Eaves: An IoT-Based Acoustic Social Distancing Assistant for Pandemic-Like Situations

Pallav Kumar Deb, *Graduate Student Member, IEEE*, Sudip Misra, *Senior Member, IEEE*, Anandarup Mukherjee, *Graduate Student Member, IEEE*, and Sukriti Shaw

Abstract—In this article, we propose an IoT-based acoustic solution – Eaves – for ensuring social distancing in public areas during pandemic-like situations. Existing solutions depend on either sensing nearby radio signals such as Bluetooth or through image processing of video frames from surveillance cameras. Such methods either mandate the need for all parties to have the same application or impose Line of Sight constraints. We overcome such restrictions by using audio to ensure social distancing. The varying amplitude of the audio signals from different distances is the crux of the proposed method. Towards this, we record audios from different distances to extract human voice-centric components and use the corresponding Mel-frequency Cepstral Coefficients (MFCC). We train multiple Machine Learning models for selecting the one that predicts the distances efficiently with minimum delay and also propose possible IoT-based architectures to overcome resource limitations. Through extensive experiments and deployment, we observe a training accuracy of 97% and prediction accuracy of almost 100% up to 2 meters.

Index Terms—Audio Processing, Internet of Things, Machine Learning, COVID-19, Social Distancing.

1 Introduction

Fatal and communicable viruses lead to lasting pandemics such as the recent Coronavirus 19 (COVID-19). Human-to-human transmission of such deadly viruses causes daunting effects on society, especially when the infected individuals are unknown. General guidelines to contain the spread of the virus involve using face masks, regular sanitization (hands and surfaces), social or physical distancing, and boosting immunity. The COVID-19 analogous viruses have a transmission range of up to 6 feet (almost 2 meters), which is why social distancing is an important aspect of the fight against pandemics. Multiple solutions for ensuring social distancing are in place, which involve state-of-the-art image/video processing (surveillance using cameras [1]) and radio signal-based processing (particularly Bluetooth) like the Aarogya Setu app [2] in India. However, these solutions either need a Line of Sight (LoS) view (surveillance camera-based social distancing) or the mandatory usage of the same application by all (Bluetooth-based solution). An alternate solution that overcomes these challenges is essential to ensure reliable social distancing.

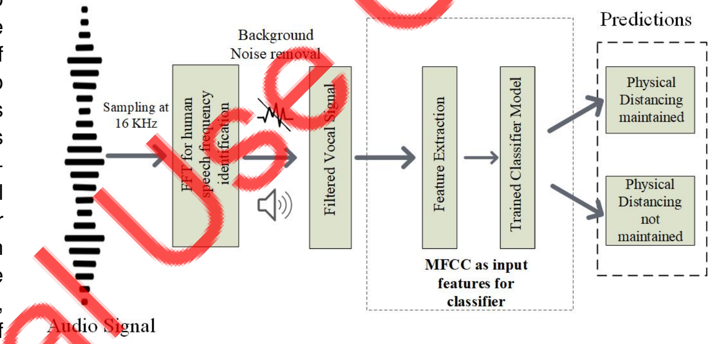
Audio Signal
Sampling at 16 KHz
FFT for human speech frequency identification
Background Noise removal
Filtered Vocal Signal
Feature Extraction
Trained Classifier Model
MFCC as input features for classifier
Predictions
Physical Distancing maintained
Physical Distancing not maintained

Figure 1: Block diagram showing flow of the proposed Eaves method

In this work, we propose Eaves (short for Eavesdropper), an audio-based social distancing method in public places. We capture the audio of the surroundings and initially probe to identify human voice. The capturing device may be a cell phone, tablet, laptop, or any other electronic device containing a microphone. On detecting the presence of human voice, we calculate their distance from the device capturing the audio signal. We achieve this by first detecting the human voice-centric frequency components in the audio and then executing noise removal routines. We then train machine learning-based models by using the Mel-frequency Cepstral Coefficients (MFCC) features from the filtered audio. The crux of our work is the power (in dB) of a specific set of frequencies (typically 85 to 155 Hz for males and 165 to 255 Hz for females) across different distances. Figure 1 depicts the information flow of the proposed system. Such audio-based solutions have two-fold advantages in comparison to existing methods: 1) It does not impose LoS constraints as when there is no direct path for the audio, it is usually because of some intermediate obstruction, which will also limit the virus transmission. 2) It is an independent application and does not mandate its usage by neighboring users. However, there exists a tradeoff that the proposed method will not be effective in case the people are silent. In the future, we plan to extend this work (audio-based or other similar methods) for developing solutions for overcoming the same. Although

