Detecting the direction of emergency vehicle sirens with microphones

Noam R. Shabtai, Eli Tzirkel

DETECTING THE DIRECTION OF EMERGENCY VEHICLE SIRENS WITH MICROPHONES

Noam Shabtai
User Experience Technologies
GM Advanced Technical Center, Israel
shabtai.noam@gmail.com

Eli Tzirkel
User Experience Technologies
GM Advanced Technical Center, Israel
eli.tzirkel@gm.com

ABSTRACT

As drivers we use both our eyes and ears as sensors whereas current autonomous vehicle sensors and decision making do not rely on sound. Sound is particularly important in cases involving Emergency Vehicle (EV) sirens, horn, shouts, accident noise, a vehicle approaching from a sharp corner, poor visibility, and other instances where there is no direct line of sight or it is limited. In this work the Direction of Arrival (DoA) of an EV is detected using microphone arrays. The decision of an Autonomic Vehicle (AV) whether to yield to the EV is then dependent on the estimated DoA.

1. INTRODUCTION

As human drivers, we are capable of using both our eyes and ears to get useful information in a traffic environment [1]. Hence, the development of AVs is challenging in that the vehicle should be able to perform no worse than a human driver (if not better), and be able to collect data from the external environment under the same conditions [2]. Rapidly moving objects such as other vehicles or bicycles, slower objects such as pedestrians, and static objects such as parked cars and barriers should be all sensed by the AV and used in algorithms for correct decision making [3]. Several sensors can be used for sensing these objects; e.g., radar [4], cameras [5], and microphones [6]. In cases where an object that emits sound is too far away or near but concealed from the car, sound recorded by microphones may be the only reliable source of information. There are vast numbers of cases where sound information is important, including EV sirens, horn, shouts, accident noise, a vehicle approaching from a sharp corner, and poor visibility.

Recently, Waymo shared a report with the US Transport Department where microphones are used as "supplemental sensors" [7]. Furthermore, Waymo has developed microphones that let its robocars hear sounds twice as far away as previous sensors while also letting them discern where the sound is coming from [8]. Moreover, a video is available on the web, where it is shown how Waymo is learning to recognize emergency vehicles in Arizona, using sound and light [9].

In this work, we focus on the DoA estimation of EVs using microphone arrays. The estimated DoA can be used to decide whether to yield to an approaching EV. In practice, an EV siren is detected prior to the estimation of its DoA; however, this is a different and easier problem and can be handled using audio signature, and therefore is not addressed in this work. The DoA is estimated using a Multiple Signal Classification (MUSIC)-based algorithm and includes time smoothing technique to improve the reliability of the estimated DoA values. For the DoA estimation using internal microphones we implement a transfer function projection. Here, the DoA can be roughly estimated to determine whether the EV is approaching from behind, in which case the decision of the AV should be to yield to the EV.

Both internal and external microphone array approaches were investigated for their performance. The rational for using an external microphone array is that the results are more reliable and free-field steering vectors can be used; however, the microphones need to be protected from wind. Internal microphones have the advantage of already being available in the car for other applications; e.g., beamforming for the enhancement of Automatic Speech Recognition (ASR) performance. Unfortunately, free-field steering vectors cannot be used and transfer functions were measured with a lower spatial resolution instead. The results showed that despite the additional cost of mounting an external microphone array, it is recommended since the estimated DoA values are far more reliable than the ones achieved using the internal array.

2. DOA ESTIMATION

In this work a MUSIC-based algorithm was used for DoA estimation. Let $s(t,f)$ be the source signal in the Short Time Fourier Transform (STFT) domain. This signal is then received at the m'th microphone as

$$x_m(t,f) = a_m(f,\theta_i)\, s(t,f)\,, \qquad (1)$$

where $a_m(\cdot,\theta_i)$ is the transfer function from a source at direction θ_i to the m'th microphone. The signal vector at

doi:10.25836/sasp.2019.22

all M microphones can be represented as

$$\begin{aligned}\mathbf{x}(t,f) &= [x_1(t,f), x_2(t,f), \ldots, x_M(t,f)]^T \\ &= \mathbf{a}(t,f)\, s(t,f), \end{aligned} \quad (2)$$

where

$$\mathbf{a}(f,\theta_i) = [a_1(f,\theta_i), a_2(f,\theta_i), \ldots, a_M(f,\theta_i)]^T \quad (3)$$

is referred to as the *steering vector* from direction θ_i at frequency f.

