Takeda, R. and Komatani, K.

Paper:

Noise-Robust MUSIC-Based Sound Source Localization Using Steering Vector Transformation for Small Humanoids

Ryu Takeda and Kazunori Komatani

The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
E-mail: {rtakeda, komatani}@sanken.osaka-u.ac.jp
[Received July 20, 2016; accepted November 2, 2016]

We focus on the problem of localizing soft/weak voices recorded by small humanoid robots, such as NAO. Sound source localization (SSL) for such robots requires fast processing and noise robustness owing to the restricted resources and the internal noise close to the microphones. Multiple signal classification using generalized eigenvalue decomposition (GEVD-MUSIC) is a promising method for SSL. It achieves noise robustness by whitening robot internal noise using prior noise information. However, whitening increases the computational cost and creates a direction-dependent bias in the localization score, which degrades the localization accuracy. We have thus developed a new implementation of GEVD-MUSIC based on steering vector transformation (TSV-MUSIC). The application of a transformation equivalent to whitening to steering vectors in advance reduces the real-time computational cost of TSV-MUSIC. Moreover, normalization of the transformed vectors cancels the direction-dependent bias and improves the localization accuracy. Experiments using simulated data showed that TSV-MUSIC had the highest accuracy of the methods tested. An experiment using real recorded data showed that TSV-MUSIC outperformed GEVD-MUSIC and other MUSIC methods in terms of localization by about 4 points under low signal-to-noise-ratio conditions.

Keywords: sound source localization, MUSIC, matrix decomposition, microphone array, robot

1. Introduction

1.1. Background

Sound source localization (SSL) is the most fundamental function for robot audition [1] because it enables robots to detect sound events and recognize sound directions. This “awareness of sound” is an essential trigger of robot action. For example, once a robot recognizes a sound event, it can take steps to identify the source through additional actions (e.g., looking around, asking someone, and/or moving toward it) even if it did not initially recognize it only through sound.

We focus on the problem of localizing soft/weak voices recorded by small humanoid robots, such as NAO [2] (Fig. 1). For example, the speech of elderly people is usually weak, and some people cannot speak loudly. Since the power of soft/weak voices is usually low, such voices are easily affected by noise, making it difficult for robots to detect the sound and localize the source. Although speech features, such as the envelope of the speech power spectrum, are also affected by speaker hoarseness, breathiness, and roughness [3], we focus here on the problems caused by the low power of soft/weak voices.

The localization of soft/weak voices is more difficult for small humanoid robots than for robots with large bodies and many microphones (typically more than eight) [4, 5]. First, the signal-to-noise ratio (SNR) for human speech when there is robot internal noise is usually low because the noise sources are close to the microphones. This holds true for all robots, and one of the most practical solutions is to use many microphones and tune their positions to improve the SSL performance. Although many microphones can usually be implemented in large robots, only a few can be implemented in small robots owing to the limited space, which makes it more difficult for small robots to localize sound sources under internal noise conditions. With NAO, the SNR is less than 0 dB when someone speaks to NAO with a typical loudness from a distance of 1.0 m owing to the decay in the speech signal, the limited distances between NAO’s microphones, and the internal noise. These difficulties must be overcome to enable small humanoid robots to reliably detect the sound of a person speaking.

Multiple signal classification (MUSIC) [6] is one of the most promising and widely used high-resolution SSL methods for dealing with internal noise in robot audi-
1. This paper extends the analysis, results and discussion presented in our previous paper [10].

1.2. Related Works on Sound Source Localization

The microphone-array SSL methods used in robot audition can be divided into two types depending on how the reference steering vectors are constructed: 1) by using geometrical information and 2) by actual measurement.

SSL methods based on geometrical information uses microphone positions and sound source positions. The traditional SSL methods [11] are based on the delay-and-sum (DS) beamformer [12], interaural phase and intensity difference [13], cross-correlation [14, 15], and subspace methods [6]. Since the resolution of a DS beamformer is poor, localizing multiple sound sources is difficult. The use of a specially designed microphone arrangement [16, 17] improves the robustness against noise and reverberation. However, calculating the SVs geometrically is not realistic for humanoid robots because the shape of a robot’s body and the uncontrollable microphone arrangement cause the arriving sound signals to have a position-dependent power decay and a time delay. Reflections from the floor also affect small humanoid robots. Therefore, measurement-based SSL is a useful way to calibrate these acoustic properties.

Measurement-based SSL methods use actual measured impulse responses for reference SVs, and measurement-based MUSIC is often used in robot audition [4]. If the environmental conditions, such as reverberation time, match the measured impulse response, robust localization can be achieved in various environments. The resolution of SSL, such as the azimuth degree, obviously depends on the number of reference points used for recording the impulse responses. The computational cost for detecting sound sources from a localization score is proportional to the number of references. A five-degree resolution is often used for robots, and calculating such impulse responses is feasible.

