INTEGRATED APPROACH OF FEATURE EXTRACTION AND SOUND SOURCE ENHANCEMENT BASED ON MAXIMIZATION OF MUTUAL INFORMATION

Yuma Koizumi[1], *Kenta Niwa*[1], *Yusuke Hioka*[2], *Kazunori Kobayashi*[1], *and Hitoshi Ohmuro*[1]

[1]: NTT Media Intelligence Laboratories, Tokyo, Japan

[2]: Department of Mechanical Engineering, University of Auckland, Auckland, New Zealand

ABSTRACT

We investigated informative acoustic feature extraction based on dimension reduction for collecting target sources on a noisy sports field. Although a Wiener filter is often used for sound source enhancement, it is difficult to accurately design the Wiener filter by simply using spatial cues because the noise on a sports field (e.g., cheering from spectators) arrives from the same direction as that of the targeted source. A statistical approach is used to estimate the Wiener filter by using pre-trained acoustic feature models. However, an *informative* acoustic feature, which provides a powerful clue for clear extraction of the target source, is unknown. For this study, we developed a method for optimizing a projection matrix for dimension reduction by maximizing the mutual information between acoustic features and the Wiener filter. Through experiments using two-directional microphones on a mock sports field, we confirmed that the proposed method outperformed previous methods in terms of both the noise reduction and quality of the recovered sound sources.

Index Terms— Microphone array, Sound source enhancement, Wiener filter, Gaussian mixture model, Mutual information

1. INTRODUCTION

Technologies providing users with immersive audio (e.g. 22.2 Multichannel Audio [1, 2], Dolby Atmos [3], SAOC [4] and MPEG-H [5]) have been extensively studied. With these technologies, such as in a free viewpoint TV [6, 7], audiences can experience as if they dived into a movie scene by flexibly and accurately controlling sound source localization. To apply this technology to real-world media (e.g., live broadcast/webcast and documentaries), sound source enhancement is required. Since a target sound source is surrounded by various interfering noise, we focused on sound recording for sports game broadcasting; thus, aimed at collecting target sources (e.g., ball sounds and/or voices of players) on a noisy sports fields.

The use of microphone arrays is a common approach for sound source enhancement (mainly speech enhancement) in noisy environments [8]. Conventional studies on microphone array techniques have been mainly focused on spatial cues, i.e., phase/amplitude differences between microphones to point the directivity beam at the sound source for isolating the target sound (i.e. beamforming). Unfortunately, these approaches require a huge number of microphones to design a sharp directivity to clearly extract a target source in a noisy environment [9, 10, 11, 12]. To boost noise reduction performance, the Wiener filter has been used as a post-filter of the beamforming in previous studies [13, 14, 15, 16].

However, since the noise on a sports field (e.g., cheering from spectators) often arrives from the same direction as that of the target source as shown in Fig. 1, it is difficult to extract the target source by using a small number of microphones [17] or a shotgun microphone [18, 19], which can be used on a sports field.

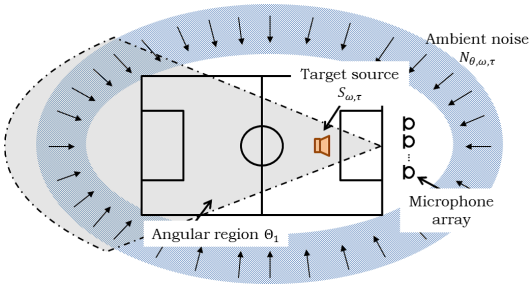

Fig. 1. Sound collection on noisy sports field based on spatial cues.

Recent studies have attempted to use pre-trained statistical models for sound source enhancement (also mainly speech enhancement) [20, 21] in which the Wiener filter is estimated from the acoustic features. Thus, *informative* acoustic features, which provide a powerful clue for clear extraction of the target source, are essential. In speech enhancement, mel-filterbank cepstrum coefficient (MFCC) and/or mel-filterbank output (MFBO) are empirically used, because a speech signal can be characterized by its spectral envelope. However, temporal sharpness is often more distinctive than the spectral envelope in a sound from typical sources observed on a sports field, e.g., kicking a ball. To support various types of target sources surrounded by ambient noise on a sports field, MFCC is not informative. Thus, a design method of informative acoustic features needs to be studied.