P. K. Deb, S. Misra, and A. Mukherjee are with the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India. e-mail: (pallv.deb,sudipm)@iitkgp.ac.in, anandarupmukherjee@ieee.org
S. Shaw is with the Department of Electronics and Communications Engineering, SRM Institute of Science and Technology, India. e-mail: sukritishaw18@gmail.com

challenging, one alternate solution may be to detect sounds other than human voice. Further, methods for implementing machine learning models on resource-constrained devices is challenging. The scope and feasibility of such solutions in IoT devices and their networking aspects may be found in [3].

Example Scenario: Consider an individual strolling in a public area and have the proposed Eaves application installed on his/her cell phone. On listening to the audio from the environment, it identifies the human-centric frequency components and predicts the distance. On detecting violations (preset to 2 meters), it alerts the users immediately. Non-LoS audio captures do not affect the performance of the proposed work and it also does not depend on the other neighboring people to use the same application.

1.1 Motivation

Applications such as in [1] and [2] have multiple dependencies. The latter involves identifying social distancing violations from videos. Such methods mandate LoS view of the area of interests and challenges such as occlusions are common. On the other hand, the latter depends on detecting radio signal interference. Further, it mandates all users to use the same application and disrupts in case the conditions are not met. Other solutions that depend on GPS information [4] summon security threats. Social distancing solutions should not depend on such constraints and this acts as a motivation for us to develop Eaves. Such audio-based social distancing has the potential to overcome all the aforementioned challenges.

1.2 Contribution

In this article, we present Eaves, an audio-based social distancing solution. The following constitute the major research contributions of this article while developing the proposed solution:

- ***Acoustic Social Distancing:*** We train a Machine Learning (ML)-based method to ensure social distancing in public areas using audio. We detect the possible breach in social distancing based on the amplitude values of the MFCC coefficients corresponding to human sound.
- ***Comprehensive Analysis:*** We provide an analysis of the recorded audios and their variations with changing distances. As mentioned earlier, we focus on human sounds and predict the possible distance.
- ***IoT-based Architectures:*** We propose possible IoT-based fog-cloud architectures for minimizing delays and offering services on-demand, for devices that do not have on-board processing hardware. This is particularly useful for resource-constrained devices.
- ***Evaluation:*** Through extensive experiments, we demonstrate the feasibility and efficiency of the proposed work.

It may be noted that audio-based methods may consume higher battery power and its remedy is beyond the scope of this article. In the future, we plan to extend this work and propose methods for optimizing battery usage. Also, we do not store the captured audios and hence do not raise privacy concerns.

2 Related Works

The current situations have led to an upsurge in research and innovations contributing to spreading awareness among the common public, controlling and managing the resources tactfully to combat the consequences of pandemic-like situations. In particular, various implementations have been done using the Internet of Things, Artificial Intelligence, Blockchain, and Wearable Technologies [5], [6]. Augmented Reality (AR)-based methods are also useful in restricting the passive spread of the virus [7].

Drones, robots, and autonomous vehicles have been used for crowd surveillance, public announcements, thermal screening of masses, spraying disinfectants, and delivery of medical and other essentials. The snag in employing drones is its integration due to the lack of government regulatory policies, unlawful activities such as hacking, cyber terrorism, and other unsafe operations making them vulnerable, and the external battery, and load capacity [8]. Tripathi and Mohapatra [9] proposed a wearable EasyBand for contact tracing and helping its user in maintaining precautions along with monitoring their health conditions. However, the challenges in deploying these wearables at large are their accessibility to the masses, battery life, and privacy and security concerns. In solutions such as GPS-based social distancing backed with cloud platform [4], challenges due to transmission delays and privacy concerns exist.