2. Music-Based Sound Source Localization

In this section, we introduce the principle of SSL based on MUSIC using standard eigenvalue decomposition (SEVD) and other matrix decomposition techniques for noise-robust localization. The sound signals presented hereafter were analyzed using short-time Fourier transformation, and all model variables are represented in the short-time frequency domain with frame index $t$ and frequency-bin index $w$ [18, 19].

2.1. SEVD-MUSIC

The sound arrival process, from $M$ sound sources to sound signals $x_w[t] = [x_{w,1}[t], \ldots, x_{w,N}[t]]^T$ received at $N$ ($M < N$) microphones embedded in a robot, is modeled as a linear time-invariant system. The observed vector $x_w[t]$ is represented as

$$x_w[t] = \sum_{m=1}^{M} a_w(r_m)s_{w,m}[t] + n_w[t],\ldots$$

where $s_{w,m}[t]$ represents the $m$-th source sound signal and $n_w = [n_{w,1}[t], \ldots, n_{w,N}[t]]^T$ is a noise signal vector. The vector $a_w(r) = [a_{w,1}(r), \ldots, a_{w,N}(r)]^T$ is a steering vector representing the transfer function from reference sound
position \( r \) to each microphone. The time and intensity difference information for an arriving signal is expressed by the amplitude and the phase of \( a_w(r) \).

MUSIC uses the observed correlation matrix \( R_w = E[x_w[n]x_w^H[n]] \) to obtain the statistical properties needed for localization. Here, the notation \( A^H \) denotes the Hermitian transpose of matrix \( A \), and \( E[\cdot] \) denotes an expectation operator. Given the de-correlation between source signals and noise signals, \( R_w \) is represented as

\[
R_w = \sum_{m=1}^{M} \sum_{n=1}^{M} a_w(r_m)E[s_{w,m}[t]s_{w,m}^H[t]]a_w^H(r_n) + E[n_w[n]n_w^H[n]],
\]

(2)

where \( * \) represents a complex conjugate, \( S_w \) represents the correlation matrix of the source signals, and \( K_w \) represents that of the noise signals.

The linear space spanning correlation matrix \( R_w \) can be divided into two orthogonal sub-spaces: signal space \( S_w \) and noise space \( S_n \). In the white noise case, i.e., \( K_w = I \sigma^2 \), with identity matrix \( I \) and a variance \( \sigma^2 \), SEVD decomposes \( R_w \) on its Hermitian property:

\[
R_w = S_w + I \sigma^2 = E_w \Lambda_w E_w^H,
\]

(4)

where \( E_w = [e_w,1,...,e_w,N] \in \mathbb{C}^{N \times N} \) and \( \Lambda_w = \text{diag}(\lambda_{w,1},...\lambda_{w,N}) \) correspond to the eigenvectors and eigenvalues, respectively. The eigenvalues are sorted in descending order. The vector \( e_w,j \in \mathbb{C}^{N} (i = 1,...,M) \) corresponds to the basis set of signal space \( S_w \) and \( e_{w,j} \in \mathbb{C}^{N} (j = M+1,...,N) \) corresponds to that of noise space \( S_n \). These vectors mean that the steering vectors in \( S_w \) can be expressed by the linear combination of signal space bases and that they are orthogonal to the noise space bases. In other words, \( a_w^H(r_m)e_{w,j} = 0 (e_{w,j} \in S_n) \) holds over the correct sound positions \( r_m (m = 1,...,M) \).

Sound position \( r \) can be estimated on the basis of the orthogonality between basis vectors \( e_{w,j} \in S_w \) and the reference steering vectors. The reference steering vectors can be obtained from the analytical transfer function model or the measured impulse responses at discrete points \( r_k (k = 1,...,K) \). The SEVD-MUSIC estimator \( P_w(r) \) is defined as

\[
P_w(r) = \frac{1}{\sum_{i=M+1}^{N} |a_w^H(r)e_{w,i}|^2},
\]

(5)

The function takes a high value when \( r \) is a true position of source \( r_m \). For a broad-band signal, broad-band estimator \( P \) is calculated by, in a simple example, averaging \( P_w \) from a lower bin \( W_l \) to an upper bin \( W_u \). Source signals are detected and localized by thresholding and peak-picking of the estimator. Of course, there are more complex techniques [20], but such a technique is not required for our whitening problem.

