Exploring a Probabilistic Approach to Vehicle Sound Source Localization in Urban Scenes

Julia Wilkins[1], Magdalena Fuentes[1], Luca Bondi[2], Shabnam Ghaffarzadegan[2],
Bea Steers[1], Ali Abavisani [2], Juan Pablo Bello[1]
[1]New York University [2]Bosch Research

Abstract

The localization of sound sources in audio-visual urban scenes is a challenging task motivated by real-world applications such as traffic monitoring and autonomous driving. In this work, we propose a method, ***P****robability* ***D****ensity* ***Loc****alization (****PDLoc****), for estimating the position of vehicles in real-world urban scenes using stereo-audio input and corresponding visual annotations of approximate vehicle location. We suggest a novel representation of bounding box annotations of vehicles in urban scenes via a summation of Gaussian distributions constrained to a true probability density function, and train a deep learning model in a regression setting to predict this representation. We show how this approach can yield successful results in a supervised setting for vehicle sound source localization and density approximation, and analyze performance of our model against a traditional beamforming approach. We explore the benefits and limitations of our proposed model and present an analysis of challenges and potential future work in sound source localization and density estimation in urban audio-visual scenes.*

1. Motivation and Related Work

The computational understanding of urban scenes is a growing area of research with major impact in both industry and the public sector, with applications such as traffic monitoring, the development of assistive devices for the hearing-impaired, and autonomous driving. Automatically understanding an urban scene requires estimating not only *which* objects are present in the scene, but also *where* they are and *how* they are moving, all in the context of real-world environments. Urban scenes are extremely complex both acoustically and visually, often containing visually occluded sound sources or sounds occurring from the same approximate position in a scene.

Previous audio-only deep learning approaches and classical signal processing methods (e.g. beamforming [9], ray-space transform [3, 4], and acoustic senseor networks [5]) for sound source localization often use synthetic datasets with multi-channel audio, where the exact position of sound sources is known. This can be difficult to translate to more realistic settings. Additionally, audio-visual approaches such [2], and [12] have shown promise in sound source localization in video via learning audio-visual correspondence, but these methods use primarily single channel audio, which removes the spatial information about the location of sounding objects.

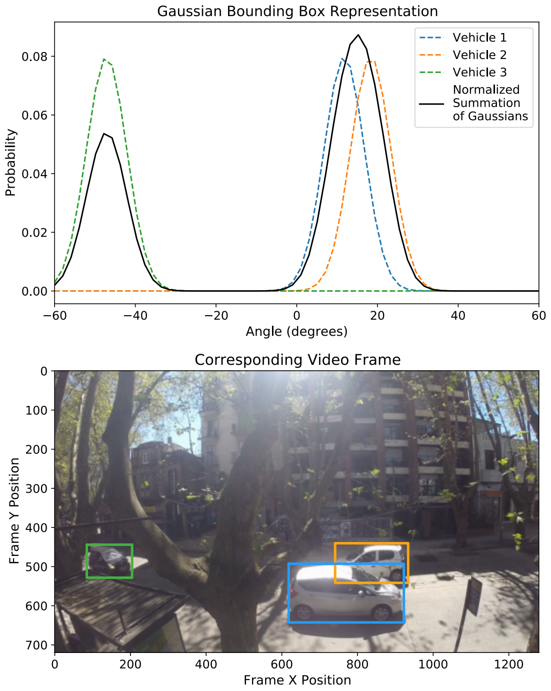

Figure 1. An example frame with bounding box annotations mapped to the proposed Gaussian ground truth representation.

Our work here builds upon the recent release of the Urbansas dataset and the proposed baseline method [6] for localization in that work, which uses a coarse binary representation of bounding boxes as the ground truth data. When multiple overlapping vehicles are present in a scene, this method lacks important information about the density of vehicles at a given position in the scene. To address this issue, we expand upon [6] in this work and propose a novel representation of bounding box annotations of vehicles in urban scenes via a summation of Gaussian distributions constrained to a probability density function.

