LCMV-based Scan-and-Sum Beamforming for Region Source Extraction

Aoto Yasue*, Benjamin Yen*, Katsutoshi Itoyama† and Kazuhiro Nakadai*
* Tokyo Institute of Technology
E-mail: yasue@ra.sc.e.titech.ac.jp
† Honda Research Institute Japan Co., Ltd.

***Abstract*—This paper addresses region source extraction (RSE), which extracts the entire sound source originating from a specified target region. To solve this problem, an LCMV-based Scan-and-Sum Beamformer (LCMV-SS-BF) that can separate a wide range of continuous regions is proposed. This method is designed by incorporating the linearly constrained minimum variance Beamformer (LCMV-BF) as a filter for continuous regions into the Scan-and-Sum Beamformer (SS-BF) as a filter for wide-area sound source separation. In addition, an incremental estimation method using the Woodbury's formula for LCMV-BF filters is proposed to improve the processing speed of the filter computation of the LCMV-BF. Numerical simulation using convolved sound datasets showed 1) the effectiveness of the LCMV-SS-BF for surface source separation and 2) the computational cost advantage of the incremental estimation method.**

I. Introduction

Source enhancement techniques such as beamforming have been widely studied, and these signal models assume point sources [1], [2]. However, in reality, ideal point sources do not exist. Therefore, model errors are inherent in the extraction of sound sources that exist within a specific area, such as non-point sound sources like waterfalls, multiple instruments in an orchestra, and the utterances of a driver whose position is constantly changing, and the expected performance cannot be achieved. In addition, when a video camera or telepresence robot is used to track and direct attention to a specific target, both audio and visual information can be used, but in most cases, image and sound are used without linking them. For example, even if you zoom in on an object in terms of images, the process of zooming in is not performed in terms of audio, and only the background noise is processed. As a result, sounds outside the target region are also recorded. To solve this problem, it is necessary to link the image and audio and control the spatial recording range so that only acoustic signals from the same range as the camera's image range are recorded (called audio-visual zooming [3]). Therefore, a new technique is required to specify a target region and extract all acoustic signals coming from within the target region. In this paper, we call this technique "region source extraction" (RSE) and treat it as a method for isolating and extracting sound sources within a specified region.

Beamforming is a technique used to extract point source in a specific direction using a microphone array and utilizing the property that each microphone has a different time delay when the direction of arrival is different [2], [4]. Beamforming is a point source extraction method that basically specifies the target direction of the source to be extracted. However, RSE requires specifying a target region instead of a target direction and extracting all sound sources within the target region, which is beyond the capabilities of conventional beamforming.

In order to perform RSE, multi-modal methods using audiovisual information are proposed [5]–[9]. These methods use a camera and microphone to synchronously record images and sound. Then assuming a strong correlation between the recorded image and audio, the audio corresponding to the objects in the image is separated and extracted using deep learning. However, the system assumes that the target object is contained in the image, and cannot be used in situations where the object is hidden or far from the camera. In addition, a large data set is required to handle various situations.

On the other hand, as methods for RSE using only audio information, there are techniques such as acoustic zooming and a Scan-and-Sum beamforming (SS-BF).

Acoustic zooming detects all sound sources in the observed sound by sound source localization and extracts each detected source using conventional beamforming. RSE is then achieved by resynthesizing only the extracted sources within the target region [10]–[12]. This method requires the number and location of sound sources as a priori information, but it is difficult to provide this information in advance for a set of sound sources such as an orchestra or a non-point source. The same problem also occurs with blind source separation [13]–[15] approaches, which can separate multiple sound sources at once, to achieve RSE.

An SS-BF [16], [17] achieves RSE by the division of the target region (each divided small region is called a sub-region) and the sum of beamformers that extract sources within the sub-region. the SS-BF specifies a target region, divides it into sub-regions, then constructs sub-beamformers (sub-BF) for all sub-regions such that the sources in the center direction of the subregion are extracted. Finally, it forms a filter for RSE by the weighting sum of the sub-BFs. This method does not require the number or location of the sources, but the extraction performance is degraded for sources located off the center of the main lobe of each sub-BF. This is due to the fact that sub-BFs are computed for the extraction of discrete points in the target region, which means SS-BF is not a filter suitable for continuous region extraction.