We previously proposed a method for automatically *selecting* acoustic features for sound source enhancement [22]. The feature selection is based on the target source detection; namely, it was assumed that discriminative acoustic features are also informative for sound source enhancement. We confirmed from experiments that the method outperformed another method that uses spatial cues of sources [15] for speech enhancement. Nevertheless, the acoustic features selected with the previous method are optimized for target source *detection* and not suitable for the Wiener filter *estimation*, which degrades the quality of the output signal. A method for designing an optimized set of acoustic features has been required for improving signal quality.

We propose a method for automatically optimizing a projection matrix to *design informative* acoustic features for collecting target sources on a sports field. This is possible by compressing a large number of acoustic feature candidates using a projection matrix. In statistical estimation, *mutual information* (MI) is commonly used to quantify the strength of dependency between two random vari-

ables since a strong dependency between input and output signals contributes to improving estimation accuracy. Thus, the projection matrix is optimized by maximizing the MI. Then the Wiener filter is calculated by modeling the relationship between the informative acoustic features and Wiener filter using a Gaussian mixture model (GMM)-based mapping function [23].

The rest of this paper is organized as follows. In Section 2, we introduce sound source enhancement using spatial cues, which the proposed method is based on. In Section 3, we describe the details of the proposed method. We explain the experimental results in Section 4 and conclude the paper with some remarks in Section 5.

2. WIENER FILTER DESIGN BASED ON POWER-SPECTRAL-DENSITY ESTIMATION IN BEAMSPACE

2.1. Sound source enhancement by Wiener filtering

Assume a problem of determining a target source surrounded by ambient noise on a sports field, as shown in Fig. 1. Sounds on the field are recorded using a microphone array with M directional microphones, such as shotgun microphones, which are often used for webcasts/broadcasts. The signal observed with the m-th microphone is expressed as

$$X_{m,\omega,\tau} = D_{m,\theta_1,\omega} S_{\omega,\tau} + \int_{\Theta} D_{m,\theta,\omega} N_{\theta,\omega,\tau} d\theta, \tag{1}$$

where $\omega = \{1, 2, ..., \Omega\}$ and $\tau = \{1, 2, ..., T\}$ denote the frequency and time indices, respectively. The term $D_{m,\theta,\omega}$ is the directivity gain of the m-th directive microphone to the angle θ, θ_1 is the location angle of the target source seen from the center of the microphone array, $S_{\omega,\tau}$ is the target source, and $N_{\theta,\omega,\tau}$ is the noise source propagating from θ (referred to as "noise"). In the following discussion, transfer functions from all sound sources to the microphones are omitted for simplicity.

The target source and all surrounding noise are assumed to be mutually uncorrelated; namely, $\mathbb{E}\left[S_{\omega,\tau} N^*_{\theta,\omega,\tau}\right]_\tau = 0$ and $\mathbb{E}\left[N_{\theta_i,\omega,\tau} N^*_{\theta_j,\omega,\tau}\right]_\tau = 0$, where $\mathbb{E}[\cdot]_\tau$ is the expectation operator for τ and * denotes the complex conjugate. The power spectral density (PSD) of the target source can be defined as $\phi_{S,\omega} = \mathbb{E}\left[|S_{\omega,\tau}|^2\right]_\tau$, and the PSD of all noise is $\phi_{N,\omega} = \int_\Theta \mathbb{E}\left[|N_{\theta,\omega,\tau}|^2\right]_\tau d\theta$. Thus, the Wiener filter is designed by

$$G_\omega = \frac{\phi_{S,\omega}}{\phi_{S,\omega} + \phi_{N,\omega}}. \tag{2}$$

In practice, to adapt to the temporal variability of the target source and noise spectra, the Wiener filter is designed frame by frame as

$$G_{\omega,\tau} \approx \frac{\phi_{S,\omega,\tau}}{\phi_{S,\omega,\tau} + \phi_{N,\omega,\tau}} = \frac{\xi_{\omega,\tau}}{1 + \xi_{\omega,\tau}}, \tag{3}$$

where $\xi_{\omega,\tau} = \phi_{S,\omega,\tau}/\phi_{N,\omega,\tau}$ is the instantaneous a priori signal-to-noise ratio (SNR), $\phi_{S,\omega,\tau} = |S_{\omega,\tau}|^2$, and $\phi_{N,\omega,\tau} = \int_\Theta |N_{\theta,\omega,\tau}|^2 d\theta$. Finally, the output signal $Y_{\omega,\tau}$ is obtained by applying the Wiener filter to one of the observed signals of the microphone array

$$Y_{\omega,\tau} = G_{\omega,\tau} X_{\omega,\tau}. \tag{4}$$

From (3) and (4), our goal for this study was achieved by estimating $\xi_{\omega,\tau}$ from the observations.

