INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

CNN-Based Processing of Acoustic and Radio Frequency Signals for Speaker Localization from MAVs

Andrea Toma, Daniele Salvati, Carlo Drioli, Gian Luca Foresti

Dept. of Mathematics, Computer Science and Physics
University of Udine
Udine, Italy
{andrea.toma, daniele.salvati, carlo.drioli, gianluca.foresti}@uniud.it

Abstract

A novel speaker localization algorithm from micro aerial vehicles (MAVs) is investigated. It introduces a joint direction of arrival (DOA) and distance prediction method based on processing and fusion of the multi-channel speech data with radio frequency (RF) measurements of the received signal strength. Possible applications include unmanned aerial vehicles (UAVs)-based reconnaissance and surveillance against intrusions and search and rescue in hostile environments. A 3-stages convolutional neural network (CNN) with a fusion layer is proposed to perform this task with the objective of augmenting the source localization from multi-channel speech signals. Two parallel CNNs process the speech and RF data, and the regression network produces predictions of the angle and distance from the source after the fusion layer. To show the performance and effectiveness of this RF-assisted method, the experimental scenario and datasets are presented and experiments are then discussed along with the results that have been obtained.

Index Terms: Speaker localization, RF-assisted multi-channel speech, direction of arrival and distance regression, multi-stage convolutional neural network, speech and RF fusion

1. Introduction

Direction of arrival (DOA) estimation from a speaker is an important task in microphone array processing [1, 2, 3] and can be applied in a number of different scenarios involving single mobile robot [4] and mobile robot sensors networks [5], or a team of drones [6]. When the speech recording is performed using microphone arrays installed on multirotor unmanned aerial vehicles (UAVs) [7, 8, 9], the processing of speech signal from acoustic sources of interest becomes especially challenging [10, 11]. Moreover, in the case of micro aerial vehicles (MAVs) of small size, the consequent constraints on the size of the microphone array [12] may lead to poor spatial resolution and range of detection issues [13]. As a matter of fact, attempts to tackle the acoustic related problems typical of multirotor aerial systems have been documented only recently [14]. When a small-size microphone array must be adopted to fit the size of a micro aerial device, a method which provides acceptable spatial resolution even with small arrays and at low signal-to-noise-ratio (SNR) is desirable.

Our work further expands the acoustic beamforming presented in [15], by investigating a novel radio frequency (RF) processing coupled with the speech DOA estimation method for an RF-assisted multi-channel speech localization where, in addition to DOA predictions, distance estimations from the source are also extracted. When the speaker localization front-end detects a speech activity originating from the direction estimated by the RF antenna components, the speaker localization can be refined and the recorded signal enhanced through an acoustic beamformer. The motivation to investigate such a strategy comes from augmenting the beamforming-based acoustic source localization for applications like UAV-based reconnaissance and surveillance against intrusions [16, 17], or in search and rescue in hostile environments [18, 19], for example in presence of adverse weather conditions. In such scenarios, fusing speech data with video could result to be infeasible, even if an array of cameras is employed as in [20], especially in environments with occlusions, fog, smoke, or dust typical of disaster areas [10] or when the target is out of the field of view (FoV). To cope with these issues, we created a semi-simulated scenario with real speech data and simulated RF data in order to validate the algorithm by analysing its behaviour over the RF parameters and antenna array setup. The proposed method is based on fusion of the two kind of data, speech and RF. We consider received signal strength (RSS) measurements [21] that can be easily collected in wireless standards such as Wi-Fi where an RSS indicator is already available. The proposed idea is based on a distributed antenna array with fixed antenna spacing where RSS measurements are collected at different space locations to the antenna positions.

The use of deep learning in speech and audio processing applications for the improvement or the new design of multichannel processing localization schemes has been explored only recently [22, 23, 24], nonetheless their use is being investigated in a variety of acoustic and speech oriented applications involving multi-channel processing like in [25] where the multi-channel spectral phase information is used as input of a convolutional neural network (CNN) for the DOA estimation. Further investigations about speaker localization using deep learning networks can be found in [26, 27, 28].