As reflected in the existing studies, the problems of conven-

tional methods include 1) the requirement of image data, 2) the requirement of the number and location of sound sources, or 3) the inability to create filters for continuous regions. To solve these problems, we propose an LCMV-based Scan-and-Sum Beamformer (LCMV-SS-BF). This method combines a linearly constrained minimum variance beamformer (LCMV-BF) with the SS-BF to achieve RSE without using image data, model training with data sets, or sound source localization as preprocessing. The LCMV-BF constructs a filter by solving a minimization problem by placing no-distortion constraints on multiple directions [18], [19]. It is commonly used to separate multiple sources at once by simultaneously providing constraints on multiple source directions. On the other hand, it has been reported that by giving each constraint point in a neighborhood, stable extraction is possible for the continuous region between the constraint points. This property is generally used to deal with direction-of-arrival mismatch [20]. This property suggests that LCMV-BF could be used as a filter for RSE. However, since the constraints of the LCMV-BF must be placed in the neighborhood to acquire the property, it is not possible to perform wide region separation. Therefore, we propose an LCMV-SS-BF, which adopts the structure of the SS-BF to LCMV-BF. To cope with the huge computational time required for the SS-BF, we also propose a method to speed up the calculation of the LCMV-BF by using a recursive least square (RLS) method to estimate the filters incrementally.

We evaluate the proposed method by numerical simulation. For the evaluation, we created self-made datasets of mixed-source speech data by convolving transfer functions to each audio data from publicly available datasets and summing them over multiple audio data. Simulation results show that LCMV-SS-BF outperforms conventional methods for RSE, indicating the superiority of the proposed method. Furthermore, the incremental estimation method achieves faster computation.

The contributions presented in this paper are as follows:

- We propose an LCMV-SS-BF as a novel RSE method that can handle a wide range of continuous regions.
- We propose the use of an incremental RLS method for LCMV-BF estimation based on the Woodbury's formula and introduce it into LCMV-SS-BF to improve its processing speed.
- We show the effectiveness of the proposed LCMV-SS-BF, including its speed-up by the incremental estimation through simulations using the self-made sound datasets.

II. Preriminary

In this section, we formulate the problem setting of RSE and the LCMV beamformer on which the proposal is based.

A. problem setting of Region Source Extraction

Fig. 1 shows the coordinate system and the definition of target sources for RSE. Let the target region be a region cut off from the spherical coordinate system with azimuth θ and elevation ϕ, where $-\mathcal{R}/2 \leq \theta, \phi \leq \mathcal{R}/2$ (the orange region in Fig. 1), and the noise region be a region, where $-\mathcal{R}_{n_{\text{out}}}/2 \leq \theta, \phi \leq \mathcal{R}_{n_{\text{out}}}/2$ and $(-\mathcal{R}_{n_{\text{in}}}/2 \leq \theta, \phi \leq \mathcal{R}_{n_{\text{in}}}/2)^c$

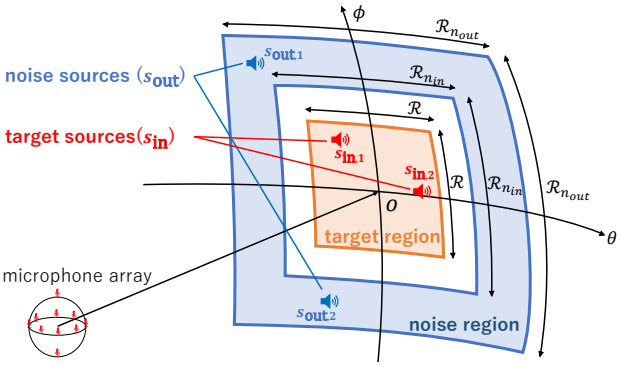

Fig. 1. A target surface region defined as a region that cuts off a portion of a spherical coordinate system.

(the blue region in Fig. 1). $(\cdot)^c$ means the residual phenomenon. These regions are parts of the spherical surface. Then target sources s_{in} are defined as the sources coming from within the target region and noise sources s_{out} are the sources coming from within the noise region (without the target region). When $z_{\omega,t,m}$ is the observed signal by the m-th microphone ($m = 1, \ldots, M$) at angular frequency ω, time frame t, the observed signal $\boldsymbol{z}_{\omega,t}$ is defined as,

$$\boldsymbol{z}_{\omega,t} = \sum_p \boldsymbol{a}_{\text{in},p,\omega,t} \cdot s_{\text{in},p,\omega,t} + \sum_q \boldsymbol{a}_{\text{out},q,\omega,t} \cdot s_{\text{out},q,\omega,t}, \quad (1)$$

$$\boldsymbol{z}_{\omega,t} = [z_{\omega,t,1}, \cdots, z_{\omega,t,M}]^T, \quad (2)$$

$$\boldsymbol{a}_{x,\omega,t} = [e^{-j\omega\tau_{1,x}}, \cdots, e^{-j\omega\tau_{M,x}}]^T, \quad (3)$$

where $\boldsymbol{a}_{x,\omega,t} \in \mathbb{C}^{M\times 1}$, $x \in [(\text{in}, p), (\text{out}, q)]$ denotes the transfer function corresponding to the sound source $s_{x,\omega,t}$, while $\cdot^T$ indicates the matrix transpose operator. $\tau_{m,x}$ indicates the time difference of arrival (TDOA) of the m-th microphone to the reference point in the x-th transfer function, $\boldsymbol{a}_{x,\omega,t}$. Now the objective of RSE can be written as extracting $\sum_p s_{\text{in},p,\omega,t}$.

