A semblance based TDOA algorithm for sound source localization

Guilherme Seidyo Imai Aldeia, Alex Enrique Crispim, Guilherme Barreto, Kaleb Alves, Henrique Ferreira, Kenji Nose-Filho

***Abstract*— In this paper we propose a new time difference delay of arrival technique based on the semblance multichannel coherency function for the problem of sound source localization. The proposed algorithm was tested on recordings from an Unmanned Aerial Vehicle (UAV) equipped with an array of 8 microphones, for estimating the azimuth and elevation angles of a speech based source. Our results shown that the semblance method has proven to have a good performance, obtaining good results regardless of the ego noise even in cases where the signal-to-noise ratio (SNR) was very low.**

***Keywords*— time difference of arrival, semblance, sound source localization**

I. Introduction

The problem of estimating the Direction of Arrival (DOA) of a propagating wave plays a fundamental role in many signal processing applications. More recently it has been of great interest for sound source localization, specially in search and rescue scenarios [1].

For the sound source localization problem, one of the main techniques employed is based on the time difference of arrival (TDOA), i.e., the delay that the propagating wave (sound) arrives at several microphones disposed in different locations [2]–[5]. However, in applications involving Unmanned Aerial Vehicle (UAV) the main problem arises from the ego noise and the fact that the sound source location and the microphones can be in movement [1].

Ego noise is the noise produced by the UAV itself. What makes it challenging is that the ego noise is non-stationary [1], changing with the velocities of the rotors, which are constantly changed very quickly to maintain the drone stabilized and allow it to move.

In this paper we propose a new time difference delay of arrival technique based on a coherency measure for multi-channel data widely used in seismic processing, the semblance coherence function [6].

This paper is organized as follows. Section II presents in details the proposed semblance based TDOA algorithm. Section III reports the methodology used to adjust the parameters and validate the algorithm. Section IV presents the results and compare them to an algorithm found in [1]. Finally, Section V summarizes the results and present future perspectives for this work.

Guilherme Seidyo Imai Aldeia is with the Center of Mathematics, Computing and Cognition, Federal University of ABC (UFABC), e-mail: guilherme.seidyo@gmail.com; Alex Enrique Crispim is with the Center of Natural Sciences and Humanities, Federal University of ABC (UFABC); Guilherme Barreto, Kaleb Alves, Henrique Ferreira and Kenji Nose-Filho are with the Center of Engineering, Modeling and Applied Social Sciences, Federal University of ABC (UFABC)

II. A semblance based TDOA algorithm

In this paper, we want to find the direction of a sound source (azimuth, elevation) using the records from an 8-channel cube-shaped microphone array embedded in a flying UAV [7].

The proposed algorithm is based on correcting the time-delay that the propagating wave arrives in each of the 8-channel microphones. Given a source at direction $\mathbf{k_d} \in \mathcal{R}^3$, that point towards a source parametrized by azimuth $\Theta_d \in [-\pi, \pi]$ and elevation $\Phi_d \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. The time delay of a microphone at location $\mathbf{m_i}$ and a reference point at the origin $\mathbf{0} = [\mathbf{0}, \mathbf{0}, \mathbf{0}]^\mathbf{T}$ is given by [5]:

$$\tau(\mathbf{m_i}) = -\frac{\mathbf{k_d} \cdot \mathbf{m_i}}{v}, \tag{1}$$

where v is the speed of sound and $\cdot$ denotes for the inner product operation.

For each value of Θ_d and Φ_d in an equally spaced grid with a distance given by the parameter Δ we correct the time-delay for each microphone and compute the semblance of the 8-channel array, given by:

$$Z_d = \frac{\sum_n \left|\sum_k \hat{s}_k(n)\right|^2}{N_r \sum_n \sum_k \left|\hat{s}_k(n)\right|^2}, \tag{2}$$

where k denotes the microphones, n denotes the time samples, N_r is the number of sensors and $\hat{s}_k(n) = s_k(n - \tau_k)$ is the signal at the time sample n of the k-th microphone after correcting the delay, τ_k, for a given Θ and Φ.

The semblance measures the level of similarity between the signals [8], so, by applying time corrections τ_k to each pair (θ ϕ) of the equally spaced grid for each microphone, the direction that maximizes the semblance value may be the sound source direction.