Practically, the signal vector $\mathbf{x}$ is received by the microphones and used to calculate $\hat{\theta}_i$, the estimation of θ_i. The autocorrelation of $\mathbf{x}$ is given by

$$\mathbf{R}_{\mathbf{x}}(t,f) = E\left[\mathbf{x}(t,f)\,\mathbf{x}^H(t,f)\right]. \quad (4)$$

Assuming full rank of $\mathbf{R}_{\mathbf{x}}$, it has M eigenvectors. The eigenvector with the largest eigenvalue is associated with the signal space, and all the rest are associated with the noise space. In general, the MUSIC algorithm is designed for any number of sources up to $M-1$, but in this application only one source was of interest. Hence, if the eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M$ are sorted in descending order, the noise space eigenmatrix is defined by

$$\tilde{\mathbf{U}}(t,f) = [\mathbf{u}_2(t,f), \mathbf{u}_3(t,f), \ldots, \mathbf{u}_M(t,f)]^T. \quad (5)$$

The MUSIC spectrum P is then calculated using the noise-space eigenmatrix $\tilde{\mathbf{U}}$ and the steering vector of the hypothetical DoA $\mathbf{a}(t,f,\theta_h)$ to form

$$P(t,f,\theta_h) = \frac{\|\mathbf{a}(f,\theta_h)\|^2}{\mathbf{a}(f,\theta_h)^H \tilde{\mathbf{U}}(t,f)\,\tilde{\mathbf{U}}(t,f)^H \mathbf{a}(f,\theta_h)}. \quad (6)$$

As a first step, the frequency for which the MUSIC spectrum is calculated is selected as the one with highest energy in the received signal at the first microphone. That is

$$f_0(t) = \arg\max_f |x_1(t,f)|, \quad (7)$$

and then the estimated DoA is given by

$$\hat{\theta}_i(t) = \arg\max_{\theta_h} P(t, f_0(t), \theta_h). \quad (8)$$

Temporal smoothing is performed to prevent the consideration of non-realistic estimated DoA values. If in Eq. (8) the raw estimated DoA value $\hat{\theta}_i(t)$ is given using the plain maximum value of the MUSIC spectrum $P(t,f,\theta_h)$, then the smoothed DoA is

$$\hat{\theta}_s(t) = \arg\max_{\theta_h} \int_{t-T}^{t} P(\tau, f_0(\tau), \theta_h)\, d\tau. \quad (9)$$

Frequency smoothing can be used to select frequencies f_0 that are near the previously selected frequency since the siren signal is essentially an ascending and decreasing chirp signal. However based on some preliminary results, it was decided not to use frequency smoothing.

3. EXTERNAL MICROPHONE ARRAY

3.1 Hardware

The external microphone array consisted of 4 Micro-electromechanical system (MEMS) microphones selected from 32, as can be seen in Fig. 1, arranged as a square of dimensions $5 \times 3\text{cm}^2$. The grid dimensions of the microphone array was taken into consideration when calculating the free-field steering vectors for the DoA estimation algorithm. The external microphone array was placed outside the car and mounted on the roof as can be seen in Fig. 2.

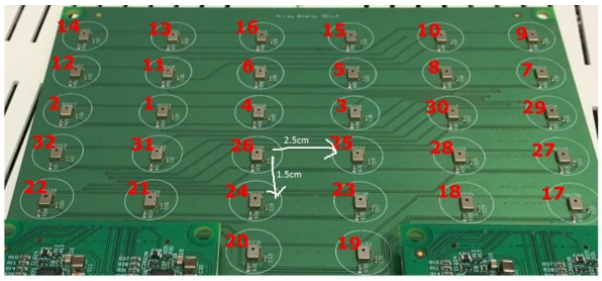


Figure 1: External microphone array.

Figure 2: External microphone array on the roof of the car.

3.2 Steering Vector

The advantage of the external microphone array is that for the DoA estimation algorithm, the steering vectors can be roughly considered like those in free field. Since the EV is far away, the incident wave form can be considered to be a plane wave, as shown in Fig. 3, where θ_i is the DoA angle, and r_m, θ_m are the distance and angle of the m'th microphone from the origin of the microphone array, respectively.