2.2. Wiener filter design by PSD estimation in beamspace

For sound source enhancement using spatial cues, we apply the *PSD estimation in beamspace* [15]. Let Θ_1 be an angular region, where the target sound source is located, and $\Theta_l, (l = 2, ..., L)$ be a set of unique $L-1$ angular regions outside Θ_1. Assume ϕ_{Θ_l} is the PSD of sound sources located within $\Theta_l (l = 1, ..., L)$ (spatial PSD), and the directivity to each angular region $|D_{m,\Theta_l,\omega}|^2$ is assumed to be constant. Then the PSD of the m-th directive microphone observation $\phi_{X_m,\omega,\tau} = |X_{m,\omega,\tau}|^2$ is rewritten as $\phi_{X_m,\omega,\tau} = \sum_{l=1}^{L} |D_{m,\Theta_l,\omega}|^2 \phi_{\Theta_l,\omega,\tau}$, where $\phi_{\Theta_1,\omega,\tau} = \phi_{S,\omega,\tau} + \int_{\Theta_1} |N_{\theta,\omega,\tau}|^2 d\theta$ and $\phi_{\Theta_l,\omega,\tau} = \int_{\Theta_l} |N_{\theta,\omega,\tau}|^2 d\theta$. These relationships are rewritten in the matrix form

$$\underbrace{\begin{bmatrix} \phi_{X_1,\omega,\tau} \\ \vdots \\ \phi_{X_M,\omega,\tau} \end{bmatrix}}_{\boldsymbol{\Phi}_{X,\omega,\tau}} = \underbrace{\begin{bmatrix} |D_{1,\theta_1,\omega}|^2 & \cdots & |D_{1,\theta_L,\omega}|^2 \\ \vdots & \ddots & \vdots \\ |D_{M,\theta_1,\omega}|^2 & \cdots & |D_{M,\theta_L,\omega}|^2 \end{bmatrix}}_{\boldsymbol{D}_\omega} \underbrace{\begin{bmatrix} \phi_{\Theta_1,\omega,\tau} \\ \vdots \\ \phi_{\Theta_L,\omega,\tau} \end{bmatrix}}_{\boldsymbol{\Phi}_{S,\omega,\tau}}. \tag{5}$$

Thus, $\boldsymbol{\Phi}_{S,\omega,\tau}$ is calculated by solving the simultaneous equation by

$$\boldsymbol{\Phi}_{S,\omega,\tau} = \boldsymbol{D}^+_\omega \boldsymbol{\Phi}_{X,\omega,\tau}, \tag{6}$$

where $^+$ denotes the pseudo inverse. The Wiener filter, which enhances the sources located within Θ_1, is calculated by

$$\tilde{G}_{\omega,\tau} = \frac{\phi_{\Theta_1,\omega,\tau}}{\sum_{l=1}^{L} \phi_{\Theta_l,\omega,\tau}} \approx \frac{\phi_{S,\omega,\tau} + \int_{\Theta_1} \mathbb{E}\left[|N_{\theta,\omega,\tau}|^2\right] d\theta}{\phi_{S,\omega,\tau} + \phi_{N,\omega,\tau}}. \tag{7}$$

As can be seen in the numerator of (7), the ideal Wiener filter cannot be obtained by only applying the PSD estimation in beamspace. In other words, the target source cannot be determined by only using spatial cues.

The noise PSD in $\phi_{\Theta_1,\omega,\tau}$ is suppressed compared to the noise PSD in $\phi_{X_m,\omega,\tau}$. This fact suggests a combination of sound source enhancement using spatial cues and feature extraction will be effective for target-source enhancement. The next section describes informative feature extraction from the spatial PSD calculated from (6).

3. PROPOSED METHOD

To collect the target source on a noisy sports field, we discuss a principle of the acoustic feature extraction for a priori SNR estimation. Informative acoustic features and the ideal prior SNR are denoted as $\boldsymbol{f}_\tau \in \mathbb{R}^{\mathsf{D}}$ and $\boldsymbol{\xi}_\tau \in \mathbb{R}^{\mathsf{B}}$, respectively. Here $\boldsymbol{\xi}_\tau$ is the prior SNR vector compressed by mel-filterbanks defined by $\boldsymbol{\xi}_\tau = (\xi^{\text{mel}}_{1,\tau}, ..., \xi^{\text{mel}}_{\mathsf{B},\tau})^{\mathsf{T}}$, where $\xi^{\text{mel}}_{b,\tau}$ is the ideal a priori SNR of the b-th mel-filterbank and $^{\mathsf{T}}$ denotes the transpose.

