MICROPHONE ARRAY OPTIMIZATION FOR AUTONOMOUS-VEHICLE AUDIO LOCALIZATION BASED ON THE RADON TRANSFORM

Ohad Barak, Nizar Sallem and Marc Fischer

Siemens Digital Industries Software Corporation
46871 Bayside Parkway
Fremont, CA 94538, USA
ohad.barak@siemens.com

ABSTRACT

Beamforming is a standard method of determining the Direction-of-Arrival (DoA) of wave energy to an array of receivers. In the case of acoustic waves in an air medium, the array would comprise microphones. The angular resolution of an array depends on the frequency of the data, the number of microphones, the size of the array relative to the wavelengths in the medium, and the geometry of the array, i.e., the positions of the microphones in relation to each other. The task of finding the right balance between the aforementioned parameters is microphone-array optimization. This task is rendered even more complicated in the particular context of sound classification and localization for self driving cars as a result of the design limitations imposed by the automotive industry. We present a microphone array optimization method suitable for designing arrays to be placed on vehicles, which applies beamforming using the Radon transform. We show how our method produces an array geometry with reasonable angular resolution for audio frequencies that are in the range of interest for a road scenario.

Index Terms— direction-of-arrival, angular resolution, audio classification, microphone array

1. INTRODUCTION

Automotive OEMs (Original Equipment Manufacturers) have recently shown interest in audio as an additional source of environmental perception for autonomous vehicles. The classic sensory setup is made up of cameras, Lidars, Radars and ultrasound. Although these devices provide a good amount of data, they remain restricted to the line of sight. On the other hand, the ambient audio wave field, which can be captured using microphones, is less restricted by lines of sight due to the longer wavelengths of acoustic waves in air and the sizes of typical obstructions in road scenarios. In addition, many traffic road participants and driving mission-critical events are accompanied by particular sound signatures, e.g., emergency vehicles' sirens, motorcycles and tire screeches. These sound signatures can be localized through the application of beamforming to acoustic wave field data recorded by microphone arrays mounted on a vehicle, and subsequently classified using a trained network.

Alternatively, it is possible to use a machine-learning based approach to determine the DoA as in [1], or even do joint classification and localization using a trained network in a data-based manner as in [2]. However, in order to do so one must have known DoA for every sound event in the training dataset. Generating a labeled training dataset containing road audio objects and their respective DoAs would be quite a challenging prospect, especially given that the recording array and the sources would be moving in an unpredictable environment.

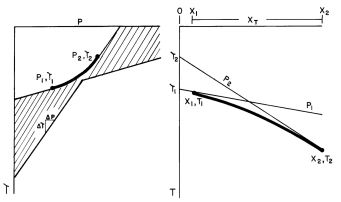

Figure 1: Figure from [7], showing a curved $x-t$ trajectory (right) and its transform to $\tau-p$ (left). For perfectly plane waves acquired with no spatial aliasing and infinite offset, a line with slope p in the $x-t$ space will transform to a point in the $\tau-p$ space.

Optimization is an inherent part of array design that has been and continues to be addressed in the literature. [3] show a comparison of array responses for existing regular geometry array designs. An example of a stochastic inversion method used for optimizing array responses which results in irregular array geometries is shown in [4]. [5] show an optimization based on a genetic algorithm for localizing sound sources in a room. [6] optimize a set of concentric spherical microphone arrays for robotics by randomly distributing microphones on the spheres. In this paper we propose an inversion scheme to optimize array responses for a vehicle-mounted planar array that would provide sufficient accuracy in terms of angular resolution and be practical and low-cost for OEMs to integrate.

Designing a microphone array suitable for mounting on a vehicle is a challenge. Indeed, a) usable surfaces on a car are restricted, b) cost is a major concern for OEMs, c) seamless integration with the car design is a must, and d) angular resolution is a critical factor for decision-making software.