The free-field steering vector can therefore be calculated in an x-y plane. Let f be the frequency of the sound that is generated by the EV. At this frequency the wave number is $k = \frac{2\pi f}{c}$ where $c = 343\frac{\text{m}}{\text{s}}$ is the speed of sound. The frequency response from the source at θ_i to the m'th microphone with regard to the origin is given by

$$a_m(f,\theta_i) = e^{-jkr_m\cos(\theta_i-\theta_m)}, \quad (10)$$

neglecting differences in amplitude attenuation from the source to the origin and to the microphones. The steering vector of the array that contains M microphones is given

doi:10.25836/sasp.2019.22

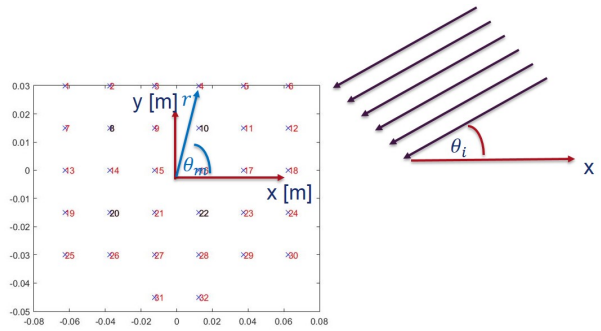


Figure 3: Plane wave propagating to the external microphone array.

by

$$\mathbf{a}\left(f,\theta_i\right)=\left[a_1\left(f,\theta_i\right),a_2\left(f,\theta_i\right),\ldots,a_M\left(f,\theta_i\right)\right]^T. \tag{11}$$

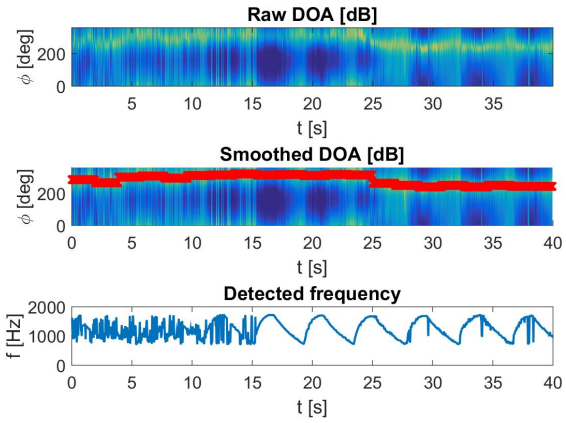
3.3 EV Experimental Results

The external microphone array was mounted on the roof of the XTS car. The car was parked near a hospital. The EVs were ambulance vehicles recorded arriving to or departing from the hospital. The parked car and the EV station can be seen in Fig. 4.

Figure 4: XTS car parked near an EV station.

The case where an EV approached from the opposite lane and made its way to the hospital is shown for example. Basically, at first the DoA comes from the frontal direction, and then switches to behind the car. Figure 5 shows the MUSIC spectrum, the estimated DoA, and the selected frequency for this case. At $t = 25$s the peak of the MUSIC spectrum shifted from values near 360° to values near 180°. It has been detected that the porches of the microphone array board reflect the sound, and therefore even though the EV was on the left side, the peak values appeared at angles that corresponded to the right side. Nevertheless, it was easy to determine when the EV was in front or behind the car. In this case, the decision of an AV should be to continue normal driving and not to yield to the EV.

(a)



(b)

Figure 5: (a) An EV is approaching from the frontal direction (b) MUSIC spectrum and estimated DoA show switching from frontal ($\sim 360°$) to back ($\sim 180°$) direction. The selected frequency matches the siren.