2.2. GEVD-MUSIC

MUSIC using generalized eigenvalue decomposition (GEVD-MUSIC) instead of SEVD [6] is robust against noise by using prior noise information. If it is assumed that the noise signals can be obtained in advance, the effect of noise correlation can be canceled by multiplying whitening matrix \( W_w^{-1/2} \) with Eq. (3) from the left and/or the right. Here, matrix \( W_w^{-1/2} \) is denoted as \( \tilde{E}_w \tilde{\Lambda}_w^{-1/2} \) with eigenvectors \( \tilde{E}_w \in \mathbb{C}^{N \times N} \) and eigenvalues \( \tilde{\Lambda}_w^{-1/2} = \text{diag}(\tilde{\lambda}_1^{-1/2},...\tilde{\lambda}_N^{-1/2}) \) of the noise correlation matrix \( K_w \).

One of the implementation of GEVD-MUSIC uses the modified eigenvectors of \( F_w = [f_{w,1},...,f_{w,N}] \in \mathbb{C}^{N \times N} \), which are defined for EVD as

\[
W_w^{-1/2}R_wW_w^{-1} = \tilde{F}_w \Gamma_w \tilde{F}_w^H,
\]

(6)

where \( \tilde{F}_w = [f_{w,1},...,f_{w,N}] \in \mathbb{C}^{N \times N} \) and \( \Gamma_w = \text{diag}(\gamma_{w,1},...\gamma_{w,N}) \) are the eigenvector and the eigenvalue matrix of the left hand side in Eq. (6), respectively. Eq. (7) modifies the effect of whitening on eigenvectors \( \tilde{F}_w \). Here, the column vectors of \( F_w \) are usually normalized before the calculation of the MUSIC spectrum. There is another implementation using Cholesky decomposition, but this EVD-based representation is intended to help us understand the whitening effect on the score of Eq. (5).

2.3. GSVD-MUSIC with CMS

GSVD-MUSIC uses generalize singular value decomposition (GSVD) to obtain the eigenvectors used in Eq. (5) after whitening the robot internal noise [5]. It reduces the computational cost because SVD is easier to use than GEVD. The correlation matrix scaling (CMS) technique [9] is applied to reduce the effect of a mismatch between the actual and the prior noise information.

GSVD decomposes the correlation matrix to the left and right singular vectors. The left singular vectors \( U_w = [u_{w,1},...,u_{w,N}] \in \mathbb{C}^{N \times N} \) are used as eigenvectors in Eq. (5), which are defined by SVD as

\[
K_w^{-p}R_w = U_w \Sigma_w V_w^H,
\]

(7)

where \( \Sigma_w = \text{diag}(\sigma_{w,1},...\sigma_{w,N}) \) represents a singular value matrix, and \( V_w = [v_{w,1},...,v_{w,N}] \in \mathbb{C}^{N \times N} \) is the right singular vectors. \( K_w^{-p} \) is defined as \( \tilde{E}_w \tilde{\Lambda}_w^{-p} \tilde{E}_w^H \). Singular vectors satisfy \( U_w \tilde{U}_w^H = I \) and \( V_w \tilde{V}_w^H = I \) respectively. Parameter \( p \) (0 ≤ p ≤ 1) is used for controlling the noise canceling effect.

2.4. Problems to Be Solved

As mentioned above, the whitening process in GEVD-MUSIC creates two problems: 1) additional computational cost and 2) direction-dependent bias in the localization score (MUSIC spectrum). Since the whitening process is an additional calculation compared with SEVD-MUSIC, its use increases the processing time of GEVD-MUSIC. The direction-dependent bias makes it difficult
to detect and localize sound sources under low SNR conditions, and the SNR of input speech signals is usually less than 0 dB owing to robot internal noise.

There is room for improvement in the noise whitening process of GEVD and GSVD. Since whitening is applied after every calculation of correlation matrix $R_w$, the same calculation is repeated if the matrix remains virtually the same. Making this calculation more efficient would thus reduce the processing time. Whitening using the whitening matrix and EVD in Eq. (6). For example, if correlated signals from other directions. The resulting unbalanced signals from certain directions, whereas it degrades sound would thus reduce the processing time. Whitening using $\hat{w}$ actually the same. Making this calculation more efficient virtually the same. The observation process equivalent to Eq. (1) after the degree of noise cancellation. Eq. (5) is then calculated instead of

$\hat{w}$

We can reduce the real-time computational cost by calculating the new reference steering vector $\hat{w}$ using

$\hat{w} = \mathbf{F}_w \hat{a}_w(r)$. 