3.1. Acoustic feature extraction based on maximization of mutual information

In statistical estimation, $\boldsymbol{\xi}_\tau$ is estimated using the conditional probability density function (PDF) of $\boldsymbol{\xi}_\tau$ given $\boldsymbol{f}_\tau$. In such an approach, the stronger dependency between $\boldsymbol{\xi}_\tau$ and $\boldsymbol{f}_\tau$ leads to a more accurate estimation result. A criterion commonly used to describe the strength of dependency is the MI between $\boldsymbol{\xi}_\tau$ and $\boldsymbol{f}_\tau$, defined as

$$I(\boldsymbol{\xi}; \boldsymbol{f}) = \iint p(\boldsymbol{\xi}, \boldsymbol{f}) \ln \frac{p(\boldsymbol{\xi}, \boldsymbol{f})}{p(\boldsymbol{\xi}) p(\boldsymbol{f})} d\boldsymbol{\xi} d\boldsymbol{f}. \tag{8}$$

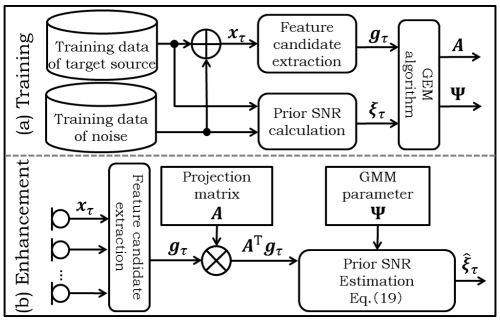

Fig. 2. Overview of the procedures in the proposed method.

Namely, the performance of sound source enhancement is increased using $\boldsymbol{f}$, which maximizes the MI.

When the informative acoustic features are unknown, a large number of potential acoustic features (e.g., MFBOs, Δ MFBOs) calculated from the spatial PSD (as in (6)) $\boldsymbol{g}_\tau \in \mathbb{R}^{\mathsf{Q}}$ is compressed to provide the informative acoustic features as $\boldsymbol{f}_\tau = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau$. This feature extraction procedure is known as dimension reduction [24, 25], and the matrix $\boldsymbol{A} : \mathbb{R}^{\mathsf{Q}} \to \mathbb{R}^{\mathsf{D}}, \mathsf{D} < \mathsf{Q}$ is called the projection matrix. If the optimization criterion of $\boldsymbol{A}$ is appropriate, the informative acoustic features are extracted from $\boldsymbol{g}_\tau$. In this paper, $I(\boldsymbol{\xi}; \boldsymbol{f})$ is maximized subject to $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} = \boldsymbol{I}_{\mathsf{D}}$ for optimizing the criterion of $\boldsymbol{A}$. Here $\boldsymbol{I}_{\mathsf{D}}$ is an identity matrix of size D.

To optimize $\boldsymbol{A}$, $I(\boldsymbol{\xi}; \boldsymbol{f})$ is reformed to an objective function $\mathcal{I}(\boldsymbol{\xi}; \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g})$, which can be maximized from a set of training data. Assuming that the function type of the joint PDF $p(\boldsymbol{\xi}, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g})$ is known, $I(\boldsymbol{\xi}; \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g})$ can be reformed approximately by replacing the expectation in (8) to the average of the training data as

$$\mathcal{I}(\boldsymbol{\xi};\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}) = \frac{1}{T}\sum_{\tau=1}^{T} \ln p(\boldsymbol{\xi}_\tau, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau) + \frac{1}{T}\sum_{\tau=1}^{T} -\ln p(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau) + C. \quad (9)$$

The first term is the average log-likelihood, the second term is the entropy of the acoustic feature, and the third term C is the entropy of the prior SNR, which is a constant value. Then $\boldsymbol{A}$ can be optimized by solving the following equation

$$\boldsymbol{A} = \arg\max_{\boldsymbol{A}} \mathcal{I}(\boldsymbol{\xi}; \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}), \quad \text{s.t. } \boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} = \boldsymbol{I}_{\mathsf{D}}. \quad (10)$$

In this maximization, the first term in (9) works to decrease the scatter of the joint vector of $\boldsymbol{\xi}_\tau$ and $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau$. On the other hand, the second term in (9) works to increase the scatter of $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau$.