B. LCMV Beamformer

When applying beamforming to the observed signal $\boldsymbol{z}_{\omega,t}$, the output signal $Y_{\omega,t}$ is defined as,

$$Y_{\omega,t} = \boldsymbol{w}_{\omega,t}^H \boldsymbol{z}_{\omega,t}, \quad (4)$$

$$\boldsymbol{w}_{\omega,t} = [w_{\omega,t,1}, \cdots, w_{\omega,t,M}]^T, \quad (5)$$

where $\boldsymbol{w}_{\omega,t} \in \mathbb{C}^{M\times 1}$ denotes the filter of a beamformer, while $\cdot^H$ indicates the matrix Hermitian transpose operator. ω and t will be omitted for simplicity thereafter.

The LCMV-BF is obtained as the solution to a constrained optimization problem that minimizes the output of the beamformer with all-pass characteristics for multiple directions as,

$$\tilde{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\text{argmin}}\, \boldsymbol{w}^H \boldsymbol{R}\boldsymbol{w} \quad (6)$$

$$\text{subject to } \boldsymbol{A}^H\boldsymbol{w} = \boldsymbol{f},$$

$$\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_k, \cdots, \boldsymbol{a}_K],$$

$$\boldsymbol{a}_k = [e^{-j\omega\tau_{1,k}}, \cdots, e^{-j\omega\tau_{M,k}}]^T,$$

$$\boldsymbol{f} = [f_1, \cdots, f_k, \cdots, f_K]^T,$$

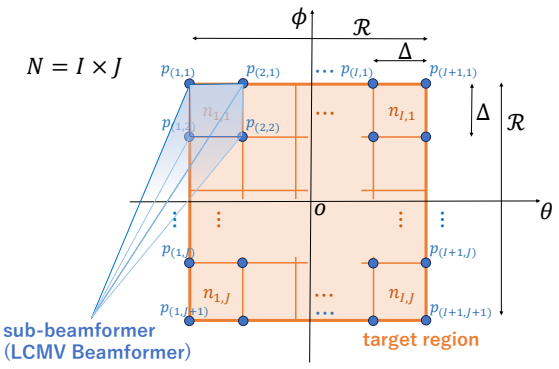

Fig. 2. LCMV-based Scan-and-Sum Beamformer (LCMV-SS-BF)

where $\boldsymbol{R} = E[\boldsymbol{z}\boldsymbol{z}^H] \in \mathbb{C}^{M\times M}$ represents the spatial correlation matrix of the observed signals. $E[\boldsymbol{z}\boldsymbol{z}^H]$ denotes the expected value of $\boldsymbol{z}\boldsymbol{z}^H$. $\boldsymbol{A} \in \mathbb{C}^{M\times K}$ is a matrix consisting of K transfer functions for constraint directions. f_k denotes the all-pass characteristics in the k-th direction.

By solving Eq. (6), the LCMV-BF is formulated as,

$$\tilde{\boldsymbol{w}} = \boldsymbol{R}^{-1}\boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{f}. \tag{7}$$

III. Proposed Method

This section proposes and formulates the LCMV-SS-BF and the incremental RLS method for LCMV-SS-BF estimation.

A. LCMV-based Scan-and-Sum Beamformer

Fig. 2 shows the abstract of the proposed LCMV-SS-BF. The target region shown in Fig. 1 is considered as an $\mathcal{R}\times\mathcal{R}$ square area as shown in Fig. 2. To construct the LCMV-SS-BF, the target region is divided by small $\Delta\times\Delta$ square regions (sub-regions) in both θ and ϕ direction to obtain N sub-regions. The sub-region (i,j) is indicated as $n_{i,j}$, and the grid point (i,j) is indicated as $p_{(i,j)}$. After that, a filter for each sub-region is computed as a sub-BF, and all sub-BFs are summed up to form the SS-BF. For the sub-BF, the LCMV-BF is selected and it is designed to separate the entire sub-region with minimal distortion by constraining the all-pass characteristics at the four edges of each sub-region. Since the target region is divided into small sub-regions, the four edges are placed in a neighborhood, which maintains the robustness to the continuous direction between each point. Thus, the filter of the proposed LCMV-SS-BF, $\boldsymbol{w}_{\text{proposed}}$ is formulated as,