In our study, we illustrate the performance of a multi-stage CNN-based algorithm where two parallel stages process intrinsic features of the speech and RF data and the third stage performs the data fusion and regression. Differently from previous works as in [15, 29], the RF-assisted algorithm proposed in this work is capable of producing joint predictions of both DOA and distance from the speaker. The small-size and low-cost hardware configuration consists of a four-acoustic sensor uniform linear array (ULA) mounted on a quadcopter MAV. In our semi-simulated scenario, an antenna array is also located on the MAV with the ULA as its centre. Both the antenna spacing and the array orientation can be chosen according to an accuracy goal. Acoustic and RF dataset are thus collected and simulated, respectively.

http://dx.doi.org/10.21437/Interspeech.2021-886

2. The speech and RF problem and data models

We address the problem of detecting the speech activity of a number of speakers positioned in front of a drone, to decide who is the active speaker among them and to estimate the DOA of the speech signal and distance of the sources. A solution is proposed where the speech DOA is paired to an RF transmission pattern analysis.

An array-based acoustic front-end is active, which is capable of speech activity detection, while the RF antenna array allows the receiver to perform an analysis of radio signal power patterns. The RF signal is emitted by transmitters located at the speakers' position. For example, modern smartphones are equipped with a number of integrated antennas like 4G or 5G (for voice and data communications), Wi-Fi, Bluetooth, global positioning system (GPS), near field communication (NFC). Wi-Fi or Bluetooth signals could then be measured for localization purposes when 5G is unavailable or in presence of an RF outage.

2.1. Speech Model

Let us refer to an ULA of M omnidirectional microphones and to a far-field model for the acoustic source wave propagation. Suppose that the acoustic wave from a speaker impinges upon the array with a direction θ. In the short-time Fourier transform domain, the data model of the multi-channel array signals can be expressed in single-source scenario as [30]:

$$\mathbf{x}(k,f) = \mathbf{a}(f,\theta)S(k,f) + \mathbf{v}^e(k,f) + \mathbf{v}(k,f) \qquad (1)$$

where k is the block time index, f is the frequency bin, $\mathbf{a}\left(f,\theta\right)$ is the array steering vector for the source direction θ, $S\left(k,f\right)$ is the source signal at the reference sensor, $\mathbf{v}^e\left(k,f\right)$ is the nonstationary drone ego-noise that is composed by multiple narrowband harmonic noise originated by the electrical engines and by the broadband aerodynamic noise induced by the propellers, and $\mathbf{v}\left(k,f\right)$ is the additive noise that is assumed to be spatially white Gaussian with zero mean and variance equals to σ^2 for all sensors.

Let $\mathbf{\Phi}\left(k,f\right) = E\left\{\mathbf{x}\left(k,f\right)\mathbf{x}^H\left(k,f\right)\right\}$ the covariance matrix of the array signal, which is symmetric and positive definite, and $E\left\{\cdot\right\}$ denotes mathematical expectation. The frequency range for the computation of $\mathbf{\Phi}\left(k,f\right)$ is (f_{min}, f_{max}). Elements of the covariance matrix denote the correlation between various microphones and their phase values provide source speaker delay information between each couple of microphones. In real-world applications, the covariance matrix $\mathbf{\Phi}\left(k,f\right)$ is unknown and has to be estimated. In general, the estimation can be computed through the averaging of the array signal blocks:

$$\widehat{\mathbf{\Phi}}(k,f) = \frac{1}{B}\sum_{k_b=0}^{B-1}\mathbf{x}(k-k_b,f)\mathbf{x}^H(k-k_b,f) \qquad (2)$$

where B is the number of snapshots in the average and $f = f_{min}, f_{min}+1, \ldots, f_{max}$.