Speech recognition and voice detection have been used in diverse applications by converting the voice signals into their corresponding MFCC. The authors in [10] used the MFCC and Linear Predictive Cepstral Coefficients (LPCC) as the features to recognize the speaker using deep learning while the authors in [11] and [12] replaced the digital domain of energy-intensive Fourier transform in the conventional method of MFCC extraction with the analog domain information processing for energy efficiency. Audio source estimation has had many applications and the methods involved in their estimation include works such as [13] and [14]. Zohourian and Martin [14] analyzed the direct-to-reverberate energy ratio (DRR) for distance calibration of the speaker at different instances. Similarly, Faraji *et al.* [13] used fuzzy algorithms for estimating the direction and location of the sound source with the help of several sensor nodes.

Synthesis: Researchers have proposed multiple solutions for combating the pandemic-like situations and spread of the virus by exploiting the features of IoT and its applications. Most of the solutions are dependent on radio signals such as Bluetooth, video surveillance-based, and GPS-based social distancing. Such methods add mandatory constraints like the same applications need to be running on all devices or need to have LoS view to avoid occlusions. Additionally, some of the methods also raise security concerns. In this article, we overcome these challenges by assuring social distancing using audio. This method does not raise any security concerns and also does not require any LoS to the subject. In the absence of LoS for audio signals, typically the virus too cannot spread.

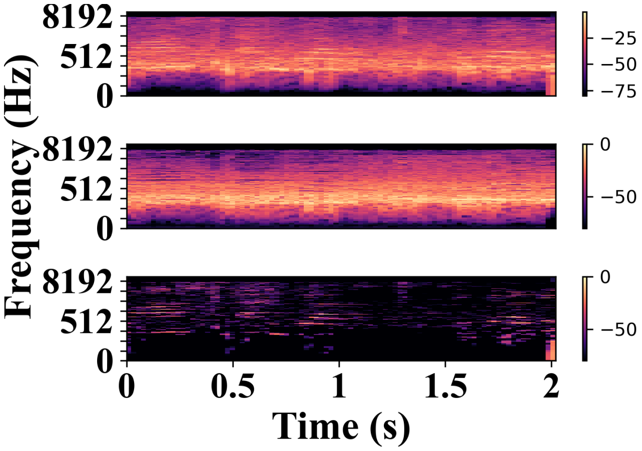

Figure 2: Spectrogram representing the raw (top), background noise (middle), and filtered (bottom) audio samples

3 System Model

The proposed method is suitable for both resource-constrained and rich devices. For instance, a handheld smartphone has an inbuilt microphone and processing configuration capable of executing trivial signal processing routines such as filtering and inferencing from pre-trained models. Similar devices like laptops, tablets, and others may operate independently, irrespective of network connectivity. On the other hand, in case of resource-constrained devices such as NodeMCU (for instance), may not be able to operate in the same fashion. These devices may require assistance from external platforms such as the cloud or fog. However, task offloading is not a trivial task, especially for real-time applications. This is because although the external platforms offer superior processing configurations, the network latency plays a big role in inducing unwanted delays, particularly in the case of the cloud. In the case of distributed architectures such as fog/edge, optimal decisions (combination of network and device states) on the node selection is important. In such cases, the device may adopt any of the computation offloading schemes such as in [15]. For Eaves, the fog nodes will have the pre-trained model stored on them and the user end devices may stream the sensed audios in real-time. The fog nodes execute the model on the incoming data and inform the users of the social distance violations accordingly. The data processing close to the users helps in minimizing delays in comparison to the cloud.

3.1 Data Gathering and Pre-Processing

The crux of the proposed social distancing method depends on the frequencies pertaining to human audio and its variation in amplitude with respect to the distance of the source. Towards this, we first record and detect the presence of human voice in each audio sample. On sensing the presence of human voice, we find corresponding MFCC coefficients and use them as the features for training the proposed model. Depending on the network architecture, this trained model may be on the resource-rich devices or stored in the fog nodes/cloud depending on the deployed architecture.

