A Spherical Microphone Array System for Traffic Scene Analysis

Zhiyun Li, Ramani Duraiswami, Elena Grassi, Larry S. Davis

Abstract—This paper describes a practical spherical microphone array system for traffic auditory scene capture and analysis. Our system uses 60 microphones positioned on the rigid surface of a sphere. We then propose an optimal design of a robust spherical beamformer with minimum white noise gain (WNG) of -6 dB. We test this system in a real-world traffic environment. Some preliminary simulation and experimental results are presented to demonstrate its performance. This system may also find applications in broader areas such as 3D audio, virtual environment, etc.

I. INTRODUCTION

Recently computer vision based traffic surveillance has been attracting many researchers [1][2][3][4]. On the other hand, the sound from a running vehicle also provides rich information about the vehicle. Researchers have used analogous methods from computer vision for vehicle sound signature recognition [5]. To extract spatial information from the auditory scene, however, we need a microphone array, such as the SAS-I system [6]. Spherical arrays of microphones are recently becoming the subject of some study as they allow omnidirectional sampling of the soundfield [8][9], and seeing potential applications in spatial soundfield analysis. In this paper, we will build a practical spherical microphone array system and use it for real-world traffic auditory scene capture and analysis.

This paper is organized into four parts: first, we briefly overview our system; we then present the theory of spherical beamformer which forms the basis of our system; third, we will design an optimal and robust beamformer for our array under the constraint of minimum -6 dB white noise gain; in the fourth part, we will demonstrate the performance of our system by simulations and real-world experiments.

II. SYSTEM OVERVIEW

Our system is as shown in Fig. 1. The main part is the spherical microphone array which consists of 60 omnidirectional microphones mounted on the rigid spherical surface of radius 10 cm. The positions of the 60 microphones are decided as the 64 nodes in [10] with four nodes removed because of the cable outlet and the mounting base. In fact, the layout also can be flexible [13]. The sound signals received by the array will flow through a 64 channel preamplifier before acquired by two 32 channel NI-DAQ cards on the computer.

I. THEORY OF SPHERICAL BEAMFORMER

The signal processing unit of our system is a spherical beamformer which is a spatial filter of 3D space. The basic principle of spherical beamformer is to make use of the orthonormality of spherical harmonics to decompose the soundfield arriving at a spherical array. Then the orthogonal components of the soundfield are linearly combined to approximate a desired beampattern.

A. Scattering from A Rigid Sphere

For a unit magnitude plane wave $k$, incident from direction $(\theta_k, \varphi_k)$, the incident field at a point $(\theta_s, \varphi_s, r_s)$ is

$$p_i = e^{ikr_s}$$

$$= 4\pi \sum_{n=0}^{\infty} \frac{j_n(kr_s)}{h_n^0(kr_s)} \times \sum_{m=-n}^{n} Y^m_n(\theta_k, \varphi_k) Y^*_m(\theta_s, \varphi_s),$$

where $j_n$ is the spherical Bessel function of order $n$, $Y^m_n$ is the spherical harmonics of order $n$ and degree $m$. * denotes the conjugation. At the same point, the field scattered by the rigid sphere of radius $a$ is [7]:

$$p_s = -4\pi \sum_{n=0}^{\infty} \frac{j_n'(ka)}{h_n^0(ka)} h_n(kr_s) \times \sum_{m=-n}^{n} Y^m_n(\theta_k, \varphi_k) Y^*_m(\theta_s, \varphi_s).$$
The total field on the surface ($r_s = a$) of the rigid sphere is:

$$p_t = (p_s + p_i)|_{r_s=a}$$

$$= 4\pi \sum_{n=0}^{\infty} i^n b_n(ka) \sum_{m=-n}^{n} Y_n^m(\theta_k, \varphi_k) Y_n^{m*}(\theta_s, \varphi_s),$$

where $b_n(ka) = \hat{f}_n(ka) - \frac{\hat{f}_n(ka)}{h_n(ka)} h_n(ka)$.

**B. Soundfield Decomposition and Beamforming**

If we assume that the pressure recorded at each point $(\theta_s, \varphi_s)$ on the surface of the sphere $\Omega_s$ is weighted by

$$W_n^{m*}(\theta_s, \varphi_s, ka) = \frac{Y_n^m(\theta_s, \varphi_s)}{4\pi i^n b_n(ka)}.$$

Then making use of orthonormality of spherical harmonics:

$$\int_{\Omega_s} Y_n^{m*}(\theta_s, \varphi_s) Y_{n'}^{m'}(\theta_s, \varphi_s) d\Omega_s = \delta_{nm} \delta_{mm'},$$

the total output from a pressure-sensitive spherical surface is:

$$P = \int_{\Omega_s} p_t W_n^{m*}(\theta_s, \varphi_s, ka) d\Omega_s = Y_n^{m*}(\theta_k, \varphi_k).$$