2.2. RF Model

The path loss model (PLM) can accurately describe the RSS data in light of sight (LOS) conditions. Specifically, the path loss of a communication channel can be defined as the ratio of the transmited power to the received power. In general the dB path loss is a non-negative number since the channel does not contain active elements, and thus can only attenuate the signal. The formulation is based on statistical analysis where an RF signal travels from a transmitting antenna to the receiving one. According to the Friis transmission equation from antenna and propagation [31], the received power is affected by the signal propagation in the wireless environment and by some channel and antenna parameters forming the PLM [32]:

$$PL_{dB}^{i,j}\left(d_{i,j}\right) = PL_{dB}^{i}\left(d_0\right) + 10\alpha\log_{10}\left(\frac{d_{i,j}}{d_0}\right) \qquad (3)$$

where $PL_{dB}^{i,j}\left(d_{i,j}\right)$ is the path loss expressed in dB with distance $d_{i,j}$ from the i-th transmitter to the j-th receiver, $PL_{dB}^{i}\left(d_0\right)$ is the path loss at the reference distance d_0 from the i-th transmitter, and α is the attenuation factor. Assuming omnidirectional antennas, the path loss at d_0 only depends on the transmitting node and is given by:

$$PL_{dB}^{i}\left(d_0\right) = -20\log_{10}\frac{\lambda_i}{4\pi d_0} \qquad (4)$$

where $\lambda_i = c/f_i$ is the wavelength corresponding to the frequency f_i of the transmitted signal.

By denoting the RF signal power (in *dBm* units) transmitted by the i-th transmitter with $P_{t,dBm}^{i}$ and the corresponding power received by the j-th antenna at distance $d_{i,j}$ from the transmitter with $P_{r,dBm}^{i,j}\left(d_{i,j}\right)$, the path loss function can be written as difference of the two powers:

$$PL_{dB}^{i,j}\left(d_{i,j}\right) = P_{t,dBm}^{i} - P_{r,dBm}^{i,j}\left(d_{i,j}\right). \qquad (5)$$

The received power can, then, be expressed as:

$$P_{r,dBm}^{i,j}\left(d_{i,j}\right) = P_{r,dBm}^{i}\left(d_0\right) - 10\alpha\log_{10}\left(\frac{d_{i,j}}{d_0}\right) + \omega^{i,j} \qquad (6)$$

where $P_{r,dBm}^{i}\left(d_0\right)$ is the received power at the reference distance d_0 from the i-th transmitter. The received power in Eq. 6 follows a log-decreasing law over the distance with the reduction rate determined by α.

Rapid variations of the received power due to multipath (fast fading) are filtered out by averaging the instantaneous power of the received signal over sufficiently short time intervals (corresponding to an average over a few wavelengths). Shadowing (slow fading) is typically modelled as a log-normal random variable with zero-mean and standard deviation σ_w.

3. Proposed CNN Model with Speech and RF Fusion

The two previous frameworks are then combined in an RF-assisted speech localization. In this section, the architecture of the proposed multi-stages CNN for regression with speech-RF fusion is described in detail. It consists of 3 networks with convolutional and fully connected layers. Two parallel CNNs are employed to extract and process intrinsic features from the multi-channel speech and the RF signals. These features discriminate the input data according to the incident angles and the distances from the source. The third network performs speech-RF fusion and regression. It outputs not only DOA $(\widehat{\theta})$ but also distance $(\widehat{d})$ predictions simultaneously. Fig. 1 shows the layer structure of the whole network in detail.

The speech-CNN processes 2-dimensional matrices consisting of phase values computed from the elements in the complex-valued estimated covariance matrix of the multi-channel acoustic signal in Eq.2. These matrices consist of $M \cdot M$ elements, with M being the number of acoustic channels. This can be represented by $\angle\,\widehat{\mathbf{\Phi}}(k,f)$ where $\angle$ denotes