2. Proposed Method

Task Definition Our goal is to create a supervised model that uses stereo-audio data from videos of urban scenes to predict the location of sounding vehicles in a scene over time. To do so, we propose a model output representation

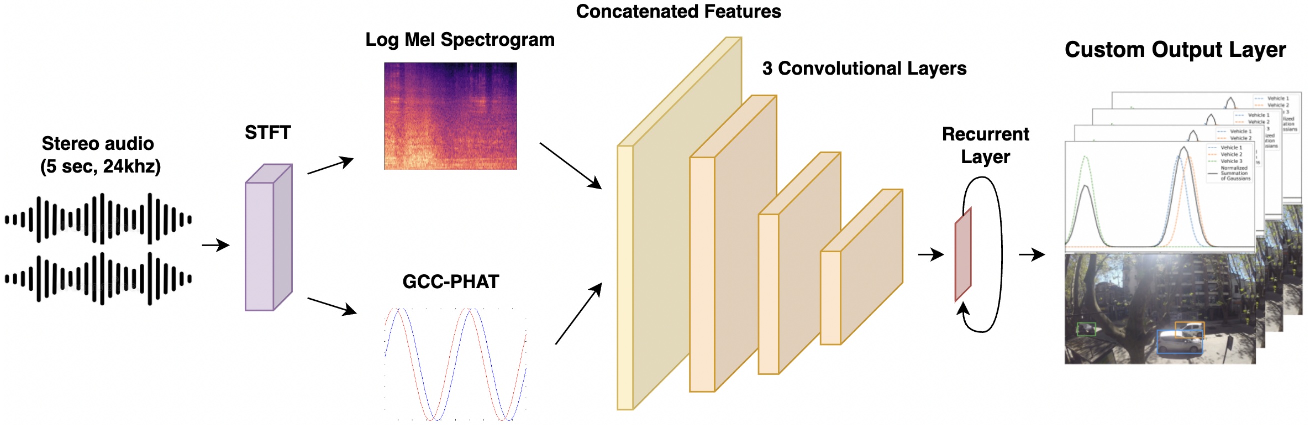

Figure 2. PDLoc system architecture. The custom output layer is comprised of a summation of Gaussian distributions constrained to a true probability density function per frame in a video clip, such that the position and density of vehicles is encoded in this representation.

that encodes both the position and density of vehicles in an urban scene at a given time. This representation offers an advantage in scenes in which multiple overlapping sound sources are present in that an approximation of the number of vehicles at that position is preserved. Additionally, we avoid mapping vehicle sound to a box or a quantized region, exploring perhaps a more realistic representation of vehicle sound production in the form of a normal distribution.

Generating a Gaussian representation of bounding box annotations Given a video of a urban scene, the position of vehicles over time is annotated via bounding boxes. We extract the center horizontal position of each vehicle from its bounding box and use this as an approximation for the Direction of Arrival (DOA) of the sound of the vehicle. We assume the microphones are centered in the Field of View (FoV) of the camera, and map the center of each bounding box from its native Cartesian coordinates in the image frame to an azimuth angle in $[-\mathrm{FoV}/2, \mathrm{FoV}/2]$, as shown in Figure 1. We assume a FoV of 120° for all videos.

For the n^{th} bounding box in a given frame, we define an angle vector $\vec{\theta}^n \in \mathbb{R}^{N_\theta}$, where $\theta_i^n \in [-60, 60]$ and $i \in [0, N_\theta - 1]$, and N_θ is the number of angles at which we will derive a traffic density estimation in each frame. We set $N_\theta = 60$. We then generate a 1D Gaussian distribution over $\vec{\theta}^n$ for each vehicle n present in a frame, using the Gaussian PDF below:

$$g(\vec{\theta}^n|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\vec{\theta}^n - \mu)^2}{2\sigma^2}}, \tag{1}$$

where μ is a scalar and the substraction $\vec{\theta}^n - \mu$ is done element-wise. We refer to $g(\vec{\theta}^n|\mu, \sigma^2)$ as $g(\vec{\theta}^n)$ for simplicity. For each $g(\vec{\theta}^n)$, we take μ to be the center-point of the bounding box in the horizontal direction, mapped to an angle as described above. We use a fixed variance for every bounding box at $\sigma^2 = 5$. We fix the standard deviation across Gaussians to simplify the learning task and minimize assumptions about the scene and vehicles within, namely street angle and loudness related to camera proximity.