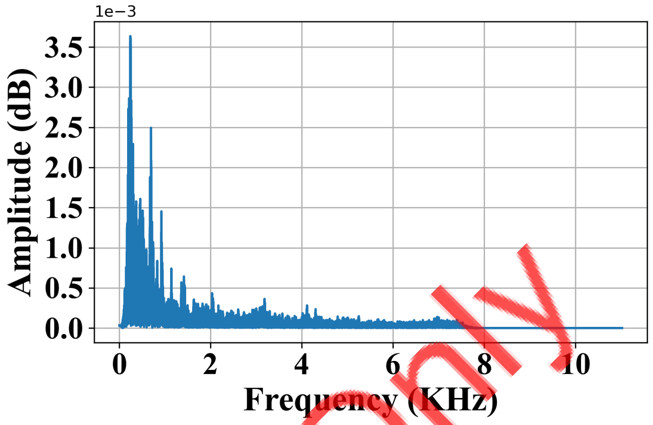

Figure 3: The Fourier transform representation depicting the frequency range dominated by human speech

We generate datasets by recording audio samples from different distances. In this work, we record the audios using a smartphone in a controlled environment (in .wav format). It may be noted that the recording may be done using any other recording device, as mentioned in Section 1. We make speeches in the same tone that we use for normal conversations. Depending on the speech, the duration in each audio sample may vary from $10 - 15$ seconds. We make the same speech with the same tone (intensity) from different distances ($0.5 - 2.5$ meters). We record the data at a fixed sampling rate and annotate the data according to the distances. For an arbitrary audio recording, we present the spectrogram in Figure 2 for a 5 seconds slice. The top image represents the signal components in its raw form. The middle image is the background noise components. On filtering the noise and extracting the voice components, we achieve the spectrogram at the bottom of Figure 2.

3.2 Feature Analysis

3.2.1 Frequency Components

Human voice frequency lies in the range of 100-180 Hz, the male voice can go up to 100 Hz whereas the female can reach up to 180 Hz. We convert the array of sampled audio signals into the frequency domain using the Fast Fourier Transform (FFT). We use FFT to identify the frequency components of the audio signal and isolate the human speech frequencies from the surrounding noise. Figure 3 represents the FFT of an arbitrary 5-second long audio signal consisting of several frequencies out of which the ones greater than 2 KHz are the noise components (negligible amplitudes). We set the sampling rate to 16 KHz and comply with the Nyquist sampling theorem, implying that the frequencies possessed by the signal are less than equal to 8 KHz (16000/2). We observe higher power in the frequencies ranging from 0 to 1 KHz (Figure 3), with even higher amplitudes in the range where the typical human speech exits. We use the magnitudes within this frequency range for further analysis of the signal. We extract the relevant features by calculating the MFCC coefficients through cepstral analysis on the Mel spectrum.

3.2.2 *Training and Selection of the Predictive Model*

We use the audio signals and MFCC coefficients mentioned earlier for training our model for finding (recording device) distances from the source. We adopt the process mentioned in Sections 3.1 and 3.2 for shaping our inputs to the training model. Among the available set of ML models, we train our work on 4 models to select the one that performs the best. We now present a brief background of the models before listing their performance.

- ***Linear Support Vector Classifier (SVC):*** Linear SVC model returns the best fit hyperplane once we feed in the data set categorizing them and resulting in the grouping of the classes for prediction. Theoretically, it has a faster convergence rate with increasing number of data points.
- ***Decision Tree:*** Decision tree is one of the fastest and the simplest models with high performance as compared to most other machine learning models in terms of accuracy. It creates a tree like structure with the nodes representing unique attributes that are keys for determining the decisions.
- ***Random Forest:*** Random forest consists of multiple decision trees which makes it robust and helps in preventing overfitting. It takes a random sub-sample of the attributes in all the recurrences and facilitates training on different samples that decrease the variance leading to higher prediction accuracy.
- ***Convolutional Neural Network (CNN):*** CNN model works on a fully connected deep learning network, based on a stack of sequential layers from the input to the output end. We include 3 dense layers between the input and the output ends and train the models with over 6 passes through the entire dataset.