The amplitude $||\hat{w}(r)||$ does not depend on the position only if the noise eigenvalues are the same, i.e., $\hat{\lambda}_{m,j} = \text{const}$. Moreover, the amplitudes of each element of $\hat{a}_w(r)$ are also a function of position $r$ owing to the power decay caused by the robot’s body. Therefore, $||\hat{w}(r)||$ changes non-linearly in accordance with position $r$ and frequency bin $w$. The amplitudes of the reference steering vectors must be normalized in accordance with Eq. (5) because they cannot be distinguished from the amplitude of the original sound sources (Eq. (10)). Without normalization, the amplitudes of the reference steering vectors directly affect the MUSIC spectrum. These un-normalized amplitudes are the source of the direction-dependent bias. Our transformed steering vectors are normalized through Eq. (5), whereas there is no procedure for direction-dependent

### Table 1. Calculation costs for frequency-bin $w$.  

<table>
<thead>
<tr>
<th>Method</th>
<th>Preprocessing</th>
<th>Real-time processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVD-MUSIC</td>
<td>$C_{\text{evd}} + O(N^2K)$</td>
<td>$-$</td>
</tr>
<tr>
<td>GEVD-MUSIC</td>
<td>$C_{\text{evd}} + O(N^3) + O(N^2K)$</td>
<td>$C_{\text{evd}}$</td>
</tr>
<tr>
<td>GSVD-MUSIC</td>
<td>$C_{\text{svd}} + O(N^2K)$</td>
<td>$C_{\text{svd}}$</td>
</tr>
<tr>
<td>TSV-MUSIC (Ours)</td>
<td>$C_{\text{svd}} + O(N^2K)$</td>
<td>$C_{\text{svd}} + O(N^2K)$</td>
</tr>
</tbody>
</table>

where $\hat{w}[t]$ is a zero-mean Gaussian noise vector that satisfies $E[\hat{w}[t]\hat{w}^H[t]] = I$. This normal form leads to the transformation of the whitening effect into a localization score, as explained in Section 3.2. The calculation costs are summarized in Table 1. $C_{\text{evd}}$, $C_{\text{gevd}}$ and $C_{\text{svd}}$ represent the calculation costs of EVD (Eq. (4)), GEVD (Eq. (6)) and SVD (Eq. (8)), respectively. The calculation cost of $\mathbf{K}^{-1/2}_w$ or $\mathbf{W}_w^{-1/2}$ is also expressed by $C_{\text{svd}}$ in the pre-processing time. $O(N^3)$ and $O(N^3)$ represent the cost of Eq. (5) and the cost of matrix multiplication in Eq. (7), respectively. Here, $K$ is the number of reference points, and $N$ is the number of microphones. If we have several patterns of prior noise information, we simply transform each steering vector and switch them in real time in accordance with each noise pattern. The results for the processing time are presented in Section 4.2.4.

### 3.2. Cancellation of Direction-Dependent Bias

Here we explain how we eliminate the direction-dependent bias in the score that degrades the localization accuracy. Although the transformation of steering vectors seems to be simply a different expression of GEVD-MUSIC, it can help us understand the direction-dependent bias of the score caused by the whitening. This bias appears when there are several noise signal sources.

The score bias is caused by the modulation of the transformed steering vectors. The new steering vector $\hat{w}(r)$ is linearly transformed using matrix $\mathbf{W}_w^{-1/2} \mathbf{F}_w \hat{a}_w(r)$. and amplitude $||\hat{w}(r)||$ obviously depends on position $r$.

$$||\hat{w}(r)|| = a_w^H(r) \hat{a}_w^{-2p} \hat{a}_w^H a_w(r) \ldots (11)$$

The amplitudes of the reference steering vectors must be normalized in accordance with Eq. (5) because they cannot be distinguished from the amplitude of the original sound sources (Eq. (10)). Without normalization, the amplitudes of the reference steering vectors directly affect the MUSIC spectrum. These un-normalized amplitudes are the source of the direction-dependent bias. Our transformed steering vectors are normalized through Eq. (5), whereas there is no procedure for direction-dependent
normalization in the original GEVD-MUSIC. Therefore, the MUSIC spectrum of our method differs from that of GEVD-MUSIC. Since the amplitudes increase as SNR decreases, their localization accuracies differ, especially at low SNR.

Figures 2 and 3 show the power spectra and log-scale eigenvalues of the four-channel noise signals of NAO. The horizontal axis denotes the time, and the vertical axis denotes the frequency in Hertz. The reference noise was a recording of NAO’s internal noise signal. The reference steering vectors were calculated on the basis of NAO’s four-channel impulse response. They were normalized before the transformation, and the vertical axis denotes the azimuth of the reference point in degrees, and the horizontal axis denotes the frequency after the transformation. Since the amplitudes of white noise are the same as \( \sigma^2 \) according to Eq. (4), the calculation of the MUSIC spectrum is mathematically equivalent to that for GEVD-MUSIC because the eigenvalues mean that there were several internal noise sources because an eigenvalue represents the intensity of the power of an internal noise source. If the eigenvalues of white noise are the same as \( \sigma^2 \) according to Eq. (4). The right singular vectors in Eq. (8) as GSVD-L, GSVD-R and that with left singular vectors as GSVD-R.