First, we describe the theory behind our beamforming method which is based on the $\tau-p$ transform (i.e., the Radon transform). Then, we show how we use the $\tau-p$ transform in an array geometry inversion that seeks to optimize the angular resolution of a given array for a range of frequencies typical for road audio events. Finally, we show the results of the inversion for a particular vehicle-mounting scenario.

2. BEAMFORMING THEORY

2.1. The $\tau - p$ transform

The $\tau - p$ transform [7, 8], or slant stacking, is a method of decomposing a recorded wavefield into its plane-wave components. It is commonly used for velocity analysis of seismic waves that are acquired by geophone arrays deployed on the surface of the Earth. The transform is defined as:

$$\Psi(\tau, p) = \int_{-\infty}^{\infty} u(\tau + px, x)dx, \tag{1}$$

where $u(t, x)$ is the wavefield recorded by a linear receiver array, at time t and at a horizontal offset x from a defined starting position. p is the ray parameter, or apparent slowness, and is defined as $\frac{1}{v}$, where v is medium velocity. In the $x-t$ space, $t = \tau+px$ represents a line with slope p and vertical intercept time τ. Figure 1 depicts a hyperbola in the $x-t$ space, and how it translates to an ellipse in the $\tau - p$ space. A perfectly plane wave propogating in the horizontal direction along the X axis will appear in the $\tau - p$ space as a single point.

Equation (1) is then generalized to the 2D case where the receivers are positioned on an X-Y plane as:

$$\Psi(\tau, p_x, p_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(\tau + p_x x + p_y y, x, y)dxdy, \tag{2}$$

where the apparent slowness vector $\mathbf{p}$ now has two components (p_x, p_y).

For the discrete case, the $\tau - p$ transform is effectively a simple summation into bins operator. The number of slowness bins are in the model space of the transform, and can be determined on application. Under the assumption that the acoustic waves are propagating along the X-Y plane in a homogeneous air medium, the ratio between the p_x and p_y slowness values for each bin will indicate the azimuthal direction of arrival of acoustic wave energy as $\phi_a = \tan^{-1}\left(\frac{p_y}{p_x}\right)$. The parameters for the number of receivers and their positions are in the data space of the transform, and are determined by the geometry of the receiver array.

2.2. Beamforming

Beamforming is done by first applying the $\tau - p$ transform (2) to the wavefield data recorded by receiver arrays. A weighting can be applied in the $\tau - p$ space to select particular p_x and p_y slownesses, which effectively translates to selecting energy by angles of arrival of the sound waves. Note, that as a result of the air medium having a near-constant acoustic velocity, this selection can be used to filter acoustic energy coming from the vertical directions, as vertically propagating arrivals will have a very low apparent horizontal slowness on a planar, horizontal array. This is a useful attribute in the autonomous driving case since the body of the car and its interaction with the road surface emit vibrations and noise.

Wind noise and the vehicle's self-noise can potentially have a strong effect on the localization accuracy. However, mitigation of these noises are not subjects of this paper.

3. ARRAY GEOMETRY OPTIMIZATION

We developed an array geometry optimization methodology for a given, constant number of microphones, which takes into account certain restrictions on microphone positioning. In comparison to other stochastic array optimization methods such as [4], our method caters to automotive OEMs by: a) splitting the singular array into multiple interconnected, small sub-arrays with identical geometry, b) taking into account the available physical installation surface, and c) the constant number of microphones. The localization is then derived from each sub-array and from the composite array defined by the sub-arrays. Attending to these issues in the array design process will have a direct effect on the total production costs.

3.1. Array optimization inversion

The optimization is a stochastic Monte Carlo inversion which also utilizes simulated annealing. The model space is the $\tau - p$ domain $g\,(\tau, p_x, p_y)$, and the data space are the recorded time series' in the x-y plane $d\,(t, x, y)$. Equation (2) defines the adjoint operator $\mathbf{F}'$ which inputs microphone array data and applies the $\tau-p$ transform: $g = \mathbf{F}'d$. The forward operation $d = \mathbf{F}g$ inputs a model of a sound event in the $\tau - p$ domain and outputs the data as it would be recorded on the x-y plane.