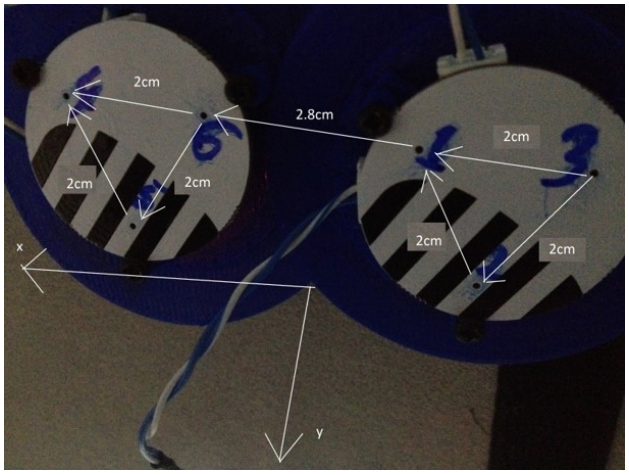
4. INTERNAL MICROPHONE ARRAY

4.1 Hardware

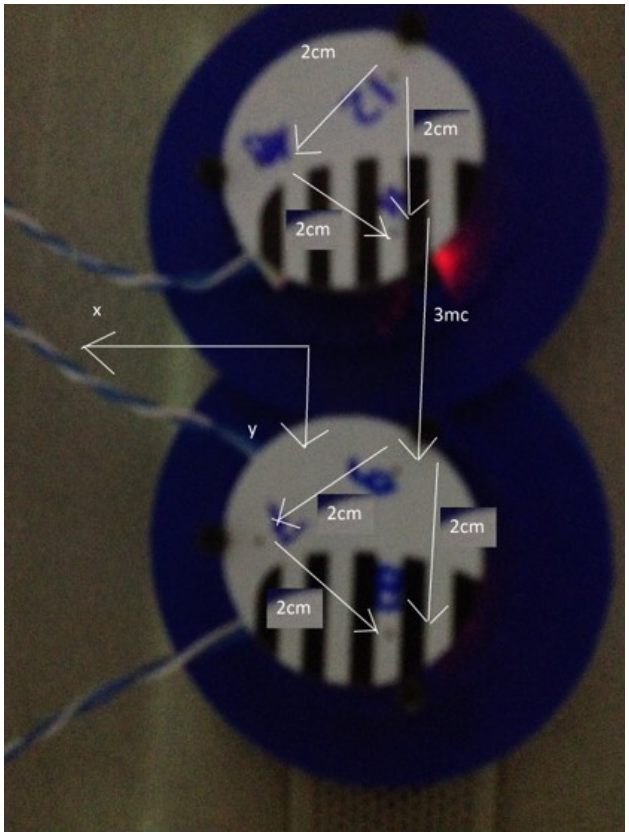
The internal microphone array consisted of different microphones but the same dedicated hardware for sound acquisition as for the external microphone array. The array contained two sub arrays with 3 MEMS each, together forming an array of 6 microphones, from which a new subset of microphone could be selected to form a different microphone array configuration.

The internal microphone array was placed inside the car and mounted either above the rear-right or the front-left passenger, corresponding to the placement of arrays for speech recognition or hands-free calls. A subset of 4 microphones can be selected to form an end-fire configuration above the rear-right passenger as presented in Fig. 6a, or a broad-side configuration above the front-left passenger, as can be seen in Fig. 6b. For the rear-right array, the distance between any pair of microphones on each sub array was 2cm, and the minimum distance between microphones from different sub arrays was 2.8cm, as shown in Fig. 6a. For the front-left array, the distance between any pair of microphones on each sub array was 2cm, and the minimum distance between microphones from differ-

ent sub arrays was 3cm, as shown in Fig. 6b.



(a)



(b)

Figure 6: Distance between microphones in the internal array, a subset of 4 microphones forms (a) an end-fire array above the rear-right passenger and (b) a broad-side array above the front-left passenger.

4.2 Steering Vector

In the case of the internal microphone array, the steering vector cannot be calculated using a free-field representation and instead, the frequency response of the car from each DoA needs to be considered. Therefore, rather than an analytic calculation of the steering vector with high spatial resolution as used by the external array, in the case of the internal array the transfer function needs to be measured in a quiet area with much lower spatial resolution. Since it is very difficult to measure the Acoustic Transfer Function (ATF) from the source to each microphone, the Relative Transfer Function (RTF) was used instead, in such a way that at each microphone the frequency response was calculated relative to the ATF at the 1st microphone.

Let h_m be the ATF from the source to the m'th microphone. The RTF is given by

$$RTF_m(f) = \frac{h_m(f)}{h_1(f)}. \tag{12}$$

If an acoustic source emits a signal $x(f)$ and assuming a noise signal $n_m(f)$ at the m'th microphone, the recorded signal at the m'th microphone is

$$\begin{aligned} y_m(f) &= h_m(f)x(f) + n_m(f) \\ &= RTF_m(f)h_1(f)x(f) + n_m(f) \\ &= RTF_m(f)\left(y_1 - n_1(f)\right) + n_m(f) \\ &= RTF_m(f)y_1(f) + \tilde{n}_m(f), \end{aligned} \tag{13}$$

where $\tilde{n}_m(f) = n_m(f) - RTF_m(f)n_1(f)$.

