INFORMATIVE ACOUSTIC FEATURE SELECTION ON MICROPHONE ARRAY WIENER FILTERING FOR COLLECTING TARGET SOURCE ON SPORTS GROUND

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 propose a Wiener filter design method for collecting target sources on a noisy sports field. Because the noise on a sports field, e.g., cheering from the audience, arrives from the same direction as that of the targeted source, it is difficult to accurately design a Wiener filter by simply using spatial cues. This study focused on a combination of spatial cues and acoustic feature modeling. The Wiener filter using our method was designed using a Gaussian-mixture-model-based mapping function with automatically selected informative acoustic features from pre-enhanced observation using spatial cues. Through experiments using two-directional microphones on a mock sports field, it was confirmed that the proposed method outperformed previous methods that used either spatial cues or acoustic feature modeling only.

Index Terms— Microphone array, Wiener filter, acoustic feature estimation, Gaussian mixture model

1. INTRODUCTION

Technologies providing users with immersive audio and images, such as free viewpoint TV, have been actively studied for webcasting/broadcasting [1, 2]. A targeted application of such technologies is recording sports games on a large field. Various studies have been conducted on image/video processing [1, 2], although there have been very few reports on recording immersive audio signals of sports games [3, 4]. The goal of this study was to collect target sources, e.g., ball sounds and/or voices of players in a noisy football game, in real time. We propose a method for collecting target sources to collect sounds that typically used to be buried in the ambient noise so that users can hear sounds that they have never experienced before.

The use of microphone arrays is a common approach for sound source enhancement in noisy environments [5]. Conventional studies on microphone array techniques have mainly focused on spatial cues, i.e., phase/amplitude differences between microphones. Sharp directivity for clearly extracting a target source can be formed using a microphone array with a huge number of microphones [6, 7, 8, 9]. With a small number of microphones, Wiener filtering has been used as a post-filter of beamforming to boost noise reduction performance [11, 12, 13, 14]. With the Wiener filter calculated from the power spectral density (PSD) estimated in beamspace [13], it has been reported that the sound sources located within about 60 degrees from the targeted angle can be segregated from other surrounding noise. However, since the noise on a sports field, e.g., cheering from the audience, often arrives from the same direction as that of the targeted sound source, as shown in Fig. 1, it is difficult to accurately design a Wiener filter by simply using the spatial cues.

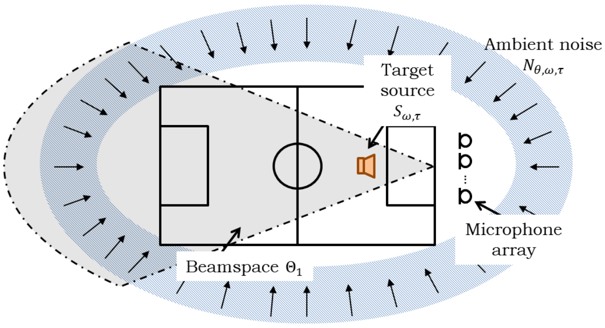

Figure 1: Sound collection on noisy sports stadium based on spatial cues.

Some recent studies have attempted to use pre-trained acoustic feature models of the target source along with microphone array signal processing [15, 16, 17, 18]. With this approach, Mel-filter cepstrum coefficients (MFCCs) or Mel-filter bank outputs (MFBOs) are used for acoustic feature modeling, and the Wiener filter is designed by calculating the similarity between the pre-trained acoustic features and the observation. However, it has been reported that MFCCs and MFBOs are not informative acoustic features for detecting/emphasizing the target source amid surrounding ambient noise [19, 20]. Hence, an alternative method is required that can effectively determine the acoustic features on a noisy sports field.

In this study, we developed a Wiener filter design method with informative acoustic features selection. This method uses a pre-enhanced signal by applying a technique based on spatial cues to select acoustic features that contain information that determines the target source from a large number of potential acoustic features. It uses some MFCCs and MFBOs or other acoustic features (e.g., ΔMFBOs) and is therefore expected to be more effective for detecting the target source than simply using full-band MFCCs and MFBOs of a pre-enhanced signal. Since the pre-enhanced signal should mainly contain the target source components, a more informative set of acoustic features for describing the target source can be selected by using this method. Once a set of informative acoustic features is selected, the model for calculating the Wiener filter is structured using, e.g., a Gaussian-mixture-model-(GMM)-based mapping function [21].