3.3. Modified GSVD-MUSIC: GSVD-R

We propose using the right singular vectors in Eq. (8) for eigenvectors for MUSIC spectrum calculation instead of the left ones [5], which are used in GSVD-MUSIC. The steering vector transformation gives us another interpretation of GSVD-MUSIC, one that enables us to improve the performance. We call our proposed GSVD-MUSIC with right singular vectors as GSVD-R and that with left singular vectors as GSVD-L.

Since GSVD-R is more robust against the bias caused by the whitening process in Eq. (8) than GSVD-L, GSVD-R can utilize the orthogonality between reference steering vectors \( \mathbf{a}_w \) and the singular vectors in the noise space. With GSVD-L, there is a risk of getting an incorrect localization score. This can be understood by considering the simplest case. We assume that the observed signals include one dominant sound source signal and no noise signals, and change the form of Eq. (8):

\[
\mathbf{K}_w^{-1} \mathbf{R}_w = \mathbf{K}_w^{-1} \mathbf{a}_w(\mathbf{r}_1) \mathbf{E}[|s_{w,1}|^2] \mathbf{a}_w^H(\mathbf{r}_1), \quad \text{(14)}
\]

\[
= \tilde{\mathbf{a}}_w(\mathbf{r}_1) \mathbf{E}[|s_{w,1}|^2] \tilde{\mathbf{a}}_w^H(\mathbf{r}_1), \quad \text{(15)}
\]

\[
= \mathbf{U}_w \Sigma_w \mathbf{V}_w^H, \quad \text{and (16)}
\]

where \( \tilde{\mathbf{a}}_w(\mathbf{r}_1) = \mathbf{K}_w^{-1} \mathbf{a}_w(\mathbf{r}_1) \) is a transformed steering vector. One of the solutions to the singular value decomposition problem in this simple case is represented by

\[
\mathbf{U}_w = \frac{\tilde{\mathbf{a}}_w(\mathbf{r}_1)}{||\tilde{\mathbf{a}}_w(\mathbf{r}_1)||}, \quad \text{and (17)}
\]

\[\Sigma_w = ||\tilde{\mathbf{a}}_w(\mathbf{r}_1)|| E[|s_{w,1}|^2] ||\mathbf{a}_w(\mathbf{r}_1)||, \quad \text{and (18)}
\]

\[
\mathbf{V}_w = \frac{\mathbf{a}_w(\mathbf{r}_1)}{||\mathbf{a}_w(\mathbf{r}_1)||}, \quad \text{and (19)}
\]

If de-correlated noise signals exist and their powers are small relative to the dominant sound source (high SNR), the right singular vectors will consist of \( \mathbf{a}_w(\mathbf{r}_1) \) and other orthogonal vectors that satisfy \( \mathbf{V}_w \mathbf{V}_w^H = \mathbf{I} \). The left sin-
ular vectors will also consist of \( \tilde{a}_w(r_1) \) and other orthogonal vectors that satisfy \( U_w U_w^H = I \). Thus, GSVD-R calculates a more correct localization score than GSVD-L because some right singular vectors are orthogonal to \( a_w(r_1) \). However, the accuracy of the orthogonality depends on the SNR between the sound source signals and the noise signals.

The advantage of GSVD-R compared with other forms of MUSIC is the good balance between computational cost and localization accuracy. The computational cost of GSVD is usually lower than that of GEVD, and the localization accuracy is between those of GSVD-L and GEVD-MUSIC. The actual performance of GSVD-R under low SNR conditions remains to be investigated experimentally.

3.4. Analysis of Prior Noise Sensitivity

The sensitivity to mismatched noise correlation is important because the exact noise correlation matrix \( K_w \) may be difficult to obtain for practical use. Of course, we can apply a method based on updating \( K_w \), such as incremental GSVD-MUSIC [9] or joint estimation of the source position and correlation matrix [21], for further improvement.

The main factor affecting sensitivity is the power difference between pre-measured noise \( A_w \) and the actual noise since the whitening error is directly amplified. The worst amplified case is when the power of the pre-measured noise is small, whereas that of the actual one is large. Since GSVD-L may match this pattern, its performance is affected by a change in noise. In contrast, GEVD-MUSIC uses steering vectors scaled by \( \tilde{A}_w^{-p} \); therefore, the whitening error is partially canceled. We can control parameter \( p \) to reduce the sensitivity. Therefore, GEVD-MUSIC is more robust than GSVD-L-MUSIC against a mismatch.

