5Q-01

A Spatial Filter Design for Surface Sound Source Separation

Zhi Zhong[1], Katsutoshi Itoyama[1], Kenji Nishida[1], Kazuhiro Nakadai[1,2]
1 Tokyo Institute of Technology   2 Honda Research Institute Japan

1. Introduction

A sound source is mainly modeled as a point source in spatial signal processing [1], though surface sources with a certain shape and size are common in the real world. Some studies have been carried out to design a spatial filter to separate non-audio or audio signals emitted from a region [2, 3, 4]. We proposed a spatial filter design for surface sound source separation (SSS), which is called the scan-and-sum beamformer [5]. It performs SSS of surface sources distributed in the azimuth angle domain. This paper introduces an optimization to the summation procedure in the previously-reported scan-and-scum filtering method. Various simulations are carried out to show that the scan-and-sum filtering improves sound source separation for a mixture of surface sound sources.

2. Scan-and-sum Beamformer

In the scan-and-sum beamformer, a point source sub-beamformer changes its focus DOA, scans with appropriate scan density the region where the target surface source exists. Then sub-beamformers are integrated to the surface beamformer, which is the scan-and-sum beamformer. An illustration is shown in Fig.2, where the scanning sub-beamformers (black) produces a broad pattern (blue) in approximation to the ideal surface pattern (red).

A uniformly-aligned M-ch linear microphone array is deployed with element distance d. Transfer function (TF) is denoted as $\boldsymbol{a}(\theta, f) \in \mathbb{C}^{M\times 1}$. f the frequency would be omitted. The direction response of a beamformer in a frequency point is the “pattern” defined as

$$p_{\phi_n}(\theta) = \boldsymbol{W}_{\phi_n}^H \times \boldsymbol{a}(\theta), \tag{1}$$

where $\boldsymbol{W} \in \mathbb{C}^{M\times 1}$ is the coefficient vector, H means Hermitian conjugate, and ϕ_n stands for a focusing direction of arrival (DOA).

The unnormalised scan-and-sum beamformer was formulated as an unweighted summation:

$$P(\theta) = \sum_{n=1}^{N} p_{\phi_n}(\theta) \tag{2}$$

where $N(=\frac{|\Theta_{\text{tar}}|}{\Delta\theta}+1)$, Θ_{tar} is the set in which target sources exist, $\Theta_{\text{tar}} = \{\phi_n : n = 1, 2, \ldots, N\}$. Θ_{itf} is the set for interference sources.

The previous research [5] focused on 2 factors: (1) an input of M-channel audio datas which is related to microphone number in an array, considered as physical cost; (2) a middle layer of sub-beamformers which is related to an average scanning stepsize $\Delta\theta$, considered as computational cost. A trade-off between performance and cost is discussed.

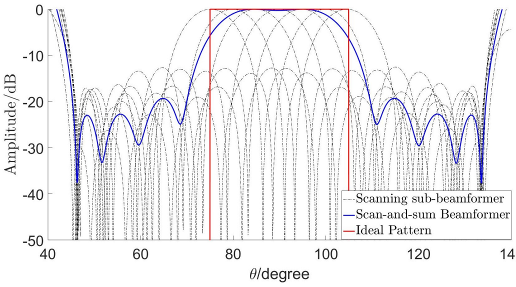

Fig.1 Illustration of the Scan-and-sum Beamformer

3. The Optimization of Summation

Suppose the focusing DOA of n-th sub-beamformer is $\phi_n = \phi_1 + (n-1)\Delta\theta$, where $n = 1, 2, \ldots, N$, ϕ_1 is the lower bound of Θ_{tar}. Concerning Eq.1, the middle layer of sub-beamformers is represented as

$$\boldsymbol{q}(\theta) = [p_{\phi_1}(\theta), p_{\phi_2}(\theta), \ldots, p_{\phi_N}(\theta)]. \tag{3}$$

A summation vector could be defined as

$$\boldsymbol{B} = [b_1, b_2, \ldots, b_N]^T \in \mathbb{C}^{N\times 1}, \tag{4}$$

for instance, the vector of unweighted summation in [5] is $\boldsymbol{B}_{\boldsymbol{u}} = [1, 1, \ldots, 1]^T$.

A normalization parameter which normalises the maximum response in a pattern to 0dB is searched as

$$\lambda_{\max} = \max\{|\boldsymbol{q}(\theta) \times \boldsymbol{B}_{\boldsymbol{u}}| \ : \ \theta \in \theta_{\text{tar}}\}. \tag{5}$$

The weighted summation of sub-beamformers is represented by

$$\frac{\boldsymbol{q}(\theta)}{\lambda_{\max}} \times \boldsymbol{B} = P(\theta), \tag{6}$$

The coefficient vector of a scan-and-sum beamformer is a weighted summation:

$$\boldsymbol{W} = \frac{\sum_{n=1}^{N} b_n \boldsymbol{W}_{\phi_n}}{\lambda_{\max}}. \tag{7}$$

To optimize the summation procedure, we define an ideal pattern $D(\theta)$ for a surface beamformer in azimuth dimension as

$$D(\theta) = \begin{cases} 1, & \theta \in \Theta_{\text{tar}} \\ 0, & \theta \in \Theta_{\text{itf}} \\ \text{no definition}, & \text{other situation} \end{cases}. \tag{8}$$