This paper is organized as follows. In section 2, sound source enhancement using spatial cues, known as the *PSD estimation in beamspace method*, is briefly explained. In Section 3, the details of the proposed method are described. The experimental results are explained in Section 4, and the paper is concluded with some remarks in Section 5.

2. WIENER FILTER DESIGN BASED ON PSD ESTIMATION IN BEAMSPACE

2.1. Wiener filtering

We assume a problem of determining a target source surrounded by ambient noise on a sports field, as shown in Fig. 1. All the sounds are observed with a microphone array with M directive microphones such as shotgun microphones often used for webcasts/broadcasts. The signal observed with the m-th microphone is expressed as

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

where ω and τ denote the frequency and time indices, respectively. The term $D_{m,\theta_n,\omega}$ is the directivity gain of the m-th directive microphone to the angleθ_n, $A_{m,\theta}$ is the transfer function from a source located at angle θ to the m-th directive microphone, $S_{\omega,\tau}$ is the target source, and $N_{\theta,\omega,\tau}$ is the noise source propagating from θ. The target source and all surrounding noise are assumed to be mutually uncorrelated; namely, $\mathbb{E}\left[S_{\omega,\tau} N^*_{\theta,\omega,\tau}\right] = 0$ and $\mathbb{E}\left[N_{\theta_i,\omega,\tau} N^*_{\theta_j,\omega,\tau}\right] = 0$, where $\mathbb{E}[\cdot]$ is the expectation operator and * denotes the complex conjugate.

The PSD of the target source can be defined as $\phi_{S,\omega,\tau} = \mathbb{E}\left[|S_{\omega,\tau}|^2\right]$, and the PSD of all noise is $\phi_{N,\omega,\tau} = \int_\Theta \mathbb{E}\left[|N_{\theta,\omega,\tau}|^2\right] d\theta$. Thus, the ideal Wiener filter is designed using

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

Output signal $Y_{\omega,\tau}$ is obtained by applying the Wiener filter to one of the observed signals

$$Y_{\omega,\tau} = G_{\omega,\tau} X_{m,\omega,\tau}. \tag{3}$$

From (2) and (3), the goal of this work was achieved by estimating $\phi_{S,\omega,\tau}$ and $\phi_{N,\omega,\tau}$ from the observations.

2.2. PSD estimation in beamspace

We use the PSD estimation in beamspace method [13, 14] to estimate the PSD of the target sound and noise. We define an angular width Θ_1 as the target beamspace, and $\Theta_l, (l = 2, ..., L)$ as a set of unique $L-1$ angular widths outside Θ_1.

Assume ϕ_{Θ_l} is the PSD of sound sources located on $\Theta_l (l = 1, ..., L)$ (spatial PSD). Then the transfer functions from the sources in each beamspace to the m-th directive microphone are the same, and the following relationships hold

$$\phi_{X_m,\omega,\tau} = \sum_{l=1}^{L} |H_{m,\theta_l,\omega}|^2 \phi_{\Theta_l,\omega,\tau}, \tag{4}$$

$$\phi_{\Theta_1,\omega,\tau} = \phi_{S,\omega,\tau} + \int_{\Theta_1} \mathbb{E}\left[|N_{\theta,\omega,\tau}|^2\right] d\theta, \tag{5}$$

$$\phi_{\Theta_l,\omega,\tau} = \int_{\Theta_l} \mathbb{E}\left[|N_{\theta,\omega,\tau}|^2\right] d\theta. \tag{6}$$

Here, $\phi_{X_m,\omega,\tau} = \mathbb{E}\left[|X_{m,\omega,\tau}|^2\right]$ is the PSD of the m-th directive microphone observation, and $H_{m,\theta,\omega} = D_{m,\theta,\omega} A_{m,\theta,\omega}$. These equations are rewritten by the matrix form

$$\underbrace{\begin{bmatrix} \phi_{X_1,\omega,\tau} \\ \vdots \\ \phi_{X_M,\omega,\tau} \end{bmatrix}}_{\boldsymbol{\Phi}_{X,\omega,\tau}} = \underbrace{\begin{bmatrix} |H_{1,\theta_1,\omega}|^2 & \cdots & |H_{1,\theta_L,\omega}|^2 \\ \vdots & \ddots & \vdots \\ |H_{M,\theta_1,\omega}|^2 & \cdots & |H_{M,\theta_L,\omega}|^2 \end{bmatrix}}_{\boldsymbol{D}_\omega} \underbrace{\begin{bmatrix} \phi_{\Theta_1,\omega,\tau} \\ \vdots \\ \phi_{\Theta_L,\omega,\tau} \end{bmatrix}}_{\boldsymbol{\Phi}_{S,\omega,\tau}}. \tag{7}$$