This process is summarized by the Algorithm 1. The algorithm returns a matrix z — called *semblance panel* — containing the semblance correlation value for each pair of angles (Θ_d, Φ_d) in the equally spaced grid, along with two lists containing the azimuth and elevation values used to create the semblance grid, as illustrated in Figure 1.

In Figure 2 we illustrate a frame of an 8-channel audio signal in a noiseless scenario before and after the alignment performed by the algorithm by selecting the values of Θ and Φ that maximizes the semblance.

Preliminary tests showed that our approach could be improved by dividing the audio in several frames and applying the Algorithm 1 for each frame, obtaining multiple semblance panels, then combining the returned panels with a method called *Max pooling* — by picking the highest value for each pair of angles (θ, ϕ) in all panels — to obtain a final panel

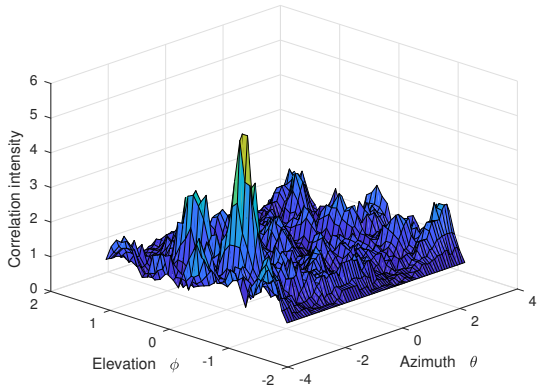

Fig. 1

3D SURFACE REPRESENTATION OF THE SEMBLANCE FUNCTION RESULT. THE HIGHER VALUES OF THE SIGNAL CORRELATION ARE REPRESENTED BY THE HIGHEST PEAKS.

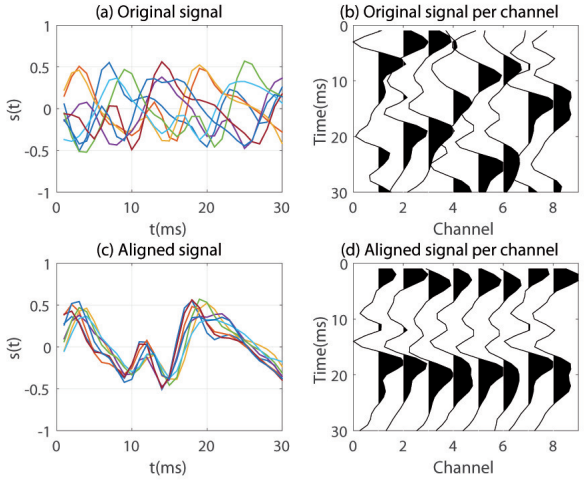

Fig. 2

FRAME OF AN 8-CHANNEL AUDIO SIGNAL IN A NOISELESS SCENARIO BEFORE AND AFTER THE ALIGNMENT PERFORMED BY SELECTING Θ AND Φ THAT PROVIDED THE MAXIMUM SEMBLANCE VALUE.

(Algorithm 2). This process is able to enhance the signal-to-noise ratio in the frames that the sound source to be localized is active, and consequently obtain better results than the global approach (using only one frame), as shown in the Section Results. The ego-noise can show high semblance values, nevertheless when the speech signals are aligned they present a stronger correlation, in a way that each frame that contains speech signals will have peaks of higher values than frames with pure ego-noise.

Finally, the local approach can be summarized as follows: first, the data is divided into frames, according to Algorithm 2; then, for each frame, is obtained a semblance panel, according to Algorithm 1; these panels are combined into one by picking the highest value for each pair of angles (θ, ϕ) in all panels (*Max pooling*) and then the pair of angles that have the highest