$$\boldsymbol{w}_{\text{proposed}} = \sum_{i=1}^{I}\sum_{j=1}^{J} b_{i,j}\cdot\boldsymbol{R}^{-1}\boldsymbol{A}_{i,j}(\boldsymbol{A}_{i,j}^H\boldsymbol{R}^{-1}\boldsymbol{A}_{i,j})^{-1}\boldsymbol{1}, \tag{8}$$

$$\boldsymbol{A}_{i,j} = [\boldsymbol{a}_{p_{(i,j)}}, \boldsymbol{a}_{p_{(i+1,j)}}, \boldsymbol{a}_{p_{(i,j+1)}}, \boldsymbol{a}_{p_{(i+1,j+1)}}], \tag{9}$$

where $\boldsymbol{A}_{i,j} \in \mathbb{C}^{M\times 4}$ includes the four corner transfer functions of $n_{i,j}$ (see Fig. 2) and $\boldsymbol{B} = [b_{1,1}, \cdots, b_{i,j}, \cdots, b_{I,J}]$ is an weight vector for each sub-region.

The LCMV-SS-BF compensates for the disadvantages of both the LCMV-BF and the SS-BF by combining them: by using the structure of the SS-BF, the constraint points of the LCMV-BFs can be placed in the neighborhood, allowing a wide range RSE. On the other hand, by using the LCMV-BF as a sub-BF, the SS-BF can construct a filter for continuous region extraction instead of the sum of filters for discrete points. So the LCMV-SS-BF can extract sound sources from a wide range and continuous region without using visual information and knowledge regarding the number and directions of sound sources included in the region.

B. Incremental RLS method for LCMV-BF estimation

The SS-BF requires repeated calculations of sub-BFs to estimate the filter, which increases the computation time. To solve this problem, we propose a method to reduce the computation time of the LCMV-BF, which is an adaptive filter and requires computationally expensive inverse matrices of $\boldsymbol{R}$ and $\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A}$. To avoid such inverse matrix computation, we propose an incremental estimation method based on recursive least square (RLS) method [21].

While the inverse of $\boldsymbol{R}_{\omega,t}$ needs to be computed for each time frame t and angular frequency ω, the RLS method can compute it incrementally using the result obtained at the previous time frame. First, $\boldsymbol{R}_{\omega,t}$ is defined in a recursive manner by omitting ω again for simplicity as,

$$\boldsymbol{R}_t = \alpha\boldsymbol{R}_{t-1} + (1-\alpha)\,\boldsymbol{z}_t\boldsymbol{z}_t^H, \tag{10}$$

where α is the forgetting coefficient. Using the Woodbury's formula [22], $\boldsymbol{R}_t^{-1}$ can be formulated as follows:

$$\boldsymbol{R}_t^{-1} = \frac{1}{\alpha}\left(\boldsymbol{R}_{t-1}^{-1} - \frac{1}{\gamma_t}\boldsymbol{R}_{t-1}^{-1}\boldsymbol{z}_t\boldsymbol{z}_t^H\boldsymbol{R}_{t-1}^{-1}\right) \tag{11}$$

$$\gamma_t = \frac{\alpha}{1-\alpha} + \boldsymbol{z}_t^H\boldsymbol{R}_{t-1}^{-1}\boldsymbol{z}_t. \tag{12}$$

Likewise, the incremental expression of $\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A}$ in Eq. (7) is formulated by using Eq. (11) as,

$$\boldsymbol{A}^H\boldsymbol{R}_t^{-1}\boldsymbol{A} = \frac{1}{\alpha}\left(\boldsymbol{A}^H\boldsymbol{R}_{t-1}^{-1}\boldsymbol{A} - \frac{1}{\gamma_t}\boldsymbol{l}_t\boldsymbol{l}_t^H\right) \tag{13}$$

$$\boldsymbol{l}_t = \boldsymbol{A}^H\boldsymbol{R}_{t-1}^{-1}\boldsymbol{z}_t. \tag{14}$$

Using the Woodbury's formula, $\boldsymbol{D}^{-1} = (\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}$ can be expressed as

$$\boldsymbol{D}_t^{-1} = \alpha\left(\boldsymbol{D}_{t-1}^{-1} + \frac{1}{\delta_t}\boldsymbol{D}_{t-1}^{-1}\boldsymbol{l}_t\boldsymbol{l}_t^H\boldsymbol{D}_{t-1}^{-1}\right) \tag{15}$$

$$\delta_t = \gamma_t - \boldsymbol{l}_t^H\boldsymbol{D}_{t-1}^{-1}\boldsymbol{l}_t. \tag{16}$$

By using Eqs. (11) and (15), the calculation of the LCMV-BF can be obtained as an incremental estimation without using inverse matrix calculations.