4. Experiments Using Simulated Data

We evaluated the performance of MUSIC-based SSL by using simulated data and the impulse responses of NAO under various SNR conditions. The accuracy of sound detection and sound localization was evaluated because these two factors are important for awareness by sound. The MUSIC spectra of EVD-, GEVD-, and TSV-MUSIC were compared to determine the effect of direction-dependent bias. We first explain the experimental setup and then present and discuss the results.

4.1. Experimental Setup

The data were generated by using impulse responses recorded in our laboratory. Such responses can be used to reconstruct the acoustic properties, such as reflections. The internal noises were recorded and added to the synthesized signals.

4.1.1. Recording Conditions and Speech Data

The impulse responses for the reference steering vectors were recorded at 16 kHz in a reverberant room with an RT20 of about 640 ms, where RT20 is the reverberation time. The room size was 3.77 m \( \times \) 6.77 m \( \times \) 2.57 m (depth \( \times \) width \( \times \) height). These dimensions are illustrated in the upper part of Fig. 5.

We denote the loudspeaker positions as (distance [cm], height [cm]). The combinations of patterns for recording impulse responses were set to P1(30, 30), P2(90, 30), and P3(90, 90) to take into account various situations in which a person might talk to a robot from different distances and heights. The resolution of the directional angle was 5° (72 directions), as shown in the lower part of Fig. 5. There were 216 recorded impulse response positions in total.

The speech data for the evaluation came from two male and two female speakers in the ASJ-JNAS corpora.\(^2\) The number of utterances was three on average, and the content consisted of phonetically balanced sentences. The utterance length ranged from 3 to 10 s. Two types of internal noise, with and without servo noise, were recorded for NAO and added to the simulated speech data in advance. Therefore, the input data for SSL included one speech signal and one recorded internal noise signal. The noises without the servo for calculating \( K_w \) were captured at different times. The noise signal without the servo was almost stationary as it was mainly caused by NAO’s fan whereas the noise signal with the servo was more dynamic. The constructed test set included speech signals for a combination of 36 directions (10° intervals) \( \times \) 3 positions \( \times \) speaker patterns.

4.1.2. Signal Analysis Parameters

The parameters for short-time Fourier transformation were the same for all methods: the Hamming window

size was 256 points (16 ms) and the window shift was 80 points (5 ms).

The same MUSIC parameters (Table 2) were used for all methods: the block size for calculating $R_w$ was 100 (500 ms), and the bandwidth $[w_l, w_h]$ for averaging $P_v$ was [570, 3000] Hz. We first used scale parameter $p = 0.1, 0.2, 0.3, 0.4, 0.5$ for TSV-MUSIC and $p = 0.1$ to 1.0 with a step size of 0.1 for SVD-MUSIC. We then used the parameter setting that maximized the result for each method. The assumed number of sound sources was one. We implemented each method by using the Octave software. We found the optimum threshold parameters for sound detection by using a grid search.

4.1.3. Evaluation Criteria and Conditions

We compared the performance in terms of the F-measure for 1) sound detection (sound event does or does not exist) and 2) estimated direction (true direction or not) at the block level (500 ms) and also in terms of the mean-absolute-error (MAE) in azimuth degrees. The F-measure was defined as the harmonic mean of precision and recall.

The F-measure values were calculated using the block-level reference calculated using SEVD-MUSIC for the same speech data without noise. The F-measure for sound detection was judged by whether the sound event was detected or not, not by the estimated position. The F-measure for the estimated direction was judged by whether the estimated direction was correct or not; the estimated distance and height of the sound source were ignored. The MAE was calculated by using the blocks for which sound was detected.

We compared the performance of SEVD, GEVD, GSVD-L, GSVD-R, and TSV-MUSIC under matched- and mismatched-noise conditions for a reference noise correlation matrix with nine different SNRs (from −40 to 40 dB) averaged over four microphones and source positions. The SNRs were different at each microphone due to NAO’s body structure and the noise source positions. The noise eigen without the servo was used as the reference noise.

The processing time for each method with different numbers of microphones was compared to determine the computational efficiency. Since NAO has only four microphones, this experiment used impulse responses recorded with a linear array. Noise signals for disturbance and reference were simulated using a similar eigenvalue distribution of the actual internal noise.

4.2. Experimental Results

We first show the F-measure with the optimum threshold for the matched-noise condition shown in Fig. 6 and then show it for the mismatched-noise condition shown in Fig. 7.

