficult since the ego-noise is directional and stronger than the target sound. GEVD-MUSIC (Generalized eigen-value decomposition MUSIC) exploits a noise correlation matrix as additional information to improve robustness to noise. Although several schemes have been proposed to blindly estimate the noise correlation matrix from the microphone signal [9, 19], the estimation is usually inaccurate due to the nonstationarity of the ego-noise. To solve this problem, supervised approaches (which need additional sensors to monitor the behavior of the MAV) were proposed to build a noise template database, from which the noise correlation matrix can be estimated corresponding to the motor rotation speed and the MAV behavior [20–22]. The accurate ego-noise estimation enables supervised approaches to improve source localization performance in low-SNR scenarios. However, the need for dedicated monitoring sensors limits the versatility and applicability of supervised approaches.

In this paper we propose an unsupervised approach for sound source localization from a microphone array mounted on an MAV. The ego-noise mainly consists of harmonic components whose energy peaks at isolated harmonic frequencies (Fig. 1). Likewise, the target sound (e.g., human speech or emergency whistle) mainly consists of harmonic components. Based on the observation that the harmonics of the ego-noise and the target sound usually occupy

Fig. 1. (a) A multirotor MAV with an array of $M = 8$ microphones capturing a target sound. (b) Time-frequency spectrum of the ego-noise and a speech signal (present during 5-7 s and indicated with a red box) recorded by microphone $r_1$ on an operational MAV. The harmonics of the ego-noise and the speech signal generally occupy different time-frequency bins. Our method exploits this time-frequency sparsity.
Local DOA estimation
Spatially informed filtering
Kurtosis
Spatial likelihood map

\(x_1(k,l)\) \(x_M(k,l)\) \(\theta_{TF}(k,l)\) \(y(k,l,\theta_1)\) \(y(k,l,\theta_D)\) \(\theta_1, \ldots, \theta_D\) \(\kappa(k,\theta_1)\) \(\kappa(k,\theta_D)\) \(\rho_{SP}(\theta)\)

\[\ldots\]

\[\ldots\]

\(R\)

Fig. 2. Block diagram of the proposed method for sound source localization with \(M\) microphones mounted on an MAV. For the definition of the variables please see the body of the paper.

2. PROPOSED METHOD

2.1. Preliminaries

Let a circular array with \(M = 8\) microphones be mounted on a multirotor MAV. Let \(r_m = [r_{mx}, r_{my}]^T\) be the location of the \(m\)-th microphone. The locations of the microphones \(\mathbf{R} = [r_1, \ldots, r_M]\) are assumed to be known. Let a target sound source located in the far field emit sound with DOA \(\theta\.\) The corresponding microphone signal, \(x(n) = [x_1(n), \ldots, x_M(n)]^T\), contains the target sound, \(s(n) = [s_1(n), \ldots, s_M(n)]^T\), and the ego-noise, \(v(n) = [v_1(n), \ldots, v_M(n)]^T\), i.e. \(x(n) = s(n) + v(n)\). This signal can be written in the short-time Fourier transform (STFT) domain as \(x(k,l) = s(k,l) + v(k,l)\), where \(k\) and \(l\) denote the frequency and frame indices, respectively. Let \(K\) and \(L\) denote the total number of frequency bins and time frames, respectively.

Since the motors and propellers are closer to the microphones than the target sound source, the MAV sound recording usually presents an extremely low SNR. We assume a low-reverberant environment without natural wind and that the MAV hovers stably while recording the sound from a static source (i.e., the locations of the microphones and the sound source are fixed).

We consider a single target sound from the noisy microphone signal and aim (i) to detect the existence of the target sound and (ii) to estimate its direction. To this end, we compute a spatial likelihood function \(\rho(\theta)\) that presents a peak value corresponding to the DOA of the target sound.

In the following, we present the proposed method, which is composed of three main steps (Fig. 2): local DOA estimation, spatially informed filtering and spatial likelihood function computation.

2.2. Local DOA estimation

Given the microphone signal \(x(k,l)\) and the microphone locations \(\mathbf{R}\), the DOA of the sound at each time-frequency bin can be estimated by building a local spatial likelihood function [25]

\[
\gamma_{\text{TF}}(k,l,\theta) = \Re \left\{ \sum_{m_1, m_2 = 1 \atop m_1 \neq m_2}^M \frac{x_{m_1}(k,l)x_{m_2}^*(k,l)}{|x_{m_1}(k,l)x_{m_2}(k,l)|} e^{j2\pi f_k \tau_{m_1, m_2, \theta}} \right\},
\]