Algorithm 1: Find semblance global (find_global)

input : Δ: interval between angles to be tested
SoS: speed of sound on the medium
Fs: sampling rate
s: matrix containing the audio of the 8-channel microphones
$micPos$: array with coordinates [x, y, z] of the microphones positions

output: z: matrix mapping correlation with angles
Θ: tested values for elevation
Φ: tested values for azimuth

step = $\Delta * \pi/180$;
$\Theta = [\ \theta \mid \theta \leftarrow [-\pi, -\pi + step, ..., \pi]]$;
$\Phi = [\ \phi \mid \phi \leftarrow [-\pi/2, -\pi/2 + step, ..., \pi/2]]$;
$\tau = [\ ]$;

for *(i, θ) in (range(Θ), Θ)* **do**
  for *(j, ϕ) in (range(Φ), Φ)* **do**
    $kd = [cos(\theta) * cos(\phi), sin(\theta) * cos(\phi), sin(\phi)]$;
    for *(k, mic) in* $(range(micPos), micPos)$ **do**
      $\tau[i, j, k] = round(((kd * mic')/SoS) * Fs)$;

for *i in range(Θ)* **do**
  for *j in range(Φ)* **do**
    for *k in range(numMic)* **do**
      $\hat{s}_k(n) = s_k(n - \tau[i, j, k])$
    z[j, i] = semblance($\hat{s}$)

return z, Θ, Φ;

semblance correlation value is the sound source direction, being this pair the predicted direction.

III. METHODOLOGY

In order to validate the proposed method, we used three clean speech audio files (recorded with the drone in a fixed position) and a file with pure ego noise (recorded with the drone fixed and with all motors at a speed of 70 rotations per second). Then we combined those signal and noise files for different SNR levels. The values for the SNR (dB) tested were from 24 to -3dB with a step size of -3 dB and from -3dB to -21dB with a step size of -1dB. This approach allow us to have precise information about the SNR, making possible to evaluate the performance in terms of the relation between the noise and speech.

As a measure of performance, we compute the great circle distance, given by Equation 3 [9], where θ_1 and θ_2 represents the predicted and truth azimuth angles, ϕ_1 and ϕ_2 the predicted and truth elevations angles, and $\Delta\theta = \theta_2 - \theta_1$.

All the files were provided by [7], with their respective correct azimuth (θ) and elevation (ϕ) angles. However, since the noise file provided is larger than the speech file, we took out the beginning of the file (where the propellers are at a transient phase) and took a cut of the same size as the speech files.

Since the proposed method has multiple hyper-parameters, we performed a *gridsearch*, a common technique in machine

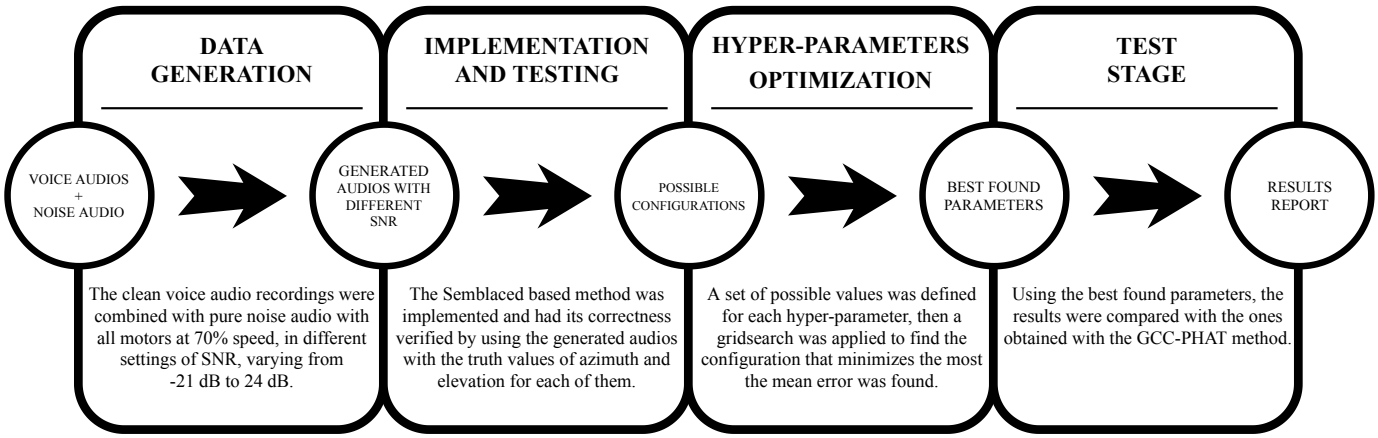

Fig. 3
METHODOLOGY FLOWCHART.