4 Results

We train the models mentioned in Section 3.2.2 and select the one that offers the best prediction accuracy. We also bias our model selection towards reducing the response time (delay). In this section, we present our observations and then demonstrate the performance of the final model. It may be noted that we train and test our models on a Dell Inspiron laptop with an i5 processor. We plan to extend this work by deploying the same on resource-constrained devices with single processor boards and study the performance.

4.1 Training Accuracy of the Models

We calculate the prediction accuracy by taking the ratio of the number of correct predictions to the total number of predictions during the validation of the training dataset. Table 1 represents the comparison of the training accuracy of the different models. We observe that the Random Forest model has the highest training accuracy of 97.73% while the Linear SVC model has the least accuracy of 61.82%. We attribute such low accuracies in SVC to the non-linearity of the input audio data. This leads to the uneven division of the data by the best fit hyperplane, causing poor predictions. Random forest classifier also exhibits low variance and due to the randomness in the repeated iteration of variables selection and demonstrates predictions with high accuracy in comparison to the other models. It does not consider all the features at once and generalizes data with more detail. With respect to the accuracy, we bias our selection towards the random forest model.

Table 1: Accuracy of the models

Sl. No.	Model	Accuracy(%)
1	Linear SVC	61.82
2	Decision Tree	81.82
3	Random Forest	97.73
4	CNN	87.5

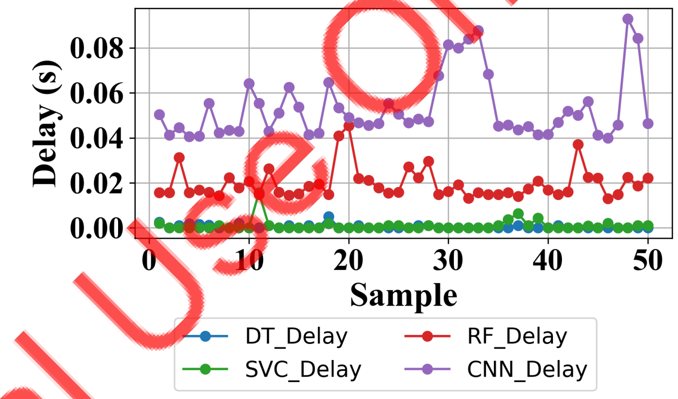

Figure 4: Plot representing the delays in predicting with the respective models

4.2 Delay/Response Time

We record the response time of the trained models and present them in Figure 4. The delay, in this case, is the time necessary for the models to predict the distance from the audio samples. We observe that the CNN model takes the maximum time with an average of 53.5 ms, followed by the Random Forest model with 19.6 ms. In contrast to the random forest, decision trees have a single tree which leads to a negligible delay of 470 μs. Further, we observe that the Linear SVC model has a slightly higher delay of 917 μs. This is due to the simplicity of the models without any extra parameters. The complex artifacts in CNN and Random Forest models are time-consuming because the presence of the neural network layers (for the former) and multiple trees (for the latter) lead to higher delays. From our discussions in 4.1 and observations in Figure 4, we select the random forest model for final deployment as it offers accuracy and response time with minimum tradeoff.

4.3 Confusion Matrix on Deployment

We deploy the random forest model and test its performance. Figure 5 represents our observations from the same with the columns depicting the actual distance of the source and rows depicting the predictions. We observe that the developed model correctly predicts all the distances, particularly from $0.5 - 2$ meters. However, the model starts to generate incorrect predictions in the case of 2.5 meters. According to the guidelines of the COVID-19-like viruses, people need to maintain a social distance of 2 meters. Under these constraints, the developed model makes the predictions with good accuracy and alerts the user in the case of violations. We comment that the proposed model works efficiently and has the potential to restrict the spread of pandemic-like viruses.