where \(f_k\) denotes the frequency at the \(k\)-th bin; \(\tau_{m_1, m_2, \theta} = \frac{\|r_{m_2} - r_\theta\|}{\|r_{m_1} - r_\theta\|}\) denotes the delay between two microphones \(m_1\) and \(m_2\) with respect to the sound coming from \(\theta\), where \(r_\theta\) is the location of the far-field sound source with DOA \(\theta\) and \(c\) is the velocity of sound. The term \(e^{j2\pi f_k \tau_{m_1, m_2, \theta}}\) represents the inter-channel phase difference, theoretically computed with the delay \(\tau\); the term \(\frac{x_{m_1}(k,l)x_{m_2}^*(k,l)}{|x_{m_1}(k,l)x_{m_2}(k,l)|}\) represents the inter-channel phase difference measured from \(x_{m_1}\) and \(x_{m_2}\), where the superscript ‘*’ denotes complex conjugation; the operator \(\Re\{\cdot\}\) denotes the real component of the argument. The spatial likelihood \(\gamma_{\text{TF}}\) tends to present a high value if these two inter-channel phase differences are consistent with each other. The DOA can thus be estimated as

\[
\theta_{\text{TF}}(k,l) = \arg \max_\theta \gamma_{\text{TF}}(k,l,\theta).
\]

2.3. Spatially informed filtering

The localization results at individual time-frequency bins can be used to construct a spatially informed filter, which extracts the sound coming from a direction \(\theta\) [24, 25]. To implement this spatial filter, we first detect the time-frequency bins that belong to the desired direction, assuming the DOAs estimated at these time-frequency bins to be Gaussian distributed with mean \(\theta\) and standard variance \(\sigma\). The detection is then performed by measuring the closeness of each time-frequency bin to the direction \(\theta\):

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

where the closeness measure \(c_d\) lies in the interval \([0, 1]\), with a higher value indicating a higher probability that the \((k, l)\)-th bin belongs to the target sound. Next, we calculate the correlation matrix of the sound signal from the direction \(\theta\) as

\[
\Phi_{ss}(k,l,\theta) = \frac{1}{L} \sum_{l=1}^L c_d^2(k,l,\theta)x^H(k,l)x(k,l),
\]

where the closeness measure \(c_d(k,l,\theta)\) indicates the contribution of the \((k, l)\)-th bin to the correlation matrix, and the superscript \(^H\) denotes the Hermitian transpose. Given this estimated target sound correlation matrix, an adaptive beamformer can be formulated easily. We use the multichannel Wiener filter [26]

\[
\mathbf{w}_{\text{TF}}(k,l,\theta) = \Phi_{xx}^{-1}(k,l)\phi_{ss}(k,l,\theta),
\]

where \(\phi_{ss}(k,l,\theta)\) denotes the first column of \(\Phi_{ss}(k,l,\theta)\), and \(\Phi_{xx}(k,l)\) is the correlation matrix of the microphone signal, which can be estimated directly using \(\Phi_{xx}(k,l) = \frac{1}{L} \sum_{l=1}^L x(k,l)x^H(k,l)\).

The sound coming from \(\theta\) is extracted as

\[
y_{\text{TF}}(k,l,\theta) = \mathbf{w}_{\text{TF}}^H(k,l,\theta)x(k,l).
\]
In this way we can construct multiple spatial filters pointing at a set of predefined candidate directions \(\{\theta_1, \ldots, \theta_D\}\) in the search space.

2.4. Spatial likelihood function

We compare the non-Gaussianity of the \(D\) spatial filtering outputs. The non-Gaussianity of a sequence can be measured with its statistical kurtosis value, where a higher kurtosis indicates a higher non-Gaussianity [27]. The kurtosis value \(\xi(k, \theta)\) is calculated at each frequency bin:

\[
\xi(k, \theta) = K(\tilde{y}_H(k, \theta)),
\]

\(\text{where } \tilde{y}_H(k, \theta)\) denotes the time sequence \([y_H(k, \cdot); \theta]\) and \(K(\cdot)\) denotes the kurtosis value of the sequence. Considering the whole frequency band, we calculate a global spatial likelihood function as

\[
\hat{\rho}_S(\theta) = \frac{1}{K} \sum_{k=1}^{K} \xi(k, \theta).
\]

We normalize (8) into the interval \([0, 1]\) by using

\[
\rho_S(\theta) = \mathcal{N}(\hat{\rho}_S(\theta)) = \max\left(\frac{\hat{\rho}_S(\theta) - \max(\rho_{TH}, \text{mean}(\hat{\rho}_S))}{\max(\hat{\rho}_S) - \min(\hat{\rho}_S)}, 0\right),
\]

where \(\min(\cdot), \max(\cdot), \text{ and } \text{mean}(\cdot)\) denote the minimum, maximum and mean values of the sequence; \(\mathcal{N}(\cdot)\) denotes the normalization procedure where \(\rho_{TH}\) is a predefined threshold. A sequence with a kurtosis value lower than \(\rho_{TH}\) will be detected as noise and hence has its spatial likelihood set to zero. We set \(\rho_{TH} = 5\), which is slightly higher than the kurtosis value of Gaussian noise [27].