4.3 Wienner Filter

The estimation of the RTF based on the Wiener filter solution that minimizes the variance of the error is performed using

$$\widehat{RTF}_m(f) = \arg\min_{RTF} E\left[\left|y_m(f) - RTFy_1(f)\right|^2\right]. \tag{14}$$

which leads to

$$\widehat{\mathbf{RTF}}_m = \left[\frac{y_1^\star(1)}{\|y_1(1)\|^2}y_m(1), \ldots, \frac{y_1^\star(F)}{\|y_1(F)\|^2}y_m(F)\right]^T. \tag{15}$$

.

4.4 Generalized Eigenvalue Decomposition (GEVD)

Defining vectors with microphone indices rather than frequencies as coordinates

$$\mathbf{y}(f) \triangleq [y_1(f), y_2(f), ..., y_M(f)]^T \tag{16}$$
$$\mathbf{n}(f) \triangleq [n_1(f), n_2(f), ..., n_M(f)]^T \tag{17}$$
$$\mathbf{h}(f) \triangleq [h_1(f), h_2(f), ..., h_M(f)]^T \tag{18}$$

yields the following vector form

$$\mathbf{y}(f) = \mathbf{h}(f)x(f) + \mathbf{n}(f) \tag{19}$$

to Eq. (13). Applying the autocorrelation operator to Eq. (19) yields

$$\mathbf{R}_\mathbf{y}(f) = \sigma_x^2(f)\mathbf{h}(f)\mathbf{h}^H(f) + \mathbf{R}_\mathbf{n}(f). \tag{20}$$

The process of GEVD of $\mathbf{R}_\mathbf{y}(f)$ with respect to $\mathbf{R}_\mathbf{n}(f)$ relates the generalized eigenvalues $\lambda_m(f)$ to the corresponding generalized eigenvectors $\mathbf{v}_m(f)$ by solving

$$\mathbf{R}_\mathbf{y}(f)\mathbf{v}_i(f) = \lambda_i(f)\mathbf{R}_\mathbf{n}(f)\mathbf{v}_i(f), \qquad i = 1, 2, \ldots, M \quad , \tag{21}$$

doi:10.25836/sasp.2019.22

assuming that the rank and the number of microphones are identical and equal to M.

Assuming that the eigenvectors are sorted in descending order

$$\lambda_1(f) \geq \lambda_2(f) \geq \cdots \geq \lambda_M(f) \quad , \tag{22}$$

The generalized eigenvector that corresponds to the largest generalized eigenvalue is a rotated and scaled form of the ATF [10]. The RTF can be calculated using

$$\widehat{\mathbf{RTF}}_i(f) = \frac{\mathbf{R_n}(f)\mathbf{v}_1(f)}{(\mathbf{R_n}(f)\mathbf{v}_1(f))_{(1)}}, \tag{23}$$

where subscript $(\cdot)_{(1)}$ indicates the first coordinate of a vector.

4.5 RTF Estimation Performance

The estimation of the RTF was evaluated using Signal to Distortion Ratio (SDR). The SDR was used to calculate the distortion between the signal recorded by a microphone in the array y_m to the signal that is generated by filtering the signal recorded from the first microphone y_1 with RTF_m:

$$SDR_m(f_1, f_2) = \frac{1}{f_2 - f_1} \int_{f_1}^{f_2} \frac{\|y_m(f)\|^2}{\left\|y_m(f) - y_1(f)\widehat{RTF}_m(f)\right\|^2} df \tag{24}$$

The SDR values are displayed in Fig. 7 and Fig. 8 for the performance evaluation of the RTF estimation process using the internal microphone array in the broad-side and end-fire configurations, respectively. The RTF was evaluated using a controlled measurement where the recording car was placed in an isolated parking spot, and another car displayed a sweep signal using a speaker mounted on its roof from different directions with a resolution of $45°$. The angle of direction is displayed on the horizontal axes, and the microphone index m is displayed on the vertical axes. The corresponding SDR value is expressed in dB units using gray levels.