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

$$\boldsymbol{\Phi}_{S,\omega,\tau} = \boldsymbol{D}^+_\omega \boldsymbol{\Phi}_{X,\omega,\tau}, \tag{8}$$

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

$$\begin{aligned} \tilde{G}_{\omega,\tau} &= \frac{\phi_{\Theta_1,\omega,\tau}}{\sum_{l=1}^{L} \phi_{\Theta_l,\omega,\tau}} \\ &\approx \frac{\phi_{S,\omega,\tau} + \int_{\Theta_1} \mathbb{E}\left[|N_{\theta,\omega,\tau}|^2\right] d\theta}{\phi_{S,\omega,\tau} + \phi_{N,\omega,\tau}}. \end{aligned} \tag{9}$$

As we can see from the numerator in (9), we cannot obtain the ideal Wiener filter by only applying the PSD estimation in beamspace. Thus, the target source cannot be determined.

3. PROPOSED METHOD

3.1. Overview of procedures

The Wiener filter can also be calculated from the prior signal-to-noise ratio (SNR) given by $\xi_{\omega,\tau} = \phi_{S,\omega,\tau}/\phi_{N,\omega,\tau}$ [10, 23, 24]. With this approach, one does not need to estimate both $\phi_{S,\omega,\tau}$, and $\phi_{N,\omega,\tau}$ directly; thus, the Wiener filter is rewritten as

$$G_{\omega,\tau} = \frac{\xi_{\omega,\tau}}{1 + \xi_{\omega,\tau}}. \tag{10}$$

To estimate the prior SNR, the minimum mean square error (MMSE) estimator is used as

$$\hat{\boldsymbol{\xi}}_\tau = \mathbb{E}\left[p\left(\boldsymbol{\xi}_\tau | \boldsymbol{q}_\tau\right)\right], \tag{11}$$

where $\boldsymbol{\xi}_\tau$ is a prior SNR vector, $\hat{\boldsymbol{\xi}}_\tau$ is an estimated prior SNR vector, and $\boldsymbol{q}_\tau$ is an informative acoustic feature for detecting the target source in noisy environments. The informative acoustic feature is calculated from a pre-enhanced signal using the spatial cue $\tilde{Y}_{\omega,\tau} = \tilde{G}_{\omega,\tau} X_{\omega,\tau}$. To avoid the curse of dimensionality, the prior SNR is compressed using MFBOs,

$$\boldsymbol{\xi}_\tau = (\xi^{\text{mel}}_{1,\tau}, ..., \xi^{\text{mel}}_{F,\tau})^T, \tag{12}$$

where $\xi^{\text{mel}}_{f,\tau}$ is the prior SNR of the f-th MFBO. The conditional probability distribution function (PDF) for MMSE estimator $p\left(\boldsymbol{\xi}_\tau | \boldsymbol{q}_\tau\right)$ can be deformed using the joint PDF $p\left(\boldsymbol{\xi}_\tau, \boldsymbol{q}_\tau\right)$ [22].

To efficiently model $p\left(\boldsymbol{\xi}_\tau, \boldsymbol{q}_\tau\right)$ from finite size training data, the relationships between $\boldsymbol{\xi}_\tau$ and $\boldsymbol{q}_\tau$ should be represented simply. To this end, (11) can be reformed depending on whether the target source is active ($z_\tau = 1$) or not ($z_\tau = 0$). Since the ideal prior SNR $\xi^{\text{mel}}_{f,\tau}$ is equal to 0 when the target source is non-active ($z_\tau = 0$), (11) is reformed by

$$\begin{aligned} \hat{\boldsymbol{\xi}}_\tau &= \mathbb{E}\left[\sum_{i=0}^{1} p(z_\tau = i | \boldsymbol{q}_\tau) p\left(\boldsymbol{\xi}_\tau | \boldsymbol{q}_\tau, z_\tau = i)\right)\right], \\ &= p(z_\tau = 1 | \boldsymbol{q}_\tau) \mathbb{E}\left[p\left(\boldsymbol{\xi}_\tau | \boldsymbol{q}_\tau, z_\tau = 1)\right)\right]. \end{aligned} \tag{13}$$