3. RESULTS

We compare the proposed spatial filtering-based algorithm (SP), with six source localization algorithms, namely the well-known SRP-PHAT algorithm (SRP) [14]; three GEVD-MUSIC algorithms: GMUSIC, which assumes the noise correlation matrix to be known [19]; MUSIC, which uses an identity matrix as the estimate of the noise correlation matrix [19]; iMUSIC, which estimates the noise correlation matrix incrementally from the microphone signal [19]; and two additional time-frequency processing-based algorithms we create based on the observations we discussed in this paper: the histogram-based algorithm (Hist) and a combined algorithm (HiSP). Hist builds a histogram using the localization results \(\{\theta_{TH}(\cdot)\}\) at all time-frequency bins [23, 28], i.e., \(\hat{\rho}_{TH}(\theta) = \mathcal{H}(\{\theta_{TH}\})\), where \(\mathcal{H}(\cdot)\) denotes building a histogram of the values in the sequence. The spatial likelihood function is obtained by normalizing the histogram with \(\rho_{TH}(\theta) = \mathcal{N}(\hat{\rho}_{TH}(\theta))\). HiSP combines Hist and SP by using \(\rho_{SP}(\theta) = \mathcal{N}(\rho_{TH}(\theta) + \hat{\rho}_S(\theta))\).

SP performs robustly in low-SNR scenarios, whereas the kurtosis difference between input and output signals is evident. However, this difference becomes less evident in high-SNR scenarios, decreasing the localization resolution. In contrast, Hist has a high localization resolution in high-SNR scenarios but degraded performance in low-SNR scenarios. HiSP combines these two complementary algorithms to achieve robust performance for different SNRs.

We evaluate the source localization performance in terms of target sound detection ability, which is measured by the normalized spatial likelihood value at the target direction, i.e., \(\rho(\theta_d)\). Lying in the interval \([0, 1]\), the higher \(\rho(\theta_d)\), the more capable the algorithm is to detect the target sound. For this reason, the spatial likelihood function obtained by each algorithm is normalized with (9) by setting \(\rho_{TH} = 0\) (except in SP \(\rho_{TH} = 5\)). At sampling rate 8 kHz, we set the STFT frame length as 1024 with half overlap. For SP, we set \(\sigma = 10^6\) in (3). The search space \(\{\theta_1, \ldots, \theta_D\}\) is set as \(-180^\circ \sim 180^\circ\) with a step of \(1^\circ\). This search space is also used as the histogram bin for Hist. The duration of the test signal is 5 s.

We built a hardware prototype with a circular microphone array consisting of eight omnidirectional lavelier microphones fixed above a 3DR IRIS quadcopter [13]. The diameter of the array is 0.2 m. The microphone signal is generated by adding the ego-noise and the target sound at different input SNRs. The ego-noise is recorded in a room with reverberation time of 200 ms [13]. The MAV remains physically static and the motor speed is varied during the recording. The target sound is simulated with the image-source method [29] in a space of size 20m\(\times\)20m\(\times\)4m and with reverberation time of 200 ms. A speech source is placed 10 m away and with a varying DOA from \(-180^\circ\) to \(180^\circ\), with an interval of \(30^\circ\).

Fig. 3 depicts the intermediate processing results by SP, Hist and HiSP for a target sound coming from \(0^\circ\) with input SNR \(-15\) dB. Fig. 3(a) depicts the time-frequency spectrum of the input signal at one microphone, where the target sound is severely masked by the ego-noise. However, as shown in Fig. 3(c), performing local DOA estimation can still detect the time-frequency bins that belong to the target sound (i.e. at DOA \(0^\circ\)). Fig. 3(d) depicts the kurtosis function \(\xi(k, \theta)\), where a high kurtosis value can be observed at DOAs around \(0^\circ\). Fig. 3(e) depicts \(\hat{\rho}_S(\theta)\), the kurtosis value averaged across the whole frequency band, and Fig. 3(f) the normalized value \(\rho_{SP}(\theta)\). A peak can be clearly observed around \(0^\circ\). Fig. 3(g) depicts \(\hat{\rho}_{TH}(\theta)\), the histogram of local DOA estimates, and Fig. 3(h) the normalized value \(\rho_{Hist}(\theta)\). A peak can be observed at \(0^\circ\), which is much weaker than the peak of the ego-noise around \(-90^\circ\). Fig. 3(i) depicts the sum of \(\rho_{SP}(\theta)\) and \(\rho_{Hist}(\theta)\), and Fig. 3(j) the normalized value \(\rho_{HiSP}(\theta)\). The peak of the target sound is also clearly observed. Additionally, Fig. 3(b) depicts the time-frequency spectrum of the spatial filtering output pointing at \(0^\circ\). The target sound is well extracted while the ego-noise is effectively suppressed.
Fig. 4. Spatial likelihood functions obtained by various source localization algorithms for a target sound with DOA $0^\circ$: (a) input SNR: $\infty$; (b) input SNR: $-15$ dB.