Fig. 2 gives an overview of the procedures of the proposed method; (a) Training phase, $\boldsymbol{A}$ and $p(\boldsymbol{\xi}_\tau, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau)$ are optimized by maximizing $\mathcal{I}(\boldsymbol{\xi}; \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g})$; (b) sound source enhancement phase, $\boldsymbol{\xi}_\tau$ is estimated using trained $\boldsymbol{A}$ and $p(\boldsymbol{\xi}_\tau, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau)$.

3.2. Optimization for joint PDF and projection matrix

To easily calculate the conditional PDF, the joint PDF $p(\boldsymbol{\xi}, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g})$ is modeled using a GMM. The joint vector $\boldsymbol{\nu}_\tau = (\boldsymbol{\xi}_\tau^{\mathsf{T}}, (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau)^{\mathsf{T}})^{\mathsf{T}}$ and its PDF trained with the GMM are expressed by

$$p(\boldsymbol{\nu}_\tau) = p(\boldsymbol{\xi}_\tau, \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau) = \sum_{k=1}^{K} w_k \mathcal{N}\left(\boldsymbol{\nu}_\tau; \boldsymbol{\mu}_k^\nu, \boldsymbol{\Sigma}_k^\nu\right), \quad (11)$$

where K is the number of mixtures, $\mathcal{N}$ is the Gaussian distribution, and w_k is the mixing weight for the k-th Gaussian distribution. The parameters of the k-th Gaussian distribution, the mean vector $\boldsymbol{\mu}_k^\nu$ and covariance matrix $\boldsymbol{\Sigma}_k^\nu$, are respectively written as

$$\boldsymbol{\mu}_k^\nu = \begin{bmatrix} \boldsymbol{\mu}_k^\xi \\ \boldsymbol{\mu}_k^f \end{bmatrix}, \quad \boldsymbol{\Sigma}_k^\nu = \begin{bmatrix} \boldsymbol{\Sigma}_k^{\xi\xi} & \boldsymbol{\Sigma}_k^{\xi f} \\ \boldsymbol{\Sigma}_k^{f\xi} & \boldsymbol{\Sigma}_k^{ff} \end{bmatrix}. \quad (12)$$

Here $\boldsymbol{\mu}_k^\xi$ and $\boldsymbol{\mu}_k^f$ are the mean vectors of the variables $\boldsymbol{\xi}_\tau$ and $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau$, respectively. Likewise, $\boldsymbol{\Sigma}_k^{\xi\xi}$ and $\boldsymbol{\Sigma}_k^{ff}$ respectively denote the covariance matrix of $\boldsymbol{\xi}_\tau$ and $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau$.

To calculate (10), the projection matrix $\boldsymbol{A}$ and the parameters of the joint PDF $\boldsymbol{\Psi} = \{w_k, \boldsymbol{\mu}_k^\nu, \boldsymbol{\Sigma}_k^\nu\}_{k=1}^K$ are optimized simultaneously. However, the marginal distribution $p(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau)$ follows a GMM whose parameters are included in $\boldsymbol{\Psi}$. Therefore, maximization of (9) on $\boldsymbol{\Psi}$ is difficult. Accordingly, in this study, $\boldsymbol{\Psi}$ was optimized only for the joint PDF. In addition, to optimize $\boldsymbol{A}$, the steepest gradient method (SGD) is used because (10) cannot be solved analytically. Overall, the proposed simultaneous optimization is an extension of the expectation-maximization (EM) algorithm for GMM training, i.e., the generalized EM (GEM) algorithm [26], which includes a quasi-optimization for $\boldsymbol{A}$ in the M-step.

First, $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau$ is rewritten as $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau = \sum_{q=1}^{\mathsf{Q}} \boldsymbol{a}_q g_{q,\tau}$. Then, each raw vector of $\boldsymbol{A}$ (i.e. $\boldsymbol{a}_{1,\ldots,\mathsf{Q}}$) is optimized using SGD as

$$\boldsymbol{a}_q \leftarrow \boldsymbol{a}_q - \epsilon \nabla \boldsymbol{a}_q, \quad (13)$$

where ϵ is the step size and $\nabla \boldsymbol{a}_q$ is calculated as

$$\nabla \boldsymbol{a}_q = \frac{1}{T}\sum_{\tau=1}^{T} g_{q,\tau} \sum_{k=1}^{K} \gamma_{k,\tau} \left\{ \boldsymbol{d}_{\xi,k,\tau}^{\mathsf{T}} \boldsymbol{\Lambda}_k^{\xi f} + \boldsymbol{d}_{f,k,\tau}^{\mathsf{T}} \boldsymbol{\Lambda}_k^{ff} \right\} - \eta_{k,\tau} \boldsymbol{d}_{f,k,\tau}^{\mathsf{T}} \left(\boldsymbol{\Sigma}_k^{ff}\right)^{-1}, \quad (14)$$