IV. Evaluation

The proposed methods are evaluated by simulations with convolved sound datasets with publically available audio datasets, in terms of extraction performance improvement and computational speed.

A. Experimental setting

As a recording device, a 10 ch spherical microphone array of radius 3 cm was selected, which has eight microphones at the equator and another two at the poles, as shown in Figure 1. The acoustic transfer functions were calculated geometrically assuming a free acoustic field and plane wave model. For the simulations, we prepared four datasets, Datasets 1 to 4. The comparison of each dataset is shown in TABLE I. Datasets 1 and 2 were used to evaluate the extraction performance when different types of sound sources were mixed together as the target and noise sources. As the target sources, we used the train-clean-100 dataset from LibriSpeech [23], which is the dataset of human speech sounds. As the noise sources, we used the background noise sources from the CHiME3 dataset [24], which is the dataset of background noise sounds. Datasets 3 and 4, on the other wards, were used to evaluate the extraction performance when the same type of sound sources were mixed together, and the dev-clean dataset from LibriSpeech was used as both target and noise source. Datasets 1 and 3 were used to evaluate the extraction performance in the case of the mixture of one target source and one noise source, while Datasets 2 and 4 were used to evaluate the extraction performance in the case of the mixture of two target sources and two noise sources, which means the multiple sources extraction. An appropriate acoustic transfer function was applied to each source so that target sources were randomly placed in the target region and noise sources were randomly placed in the noise region. And by mixing them, we created 100 6-second-long synthetic sounds for each Dataset, each $\mathcal{R}$ size to be evaluated. The setting of $\mathcal{R}$ were conducted on 20°, 40°, and 60°. $\mathcal{R}_{n_{\text{in}}}$, $\mathcal{R}_{n_{\text{out}}}$ were set to $\mathcal{R} + 20°, \mathcal{R} + 100°$.

TABLE I
SETTING OF EACH DATASET

Dataset	the number of		the kind of	
	Target Sources	Noise Sources	Target Sources	Noise Sources
Dataset 1	1	1	Speech	Background
Dataset 2	2	2	Speech	Background
Dataset 3	1	1	Speech	speech
Dataset 4	2	2	Speech	speech

Six separation methods were compared, including the proposed LCMV-SS-BF (LCMV-SS), the Minimum Variance Distortionless Response beamformer (MVDR-BF, MVDR), the LCMV-BF (LCMV), the SS-BF (SS), Independent Low-Rank Matrix Analysis (ILRMA) [13], [14], and Neural Network based Generalized Eigenvalue Beamformer (NN-GEV) [1]. The first four methods, including the proposed method, are beamformer-based approaches. MVDR-BF is the generally used beamforming method for point source extraction. ILRMA is a kind of blind source separation method, which separates the mixture of sound without sound source or system information. It specifies the number of sound sources in the sound mixture and separates the sound mixture into the specified number of sound sources based on the sound source model. When used as RSE, the number of sound sources in the sound mixture must be known, and it is further necessary to determine whether the separated sound sources are contained within the target region. NN-GEV is a mask-based method with a neural network. It generates masks for the target and noise sources using neural network and calculates the GEV beamformer [25], [26] using the spatial correlation matrix computed from them.

For evaluating the computation time of LCMV-BF, three methods, that is, i) the proposed method (RLS2) that applies the RLS method to $\boldsymbol{R}^{-1}$ and $(\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}$, ii) the Block-wise method (Block) that calculates the inverse matrix, and iii) RLS1 that applies the RLS method only to $\boldsymbol{R}^{-1}$ were compared in terms of the processing speed for computing one LCMV sub-BF.

The signal-to-distortion ratio (SDR) [27] and the short-time objective intelligibility measure (STOI) [28] were used as the objective and subjective metric of sound source extraction performance, respectively. The real-time factor (RTF) was used to evaluate the performance of the processing speed, and it was measured for computing a single LCMV sub-BF. The measurements were performed using MATLAB on a PC equipped with an Intel(R) Core(TM) i7-1185G7 @ 3.00 GHz.