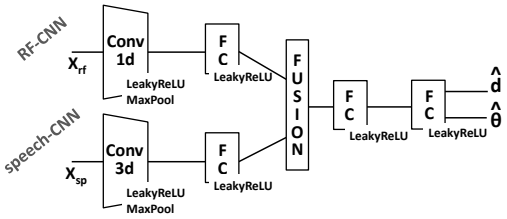

Figure 1: *The architecture of the 3-stages CNN with a fusion layer*

the element-wise phase of the matrix $\widehat{\mathbf{\Phi}}(k,f)$. By grouping S frames of such matrices, a 3-dimensional $M \cdot M \cdot S$ array denoted with $\mathbf{X}_{sp}$ is obtained and form 1 input channel of the Conv3d layer. The RF-CNN processes RSS values related to the received power of RF signals from multiple transmitters to multiple receivers, namely $P^{i,j}_{r,dBm}(d_{i,j})$ with $i = 1,\dots N_{tx}$ and $j = 1,\dots N_{rx}$. The number of elements is $N = N_{tx} \cdot N_{rx}$ with N_{tx} the number of transmitters and N_{rx} the number of receivers. They are processed by N input channels in the convolutional layer. S frames are processed at time so that each of the N input channels of the Conv1d consists of 1-dimensional array with S elements. This input of N channels with S elements each is denoted as $\mathbf{X}_{rf}$. In both the CNNs, a convolutional layer kernel operates filtering and activation on the input data. The trained kernel is computed through an optimization method to minimize the loss function intended as a measure of distance between the CNN prediction and the target. The activation function generates the output of the convolutional layer. In our network, the convolution layers with 64 convolutional filters learn intrinsic features in the corresponding input data. The kernel size is n in Conv1d and (n_1, n_2, n_3) in Conv3d where different values can be investigated, and the activation is leaky-ReLU in both of the layers. This is followed by a MaxPool layer that performs a dimensionality reduction, and a fully connected linear layer with leaky-ReLU activation in both of the branches. They output a feature sample with 64 elements.

After fusion of the two feature vectors, regression is performed in the prediction network by two fully connected linear layers with 64 input feature sample size and leaky-ReLU activations are employed. In the regression process, linear activations are used to predict values in the continuous range. In particular, the last linear layer produces two outputs for DOA $(\widehat{\theta})$ and distance $(\widehat{d})$ predictions.

4. Experimental Scenario and Datasets

The speech data was collected from a MAV equipped with a compact ULA of four microphones. The MAV used was a Parrot Bebop 1 quadcopter shown in Fig. 2, with a 250 mm frame type, 400 g weight, and overall dimensions of 280x320 mm. The microphone array from a PlayStation Eye USB device was used as the audio front-end. This device provides a four-microphone uniformly spaced linear array with total size of 6 cm (the inter-microphone spacing being of 2 cm). The microphone array was fixed on the top of the MAV, centred with respect to the four propellers. A foam rectangular shield was put below the microphones.

Concerning the RF data, we have simulated a receiving antenna array placed on the MAV with the ULA as its centre, Fig. 2. The array can be built by connecting USB Wi-Fi antenna devices with spacing τ to a single board computer (SBC). The geometric frontal axis of the antenna array is rotated with respect to the frontal direction of the drone with angle Φ, in order to avoid ill-posed situations from RF point of view. Each speaker wears a transmitting antenna.

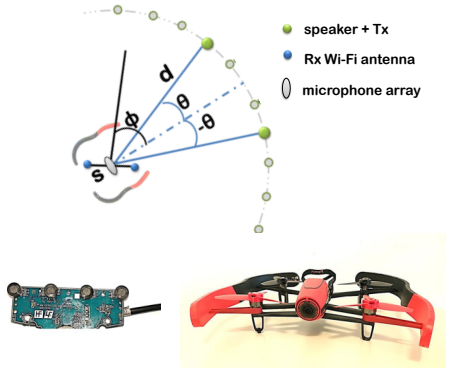

Figure 2: *Simulated RF-assisted scenario (top) and two system components: the microphone ULA (bottom-left) and the Parrot Bebop (bottom-right)*

5. Experiments and Results

Experiments have been conducted to validate and demonstrate the proposed method. A dataset was built, featuring two subjects lying frontally with respect to the drone and speaking one at a time at different positions. The two speakers were positioned symmetrically with respect to the frontal direction of the drone, uttering the same sentence one at the time. The speech signal characteristics are like in [15]. A set of 15 positions was used according to the following distances (d) and angles (θ): 2, 3, and 4 meters at $[-5, 5]$, $[-10, 10]$, $[-15, 15]$, $[-20, 20]$, and $[-25, 25]$ degrees.