This shows the gain of the plane wave coming from $(\theta_k, \varphi_k)$, for a continuous pressure-sensitive spherical microphone is $Y_n^{m*}(\theta_k, \varphi_k)$. Since an arbitrary real function $F(\theta, \varphi)$ can be expanded in terms of complex spherical harmonics, we can implement arbitrary beampatterns. For example, an ideal beampattern directed at a direction $(\theta_0, \varphi_0)$:

$$F(\theta, \varphi) = \begin{cases} 1, & (\theta, \varphi) = (\theta_0, \varphi_0) \\ 0, & \text{otherwise} \end{cases}$$

can be expanded into:

$$F(\theta, \varphi) = 2\pi \sum_{n=0}^{\infty} \sum_{m=-n}^{n} Y_n^{m*}(\theta_0, \varphi_0) Y_n^{m}(\theta, \varphi).$$

Apparently, this system can be steered into any 3D directions digitally with the same beampattern. This provides the advantage to process sound from every directions unbiasedly.

**IV. OPTIMAL DESIGN OF ROBUST SPHERICAL BEAMFORMER**

To have an ideal beampattern for our array with unbalanced microphone layout, we need:

$$A \times W = B$$

where $A$ is the coefficients of the spherical harmonics expansion of the sound field in (5):

$$A = [A_1 \ A_2 \ \cdots \ A_S],$$

and $W$ is the complex weight to be assigned to each microphones.

$$W = \begin{bmatrix} W(\theta_1, \varphi_1) \\ W(\theta_2, \varphi_2) \\ \vdots \\ W(\theta_S, \varphi_S) \end{bmatrix}.$$ 

$B$ is the coefficient of the ideal beampattern steering at $(\theta_0, \varphi_0)$ in (11):

$$B = \begin{bmatrix} Y_0^0(\theta_0, \varphi_0) \\ Y_1^{1*}(\theta_0, \varphi_0) \\ \vdots \\ Y_N^{N*}(\theta_0, \varphi_0) \\ \vdots \end{bmatrix}.$$ 

In practice, we can only achieve the beampattern to limited orders with 60 microphones. In that case, (12) becomes an over- or under-determined linear system with respect to $W$.

A robust beamformer requires the minimum WNG is $-6$ dB [12]. So to design an optimal and robust spherical beamformer with limited microphones, we need to optimize the following:

$$\text{minimize} \ |A \times W - B|^2$$

subject to:

$$\text{WNG} = 10 \log_{10} \left( \frac{|W^H d|^2}{|W^H W|^2} \right) \geq -6 \text{ dB}.$$
where \( \mathbf{d} \) is the vector of complex pressure on the microphone positions produced by the unit magnitude plane wave from \((\theta_0, \varphi_0)\). However, the WNG is also caused by the fast decay of \( b_n(ka) \) with increasing \( n \) as shown in Fig. 2.

Consider the following equation:

\[
\begin{bmatrix}
iN b_N(ka)Y_N^{\ast}(\theta_1, \varphi_1) \\
iN b_N(ka)Y_N^{\ast}(\theta_2, \varphi_2) \\
\vdots \\
iN b_N(ka)Y_N^{\ast}(\theta_S, \varphi_S)
\end{bmatrix}
\begin{bmatrix}
W(\theta_1, \varphi_1) \\
W(\theta_2, \varphi_2) \\
\vdots \\
W(\theta_S, \varphi_S)
\end{bmatrix}
= Y_N^\ast(\theta_0, \varphi_0)
\]

(19)

Using Cauchy's Inequality, we have:

\[
\left| \sum_{s=1}^{S} iN b_N(ka)Y_N^{\ast}(\theta_s, \varphi_s)W(\theta_s, \varphi_s) \right|^2 \\
\leq \left( \sum_{s=1}^{S} \left| iN b_N(ka)Y_N^{\ast}(\theta_s, \varphi_s) \times |W(\theta_s, \varphi_s)| \right|^2 \right)^{1/2}
\]

\[
\leq \sum_{s=1}^{S} |b_N(ka)|^2 \sum_{s=1}^{S} |Y_N^{\ast}(\theta_s, \varphi_s)|^2 \sum_{s=1}^{S} |W(\theta_s, \varphi_s)|^2
\]

So:

\[
\sum_{s=1}^{S} |W(\theta_s, \varphi_s)|^2 \geq \frac{|Y_N^{\ast}(\theta_0, \varphi_0)|^2}{\sum_{s=1}^{S} |b_N(ka)|^2 \sum_{s=1}^{S} |Y_N^{\ast}(\theta_s, \varphi_s)|^2}.
\]

(20)

Therefore, to have an optimal solution of \( \mathbf{W} \), we just set a constraint:

\[
b_n(ka) \geq -30 \text{ dB}.
\]

(21)

The linear system (12) then becomes:

\[
\mathbf{A} \times \mathbf{W} = \widetilde{\mathbf{B}},
\]

(22)

where \( \widetilde{\mathbf{B}} \) is \( \mathbf{B} \) modulated by a step function \( \mathcal{U}_n(ka) \):

\[
\widetilde{\mathbf{B}} = \begin{bmatrix}
Y_0^{\ast}(\theta_0, \varphi_0) \times \mathcal{U}_0(ka) \\
Y_1^{\ast}(\theta_0, \varphi_0) \times \mathcal{U}_1(ka) \\
\vdots \\
Y_{\infty}^{\ast}(\theta_0, \varphi_0) \times \mathcal{U}_{\infty}(ka)
\end{bmatrix}.
\]

(23)

\[
\mathcal{U}_n(ka) = \begin{cases}
1 & b_n(ka) \geq -30 \text{ dB} \\
0 & \text{otherwise}
\end{cases}
\]

(24)

Because of the convergence of the sound field expansion (5), the modulated linear system (22) actually contains finite equations.