B. Parameter settings of each method

For LCMV-SS-BF, Δ, was set to $10°$, and a uniform weight vector was used for $\boldsymbol{B}$. For MVDR-BF, the target direction was set to the central direction of the target region, that is, $(\theta, \phi) = (0°, 0°)$, because only a single direction can be specified in the target region. For LCMV-BF, the constraint points are given at the four edges of the target region. For SS-BF, Δ was set to $4°$ and MVDR-BF oriented toward the center of each sub-region was used as the sub-BF. For ILRMA, the number of sound sources must be given in advance. Therefore, sound source separation was first performed with the number of sound sources set equal to or lower than the actual number of sound sources. Although it is also necessary to determine whether the sound sources after separation are included in the target region, this time we ignored this. And selected the one with the best separation performance among the combinations of the sums of the separated sound sources as the result. For NN-GEV, the neural network was trained to estimate a mask that separates speech from speech mixed with environmental noise. This condition is favorable for Datasets 1 and 2, while unfavorable for Datasets 3 and 4. This method is used for comparison but is not an method for RSE because it depends on the types of sound sources.

For the calculation methods, RLS1 was initialized with $\boldsymbol{R}_0^{-1} = \boldsymbol{I} \times 10^5$ and $\alpha = 0.995$. In RLS2, the first 100 frames are calculated with RLS1, and then $\boldsymbol{R}^{-1}$ and $(\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}$ are initialized using the values at that time, and start estimation.

C. Results and discussions

TABLE II and III show the average SDR and STOI for each method, respectively. TABLE IV shows the average RTF per LCMV for each calculation method. The results of the method with the best average performance in the same dataset

TABLE II
COMPARISON IN SDR [DB]

Dataset	$\mathcal{R}$	Method					
		LCMV-SS (proposed)	MVDR	LCMV	SS	ILRMA	NN-GEV
Dataset 1	20°	10.51 ± 3.91	2.73 ± 3.98	8.44 ± 3.91	2.94 ± 2.40	**12.45 ± 7.03**	9.01 ± 3.13
	40°	9.99 ± 3.37	−0.81 ± 4.52	4.62 ± 4.54	0.41 ± 2.93	**13.74 ± 6.53**	9.05 ± 3.15
	60°	8.96 ± 2.96	−1.81 ± 4.49	0.55 ± 4.19	−1.15 ± 3.63	**13.60 ± 7.15**	9.23 ± 3.20
Dataset 2	20°	**9.85 ± 3.69**	3.22 ± 3.02	8.03 ± 3.68	4.32 ± 2.06	−0.74 ± 4.61	6.24 ± 3.38
	40°	**10.04 ± 2.86**	0.44 ± 2.84	4.98 ± 3.77	2.36 ± 2.36	−0.75 ± 4.54	5.27 ± 3.42
	60°	**8.60 ± 2.39**	−0.57 ± 4.00	0.97 ± 3.90	1.09 ± 2.84	−0.80 ± 4.55	3.98 ± 3.47
Dataset 3	20°	**13.32 ± 3.84**	4.45 ± 3.76	10.31 ± 4.46	4.60 ± 1.69	•10.80 ± 10.61	−3.48 ± 6.78
	40°	•12.66 ± 2.86	0.92 ± 4.19	7.37 ± 4.34	1.57 ± 1.93	**13.30 ± 9.95**	−2.68 ± 6.76
	60°	•10.33 ± 3.37	−2.14 ± 4.14	3.54 ± 4.72	−0.83 ± 2.99	**11.42 ± 10.42**	−3.09 ± 6.77
Dataset 4	20°	**10.40 ± 3.69**	4.76 ± 3.39	8.70 ± 3.38	6.31 ± 2.41	2.44 ± 3.70	−5.34 ± 5.31
	40°	**10.67 ± 3.32**	1.54 ± 3.23	6.27 ± 3.42	4.59 ± 2.33	2.40 ± 4.18	−4.55 ± 4.84
	60°	**9.46 ± 3.06**	−0.16 ± 3.96	2.88 ± 3.34	3.21 ± 2.54	2.24 ± 4.13	−6.30 ± 5.88

TABLE III
COMPARISON IN STOI [%]