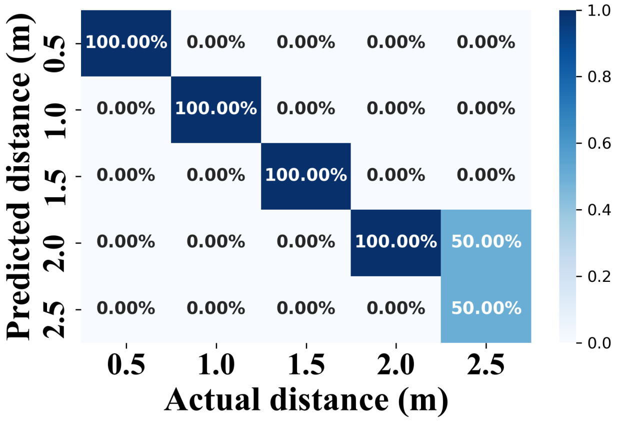

Figure 5: Confusion matrix of the chosen Random Forest model representing the accuracy percentage of individual classes

5 Conclusion

In this article, we proposed Eaves, an IoT-based acoustic solution for ensuring social distancing in public areas during pandemic-like situations. While most state-of-the-art solutions are image, radio signal-based, or GPS-based methods, they mandate the need for both parties to have the applications for sensing neighboring devices and induce security concerns. The proposed Eaves system, which calculates distance based on sound intensities, does not impose such limitations and has the potential to work independently. We performed an analysis of the captured data and described its variation with distance. We trained multiple models and presented their performances with respect to accuracy and delays, and described our bias for selecting random forest as our final model. Eaves has the potential to be used on public areas and also surveillance systems where visual clarity is challenging.

In the future, we plan to extend this work by enhancing its capabilities. The Eaves system (in this work) has been tested in controlled environments and an extensive study on uncontrolled ones is important. Further, we plan to deploy Eaves on resource-constrained systems with single board processors coupled with minimal energy consumption.

6 Acknowledgement

Ms. Sukriti Shaw has contributed to this work as part of her internship at SensorDrops Networks Pvt. Ltd. (*www.sensordropsnetworks.com*).

References

[1] M. Shorfuzzaman, M. S. Hossain, and M. F. Alhamid, "Towards the Sustainable Development of Smart Cities Through Mass Video Surveillance: A Response to the COVID-19 Pandemic," *Sustainable cities and society*, vol. 64, p. 102582, 2021.

[2] "Aarogya Setu," https://aarogyasetu.gov.in, accessed: 30 Jan. 2021.

[3] X.-L. Huang, X. Ma, and F. Hu, "Machine Learning and Intelligent Communications," *Mobile Networks and Applications*, vol. 23, no. 1, pp. 68–70, 2018.

[4] A. Ksentini and B. Brik, "An Edge-Based Social Distancing Detection Service to Mitigate COVID-19 Propagation," *IEEE Internet of Things Magazine*, vol. 3, no. 3, pp. 35–39, 2020.

[5] V. D. . T. Y. W. Daniel Shu Wei Ting, Lawrence Carin, "Digital technology and COVID-19," *Nature Medicine*, vol. 26, p. 459–461, 2020.

[6] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing Its Impact," *IEEE Access*, vol. 8, pp. 90 225–90 265, 2020.

[7] T. Amano, H. Yamaguchi, and T. Higashino, "Connected AR for Combating COVID-19," *IEEE Internet of Things Magazine*, vol. 3, no. 3, pp. 46–51, 2020.

[8] G. Yang, B. Nelson, R. Murphy, H. Choset, H. Christensen, S. Collins, P. Dario, K. Goldberg, K. Ikuta, N. Jacobstein, D. Kragic, R. Taylor, and M. McNutt, "Combating COVID-19-The Role of Robotics in Managing Public Health and Infectious Diseases," *Science Robotics*, vol. 5, no. 40, 2020.

[9] A. Tripathy, A. Mohapatra, S. Mohanty, E. Kougianos, A. Joshi, and G. Das, "EasyBand: A Wearable for Safety-Aware Mobility during Pandemic Outbreak," *IEEE Consumer Electronics Magazine*, vol. 9, no. 5, pp. 57–61, 2020.