The sampling of the slownesses p_x and p_y in the model space are specified when running beamforming. However, the sampling of the data space is determined by the microphone positions in the array. We can represent the microphone geometry of the array as a spatial sampling operator applied to the discretized data space as $\mathbf{S}d$. The matrix $\mathbf{S}$ has ones at microphone coordinates and zeros elsewhere. The adjoint operation is a cascade of two operators $\tilde{g} = \mathbf{F}'\mathbf{S}d$, where $\tilde{g}$ is the estimated model given the sampling of operator $\mathbf{S}$.

For each sound arrival angle ϕ, we define an idealized, optimal array response model in the $\tau - p$ domain as g_0. We begin with an initial array geometry determined by an initial sampling operator $\mathbf{S}$. The core of our methodology is in applying the forward and adjoint operator to the optimal response model g_0:

$$\tilde{g} = \mathbf{F}'\mathbf{S}d = \mathbf{F}'\mathbf{S}\mathbf{F}g_0. \tag{3}$$

To resolve the DoA of acoustic energy we compute the RMS in the $\tau - p$ domain along the τ axis:

$$\begin{aligned} m_0 &= \sqrt{\frac{1}{N_\tau}\sum_{\tau=0}^{N_\tau} g_0(\phi, \tau, p_x, p_y)^2}\,, \\ \tilde{m} &= \sqrt{\frac{1}{N_\tau}\sum_{\tau=0}^{N_\tau} \tilde{g}(\phi, \tau, p_x, p_y)^2}. \end{aligned} \tag{4}$$

where m_0 is our "desired" model and $\tilde{m}$ is the estimated model given the sampling $\mathbf{S}$. The model dimensions are effectively $\phi_s \times \phi_a$, where ϕ_s is the number of source angles we test for, and ϕ_a is the number of angular bins we predetermine for beamforming. Figure 2d shows the desired model m_0, which effectively states our desired angular resolution for each incidence angle.

The objective function that we wish to minimize is the L1 norm of the difference between the estimated model and the desired model. Given the desired model m_0, defining the objective function this way reduces angular localization error in practice. We also apply a model-weighting matrix $\mathbf{W}_{\phi_s \times \phi_a}$ to the objective function, which enables us to prioritize resolution for some angles of arrival at the expense of others:

$$J = \left\|\mathbf{W}\left(\tilde{m} - m_0\right)\right\|_1. \tag{5}$$

It is common to measure the main Beam Width (BW) and the Maximum Sidelobe Level (MSL) when estimating beamformer performance. However, defining the objective function as in (5) enables us to specify the both BW and MSL we wish to achieve ahead of time, and integrate them into a single measure without explicitly measuring these parameters at each iteration.

In the inversion process, the sampling operator $\mathbf{S}$ that specifies the microphone positions is randomly modified at each iteration. Certain restrictions to what random permutations are allowed: sub-array size and permissible locations, the minimum distance between microphones, and the maximum distance a sub-array may deviate from its original position.

For each array geometry permutation (i.e., changes to $\mathbf{S}$), we apply the forward and adjoint operator as in (3) calculate a new model $\tilde{m}$ as in (4), and subsequently a new value for the objective function as in (5). The inversion seeks an array geometry that minimizes the objective function. The inversion's outputs are the $\tau - p$ model that has the minimum value of J and the sampling operator $\mathbf{S}$ that produced this minimal value.

The objective function in (5) can be expanded using (3) as:

$$J = \left\| \mathbf{W} \left(\mathrm{RMS}_\tau \left(\mathbf{F}' \mathbf{S} \mathbf{F} g_0 \right) - \mathrm{RMS}_\tau \left(g_0 \right) \right) \right\|_1 , \qquad (6)$$

where RMS_τ indicates a root-mean-square operation along the τ axis.