Algorithm 2: Find semblance local (find_local)

input : $frameSize$: size of the frames
$overlap$: overlap between frames
Δ: interval between angles to be tested
SoS: speed of sound on the medium
Fs: sampling rate
s: the data of the 8-channel
$micPos$: array with coordinates [x, y, z] of the microphones positions

output: z: matrix mapping correlation with angles
Θ: tested values for elevation
Φ: tested values for azimuth

$sTotal = length(x)$; // total samples
$sSize = round(frameSize * Fs)$; // sample size
$sOverlap = round(overlap * sSize)$; // sample overlap
$nFrames = ceil((sTotal - sSize)/(sSize - sOverlap)) + 1$;
$painels = [\,]$;
for *i in range(nFrames)* **do**
  $begFrame = i * (sSize - sOverlap)$;
  $endFrame = begFrame + sSize$;
  $sFrame = s[begFrame : endFrame, :]$;
  $painels[i] = find\_global(\Delta, SoS, Fs, sFrame)$;

return $pooling(painels), \Theta, \Phi$;

learning field for hyper-parameter tuning, that consists on an exhaustive search for all possible combinations for each hyper-parameter in a given set. The search creates all combinations of hyper-parameters, then executes the algorithm for the entire dataset with different SNR configurations, returning the errors. This way, the best values for the hyper-parameters obtained are those who presented the smallest mean error for all audios (3 audios with 26 different SNR configurations, totalizing 78 measures of error). To clarift the steps to obtain the results shown in section IV, Figure 3 presents a simple flowchart diagram of the methodology. The hyper-parameters to be tuned are: *(i)* overlap between consecutive frames, *(ii)* the *frameSize* and *(iii)* Δ, with the following possible values:

(*i*) overlap = [0, 0.1, 0.2, 0.3, 0.4, 0.5]
(*ii*) Δ = [17.5, 15, 12.5, 10, 7.5, 5]
(*iii*) frameSize = [0.064, 0.128, 0.256, 0.512, 1.024]

It should be noticed that the *gridsearch* is not a step of the proposed algorithm, but a technique used to adjust its hyper-parameters, due to vast gamma of combinations that can be made.

In order to illustrate the effect of the parameters (Δ, *frameSize* and overlap) we present some curves by fixing two parameters and varying only one. The Figure 4 shows how different hyper-parameters affects the error for the tested SNR configurations, where each column represents one (out of the three different hyper-parameters) varying for the tested values, while the other two parameters used are fixed in the best value found by the *gridsearch*. In this figure, we vary the parameter Δ for a *frameSize* of 0.064s and an overlap of 20% (first column); in the second column we vary the *frameSize* and set Δ equal to 10 degree for an overlap of 20%; and finally, in the third column, we vary the overlap and set Δ equal to 10 degree and use a frame size of 0.064s.

We defined a correctness guess threshold as having an error smaller than $10°$, represented by the dotted lines in the results. This error leads to a solid angle of

$$4\pi \sin^2(\theta/2) \approx \pi\theta^2 \approx 0.1 \text{ sr},$$

from which we obtain a relative error of

$$\frac{0.1 \cdot r^2}{r} = \frac{r}{10},$$

where r is the distance between the sound source and the microphone array. This means that if $r = 10$ meters, an error of $10°$ implicates in only one meter away from the original source, which is pretty acceptable.

For the grid size (parameter Δ) it is possible to observe that for almost all the cases a grid size smaller than 10 degree does

$$\Delta\sigma = \arctan \frac{\sqrt{(\cos\phi_2 \sin(\Delta\theta))^2 + (\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2 \cos(\Delta\theta))^2}}{\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\Delta\theta)}. \quad (3)$$