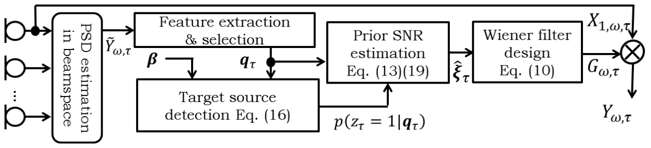
PSD estimation in beamspace
$\tilde{Y}_{\omega,\tau}$
Feature extraction & selection
$\boldsymbol{\beta}$
$\boldsymbol{q}_\tau$
Target source detection Eq. (16)
Prior SNR estimation Eq. (13)(19)
$p(z_\tau = 1|\boldsymbol{q}_\tau)$
$\hat{\boldsymbol{\xi}}_\tau$
Wiener filter design Eq. (10)
$X_{1,\omega,\tau}$
$G_{\omega,\tau}$
$Y_{\omega,\tau}$

Figure 2: Overview of proposed procedures.

Fig. 2 shows an overview of the procedures of the proposed method. To estimate $\boldsymbol{\xi}_\tau$ using (13), the following two problems need to be solved: i) selection of informative acoustic features to detect the target source accurately from noisy environments (explained in Section 3.2), and ii) modeling of the joint PDF $p\left(\boldsymbol{\xi}_\tau, \boldsymbol{q}_\tau | z_\tau = 1\right)$, which appeared in (13) (explained in Section 3.3).

3.2. Grouped LASSO logistic regression for acoustic feature selection

With the feature-selection procedure, it is assumed that a large number of potential acoustic features (e.g., MFBOs, Δ MFBOs) are listed in advance. Some of the potential acoustic features would be more informative for segregating the target source from other noise. However, it would be better to use grouped acoustic features (e.g., several MFBOs, several Δ MFBOs). Thus, the acoustic features are manually categorized into $\mathcal{G}$ groups $\boldsymbol{o}_\tau = (\boldsymbol{o}_{\tau,1}, ..., \boldsymbol{o}_{\tau,\mathcal{G}})$. We used grouped least absolute shrinkage and selection operator (LASSO) logistic regression [27] to select informative acoustic feature groups.

Logistic regression is commonly used for classifying a posteriori probability models. Based on the logistic regression, the conditional probability $p(z_\tau = 1|\boldsymbol{o}_\tau)$ can be expressed by

$$p(z_\tau = 1|\boldsymbol{o}_\tau) = \frac{1}{1 + \exp\left(-\sum_{g=1}^{\mathcal{G}} \boldsymbol{o}_{\tau,g}^T \boldsymbol{\beta}_{o,g}\right)}, \tag{14}$$

where $\boldsymbol{\beta}_{o,g}$ denotes the regression coefficient vector, which is calculated as

$$\boldsymbol{\beta}_o = \arg\min_{\boldsymbol{\beta}_o} \left\{ -\mathcal{L}(\boldsymbol{\beta}_o) + \lambda \sum_{g=1}^{\mathcal{G}} \sqrt{|g| \boldsymbol{\beta}_{o,g}^T \boldsymbol{\beta}_{o,g}} \right\}. \tag{15}$$

Here, $\mathcal{L}(\boldsymbol{\beta}_o)$, λ, and $|g|$ denote the log-likelihood function of (14), penalty parameter, and dimension of the g-th acoustic feature group, respectively. From (15), informative acoustic feature groups will be selected since the regression coefficients of uninformative features are shrunk to zero. When the selected acoustic features and corresponding regression coefficients are respectively denoted as $\boldsymbol{q}_\tau$ and $\boldsymbol{\beta}$, the conditional probability $p(z_\tau = 1|\boldsymbol{q}_\tau)$ is expressed by

$$p(z_\tau = 1|\boldsymbol{q}_\tau) = \frac{1}{1 + \exp\left(-\boldsymbol{q}_\tau^T \boldsymbol{\beta}\right)}. \tag{16}$$