4.2.1. Results for Sound Detection

As shown in the left graph in Figs. 6 and 7, the F-measure for sound detection for the matched-noise condition differed between the five methods. TSV-MUSIC had the best performance (by 0.1 point) at −20 and −10 dB. The cancellation of the biased score improved the detection rate of TSV-MUSIC, whereas GEVD failed to detect sound sources for some blocks. SEVD-MUSIC had the worst performance owing to the lack of prior noise information. The differences in the F-measures were larger for the mismatched-noise condition. The performance of GEVD-MUSIC was worse for the mismatched-noise condition than for the matched-noise condition and was similar to that of SEVD-MUSIC for both conditions. Some of the results of the mismatched-noise condition, such as those for an SNR of −40 dB, were better than those for the matched-noise condition. The mismatch in such cases might have occasionally enhanced the localization scores. Although GSVD-L-MUSIC did not work correctly, its performance was slightly better than that of SEVD-MUSIC. Its incorrect performance was caused by the unequal eigenvalues, as illustrated in Fig. 3. Whereas GEVD-MUSIC modified the eigenvectors in accordance with the noise eigenvalues $A_n$, GSVD-L has no such mechanism and, thus, amplified the effect of the noise eigenvalues. The noise-mismatched information apparently made it worse. In previous studies, the eigenvalues of the noise might have become almost the same; therefore, the performance was good. The F-measure of GSVD-R-MUSIC was better than those of GSVD-L and GEVD for the mismatched-noise condition, indicating that GSVD-R-MUSIC has better detection ability than GSVD-L-MUSIC.

4.2.2. Results for Sound Localization and MAE

As shown in the middle graph in Figs. 6 and 7, the F-measure for sound localization had the same tendency as that for sound detection, with only slight differences between the methods. Note that only the directional angle, not the distance or height, was considered in this evaluation. GEVD-MUSIC had the best performance in the high SNR area for both the matched- and the mismatched-noise conditions.
There was a clear difference in the F-measure for the matched-noise condition between TSV-MUSIC and the other methods in the low SNR area, with a difference of more than 0.05 point between GEVD-MUSIC and TSV-MUSIC. The difference was even larger for the mismatched-noise condition. However, there was a difference of more than 0.15 points in the F-measure for TSV-MUSIC between detection and localization. This means that sounds were detected for many blocks but not localized correctly by TSV-MUSIC.

As shown in the right graph in Figs. 6 and 7, for all methods for both conditions, the error increased as the SNR decreased. Although the MAEs for TSV-MUSIC were almost the same as those for GEVD-MUSIC and GSVD-R-MUSIC for the matched-noise condition, they were worse for the mismatched-noise condition. This was due to the performance gap of TSV-MUSIC between sound detection and sound localization, as mentioned above. The MAE is usually smaller when the precision of sound detection and localization is higher, not when the recall of sound detection is higher. Since the size of the MAE can be controlled by selecting the appropriate threshold parameters, a large MAE does not necessarily mean poor localization accuracy. Sequential filtering could be used to model the sound activity and MUSIC spectrum patterns and thereby further improve the accuracy.

The F-measures for sound detection and localization for an MAE of 5.0 and an SNR of −10 dB are summarized in Table 3 for the five methods. These values matched the actual conditions for SSL. The threshold parameters were selected so that the given MAE was achieved. The F-measure for TSV-MUSIC was the highest for both detection and localization for both conditions. The improvement in the F-measure for detection means that TSV-MUSIC found sound sources that GEVD did not find. The reason the F-measure for GSVD-L-MUSIC was low is that the MAE of 5.0 is a strict condition for GSVD-L-MUSIC.

### Table 3. F-measures for detection and localization with an MAE of 5.0 and at SNR of −10 dB.

<table>
<thead>
<tr>
<th>Type</th>
<th>Prior noise</th>
<th>GEVD</th>
<th>GSVD-L (p = 0.1)</th>
<th>GSVD-R (p = 0.9)</th>
<th>TSV (p = 0.4)</th>
<th>TSV (p = 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>detection</td>
<td>matched</td>
<td>0.82</td>
<td>0.07</td>
<td>0.72</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>mis-matched</td>
<td>0.76</td>
<td>0.12</td>
<td>0.61</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>localization</td>
<td>matched</td>
<td>0.69</td>
<td>0.04</td>
<td>0.55</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>mis-matched</td>
<td>0.64</td>
<td>0.08</td>
<td>0.45</td>
<td>0.78</td>
<td>0.74</td>
</tr>
</tbody>
</table>

4.2.3. Results for MUSIC Spectrum

Figure 8 shows the spectra for SEVD-, GEVD-, and TSV-MUSIC in decibels. The horizontal axis denotes the number of blocks, and the vertical axis denotes the azimuth IDs: IDs 0–71 are for the azimuths at 30 cm of distance and 30 cm of height, 72–143 are for the azimuths at 90 cm of distance and 30 cm of height, and 144–215 are for the azimuths at 90 cm of distance and 90 cm of height.
The contiguous azimuth IDs corresponded to the azimuth angles from 0° to 355° with a step size of 5°. For example, IDs 0, 72 and 144 represented 0°, and IDs 1, 73 and 145 represent 5° (see also Fig. 5). The spectrum on the left is for SEVD-MUSIC at an SNR of 20 dB. The other spectra (from the left) are for SEVD-MUSIC, GEVD-MUSIC, and TSV-MUSIC at an SNR of −20 dB. A female utterance was used to calculate these spectra, and four-channel impulse responses at 90 cm at 90 cm of distance and height were used to simulate room reverberations.