Dataset	$\mathcal{R}$	Method					
		LCMV-SS (proposed)	MVDR	LCMV	SS	ILRMA	NN-GEV
Dataset 1	20°	92.77 ± 7.98	87.30 ± 8.86	89.41 ± 9.08	89.91 ± 7.84	•93.23 ± 9.01	**94.11 ± 7.11**
	40°	•92.74 ± 6.90	82.54 ± 9.31	82.63 ± 10.96	88.38 ± 7.42	**95.06 ± 6.99**	•94.75 ± 6.58
	60°	91.57 ± 7.11	79.22 ± 9.94	73.36 ± 12.36	86.40 ± 8.13	•93.52 ± 9.63	**95.13 ± 5.72**
Dataset 2	20°	**95.14 ± 3.38**	82.67 ± 7.08	91.08 ± 5.29	90.47 ± 4.46	69.63 ± 12.35	•94.68 ± 4.20
	40°	**94.57 ± 2.93**	74.11 ± 7.87	83.71 ± 9.22	85.98 ± 6.29	68.60 ± 12.46	•94.19 ± 3.80
	60°	**92.45 ± 3.55**	70.90 ± 10.36	71.78 ± 12.77	84.36 ± 7.66	68.95 ± 12.00	•92.01 ± 5.16
Dataset 3	20°	**95.02 ± 4.57**	90.62 ± 4.86	91.92 ± 6.13	92.84 ± 3.48	86.18 ± 13.21	62.42 ± 19.39
	40°	**95.10 ± 2.88**	84.63 ± 8.14	88.32 ± 7.14	89.89 ± 4.61	89.77 ± 12.53	63.62 ± 19.37
	60°	**93.46 ± 3.90**	80.75 ± 8.08	81.40 ± 10.37	88.33 ± 5.20	86.98 ± 13.26	61.11 ± 19.76
Dataset 4	20°	**91.80 ± 5.79**	83.10 ± 8.90	89.64 ± 6.00	•90.25 ± 5.02	73.84 ± 10.80	55.22 ± 16.90
	40°	**91.91 ± 5.29**	73.60 ± 10.26	84.95 ± 7.50	85.90 ± 5.87	73.33 ± 11.11	54.77 ± 16.37
	60°	**90.92 ± 5.01**	69.06 ± 11.44	76.50 ± 9.48	82.37 ± 7.76	71.79 ± 12.45	50.10 ± 18.33

are shown in bold. In addition, P-tests were performed between the method that produced the best average performance (the method shown in bold) and the other methods to verify the difference in superiority between methods. This was done for each dataset and each $\mathcal{R}$. If the P value is greater than 0.01, there is no superiority difference. Results that were determined to have no superiority difference (equivalent to the best-performing method) are indicated by a black circle (•) in front of the number.

From TABLE II and III, In Dataset 2, Dataset 3, and Dataset 4, the proposed method is found to have the best performance with a superiority difference (although it was determined to be equivalent to ILRMA for SDR in Dataset 3). The result of the proposed method on Dataset 1 also showed performance that was second to the best method, although a difference in superiority occurred between it and the best method. This means that the proposed method has the best RSE performance among these methods. Dataset 1 is the most typical situation of one speaker and one background noise, where enhancement can be achieved with various source enhancement techniques that are not specific to RSE. In particular, IRLMA omits sound source localization, so experiments are conducted under favorable conditions.

In addition, what is noteworthy about the proposed method is its robustness to various situations. Observing the performance over varying $\mathcal{R}$, it is shown that the performance degradation of the proposed method from $\mathcal{R} = 20°$ to $\mathcal{R} = 60°$ is limited to 1.80 dB in SDR and 1.58 % in STOI. This is the smallest degradation among the beamforming-based methods. This indicates that the problem that RSE over a wide region was difficult with a single LCMV-BF has been greatly relaxed. For the variation in the number of sound sources in the target region, cases where one target source (Datasets 1 and 3) and two target sources (Datasets 2 and 4) are compared. The proposed method maintains consistent performance between Datasets 1, 3 and 2, 4, with an SDR difference of 1.12 dB and STOI difference of 0.65 %, regardless of the number of the target sources. This can be said to be a property of beamforming-based methods that are based on spatial models. On the other hand, IRLMA significantly deteriorated as the number of sources increased. This is because as the number of sound sources increases, more complex models for generating sound sources become necessary. To observe the performance over types of sound sources mixed, source separation of the mixture of different types (Datasets 1 and 2) and that of the same type (Datasets 3 and 4) are compared. The results showed that the performance of the proposed method was independent of the source type, with an SDR difference of 1.48 dB and STOI difference of 0.17 % between Datasets 1, 2 and 3, 4, while NN-GEV showed significant performance degradation under unfavorable conditions. These results suggest the robustness of the proposed method against the changes in the region

TABLE IV
COMPARISON IN RTF

Method	Block	RLS1	RLS2
RTF	1.77	0.73	**0.62**

size, number of sources, and types of sources. The proposed method can be said to be the most suitable method for RSE.

From TABLE IV, the proposed method showed the best RTF and achieved real-time processing. The proposed method was 2.85× and 1.18× faster than Block and RLS1, respectively. This demonstrates the effectiveness of the proposed method also in computational time.

V. CONCLUSION

This paper presented surface source separation for a wide range of continuous regions at high speed and proposed LCMV-SS-BF by integrating LCMV-BF and SS-BF with an incremental RLS method for LCMV-BF estimation. Simulation results showed that the proposed method outperformed other methods such as beamforming, blind separation, and mask-based separation using a neural network, and the robustness for the changes in region size, number of sources, and types of sources. In terms of processing speed, it worked in real time and was 2.85 times faster than a conventional method with inverse matrix calculations. Future work includes validation in a reverberation environment.