The optimization can be numerically solved with MATLAB function \texttt{fmincon}.

V. RESULTS

We demonstrate our system by simulations and experiments.

We first plot the beampatterns of order five at 2 kHz. We steer the beampattern in \((\theta_0, \varphi_0) = (\pi/4, \pi/4)\) direction. Fig. 3 shows the distorted beampattern without optimization. The WNG is about \(-12.73 \text{ dB}\) which damages the robustness. The optimal beampattern is shown in Fig. 4. It optimally approximates the standard beampattern of order five in the least square sense. This minimizes the irregularities inherent with the system. And it is under the WNG constraint of \(-6 \text{ dB}\). As an example, Fig. 5 shows the directivity index (DI) and WNG of an optimal spherical beamformer pointing to the same direction in a wide frequency range. Here the DI is defined as [12]:

\[
DI = 10 \log_{10} \left( \frac{1}{\frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} |H(\theta, \varphi)|^2 \sin(\theta) \sin(\varphi) \, d\theta \, d\varphi} \right).
\]

(25)
Fig. 5. Optimal spherical beamformer with minimum WNG of $-6$ dB.

here $H(\theta, \varphi)$ is the optimized beampattern. The optimal spherical beamformer for all angles in 3D space can be precomputed and stored in a lookup table.

We then took the spherical microphone array system to roadside to record real-world traffic auditory scenes. A simple scenario is as shown in Fig. 6 where a car is moving from the left side of the array to the right side along the street. To track the car in this scenario from the recordings, for each frame, we simply steer the spherical beamformer to every 3D directions to search the peaks. We assume the sound from the car is stronger than the sound from all other directions. So the peaks indicate the car locations. Since the spectrum of the sound from the car is not constant, the localization is performed at multiple frequencies. We will extend this to multiple sound source localization later in this section. The tracking path is shown in Fig. 7 where the left and right sides of the array correspond to azimuth angle $\pi$ and 0, respectively. The jitters on the tracking path may be caused by several factors:

1) the sound spectrum from the vehicle is not constant across the recording time;
2) the sound source is not simply a point source since many parts of a running vehicle can make sound;
3) the environment itself is very noisy with wind blowing, people walking and talking, a building being constructed on the other side of the street, the reflections from walls and the ground surface in addition to self noise of the system.

When the car is far away from the array, its sound fades into the environmental noise. So the jitters tend to be more salient at the two endpoints of the path. We will see this more clearly in the two-vehicle case later.

To show the actual 3D localization performance of the spherical beamformer, we pick three typical frames of the recorded sound as numbered 1-3 in Fig. 6. The localization result is a 3D surface, where each surface point represents a localization angle and its distance from the origin represents the amplitude of the beamformed signal. In an ideal free space environment under plane wave assumption, using a standard beampattern, the localization surface will have the same shape of this beampattern pointing to the plane wave direction. In a real-world environment, however, those factors listed above will affect the final result which may in turn provide some spatial cues about the sound source and the environment. For clarity, we only show the localizations using a single frequency at about 2 kHz, and plot in normalized linear scale instead of dB scale. We provide two different viewpoints for each figure, the left plot is from the same viewpoint as in Fig. 6 while the right one is from an appropriate 3D viewpoint. The first 3D localization is shown in Fig. 8. It has a spindle-like shape. The localization for the second chosen frame is shown in Fig. 9, which indicates the effect of a spreading sound source at this frequency. In addition, the irregular sidelobes are likely caused by all kinds of noise. Fig. 10 shows the 3D localization for the third chosen frame. It clearly shows a second sound source at this frequency, real or image, but weak. We can also see this sound is from the lower part of the car, possibly from the running wheels or brakes.

We also tested our system in a more challenging scenario. In another experiment, we recorded two cars moving from right to left successively in different speed. Fig. 11 shows the tracks of two cars.
VI. SUMMARY AND FUTURE WORK

We have built a practical spherical microphone array system using 60 microphones. Our system achieves optimal approximation of the standard beampattern and robustness with the minimum white noise gain of $-6$ dB. Just like camera in vision-based traffic surveillance, our spherical microphone array is an analogous tool in sound-based surveillance. We evaluated our system in a real-world traffic environment and demonstrated its performance.

Future work includes visual-audio joint traffic surveillance, complex traffic auditory scene analysis, etc.

REFERENCES