3.3. GMM-based prior SNR estimation

To calculate the conditional expectation defined in (13), the joint PDF $p\left(\boldsymbol{\xi}_\tau, \boldsymbol{q}_\tau | z_\tau = 1\right)$ is modeled using a GMM. The joint vector $\boldsymbol{\nu}_\tau = (\boldsymbol{\xi}_\tau, \boldsymbol{q}_\tau)$ and its PDF trained with the GMM are expressed by

$$\boldsymbol{\nu}_\tau \sim \sum_{k=1}^{K} w_k \mathcal{N}\left(\boldsymbol{\nu}_\tau | \boldsymbol{\mu}_k^{(\nu)}, \boldsymbol{\Sigma}_k^{(\nu)}\right), \tag{17}$$

where the mean vector $\boldsymbol{\mu}_k^{(\nu)}$ and covariance matrix $\boldsymbol{\Sigma}_k^{(\nu)}$ are respectively written as

$$\boldsymbol{\mu}_k^{(\nu)} = \begin{bmatrix} \boldsymbol{\mu}_k^{(\xi)} \\ \boldsymbol{\mu}_k^{(q)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_k^{(\nu)} = \begin{bmatrix} \boldsymbol{\Sigma}_k^{(\xi\xi)} & \boldsymbol{\Sigma}_k^{(\xi q)} \\ \boldsymbol{\Sigma}_k^{(q\xi)} & \boldsymbol{\Sigma}_k^{(qq)} \end{bmatrix}. \tag{18}$$

Here $\boldsymbol{\mu}_k^{(\xi)}$ and $\boldsymbol{\mu}_k^{(q)}$ are the mean vectors of variables $\boldsymbol{\xi}_\tau$ and $\boldsymbol{q}_\tau$ that follow the k-th Gaussian. Likewise, $\boldsymbol{\Sigma}_k^{(\xi\xi)}$ and $\boldsymbol{\Sigma}_k^{(qq)}$ respectively denote the covariance matrix of $\boldsymbol{\xi}_\tau$ and $\boldsymbol{q}_\tau$.

The conditional expectation of $\boldsymbol{\xi}_\tau$ from the trained GMM is calculated as

$$\mathbb{E}\left[p\left(\boldsymbol{\xi}_\tau | \boldsymbol{q}_\tau, z_\tau = 1\right)\right] = \sum_{k=1}^{K} p(k|\boldsymbol{q}_\tau, \boldsymbol{\mu}_k^{(q)}, \boldsymbol{\Sigma}_k^{(qq)}) \boldsymbol{\Upsilon}_{k,\tau}, \tag{19}$$

where

$$p(k|\boldsymbol{q}_\tau, \boldsymbol{\mu}_k^{(q)}, \boldsymbol{\Sigma}_k^{(qq)}) = \frac{w_k \mathcal{N}\left(\boldsymbol{q}_\tau | \boldsymbol{\mu}_k^{(qq)}, \boldsymbol{\Sigma}_k^{(qq)}\right)}{\sum_{j=1}^{K} w_j \mathcal{N}\left(\boldsymbol{q}_\tau | \boldsymbol{\mu}_j^{(qq)}, \boldsymbol{\Sigma}_j^{(qq)}\right)}, \tag{20}$$

$$\boldsymbol{\Upsilon}_{k,\tau} = \boldsymbol{\mu}_k^{(\xi)} + \boldsymbol{\Sigma}_k^{(\xi q)} \left(\boldsymbol{\Sigma}_k^{(qq)}\right)^{-1} \left(\boldsymbol{q}_\tau - \boldsymbol{\mu}_k^{(q)}\right). \tag{21}$$

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 the spline interpolation.

4. EXPERIMENTS

4.1. Experimental conditions

Experiments were conducted on a mock sports field to evaluate the proposed method's performance in collecting target sources. The target sources and microphone array were surrounded by cheering noise emitted from seven loudspeakers, as shown in Fig. 3. The target sources consisted of the sound of a tee ball being batted (baseball), the sound of a ball being kicked (football), and the shout of a goalkeeper (shout). Each target source was evaluated with ten samples that were different from the training data. 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 so their directivity beams were pointing at opposite angles to each other and were as close as possible. The noise level was adjusted to 100 dB sound pressure level (SPL) at the center of the microphone array.