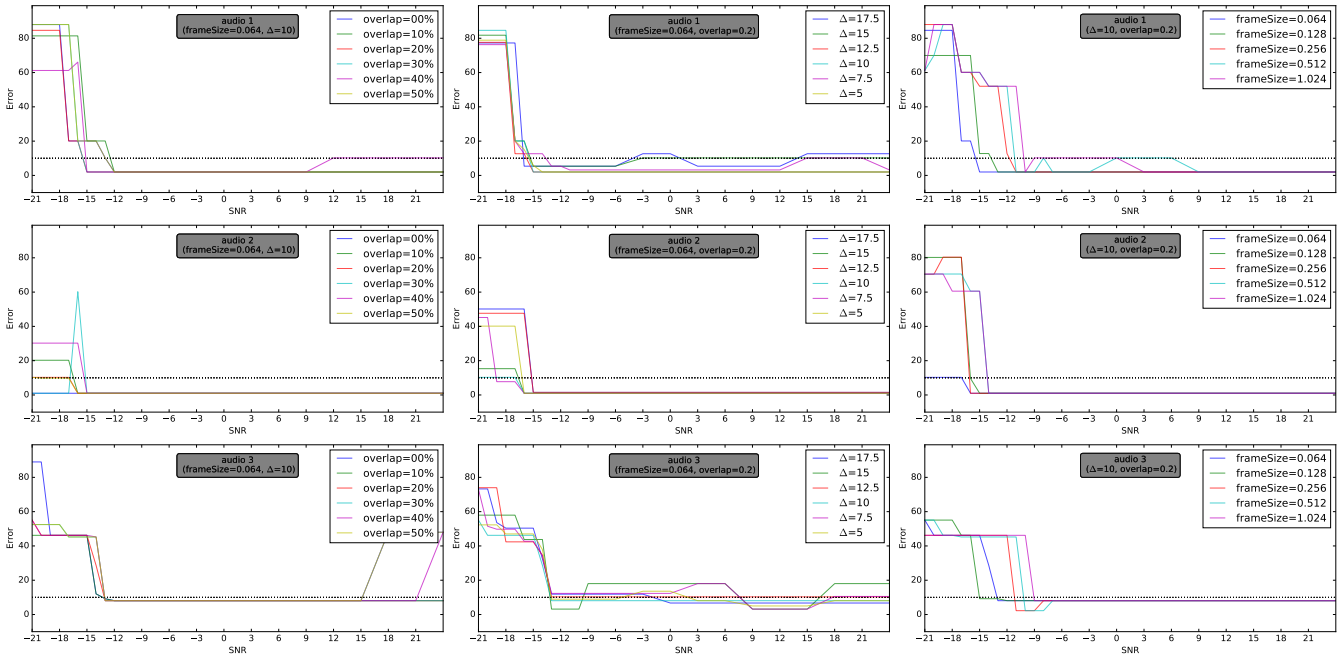

Fig. 4

ERROR VERSUS SNR FOR DIFFERENT CONFIGURATIONS. EACH COLUMN REPRESENTS THE VARIATIONS OF ONE HYPER-PARAMETER FOR THE THREE AUDIOS SEPARATELY (THE FIRST BEING THE OVERLAP, THE SECOND BEING THE Δ AND THE THIRD BEING THE *frameSize*), AND EACH LINE REPRESENTS ONE AUDIO. THE DASHED LINES DENOTE A THRESHOLD FOR THE ERROR OF 10°

not significantly change the results. For the overlap we have observed that a certain degree of overlap (e.g. 20%) may be good and for the frame size, the smaller the better (except for audio three).

In [1] the authors analyzed variations of state-of-art methods on DOA task such as the GCC-PHAT and MUSIC-based methods. However the authors pointed that the MUSIC-based methods were about twenty times slower when compared with the GCC-PHAT method. The reason why we compare our results with the GCC-PHAT method only.

IV. RESULTS

The best results (for the three audios) were obtained for $\Delta = 10$, a frame size of 0.064s and 20% of overlap. We compare it with the ones obtained by the so called generalized cross correlation phase-transform method, namely GCC-PHAT [1]–[3] with a grid space of $10°$, a default FFT window of 0.064s, and two different pooling methods (Max and Sum). These results are presented by the curves of Figure 5.

In Figure 5 we can see the errors for the proposed method using the global and the local approach (with the best found parameters) and the GCC-PHAT using the two available pooling methods [1]. Our method outperforms the GCC-PHAT for audios 1 and 2. The global approach has an intermediate performance, between the GCC-PHAT with the Max and Sum pooling methods. Even with the acceptable margin of error defined as $< 10^{\circ}$, in most of the cases the predicted azimuth and elevation angles returned by the algorithm presented an error of $\approx 2^{\circ}$.