$$\boldsymbol{d}_{\xi,k,\tau} = \boldsymbol{\xi}_\tau - \boldsymbol{\mu}_k^\xi, \quad \boldsymbol{d}_{f,k,\tau} = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau - \boldsymbol{\mu}_k^f, \quad (15)$$

$$\gamma_{k,\tau} = \frac{w_k \mathcal{N}\left(\boldsymbol{\nu}_\tau; \boldsymbol{\mu}_k^\nu, \boldsymbol{\Sigma}_k^\nu\right)}{\sum_{j=1}^{K} w_j \mathcal{N}\left(\boldsymbol{\nu}_\tau; \boldsymbol{\mu}_j^\nu, \boldsymbol{\Sigma}_j^\nu\right)}, \quad (16)$$

$$\eta_{k,\tau} = \frac{w_k \mathcal{N}\left(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau; \boldsymbol{\mu}_k^f, \boldsymbol{\Sigma}_k^{ff}\right)}{\sum_{j=1}^{K} w_j \mathcal{N}\left(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{g}_\tau; \boldsymbol{\mu}_j^f, \boldsymbol{\Sigma}_j^{ff}\right)}, \quad (17)$$

$$\left(\boldsymbol{\Sigma}_k^\nu\right)^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_k^{\xi\xi} & \boldsymbol{\Lambda}_k^{\xi f} \\ \boldsymbol{\Lambda}_k^{f\xi} & \boldsymbol{\Lambda}_k^{ff} \end{bmatrix}. \quad (18)$$

Finally, $\boldsymbol{A}$ is orthogonalized in each step by [27].

The following is a summary of the proposed method (with the GEM algorithm).

E-step :

1. Update weights $\gamma_{k,\tau}$ calculated using (16).

M-step :

1. $\boldsymbol{\Psi}$ is updated in the M-step of the EM algorithm for GMM training [26].
2. $\boldsymbol{A}$ is updated using SGD.
   2-1. Each raw vector is updated using (13).
   2-2. $\boldsymbol{A}$ is orthogonalized.
   2-3. If the update is not converged, return to 2-1.

3.3. GMM-based prior SNR estimation and Wiener filter design

The estimator of the prior SNR acquired by GMM mapping is

$$\hat{\boldsymbol{\xi}}_\tau = \sum_{k=1}^{K} \eta_{k,\tau} \left(\boldsymbol{\mu}_k^\xi + \boldsymbol{\Sigma}_k^{\xi f} \left(\boldsymbol{\Sigma}_k^{ff}\right)^{-1} \boldsymbol{d}_{f,k,\tau} \right). \quad (19)$$

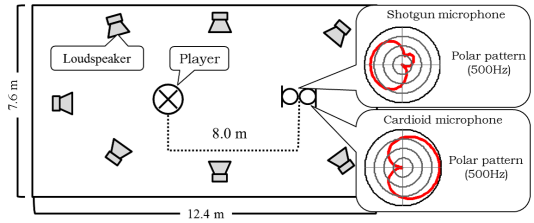

Fig. 3. Arrangement of microphone array and loudspeakers for reproducing target sound and cheering noise. The target sound source was located at x–mark.

Equation (19) is the conditional expectation of $\boldsymbol{\xi}_\tau$ from the trained GMM, and it is known as the minimum mean square error (MMSE) estimation.

Since $\hat{\boldsymbol{\xi}}_\tau$ is composed of the estimated prior SNR for each MFBO, it is transformed into the prior SNR for (linear) frequency bins by spline interpolation. The Wiener filter is then designed using (3) and the target source is extracted using (4).

4. EXPERIMENTS

4.1. Experimental conditions

Experiments were conducted on a mock sports field to evaluate the performance of the proposed method. A loudspeaker playing the target source and the microphone array were located in the middle of the field surrounded by seven loudspeakers reproducing cheering noise, which was recorded from an actual football game, as shown in Fig. 3. The target sources consisted of sound files of kicking (*football*) and hitting (*baseball*) a ball, and a shout of a goalkeeper (*shout*). The microphone array consisted of two different microphones: a shotgun microphone for creating the target beamspace and a cardioid microphone for creating the noise beamspace. The two microphones were positioned as close as possible and their directivity beams pointed at angles opposite each other. To evaluate the proposed method under various noise conditions, the noise level was adjusted to either 80, 90 or 100 dB sound pressure level (SPL) measured at the center of the microphone array.