ACKNOWLEDGMENT

This work was supported by KAKENHIJP22F22769 and JP22KF0141, and F-REI (JPFR23010102).

REFERENCES

[1] J. Heymann *et al.*, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016, pp. 196–200. DOI: 10.1109/ICASSP.2016.7471664.

[2] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[3] A. A. Nair *et al.*, "Audiovisual zooming: What you see is what you hear," in *ACM Multimedia*, 2019, pp. 1107–1118.

[4] J. Benesty *et al.*, *Microphone array signal processing*. Springer Science & Business Media, 2008.

[5] H. Zhao *et al.*, "The sound of pixels," in *ECCV*, 2018, pp. 570–586.

[6] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, 2018, Article No. 112, ISSN: 0730-0301. DOI: 10.1145/3197517.3201357.

[7] Y. Ji *et al.*, "Self-supervised fine-grained cycle-separation network (fscn) for visual-audio separation," *IEEE Transactions on Multimedia*, vol. 25, pp. 5864–5876, 2023. DOI: 10.1109/TMM.2022.3200282.

[8] R. Lu *et al.*, "Listen and look: Audio–visual matching assisted speech source separation," *IEEE Signal Processing Letters*, vol. 25, no. 9, pp. 1315–1319, 2018. DOI: 10.1109/LSP.2018.2853566.

[9] Y. Gao *et al.*, "Audio compensation network: To improve the quality of low-energy audio in visual sound separation," in *ICCECE*, 2021, pp. 727–732. DOI: 10.1109/ICCECE51280.2021.9342383.

[10] O. Thiergart *et al.*, "An acoustical zoom based on informed spatial filtering," in *IWAENC*, 2014, pp. 109–113. DOI: 10.1109/IWAENC.2014.6953348.

[11] F. Rund *et al.*, "Objective quality assessment for the acoustic zoom," in *TSP*, 2015, pp. 392–396. DOI: 10.1109/TSP.2015.7296290.

[12] W. Ruochen *et al.*, "Acoustic zooming based on real-time metadata control," in *IC-NIDC*, 2014, pp. 338–342. DOI: 10.1109/ICNIDC.2014.7000321.

[13] D. Kitamura *et al.*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016. DOI: 10.1109/TASLP.2016.2577880.

[14] D. Kitamura *et al.*, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, pp. 1–25, 2018.

[15] A. Ozerov *et al.*, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *ICASSP*, 2011, pp. 257–260. DOI: 10.1109/ICASSP.2011.5946389.

[16] Z. Zhong *et al.*, "Design and assessment of a scan-and-sum beamformer for surface sound source separation," in *IEEE/SICE SII*, 2020, pp. 808–813.

[17] Z. Zhong *et al.*, "Assessment of a beamforming implementation developed for surface sound source separation," in *2021 IEEE/SICE International Symposium on System Integration (SII)*, 2021, pp. 369–374.

[18] O. Thiergart *et al.*, "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," in *ICASSP*, 2013, pp. 659–663. DOI: 10.1109/ICASSP.2013.6637730.

[19] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972. DOI: 10.1109/PROC.1972.8817.

[20] T. Chakrabarty *et al.*, "Performance investigation of robust linearly constrained minimum variance beamforming for direction of arrival mismatch," in *ICECTE*, 2019, pp. 149–152.

[21] T. Nakatani and K. Kinoshita, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer.," in *Interspeech*, 2019, pp. 111–115.

[22] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Review*, vol. 23, no. 1, pp. 53–60, 1981.

[23] V. Panayotov *et al.*, "LibriSpeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[24] J. Barker *et al.*, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *ASRU*, 2015, pp. 504–511. DOI: 10.1109/ASRU.2015.7404837.

[25] T. Kagoshima, N. Ding, and H. Fujimura, "Adaptive noise suppression for wake-word detection by temporal-difference generalized eigenvalue beamformer," in *APSIPA ASC*, 2020, pp. 805–809.

[26] C. Boeddeker, P. Hanebrink, L. Drude, J. Heymann, and R. Haeb-Umbach, "Optimizing neural-network supported acoustic beamforming by algorithmic differentiation," in *ICASSP*, 2017, pp. 171–175. DOI: 10.1109/ICASSP.2017.7952140.

[27] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006. DOI: 10.1109/TSA.2005.858005.

[28] C. H. Taal *et al.*, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011. DOI: 10.1109/TASL.2011.2114881.