The proposed method was compared with two conventional methods: Source enhancement based on PSD estimation in beamspace [13] (#1(SP)) and acoustic-model-based source enhancement by extracting acoustic features (#2(GMM+LASSO)). The log spectral distortion (LSD) and noise level (NL) [10] were used as the evaluation measures. The LSD was used to evaluate the distortion of the target source, and the NL was used to evaluate noise reduction performance.

$$\mathrm{LSD} = \mathrm{Median}_{\tau \in \mathcal{H}} \left\{ \sqrt{\frac{1}{|\Omega|} \sum_{\omega} \Psi_{\omega,\tau}^2} \right\}, \tag{22}$$

$$\mathrm{NL} = \mathrm{Median}_{\tau \in \bar{\mathcal{H}}} \left\{ \sum_{\omega} 20 \log_{10} |Y_{\omega,\tau}| \right\} - C, \tag{23}$$

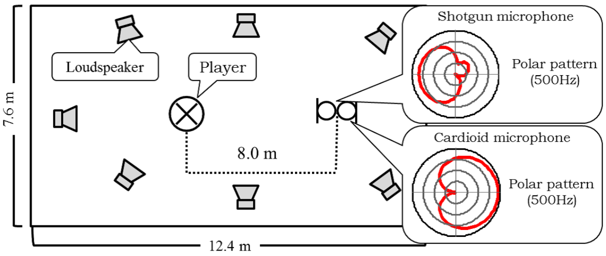

Figure 3: Arrangement of microphone array and loudspeakers for reproducing cheering noise. The athlete plays sports at the x–mark.

Table 1: Experimental conditions.

Sampling rate	48.0 kHz
FFT length	512 pts
FFT shift length	256 pts
Total dimensions of potential acoustic feature	149
Number of Groups $\mathcal{G}$	31
Number of Mel-filterbanks F	30
Number of Gaussian mixtures K	8

Table 2: List of potential acoustic features.

Acoustic feature	Groups
MFBO (target beamspace)	7
MFBO (noise beamspace)	7
ΔMFBO (target beamspace)	7
ΔMFBO (noise beamspace)	7
MFCC (target beamspace) (14 dimensions)	1
MFCC (noise beamspace) (14 dimensions)	1
Spectral entropy[28] (target beamspace)	1

where $\mathcal{H}$ denotes the interval that contains the target source, $\bar{\mathcal{H}}$ denotes the noise interval, C denotes the observed noise level, and $\Psi_{\omega,\tau} = 20\log_{10}|S_{\omega,\tau}| - 20\log_{10}|Y_{\omega,\tau}|$. A low LSD score indicates high performance of target source reproducibility and a low NL score indicates high noise reduction performance.

A training dataset for the proposed method and #2(GMM+LASSO) was generated by adding a clean target source and a noise source that had been recorded individually. The priori SNR data were also generated from these training data. Probability models were trained with 100 samples of the target source data. Penalty parameter λ was determined with cross validation. Other conditions are summarized in Table 1.

4.2. Experimental results

Fig. 4 shows the LSD and NL scores, and Figs. 5 (a)–(e) show the waveforms of the target source, observation signal, and output signals. With the proposed method, the LSD and NL scores were lower than those of the conventional methods for all targets. The NL score with #2 (GMM+LASSO) was significantly lower than #1 (SP). Modeling of the target source with PDF is thought to be effective in detecting the target source. Because acoustic features were extracted from the pre-enhanced signal, both target source detection and Wiener filter design were accurate. Thus, the results indicate that pre-enhancement based on the target source is effective for collecting target sources. The selected acoustic features for each target can be considered reasonable. For example, in the football results, a delta MFBO (target beamspace) from 4 to 16 kHz was mainly selected. From these results, we confirmed that the proposed method is effective in collecting target sources even in noisy environments.

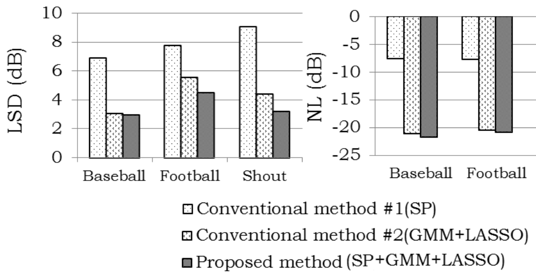

Figure 4: Experimental results.