[10] H. Yang, Y. Deng, and H.-A. Zhao, "A Comparison of MFCC and LPCC with Deep Learning for Speaker Recognition," in *Proceedings of the 4th International Conference on Big Data and Computing*, ser. ICBDC 2019. Association for Computing Machinery, 2019, p. 160–164.

[11] Q. Li, Y. Yang, T. Lan, H. Zhu, Q. Wei, F. Qiao, X. Liu, and H. Yang, "MSP-MFCC: Energy-Efficient MFCC Feature Extraction Method with Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications," *IEEE Access*, vol. 8, pp. 48 720–48 730, 2020.

[12] Q. Li, H. Zhu, F. Qiao, Q. Wei, X. Liu, and H. Yang, "Energy-Efficient MFCC Extraction Architecture in Mixed-Signal Domain for Automatic Speech Recognition," in *Proceedings of the 14th IEEE/ACM International Symposium on Nanoscale Architectures*, ser. NANOARCH '18. Association for Computing Machinery, 2018, p. 138–140.

[13] M. Faraji, S. Shouraki, E. Iranmehr, and B. Linares-Barranco, "Sound Source Localization in Wide-Range Outdoor Environment Using Distributed Sensor Network," *IEEE Sensors Journal*, vol. 20, no. 4, pp. 2234–2246, 2019.

[14] M. Zohourian and R. Martin, "Binaural Direct-to-Reverberant Energy Ratio and Speaker Distance Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 92–104, 2019.

[15] P. K. Deb, C. Roy, A. Roy, and S. Misra, "DEFT: Decentralized Multiuser Computation Offloading in a Fog-Enabled IoV Environment," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 978–15 987, 2020.

Biographies

Pallav Kr. Deb is a Ph.D. Research Scholar in the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India. He received his M.Tech degree in Information Technology from Tezpur University, India in 2017. Prior to that, he has completed the B. Tech degree in Computer Science from the Gauhati University, India in 2014. The current research interests of Mr. Deb include UAV swarms, THz Communications, Internet of Things, Cloud Computing, Fog Computing, and Wireless Body Area Networks. Further details about him are available at https://pallvdeb.github.io/

Sudip Misra (M'09 — SM'11) is a Professor and Abdul Kalam Technology Innovation National Fellow in the Department of Computer Science and Engineering at the Indian Institute of Technology Kharagpur. He received his Ph.D. degree in Computer Science from Carleton University, in Ottawa, Canada. His current research interests include Wireless Sensor Networks and Internet of Things. Dr. Misra has been serving as the Associate Editor of different journals such as the IEEE Transactions on Mobile Computing, IEEE Transactions on Vehicular Technology, IEEE Transactions on Sustainable Computing, IEEE Network, and IEEE Systems Journal. He is the Fellow of the National Academy of Sciences (NASI), India, the Institution of Engineering and Technology (IET), UK, British Computer Society (BCS), UK, Royal Society of Public Health (RSPH), UK, and the Institution of Electronics and Telecommunications Engineering (IETE), India. Professor Misra is the distinguished lecturer of the IEEE Communications Society. Further details about him are available at http://cse.iitkgp.ac.in/ smisra/.

Anandarup Mukherjee is a Ph.D. research scholar at the Smart Wireless Networking and Applications (SWAN) Laboratory, Department of Computer Science and Engineering at the Indian Institute of Technology, Kharagpur (IIT Kharagpur). He is also the Director and Co-Founder of the IoT startup, SensorDrops Networks Private Limited (http://www.sensordropsnetworks.com). His research interests include, but are not limited to, networked robots, unmanned aerial vehicle swarms, Internet of Things, Industry 4.0, 6G and THz Networks, and enabling deep learning for these platforms for controls and communications. His detailed profile can be accessed at http://www.anandarup.in

Sukriti Shaw is a B.Tech graduate of Electronics and Communications Engineering, SRM Institute of Science and Technology, India. She has interned and worked on projects based on the Internet of Things with Machine Learning at the School of Computing, National University of Singapore, Singapore in winter 2019. Some of her areas of interest and works include Wearable Technology, Internet of Things, Edge and Cloud Computing, Machine Learning, Building Models, and Recommendation Engines.