The proposed method was compared with three conventional methods: sound source enhancement based on PSD estimation in beamspace [15] (SPC), Principal-component-analysis-based projection matrix design (PCA), and target-source-detection-based feature selection [22] (AED). The SNR defined below was used as the metric to evaluate sound source enhancement performance

$$\mathrm{SNR} = \frac{1}{|\mathcal{H}|} \sum_{\tau \in \mathcal{H}} 10 \log_{10} \frac{\sum_{\omega=1}^{\Omega} |S_{\omega,\tau}|^2}{\sum_{\omega=1}^{\Omega} \left(|S_{\omega,\tau}| - |Y_{\omega,\tau}|\right)^2} \qquad (20)$$

where $\mathcal{H}$ denotes the interval that contains the target source. In addition, the perceptual evaluation of speech quality (PESQ) [28] was used for quantitatively evaluating the quality of the output sound.

A training and test datasets were generated by adding a clean target source and a noise source that had been recorded individually at a 48-kHz sampling rate. The training dataset consisted of 300 samples (combinations of 100 target sources and 3 noise levels). The test dataset consisted of 30 samples (combinations of 10 target sources and 3 noise levels). Since the training dataset was small, the model parameters were set to small values, $\mathsf{B} = 32$, $\mathsf{D} = 24$ and $K = 12$, to avoid overfitting. The potential candidates of acoustic features were: MFCCs with 23 filter banks calculated from $\phi_{\Theta_1,\omega,\tau}$ and $\phi_{\Theta_2,\omega,\tau}$, MFBOs and its Δ with 32 filter banks calculated from $\phi_{\Theta_1,\omega,\tau}$ and $\phi_{\Theta_2,\omega,\tau}$, Spectral centroid/spread/flatness calculated from $\phi_{\Theta_1,\omega,\tau}$, and MFBOs with 32 filter banks calculated from observation of shotgun microphone. Overall, the total number of candidates was $\mathsf{Q} = 255$. The frame size of the short-term Fourier transform (STFT) was 2048 and was shifted by 1024 samples.

Table 1. Approximated MI $\mathcal{I}(\boldsymbol{\xi}; \boldsymbol{A}^\mathsf{T}\boldsymbol{f})$

	Football	Baseball	Shout
PCA	5.0	5.0	4.6
AED-based [22]	5.2	6.4	5.7
Proposed method	**5.9**	**8.5**	**8.2**

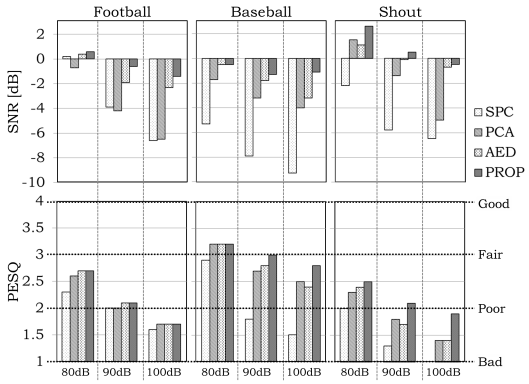

Fig. 4. Evaluation results of SNR (Top) and PESQ (Bottom).

4.2. Experimental results

Table 1 shows the approximated MI defined by (9) using different methods. The proposed method increased the approximated MIs compared to the previous methods. Fig. 4 shows the SNR and PESQ score of the target source emphasized using different methods. Overall, the proposed method exhibited higher SNR and PESQ scores compared to the other methods for different target sources and noise levels. These results suggest that the proposed method is effective in collecting target sources while maintaining the quality of the target sources even under noisy environments.

5. CONCLUSIONS

We proposed a method for automatically optimizing a projection matrix to collect target sources on a sports field. The proposed method extracts informative acoustic features by reducing the dimensions of acoustic feature candidates. A projection matrix for dimension reduction was optimized by maximizing the mutual information between the acoustic feature and Wiener filter. The Wiener filter was then designed using a GMM-based mapping function from the selected acoustic features for sound source enhancement. The experimental results on a mock sports field revealed that the proposed method outperformed previously proposed methods.

Further experiments have to be conducted on various actual outdoor sports fields for validating the practicality of the proposed method. The quality of the collected target sources should also be evaluated using subjective listening tests.

6. REFERENCES

[1] K. Hamasaki, K. Hiyama and R. Okumura, "The 22.2 Multichannel Sound System and Its Application," AES 118th Convention, 2005.