Concerning the simulated RF part, the channel attenuation factor α, the noise standard deviation σ_w, and the received power at the reference distance $P^i_{r,dBm}(d_0)$ are assumed to be independent from the transmitter-receiver paths and values that can be observed in real shadowed areas with LOS propagation and Wi-Fi communication have been considered for investigation. The mean value over 20 samples of RSS measurements is taken.

The 3-stages CNN is implemented by using Pytorch. The optimizer is Adam with learning rate 0.01 and the loss function is MSELoss. Number of epochs: 5. The number of training samples (off-line) is 76460, 67% of the whole dataset and number of testing samples (real-time) is 37660, 33% of the whole dataset. Each training and testing sample has $S = 50$ frame length. The speech angle matrix consists of 4×4 values from the estimated covariance matrix, while the RSS array consists of 4 values. Batch size is 64, for both speech and RF inputs, the kernel size is 3 in Conv1d and $(1, 1, 3)$ in Conv3D.

The performance of the proposed CNN-based speaker localization with fusion of speech and RF signals can be analysed after computing the empirical Cumulative Distribution Function (ECDF) of the distance and angle errors of the predictions obtained during testing.

In Fig. 3, results obtained by setting $\alpha = 2.3$, $\omega_w = 3$ or 6 dB, and $P^i_{r,dBm}(d_0) = -23$ or -30 dBm are shown. The receiving antenna spacing τ is 0.5m while the frontal axis of the antenna array with respect to the frontal direction of the drone is rotated by $\Phi = 25$ degrees. In the figure, the angle errors are at most 2.1 degrees in half of the cases, while the distance

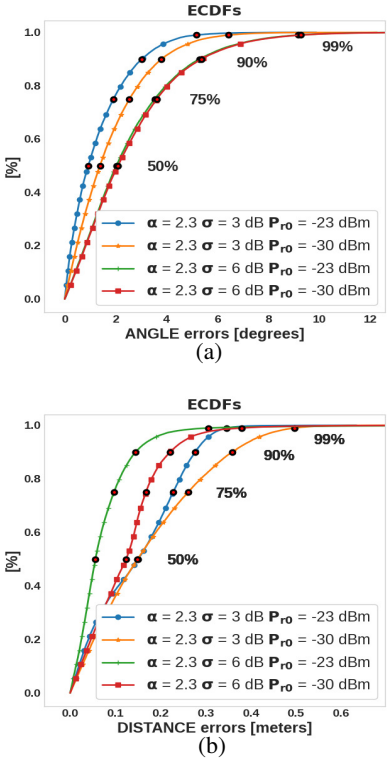

Figure 3: *ECDF at different RF channel conditions for a) angle predictions and b) distance predictions*

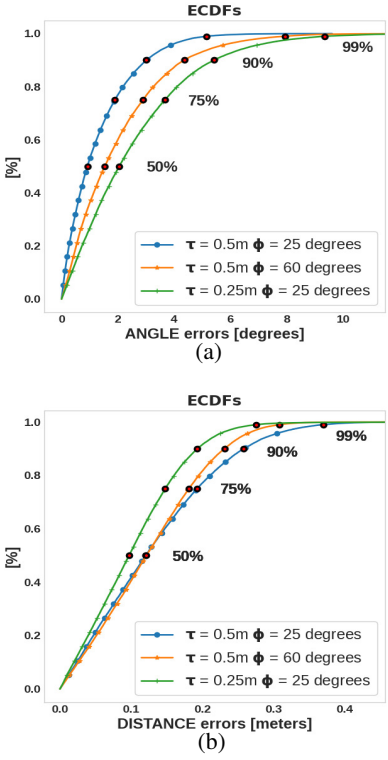

Figure 4: *ECDF at different antenna array setup for a) angle predictions and b) distance predictions*

errors are less than 15 cm in half of the cases. The 90-percentile values are less than 5.4 degrees for angle errors and less than 36 cm for distance errors.

In Fig. 4, the antenna spacing τ is set to 0.25 or 0.5m while the angle Φ between the frontal axis of the antenna array and the frontal direction of the drone is 25 or 60 degrees with fixed $\alpha = 2.3$, $\sigma_w = 3$ dB, and $P^i_{r,dBm}(d_0) = -23$ dBm. In the case of different antenna array setup, the angle errors and the distance errors show only small variations with respect the case seen at different RF channel conditions with comparable values of around 2 degrees and 13 cm as median values. Indeed, in these experiments the angle and distance predictions resulted to be robust under different RF channel conditions and antenna array setup.