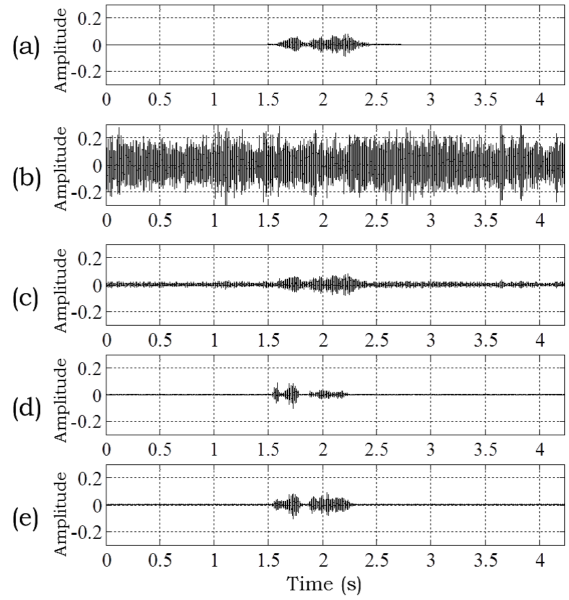

Figure 5: Waveform of (a) original target source, (b) observed signal at shotgun microphone, (c) output signal with #1 (SP), (d) output signal with #2 (GMM+LASSO), and output signal with proposed method.

5. CONCLUSIONS

We proposed a method for collecting target sounds on a noisy sports field with selecting acoustic features of a target source using grouped LASSO logistic regression. A Wiener filter is then structured using a GMM-based mapping function from the selected acoustic features. The experimental results on a mock sports field revealed that the proposed method outperformed conventional methods that used either spatial cues or acoustic features only.

Further experiments on various actual outdoor sports field are necessary to validate the practical performance of the proposed method. In addition, to evaluate quality of collected target sounds, not only quantitative evaluation but subjective evaluation should be conducted. It should also be noted that developing an automated design method of more informative acoustic features, which will prevent having to pre-define the acoustic feature candidates, is necessary.

6. REFERENCES

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

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] K. Kobayashi, K. Furuya, A. Kataoka, "A Talker-Tracking Microphone Array for Teleconferencing," *113th AES convention*

[9] K. Niwa, T. Kako, K. Kobayashi, "Microphone Array for Increasing Mutual Infomation between Sound Source and Observation Signals," in *Proc. ICASSP*, 2015.

[10] J. Benesty, S. Makino and J. Chen(Eds.), "*Speech Enhancement*," Signals and Communication Technology, Springer, 2005.

[11] 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.

[12] T. Wolff and M. Buck, "A Generalized view on microphone array postfilters," in *Proc. International Workshop on Acoustic Signal Enhancement*, 2010.

[13] 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.

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

[15] R. Chakraborty and C. Nadeu, "Sound-model-based acoustic source localization using distributed microphone arrays," in *Proc. ICASSP*, pp. 619–623, Apr. 2014.

[16] M. Souden, S. Araki, K. Kinoshita, T. Nakatani and H. Sawada, "A Multichannel MMSE-Based Framework for Speech Sources Separation and Noise Reduction," *IEEE Trans. Audio, Speech and Language Processing*, vol.21, No.9. pp. 1913 - 1928, 2013.

[17] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. ICASSP*, vol. 1, pp. 41–44, Apr. 2007.

[18] 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, No. 12, pp.2516-2531, Dec. 2013.

[19] C.V.Cotton and D.P.W. Ellis, "Spectral vs. Spectro-temporal features for Acoustic Event Detection," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011.

[20] J. Schroder, N. Moritz, M. Schadler, B. Cauchi, K. Adiloglu, J. Anemuller, S. Doclo, B. Kollmeier and S. Goetze, "On the use of spectro-temporal features for the IEEE AASP challenge ' detection and classification of acoustic scenes and events ' ," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[21] Y. Stylianou, O. Cappe and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Trans. Speech, Audio Processing*, Vol. 6, No. 2, pp.131–142 ,1998.

[22] T. Hastie, R. Tibshirani, J. Friedman, "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*," Springer, 2009.

[23] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc ICASSP*, pp.7092–7096, 2013.

[24] B. Xia and C. Bao, "Wiener Filter based Speech Enhancement with Weighted Denoising Auto-encoder and Noise Classification," *Speech Communication*, pp.13–29, 2014.

[25] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 68, pp 49–67, 2006.

[26] S. K. Shevade1 and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression," *Bioinformatics*, Vol. 19, Issue 17, pp.2246–2253, 2003.

[27] L. Meier, S. V. D. Geer and P. Buhlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 70-1, pp 53–71, 2008.

[28] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Proc EUROSPEECH*, 2001.