This implementation was done in Python and the method took about 6.8 seconds for an audio with duration of 5.0s to run on a quad-core i7 processor @1.3GHz for $\Delta = 10$, a frame size of 0.064s (the sampling rate was set in 44.100 kHz) and 20% of overlap. Even though Python is not the fastest programming language for heavy processing, our code relies on the *numpy* and *scipy.signal* libraries, which have their back-end running in C. The time consumption and the computational complexity is out of scope of this paper, but this can be improved by implementing an heuristic based on tree search algorithms, starting using higher Δ values and gradually decreasing them, applying the algorithm on the subspace that showed higher correlation, until some stop criteria is reached. While the global approach utilizes one single core to find the semblance panel, the local approach is parallelized into all available cores.

V. CONCLUSIONS

In this paper we present a new time difference delay of arrival technique based on the semblance multichannel coherency function for the problem of sound source localization.

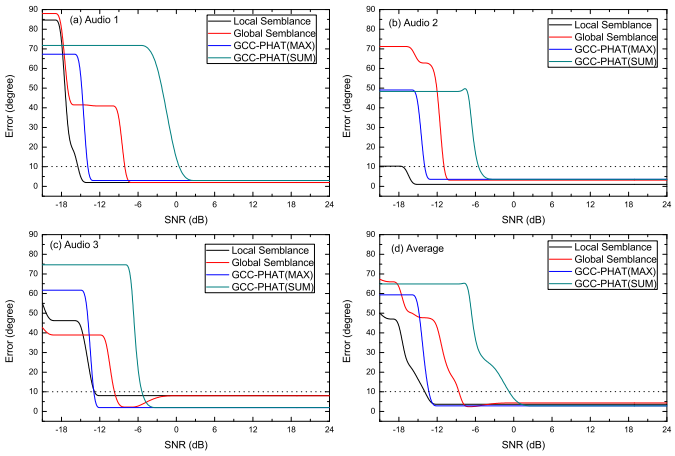

Fig. 5

COMPARISON BETWEEN OUR METHOD AND THE BEST PROPOSED METHOD IN [1]. THE DASHED LINES DENOTES A THRESHOLD OF 10° FOR THE ERROR.

The algorithm was tested for estimating the direction of a speech source (azimuth, elevation) using three audio records from an 8-channel cube-shaped microphone array embedded in a flying UAV [7] combined with an ego noise with different SNR levels. Despite the performance with the audio 3, the obtained results showed that the proposed method presents a good performance, making able to retrieve the source location (within an error of $10°$) in cases where the signal-to-noise ratio (SNR) was of -16 dB. This is aligned with the results obtained in [10], where methods of state-of-art had their performance reported.

The future perspectives for this work are: *i)* perform a complexity analysis, *ii)* apply filtering methods, and *iii)* optimize the search, to make this able to be applied in real-time situations.

REFERENCES

[1] Martin Strauss, Pol Mordel, Victor Miguet, and Antoine Deleforge, "DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018)*, Madrid, Spain, Oct. 2018, pp. 5735–5742, IEEE.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[3] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1997, vol. 1, pp. 375–378 vol.1.

[4] K. Varma, "Time delay estimate based direction of arrival estimation for speech in reverberant environments," M.S. thesis, Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 2002.

[5] J. Nikunen and T. Virtanen, "Time-difference of arrival model for spherical microphone arrays and application to direction of arrival estimation," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 1255–1259.

[6] T. T. L. Barros, "Implementation aspects of eigendecomposition-based high-resolution velocity spectra," M.S. thesis, School of Electrical and Computer Engineering of the University of Campinas, Campinas, Brazil, 2012.

[7] IEEE Technical Commitee for Audio and Acoustic Signal Processing and IEEE Autonomous System Initiative, *2019 IEEE Signal Processing Cup: Search and Rescue with Drone-Embedded Souns Source Localization*, 2019.

[8] N. S. Neidell and M. Turhan Taner, "Semblance and other coherency measures for multichannel data," *GEOPHYSICS*, vol. 36, no. 3, pp. 482–497, 1971.

[9] T. Vincenty, "Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations," *Survey Review*, vol. 23, no. 176, pp. 88–93, 1975.

[10] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors Journal*, vol. 17, no. 8, pp. 2447–2455, April 2017.