These results show that predictions in the order of a few degrees (for angle) and centimetres (for distance) can be obtained thanks to the capability of the CNNs to extract feature information, from both of the kind of data related to the ground truth of angle and distance. The features are then used to perform the regression on the data. These results, in comparison with previous work in [15] where root mean square error (RMSE) values (degree) and the variance of the localization performance were found to increase with the angle of the incident speech wave and no distance predictions were computed, motivate us to futher investigate the proposed augmented data-driven method for future applications to the MAVs-based speech source localization.

6. Conclusions and Future Work

An RF-assisted acoustic target localization has been proposed. The idea consisted on pairing the microphone array for multi-channel speech detection with a distributed antenna array placed on a drone to collect RSS measurements from the target. The proposed method is based on a three-stage CNN with two parallel networks for processing the RF and the speech data separately, an RF and speech fusion layer to combine them, and a regression layer.

Predictions of DOA of the speech signal and distance from the acoustic source are produced simultaneously by the deep network. The proposed approach has been validated in a semi-simulated scenario where two speakers lie symmetrically with respect to the frontal direction of a MAV and speak one at a time at different positions. The RF data is obtained from the PLM of RSS signal. The experiments have demonstrated the performance of the proposed method with promising results for predictions of both DOA and distance from the source and its robustness under different RF channel conditions and antenna array setup.

An experimental Wi-Fi based sensor data streaming architecture will be built in future work where high complexity processing is decentralized on a ground station that receives the data collected on-board by the drone in real-time. The proposed approach will be then validated in a fully real scenario and under different levels of speech signal degradation.

7. Acknowledgements

This research was partially supported by Italian MoD project a2018-045 "A proactive counter-UAV system to protect army tanks and patrols in urban areas" (Proactive Counter UAV), and by the ONRG project N62909-20-1-2075 "Target Re-Association for Autonomous Agents" (TRAAA).

8. References

[1] C. Veibäck, M. A. Skoglund, F. Gustafsson, and G. Hendeby, "Sound Source Localization and Reconstruction Using a Wearable Microphone Array and Inertial Sensors," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–8.

[2] T. Chen, Q. Huang, L. Zhang, and Y. Fang, "Direction of Arrival Estimation using Distributed Circular Microphone Arrays," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, 2018, pp. 182–185.

[3] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A Survey of Sound Source Localization Methods in Wireless Acoustic Sensor Networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–24, 2017.

[4] J. M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust Sound Source Localization using a Microphone Array on a Mobile Robot," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, vol. 2, 2003, pp. 1228–1233 vol.2.

[5] R. Levorato and E. Pagello, "DOA Acoustic Source Localization in Mobile Robot Sensor Networks," in *2015 IEEE International Conference on Autonomous Robot Systems and Competitions*, 2015, pp. 71–76.

[6] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-Board Relative Bearing Estimation for Teams of Drones Using Sound," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 820–827, 2016.

[7] V. Didkovskyi, S. Kozeruk, and O. Korzhik, "Simple Acoustic Array for Small UAV Detection," in *2019 IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO)*, 2019, pp. 656–659.

[8] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Sound Source Tracking by Drones with Microphone Arrays," in *2020 IEEE/SICE International Symposium on System Integration (SII)*, 2020, pp. 796–801.

[9] C. Drioli, G. Giordano, D. Salvati, F. Blanchini, and G. L. Foresti, "Acoustic Target Tracking Through a Cluster of Mobile Agents," *IEEE Transactions on Cybernetics*, vol. 51, no. 5, p. 2587–2600, 2021.

[10] Y. Yamazaki, C. Premachandra, and C. J. Perea, "Audio-Processing-Based Human Detection at Disaster Sites With Unmanned Aerial Vehicle," *IEEE Access*, vol. 8, pp. 101 398–101 405, 2020.

[11] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic Source Localization From Multirotor UAVs," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, pp. 8618–8628, 2020.