From this we observe that the role of the inversion is to produce a sampling operator $\mathbf{S}$ that diagonalizes the forward and adjoint operation, such that $\mathbf{F}' \approx (\mathbf{S}\mathbf{F})^{-1}$.

4. ARRAY OPTIMIZATION FOR ROAD AUDIO

In this section, we address a particular scenario of mounting a microphone array on the roof of a car for the purpose of localizing road audio.

4.1. Optimization setup

Before running the optimization, we first define the road objects we wish to localize. Specifically, we are interested in localizing objects such as emergency vehicles' sirens and motorcycles, since audio-based information regarding their positions on the road can have an added value for autonomous vehicles, particularly in cases where there is no line of sight.

Figure 2a is the mean spectra of road audio we observe in an independently curated dataset containing traffic noise, motorcycles and emergency vehicles' sirens. In order to determine the DoA of sounds from such road objects, the array must have a reasonable angular resolution for a wide frequency band.

We ran the inversion where the input data frequency was as shown in the blue curve in Figure 2b. Note that the input data have a high-frequency bias, to account for the low-pass response of the $\tau - p$ operator. After application of the $\tau - p$ operator during the inversion process, the spectrum is shown by the red curve in Figure 2b, which encompasses most of the frequency band we observed in our road-object dataset in Figure 2a.

The model weighting function is shown in 2c. This weighting prioritizes resolution for audio events coming from 120^o cones in the front and rear of the vehicle.

Figure 2d is the desired model m_0 for the array response. The horizontal axis is the sound arrival angle, while the vertical axis is the angle resolved by beamforming. This figure represents the result we wish the inversion to lead to, namely an array response where

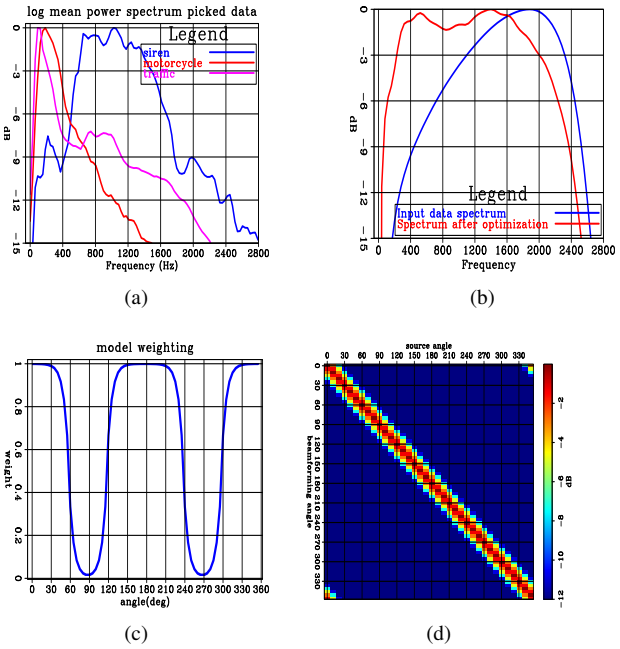

Figure 2: a) Spectrum of wavelet (blue curve) used in array optimization inversion. Note that the low frequencies are purposefully damped since the $\tau - p$ transform behaves as a low-pass filter (red curve). b) Mean spectra of road audio objects from independently curated dataset. c) Model weighting, where we prioritize angular resolution for sound arrivals coming from a 120^o cone in the front and rear of the vehicle. d) The desired model m_0, which specifies the desired angular resolution for all sound source arrival angles ϕ_a.

the main-lobe width down to the -3dB point is about 20^o. and where there are no sidelobes.