Fig. 4 depicts the spatial likelihood function obtained by the considered algorithms for a target sound coming from $0^\circ$, with input SNRs $\infty$ and $-15$ dB, respectively. In Fig. 4(a) all the algorithms can estimate the DOA of the target sound correctly when SNR$_{\text{in}} = \infty$, except that the localization resolution of SP is much lower. In Fig. 4(b) the considered algorithms perform differently for SNR$_{\text{in}} = -15$ dB. With prior knowledge of the noise correlation matrix, GMUSIC shows a sole peak at the target direction. In contrast, MUSIC and iMUSIC do not show evident peaks at the target direction, due to inaccurate estimation of the noise correlation matrix. SRP and Hist show strong peaks at ego-noise directions but weak peaks at the target direction. SP shows a sole peak at the target direction. HiSP shows two peaks, including one at the target direction.

Fig. 5(a) depicts the spatial likelihood values at the target direction, $\rho (\theta_d)$, obtained by the considered algorithms when the target sound comes at $\theta_d = 0^\circ$, with a varying input SNR from $-25$ dB to $-5$ dB. For all algorithms the obtained $\rho (\theta_d)$ rises with increasing SNR$_{\text{in}}$. GMUSIC performs the best among all the algorithms, achieving a $\rho (\theta_d)$ which is close to 1 for all SNR$_{\text{in}}$. SRP performs the worst. Hist, SP and HiSP evidently outperform MUSIC and iMUSIC. SP obtains a $\rho (\theta_d)$ which is higher than other algorithms, especially when SNR$_{\text{in}} < -10$ dB. Hist obtains a $\rho (\theta_d)$ which is close to 1 when SNR$_{\text{in}} \geq -10$ dB, but drops quickly with decreasing SNR$_{\text{in}}$ when SNR$_{\text{in}} < -10$ dB. HiSP obtains a $\rho (\theta_d)$ which is between SP and Hist when SNR$_{\text{in}} < -10$ dB.

Fig. 5(b) depicts the spatial likelihood values at the target direction obtained by various source localization algorithms for a target sound with DOA $\theta_d = 0^\circ$. (a) Simulated target sound with SNR $\rho = 0^\circ$; (b) Real-recorded target sound with DOA $\theta_d = 160^\circ$.

Fig. 5(b) depicts the evaluation results. In comparison to Fig. 5(a), GMUSIC shows degraded performance for real-recorded speech, primarily because the microphones were not calibrated. This also leads to degraded performance of MUSIC and iMUSIC. SRP performs the worst. SP outperforms all the other algorithms except GMUSIC and it is followed by HiSP.

Fig. 6. Spatial likelihood values $\rho (\theta_d)$ at the target direction obtained by various source localization algorithms for a target sound with a varying DOA $\theta_d$. (a) input SNR: $-20$ dB; (b) input SNR: $-10$ dB.

Finally, we consider a scenario where the ego-noise and the target sound are recorded separately and added at different input SNRs. A loudspeaker is placed 3 m away from the MAV at the direction $160^\circ$, playing speech signals as the target sound [13]. Fig. 5(b) depicts the evaluation results. In comparison to Fig. 5(a), GMUSIC shows degraded performance for real-recorded speech, primarily because the microphones were not calibrated. This also leads to degraded performance of MUSIC and iMUSIC. SRP performs the worst. SP outperforms all the other algorithms except GMUSIC and it is followed by HiSP.

4. CONCLUSION

We proposed a time-frequency processing approach that exploits the time-frequency sparsity of the MAV sound recording for source localization. By estimating the DOA locally at individual time-frequency bins, the proposed spatially informed filter extracts the sound from a desired direction and suppresses other directions effectively. The output signal tends to show a high kurtosis value when the spatial filter corresponds to the target direction. Experimental results demonstrate the advantage of the proposed method in the presence of strong ego-noise.

In our future work we will combine the spatial likelihood value and the mobility of the MAV within a tracker, and investigate the performance of the proposed algorithms in real environments where multiple sound sources as well as natural and motion-induced wind pose additional challenges.

Acknowledgement: This work was supported in part by the ARTEMIS-JU and the UK Technology Strategy Board (Innovate UK) through the COPCAMS Project, under grant 332913.
5. REFERENCES