For the case examined in this work, the most interesting directions are $0°$ and $180°$, which correspond to the frontal and back directions, respectively. For these directions, the RTF was estimated better for the broad-side configuration than for the end-fire configuration. Comparing Fig. 7a to Fig. 7b, and also comparing Fig. 8a to Fig. 8b, shows that the RTFs were estimated better using the LS method than when using the GEVD method for all directions and all microphones.

4.6 EV Experimental Results

Only front and back RTFs were used as steering vectors. Figure 9 shows the MUSIC spectrum and DoA estimation results for the case where an EV is approaching the car from the opposite lane. The results in the figures show that using the end-fire array it is impossible to determine whether the EV was behind or in front of the car. The DoA was estimated better using the broad-side array. This result may appear surprising, since one would expect that the

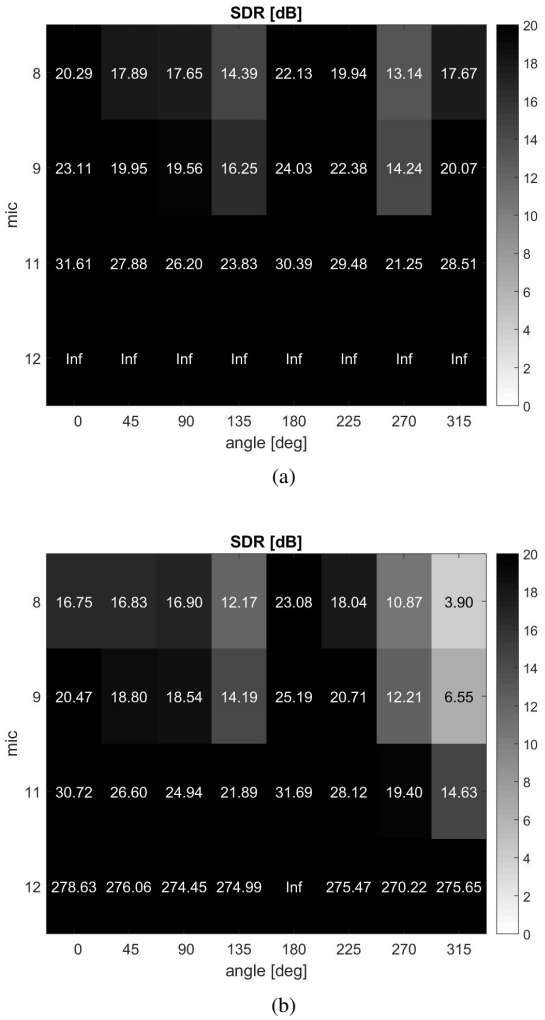


Figure 7: Evaluation of the RTF estimation using SDR for the internal array in the broad-side configuration using the (a) LS and (b) GEVD estimation methods.

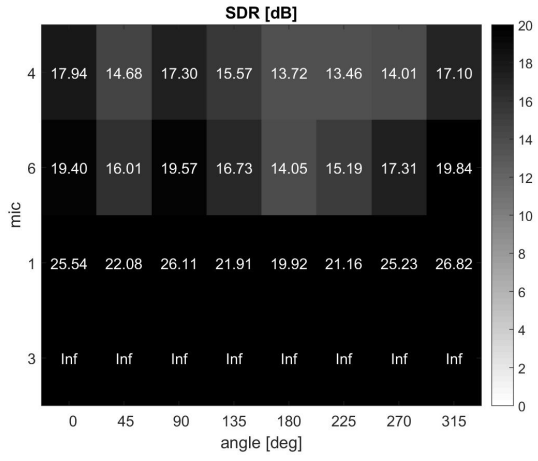
symmetry of the broad-side array around the driving direction would have caused an ambiguity for waves approaching from the front or from the back. However, as explained in the previous section, the RTFs are estimated using each microphone with less distortion using the broad-side array than using the end-fire array. Regardless, the difficulty of estimating the DoA and the lower angular resolution was greater in the case of the internal microphone array than in the case of the external one.