4.2. Array geometry for vehicle roof mounting given automotive design limitations

One emerging industry trend for installing sensory equipment on autonomous vehicles is to place an additional enclosure on the edges of the vehicle roof for housing the sensors. Therefore, we defined 20 cm wide strips along the edges of a vehicle roof where we would permit microphones to be situated, as shown in Figure 3a.

We defined four sub-arrays of three microphones each rather than one large array. The reasoning behind using four sub-arrays is because of the specific, low-cost recording hardware we intend to use in our final product, Also, this arrangement enables a compromise between low-frequency angular resolution (which requires larger distances between microphone) and high-frequency angular resolution (which requires smaller distances). We also enforce a rule that the sub-arrays must remain identical to each other in shape, though they may rotate and translate independently within their assigned areas. This was done to reduce eventual microphone array production costs, as it is simpler and cheaper to mass produce a single sub-array design.

The initial array geometry is shown in Figure 3b, and the optimized one in Figure 3c. Note that the coordinates in these Figures are relative to the roof center shown in Figure 3a.

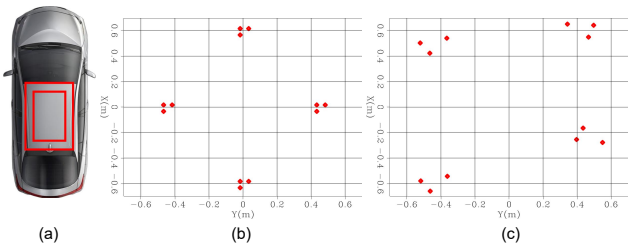

Figure 3: a) Vehicle roof-edges mounting regions. Microphone sub-arrays are permitted to be only between the larger and smaller red rectangles, in 20 cm wide strips inward from the roof edges. b) Initial array geometry. c) Optimized array geometry.

Figure 4a is the array response for the initial array geometry, while Figure 4b is the optimized response. For each frequency range we see a definite improvement in resolution, with narrower main lobes and lower-amplitude sidelobes. The higher frequencies have narrower main lobes than the low-frequencies, as we would expect given the size limitations of the array.

Figures 5a and 5b are the summary of the frequency-dependent main-lobe widths and sidelobe amplitudes, respectively. The main-lobe width is measured in degrees down to the -3dB point from maxima of the main-lobe. We observe that for a frequency of 1000 Hz (average frequency for sirens), the nominal main-lobe width is 60^o. The higher values are located between arrival angles $60^o - 120^o$ and $240^o - 300^o$, which are the low-priority directions we specified in the weighting function shown in Figure 2c. The sidelobes are at least -2dB lower than the main lobe, even for the low frequencies.

The improvement in angular resolution of the array shown in Figure 3c vs the initial geometry in Figure 3b is due to the irregularity of the microphone positions in each sub-array, combined with the irregularity of the array as a whole. The irregularity of geometry effectively means that there is a greater variance of distances between microphone pairs, and thus more wavelengths can be spatially sampled by the array without aliasing. However, note that irregularity was not explicitly imposed by the inversion, but rather it is a consequence of our objective function that optimizes for angular resolution.

5. CONCLUSION

In this article we used the Radon transform to solve a practical issue of an optimal microphone array implementation on a vehicle. The Radon transform is linear, which enabled us to combine the beamforming responses of more manageable sub-arrays while preserving the abilities of a large microphone array. Our experiment shows that such an approach provides the desirable result while appealing to OEMs through reproducible small sub-arrays. The objective function we used enables specification of the desired angular resolution for each source angle, therefore it is possible to prioritize certain DoAs according to the desired application. The use-case we envision for vehicle-mounted microphone array is for classification and localization of road audio as part of the environmental inputs of autonomous vehicles.