[12] S. Mischie and G. Găşpăresc, "On Using ReSpeaker Mic Array 2.0 for Speech Processing Algorithms," in *2020 International Symposium on Electronics and Telecommunications (ISETC)*, 2020, pp. 1–4.

[13] M. B. Andra, B. P. A. Rohman, and T. Usagawa, "Feasibility Evaluation for Keyword Spotting System Using Mini Microphone Array on UAV," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 2264–2267.

[14] S. Oh, Y.-J. Go, J. Lee, and J.-S. Choi, "Sound Source Positioning using Microphone Array Installed on a Flying Drone," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3422–3422, 2016. [Online]. Available: https://doi.org/10.1121/1.4971007

[15] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Beamforming-Based Acoustic Source Localization and Enhancement for Multirotor UAVs," in *2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy*, 2018, pp. 987–991.

[16] J. Zhang and Y. Zhang, "A Method for UAV Reconnaissance and Surveillance in Complex Environments," in *2020 6th International Conference on Control, Automation and Robotics (ICCAR)*, 2020, pp. 482–485.

[17] H. Huang and A. Savkin, "Reactive 3D Deployment of a Flying Robotic Network for Surveillance of Mobile Targets," *Comput. Networks*, vol. 161, pp. 172–182, 2019.

[18] K. Nakadai, M. Kumon, H. G. Okuno, K. Hoshiba, M. Wakabayashi, K. Washizaki, T. Ishiki, D. Gabriel, Y. Bando, T. Morito, R. Kojima, and O. Sugiyama, "Development of Microphone-Array-Embedded UAV for Search and Rescue Task," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 5985–5990.

[19] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, and L. Marcenaro, "Audio-Based Search and Rescue With a Drone: Highlights From the IEEE Signal Processing Cup 2019 Student Competition [SP Competitions]," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 138–144, 2019.

[20] H. Liu, Z. Wei, Y. Chen, J. Pan, L. Lin, and Y. Ren, "Drone Detection Based on an Audio-Assisted Camera Array," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, 2017, pp. 402–406.

[21] F. Bandiera, A. Coluccia, G. Ricci, and A. Toma, "RSS-based Localization in Non-homogeneous Environments," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4214–4218.

[22] J. Choi and J. Chang, "Convolutional Neural Network-based Direction-of-Arrival Estimation using Stereo Microphones for Drone," in *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, 2020, pp. 1–5.

[23] W. He, P. Motlicek, and J. M. Odobez, "Neural Network Adaptation and Data Augmentation for Multi-Speaker Direction-of-Arrival Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2021.

[24] L. Wang and A. Cavallaro, "Time-Frequency Processing for Sound Source Localization from a Micro Aerial Vehicle," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 496–500.

[25] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNs for Improving Acoustic Source Localization in Noisy and Reverberant Conditions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 103–116, 2018.

[26] V. Varanasi, H. Gupta, and R. M. Hegde, "A Deep Learning Framework for Robust DOA Estimation Using Spherical Harmonic Decomposition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1248–1259, 2020.

[27] S. Chakrabarty and E. A. P. Habets, "Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained With Noise Signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[28] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust Speaker Localization Guided by Deep Learning-Based Time-Frequency Masking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, 2019.

[29] D. Salvati, C. Drioli, A. Gulli, G. L. Foresti, F. Fontana, and G. Ferrin, "Audiovisual Active Speaker Localization and Enhancement for Multirotor Micro Aerial Vehicles," in *Proceedings of the 23rd International Congress on Acoustics: integrating 4th EAA Euroregio 2019, Aachen, Germany*, Sep 2019.

[30] D. Salvati, C. Drioli, and G. L. Foresti, "Localization and Tracking of an Acoustic Source using a Diagonal Unloading Beamforming and a Kalman Filter," in *Proceedings of the LOCATA Challenge Workshop, Tokyo, Japan*, 12 2018.

[31] A. Goldsmith, *Wireless Communications*. Cambridge University Press., 2005.

[32] W. Stallings, *Wireless Communications and Networks, 2nd ed.* Prentice-Hall, 2005.