Our model uses 5-second chunks of stereo audio as input data and is trained to predict data matrix $\mathbf{P} \in \mathbb{R}^{N_f \times N_\theta}$, where N_f is the number of time frames in the corresponding video clip and N_θ as defined above. We derive the value $\mathbf{P}[f, i]$ for $f \in [0, N_f - 1], i \in [0, N_\theta - 1]$ at frame f and angle index i using a Gaussian method described below and aim to predict the $N_f \times N_\theta$ matrix $\mathbf{P}$ as the output of our model in a regression paradigm.

Many urban scenes contain multiple vehicles in the same frame (i.e. at the same time), with some vehicles perhaps overlapping directly in horizontal position. To accommodate this, after generating $g(\vec{\theta}^n)$ for each bounding box n in a given frame f, we sum these distributions at the frame level. We constrain this summation of Gaussians to be a true probability density function, such that it satisfies the following properties: (1) all $g(\vec{\theta}^n)$ must be non-negative, and (2) $\sum_{i=0}^{N_\theta - 1} P[f, i] = 1$ for a given frame f. To do this, we normalize the sum of the Gaussians in each frame f over all the angles by performing element-wise division of $g(\vec{\theta}^n)$ by the total sum of distributions in that frame:

$$\mathbf{P}[f, *] = \frac{\sum_{n=0}^{V_f - 1} g(\vec{\theta}^n)}{\|\sum_{n=0}^{V_f - 1} g(\vec{\theta}^n)\|_1}, \tag{2}$$

where V_f is the number of sounding vehicles in a frame f, and $\mathbf{P}[f, *]$ represents f-th row of matrix $\mathbf{P}$. The final representation of a single clip with N_f frames and N_θ angles is defined as matrix $\mathbf{P} \in \mathbb{R}^{N_f \times N_\theta}$. This method is applied to both the creation of the ground truth data and to the output of the model as we regress on the values of $\mathbf{P}$ directly. Note that this ground truth representation presents an interesting challenge in representing frames that contain no bounding box annotations, as the distributions for these frames must still sum to 1. For this we generate a uniform distribution $\mathbf{P}[f, *] = \vec{u}$ where $\vec{u} \sim \mathcal{U}(-60, 60)$ for any empty frames.

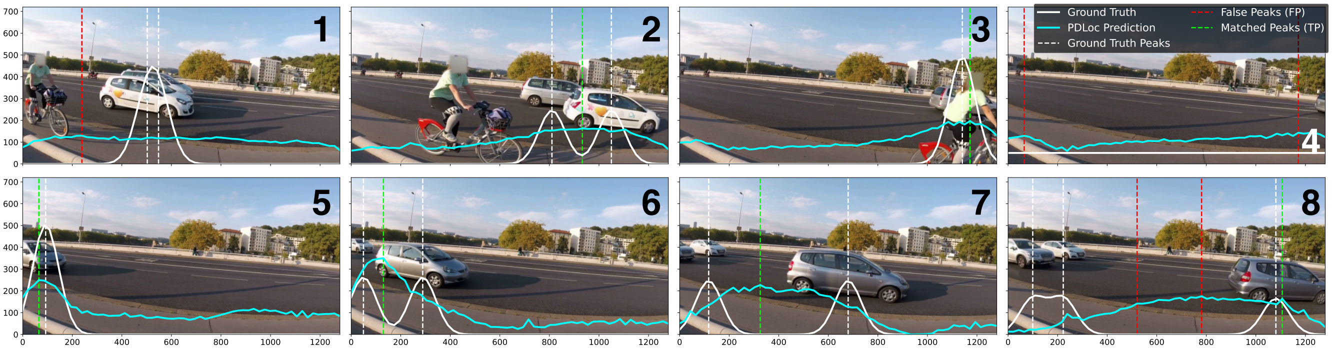

Figure 3. Selected frames from *street_traffic-lyon-1029-43289.mp4* in Urbansas, overlayed with the Gaussian ground truth representation (white) and PDLoc model predictions (blue), with ground truth peaks (dashed-white), false positive peaks (dashed-red), and true positive peaks (dashed-green). Note that here the ground truth and predicted distributions are scaled for viewing purposes.