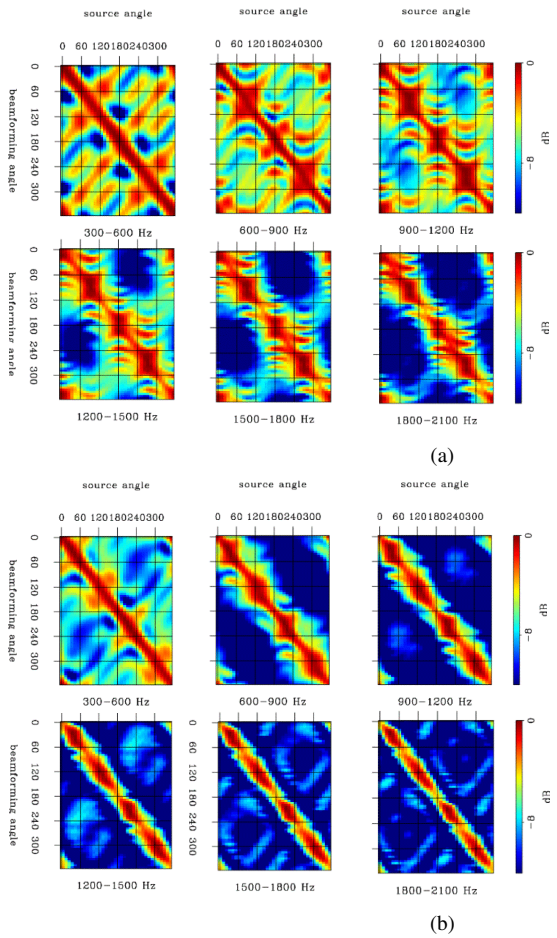

Figure 4: Array response in $\tau - p$ space for several frequency ranges. Horizontal axis is the angle of the source sound arrival, while the vertical axis is the beamforming angle achieved by applying the $\tau - p$ operator using the microphone array's geometry. a) Array response with initial geometry. b) Array reponse with optimized geometry.

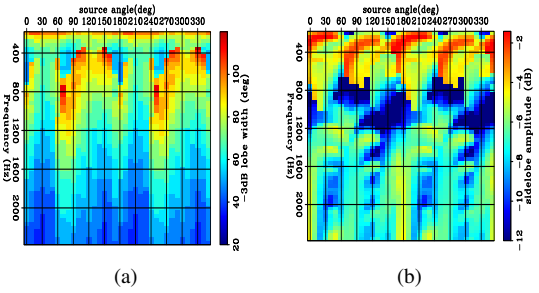

Figure 5: Summary of optimized array response in terms of main-lobe width and relative sidelobe amplitude for the frequency range of the data. a) Main-lobe width down to the -3dB point in degrees. b) Sidelobe amplitude relative to the main-lobe amplitude (i.e., -2dB in the scalebar means 2 dB lower than the main lobe's peak amplitude)

6. REFERENCES

[1] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[2] S. Adavanne, A. Politis, and T. Virtanen, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," *Proceedings of the Detection and Classification of Acoustic Scenes and Events workshop*, 2019.

[3] Z. Prime and C. Doolan, "A comparison of popular beamforming arrays," *Proceedings of the Australian Acoustical Society AAS2013 Victor Harbor*, vol. 1, p. 5, 2013.

[4] M. Bjelić, M. Stanojević, D. Šumarac Pavlović, and M. Mijić, "Microphone array geometry optimization for traffic noise analysis," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3101–3104, 2017.

[5] J. Yu and K. D. Donohue, "Performance for randomly described arrays," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 269–272.

[6] Y. Sasaki, M. Kabasawa, S. Thompson, S. Kagami, and K. Oro, "Spherical microphone array for spatial sound localization for a mobile robot," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 713–718.

[7] P. L. Stoffa, B. Peter, J. B. Diebold, and W. Friedemann, "Direct mapping of seismic data to the domain of intercept time and ray parameter—a plane-wave decomposition," *Geophysics*, vol. 46, no. 3, pp. 255–267, 1981.

[8] C. Chapman, "Generalized radon transforms and slant stacks," *Geophysical Journal International*, vol. 66, no. 2, pp. 445–453, 1981.