[2] K. Matsui and A. Ando, "Binaural Reproduction of 22.2 Multichannel Sound with Loudspeaker Array Frame," AES 135th Convention, 2013.

[3] C. Q. Robinson, S. Mehta and N. Tsingos, "Scalable Format and Tools to Extend the Possibilities of Cinema Audio," in *Proc. SMPTE*, pp. 1–12, 2012.

[4] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Holzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding," AES 124th Convention, 2008.

[5] J. Herre, J. Hilpert, A. Kuntz and J. Plogsties, "MPEG-H 3D Audio - The New Standard for Coding of Immersive Spatial Audio," *IEEE J. Sel. Top. Signal Process*, Vol.9, Issue 5, pp.770–779, 2015.

[6] M. Tanimoto, "FTV: Free-viewpoint Television," *Image Communication*, Vol. 27, No. 6, pp. 555-570, 2012.

[7] A. Hilton, J. Y. Guillemaut, J. Kilner, O. Grau and G. Thomas, "3D-TV Production From Conventional Cameras for Sports Broadcast," *IEEE Trans. on Broadcasting*, Vol. 57, pp.462–476, 2011.

[8] M. Brandstein, D. Ward (Eds.), "*Microphone Arrays*," Digital Signal Processing, Springer, 2001.

[9] J. L. Flanagan, J. D. Johnston, R. Zahn and G. W. Elko, "Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms," *J. Acoust. Soc. Am.*, vol. 78, pp. 1508-1518, Nov. 1985.

[10] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West, M. M. Sondhi, "Autodirective Microphone Systems," *Acta Acustica united with Acustica*, Vol. 73, No. 2, pp. 58-71, 1991.

[11] K. Kobayashi, K. Furuya, A. Kataoka, "A Talker-Tracking Microphone Array for Teleconferencing," AES 113th Convention, 2002.

[12] K. Niwa, Y. Hioka, K. Furuya and Y. Haneda, "Diffused Sensing for Sharp Directive Beamforming," *IEEE Trans. Audio, Speech and Language Processing*, Vol.21, pp.2346–2355, 2013.

[13] C. Marro, Y. Mahieux, K. U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering," *IEEE Trans. Speech, Audio Processing*, pp. 240–259, 1998.

[14] T. Wolff and M. Buck, "A Generalized view on microphone array postfilters," in *Proc. IWAENC*, 2010.

[15] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. Audio, Speech and Language Processing*, pp.1240–1250, 2013.

[16] K. Niwa, Y. Hioka and K. Kobayashi, "Post-filter design for speech enhancement in various noisy environments," in *Proc. IWAENC*, pp. 35–39, 2014.

[17] A. Farina, A. Capra, L. Chiesi and L. Scopece, "A Spherical Microphone Array for Synthesizing Virtual Directive Microphones in Live Broadcasting and in Post Production," AES 40th International Conference, 2010.

[18] H. Wittek , C. Faller, A. Favrot, C. Tournery, C. Langen, "Digitally Enhanced Shotgun Microphone with Increased Directivity," AES 129th Convention, 2010.

[19] R. Oldfield, B. Shirley and J. Spille, "Object-based Audio for Interactive Football Broadcast," *Multimedia Tools and Applications*, Vol. 74, pp.2717–2741, 2015.

[20] M. Fujimoto, S. Watanabe and T. Nakatani, "Frame-wise model re-estimation method based on Gaussian pruning with weight normalization for noise robust voice activity detection," *Speech communication*, vol. 54, pp.229–244, 2012.

[21] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance Based Integration of Spatial and Spectral Features for Speech Enhancement," *IEEE Trans. Audio, Speech and Language Processing*, Vol. 21, pp.2516–2531, Dec. 2013.

[22] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi and Hitoshi Ohmuro, "Informative Acoustic Feature Selection on Microphone Array Wiener Filtering for Collecting Target Source on Sports Ground," in *Proc. WASPAA*, 2015.

[23] Y. Stylianou, O. Cappe and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Trans. Speech, Audio Processing*, Vol.6, pp.131–142, 1998.

[24] H. Hino and N. Murata, "A conditional entropy minimization criterion for dimensionality reduction and multiple kernel learning," *Neural Computation*, vol. 22, pp.2887–2923, 2010.

[25] T. Suzuki and M. Sugiyama, "Sufficient dimension reduction via squared-loss mutual information estimation," in *Proc. AISTATS*, pp.804–811, 2010.

[26] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2007.

[27] A. Hyvarinen, et al., "Independent Component Analysis, " J. Wiley, New York, 2001.

[28] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.