5. CONCLUSION

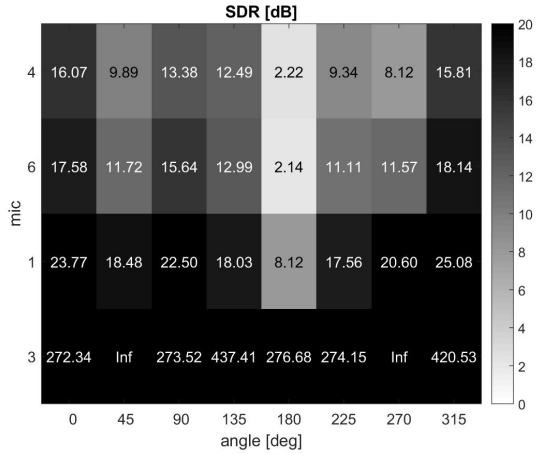
The feasibility of detecting the direction of an approaching EV was validated using an external microphone array equipped with 4 microphones. An algorithm for using internal microphones was developed in order but found to be inferior to an external array.

6. REFERENCES

[1] M. Sivak, "The information that driver use: is it indeed 90% visual?," *Perception*, vol. 25, no. 9, pp. 1081–

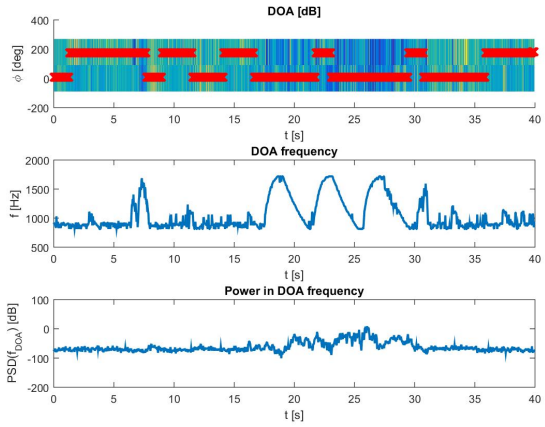


(a)

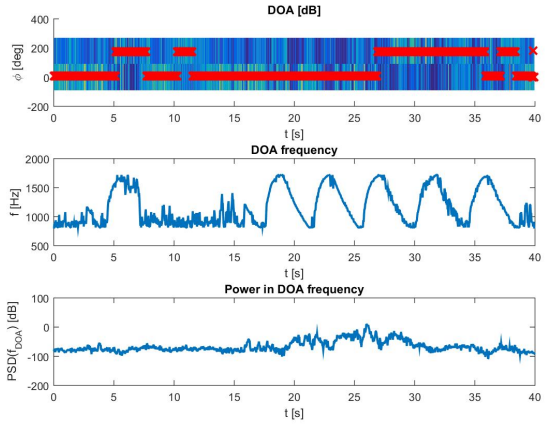


(b)

Figure 8: Fig. 7 repeated for end-fire configuration.



(a)



(b)

Figure 9: MUSIC spectrum and estimated DoA for an EV approaching from the frontal ($\sim 360°$) to back ($\sim 180°$) direction. The selected frequency matches the siren.

1089, 1996.

[2] Y. W. Seo and C. Urmson, "A perception mechanism for supporting autonomous intersection handling in urban driving," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1830–1835, 2008.

[3] A. Schaub, D. Baumgartner, and . Burschka, "Reactive obstacle avoidance for highly maneuverable vehicles based on a two-stage optical flow clustering," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2137–2152, 2016.

[4] I. Ruiz, D. Aufderheide, and U. Witkowski, "Radar sensor implementation into a small autonomous vehicle," in *Advances in Autonomous Mini Robots* (U. Rückert, S. Joaquin, and W. F. W., eds.), Berlin, Heidelberg: Springer, 2012.

[5] M. Pereira, D. Silva, V. Santos, and P. Dias, "Self calibration of multiple lidars and cameras on autonomous vehicles," *Robotics and Autonomous Systems*, vol. 83, pp. 326–337, 2016.

[6] D. I. Ferguson and J. Zhu, "Controlling autonomous vehicle using audio data," Mar. 2014. US Patent 8,676,427 B1.

[7] S. Collie, "Waymo's driverless cars make thousands of decisions every second to keep you alive," 2017. https://www.caradvice.com.au/591773/waymo-safety-report-sheds-light-on-the-life-of-a-self-driving-car/.

[8] J. Stewart, "Driverless cars need ears as well as eyes," 2017. https://www.wired.com/story/driverless-cars-need-ears-as-well-as-eyes/.

[9] M. McFarland, "Waymo's self-driving vans learn how to drive near police cars," 2017. https://money.cnn.com/2017/06/30/technology/future/waymo-chandler-arizona/index.html.

[10] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1071–1086, Aug 2009.

doi:10.25836/sasp.2019.22