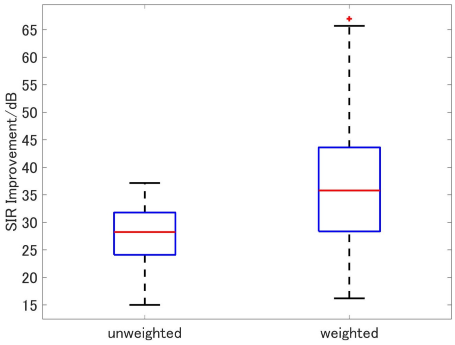

Fig.2 The Box Plot of the SIR Improvement by Unweighted and Optimized Version in 100 Times of Simulation

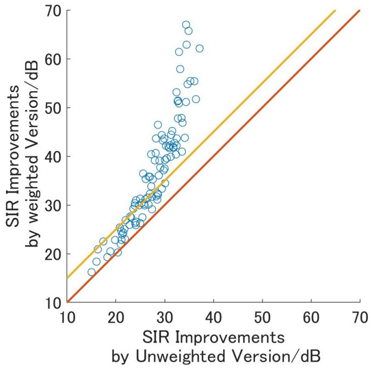

Fig.3 The Scatter Plot of the SIR Improvement by Unweighted and Optimized Version in 100 Times of Simulation

Then $\boldsymbol{q}(\theta)$ and $D(\theta)$ are extended into

$$\boldsymbol{q} = [\boldsymbol{q}(\theta_1), \ldots, \boldsymbol{q}(\theta_{|\Theta_{\text{itf}} \cup \Theta_{\text{tar}}|})]^T, \tag{9}$$

$$\boldsymbol{D} = [D(\theta_1), D(\theta_2), \ldots, D(\theta_{|\Theta_{\text{itf}} \cup \Theta_{\text{tar}}|})]^T, \tag{10}$$

where $\theta \in \Theta_{\text{itf}} \cup \Theta_{\text{tar}}$.

A least square optimization problem is constructed as

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \|\frac{\boldsymbol{q}}{\lambda_{\max}} \times \boldsymbol{B} - \boldsymbol{D}\|^2, \tag{11}$$

which could be solved as

$$\hat{\boldsymbol{B}} = (\frac{\boldsymbol{q}}{\lambda_{\max}})^+ \times \boldsymbol{D}; \tag{12}$$

where + is the Moore-Penrose pseudo inverse, implemented in Matlab as “pinv”.

4. Evaluation

In this evaluation, signals input as plane waves, $M = 20$ omnidirectional microphones are deployed and element distance is $d = 2\text{cm}$, $\Delta\theta = 0.92°$.

We prepared a dataset containing 100 different sound source distributions. In each case, 6 points in $\{\theta : 0° \leq \theta \leq 180°\}$ are random generated through the uniform distribution, and are sorted from small value to the large. points 1,2 forms surface source region no.1, points 3,4 forms region no.2, and points 5,6 forms region no.3. One of the 3 surface source region is randomly assigned as the target. A surface source is represented by point sources intensively distributed within the region at every 1°. All point sources are independent with power 1 in every frequency point. FFT length Lfft is set to 64.

SIR is the ratio between the power of target sources and the power of interferences[6]. The difference in SIR before and after a separation algorithm could be used to measure the separation rate.

Fig.2 shows the box plot of SIR improvements by the unweighted scan-and-sum beamformer and the optimized version. In this dataset, the unweighted version provides an average SIR improvements of 28dB while the optimized version achieves an average of 35dB. Fig.3 shows the scatter plot between the unweighted and optimized beamformer, in which the optimized version has better SIR performance except 1 case. In 65% cases, the optimized version improves SIR at least 5dB more compared to the unweighted version.

5. Conclusion

This paper describes a spatial filter design for surface sound source separation, which is called a scan-and-sum beamformer. Compared to previous works about an unweighted scan-and-sum beamformer, an optimized summation is introduced, bringing significant improvements in some cases. Simulations show that the scan-and-sum beamformer is applicable to various sound source distributions, improving SIR for mixtures of three surface sound sources.

Acknowledgement

This work was supported by KAKENHI 16H02884, 17K00365, and 19K12017.

References

[1] F. Asano: “Array Signal Processing for Acoustics – Localization, tracking and separation of sound sources,” The Acoustical Society of Japan, 2011. (in Japanese)

[2] H. L. van Trees, Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory. John Wiley & Sons, Inc, 2002.

[3] P. M. Woodward: “A Method of Calculating the Field over a Plane Aperture Required to Produce a Given Polar Diagram,” J. IEEE (London) Pt. IIIA, 93(10), pp. 1554–1558, 1946.

[4] M. Taseska, E. A. P. Habets: “Spotforming Using Distributed Microphone Arrays,” IEEE WASPAA, 2013.

[5] Z. Zhong, *et al.*: “Design of a Scan-and-sum Beamformer for Surface Sound Source Separation ,” the 37th Annual Conference of the Robotics Society of Japan (RSJ2019), 2019

[6] E. Vincent, *et al.*: “Performance Measurement in Blind Audio Source Separation,” IEEE Trans Audio Speech Lang Process, 14(4), pp. 1462–1469, 2006.