Dataset We use the Urbansas [6] dataset, which is comprised of 1080 10-second videos of urban traffic scenes with corresponding stereo audio. The videos have bounding box annotations of moving vehicles in the scene at 2 frames per second, for a total of 20 frames per clip. Urbansas also features audio annotations that indicate when a vehicle is clearly audible and even when it is not (i.e. ”unidentifiable sound” label). These audio annotations in combination with the bounding boxes allow us to filter to only use scenes with reasonably audible vehicles and localize them. The dataset contains data from two distinct sources: TAU [11], in which the original audio was recorded using binaural equipment, and MAVD [13], in which the audio was acquired with a stereo recorder. The resulting Urbansas dataset contains a variety of scene locations, camera angles, and scene difficulty, from very few to multiple concurrent vehicles. Additionally, note that the Urbansas annotations contain class labels, and for this problem formulation we ignore class labels, operating in a vehicle class-agnostic paradigm.

Network Architecture Our model architecture, shown in Figure 2, is drawn from the baseline system proposed in the DCASE 2019 Sound Event Localization and Detection task [1]. We expand this system to fit our use case, adding Generalized Cross Correlation with Phase Transform (GCC-PHAT) [7] as an audio feature and introducing our custom probabilistic output layer (Figure 1).

We use 5-second chunks of stereo audio as input and feed each segment through an STFT, which is then used to generate Log Mel Spectrogram and GCC-PHAT features for each audio segment. The model contains three convolutional layers followed by two bi-directional gated recurrent unit layers. We design a custom final layer to constrain the output to a true probability density function, by (1) enforcing positivity (2) replacing empty frames with a uniform distribution across angles (3) normalizing each frame by the sum of that frame at model output time. We use KL divergence averaged over frames as the loss function in training. We open source the code for training and evaluation of this model.

3. Results and discussion

PDLoc vs. Beamforming We first evaluate PDLoc against traditional beamforming approaches for the localization of sounding vehicles in urban scenes. To provide a baseline system for sound density estimation, we use a classical beamforming technique, using the implementation available in Acoular [10]. This process is based on traditional signal processing and does not include any iterative learning. The goal is to obtain a heatmap representing the sound pressure level at each angle and each point in time such that we can compare with the output of PDLoc. We separated the comparison into (1) sound source localization and (2) traffic activity detection.

Sound Source Localization For the sound source localization task, we consider only frames that contain vehicles (i.e. ”active” frames), and perform peak picking on the ground truth representation, and PDLoc and beamforming model outputs, to find local maxima above a threshold for each framewise distribution, indicating angles at which peaks are present. We use an angle matching tolerance $\hat{a}_{\delta} = 20^{\circ}$, and for each ground truth peak angle a_i, count one true-positive (TP) if exists *at least one* predicted angle $\hat{a}_j$ such that $|\hat{a}_j - a_i| \leq a_{\delta}$, otherwise count one false-negative (FN). We count one false-positive (FP) if it does not exist any i such that $|\hat{a}_j - a_i| \leq a_{\delta}$. At a recall of 0.4, PDLoc achieved an F1-score of 0.51 vs. the beamforming method at F1-score = 0.42 across all data in Urbansas. Digging deeper into the component binaural and stereo portions of Urbansas, we found the most significant difference between PDLoc and beamforming in the stereo (MAVD) data (PDLoc F1-score 0.54 vs. beamforming 0.38). We think this can be attributed to the stereo recording configuration in MAVD; the distance between the two microphone capsules here is only 1.5cm, making the directionality of the beamforming approach practically null in the range of frequencies (50Hz to 4000Hz) where vehicles sounds are mostly present [8]. **Traffic Activity Detection** We binarized the output of both PDLoc and the beamforming model to compare these methods on the task of detecting whether a given frame in a video contains *any* vehicles. PDLoc outperforms beamforming across the board, with a precision of 0.76 (at recall = 0.4) vs. 0.57 for the beamforming method. The gap in performance between the two likely lies in PDLoc being trained to discriminate traffic sounds from non-traffic sounds, while the beamforming baseline does not discriminate between the type of sound source active in the scene.

Qualitative Example We explore the performance of PDLoc on a sample video as shown in Figure 3. We observe that PDLoc is very conservative in general in predicting the *height* of the peaks in the distribution. Recall that we constrain each distribution to a PDF, so when more vehicles are present, peak height is naturally lower. For example, in frames 3 and 5, we are able to *localize* a single vehicle very accurately, but struggle to predict the height of a single, taller peak. In frames 2, 6, and 7, when two vehicles are present at the same time and slightly overlapping, we lose fine-grain localization but capture peak height with more success. We believe the challenge in accurately predicting peak height is due to a high presence of frames *without* vehicles in our training data. We constrain these distributions to a true PDF, resulting in a uniform distribution, which likely biases model output especially in moments of uncertainty.

This example illustrates potential benefits of the probabilistic framework of PDLoc over previous methods, namely information retained in the presence of multiple overlapping sound sources while maintaining moderate localization accuracy, and easily interpretable results.

How does performance vary depending on the number of vehicles in a scene? We also analayze the Mean Absolute Error (MAE) across frames based on the number of vehicles present in a scene. The dataset contains significantly more frames containing a single vehicle than those containing more than one. We observe that the frame-wise MAE is lower on average for the few frames that have significantly more (i.e. ≥ 5) vehicles. While counterintuitive initially, this is likely due to the presence and uniform modeling of empty frames in training, as this incentivizes the model towards flatter predictions, which reflects better in the MAE calculation for frames containing more vehicles.

4. Conclusions and Future Work

We propose a method for estimating the position of vehicles in real-world urban scenes using stereo-audio input and corresponding visual annotations using a novel task definition based on a probabilistic framework. Our method yields promising results in a supervised setting for vehicle sound source localization and density approximation, outperforming the beamforming baseline. As future work, to mitigate the biases of inactive frames in the learning of our model we plan to divide the task into two stages: detection followed by localization. Lastly, we plan to explore the task formulation as a fully parametric density estimation (e.g estimating a Gaussian mixture model per frame).

References

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018. 3

[2] R. Arandjelovic and A. Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 1

[3] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro. The Ray Space Transform: A New Framework for Wave Field Processing. *IEEE Transactions on Signal Processing*, 64(21):5696–5706, Nov. 2016. 1

[4] F. Borra, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, S. Tubaro, and A. Sarti. A Fast Ray Space Transform for Wave Field Processing using Acoustic Arrays. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 186–190, Amsterdam, ND, Jan. 2021. IEEE. 1

[5] A. Brendel, S. Gannot, and W. Kellermann. Localization of Multiple Simultaneously Active Speakers in an Acoustic Sensor Network. In *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 450–454, 2018. 1

[6] M. Fuentes, B. Steers, et al. Urban sound and sight: Dataset and benchmark for audio-visual urban scene understanding. In *ICASSP*, 2022. 1, 3

[7] R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, and H.-M. Park. Sound source localization based on gcc-phat with diffuseness mask in noisy and reverberant environments. *IEEE Access*, 8:7373–7382, 2020. 3

[8] D. Y. Levin, E. A. P. Habets, and S. Gannot. On the Average Directivity Factor Attainable With a Beamformer Incorporating Null Constraints. *IEEE Signal Processing Letters*, 22(11):2122–2126, 2015. 4

[9] E. Sarradj. Three-dimensional acoustic source mapping with different beamforming steering vector formulations. *Advances in Acoustics and Vibration*, 2012. 1

[10] E. Sarradj, G. Herold, A. Kujawski, T. Gensch, S. Jekosch, M. Czuchaj, and A. Pelling. *Acoular*, 2015. 3

[11] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen. A curated dataset of urban scenes for audio-visual scene analysis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2021. 3

[12] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello. Wav2clip: Learning robust audio representations from clip, 2021. 1

[13] P. Zinemanas, P. Cancela, and M. Rocamora. Mavd: A dataset for sound event detection in urban environments. *Detection and Classification of Acoustic Scenes and Events, DCASE 2019, New York, NY, USA, 25–26 oct, page 263–267*, 2019. 3