Whereas the spectrum for SEVD-MUSIC at an SNR of 20 dB was clear, that at an SNR of −20 dB had false or ambiguous peaks. The spectrum for SEVD-MUSIC had obviously false peaks at ID 215 or 143. While some peaks in the GEVD-MUSIC spectrum matched true sound position ID, there were false peaks in the no-sound-source region. These false peaks, which were caused by the direction-dependent bias, made it difficult to judge sound existence at any threshold. The spectrum for TSV-MUSIC had no direction-dependent bias and had some peaks at the true sound position.

Some false peaks were also found in the spectra for SEVD-MUSIC at an SNR of 20 dB and for TSV-MUSIC at an SNR of −20 dB. The peaks in the P1 and P2 regions of the SEVD-MUSIC spectrum were not strange because the time differences (steering vectors) of an arriving sound from the same directions are usually similar to each other. We speculated that the false peaks in the P1 region of the TSV-MUSIC spectrum were mainly caused by the whitening process and the fluctuation of the power of the internal noise signals. Since the envelope of the spectrum for TSV-MUSIC was partly similar to that for GEVD-MUSIC, the whitening itself affects the MUSIC spectrum and gave a higher score to some false positions. The fluctuation of noise power (outlier) sometimes generated some strong false peaks owing to the power mismatch between the prior and the observed noise signals as discussed in Section 3.4. Note that the noise correlation matrix differed in each segment of the same noise signals because of its non-stationarity.

4.2.4. Results for Processing Time

Table 4 shows the computer specifications, and Table 5 shows the processing time ratios for three methods. The ratios were calculated by dividing the processing time for each method by that for SEVD-MUSIC. Since SEVD-MUSIC was the baseline and, thus, had a ratio of 1.0, a smaller ratio meant more efficient processing.

The processing time ratios of GEVD-MUSIC were 7%–8% larger than those of SEVD-MUSIC, and those of GSVD-L/R-MUSIC were about 40% lower. The processing time of TSV-MUSIC was 1–2 points lower than that of GEVD-MUSIC because the latter requires one matrix multiplication of \( W_w \) before the localization.

5. Experiment Using Real Data

5.1. Experimental Setup

We also investigated the performance of MUSIC-based SSL and a vendor-supplied localization API implemented in NAO’s main control software (NAOqi) by using real recorded data. The sound source localization of NAOqi was based on the estimated time differences of arrival for different microphone pairs.

The configurations were almost the same as those in the experiments using simulated data. The test set dif-

![Fig. 8. Spectra of SEVD-MUSIC (left), SEVD-MUSIC (middle-left), GEVD-MUSIC (middle-right) and TSV-MUSIC (right) at an SNR of 20, −20, −20 and −20 dB, respectively.](image-url)

Table 4. Computer specifications.

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of microphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>Ubuntu 14.04.2 LTS</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel Core i5-4690 3.50 GHz</td>
</tr>
<tr>
<td>Memory</td>
<td>32 GB</td>
</tr>
<tr>
<td>Cache</td>
<td>6 MB</td>
</tr>
</tbody>
</table>

Table 5. Processing time ratios for three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of microphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEVD-MUSIC</td>
<td>1.070 1.075 1.087</td>
</tr>
<tr>
<td>GSVD-(L/R)-MUSIC</td>
<td>0.589 0.606 0.627</td>
</tr>
<tr>
<td>TSV-MUSIC (ours)</td>
<td>1.054 1.071 1.075</td>
</tr>
</tbody>
</table>

Table 6. Results for real recorded data.

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR = −5.3 dB</th>
<th>SNR = −14.2 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R (= S)$ MAE</td>
<td>$R (= S)$ MAE</td>
</tr>
<tr>
<td>NAOqi</td>
<td>0.12</td>
<td>4.83</td>
</tr>
<tr>
<td>SEVD</td>
<td>0.297</td>
<td>6.37</td>
</tr>
<tr>
<td>GEVD</td>
<td>0.434</td>
<td>4.13</td>
</tr>
<tr>
<td>GSVD-L ($p = 0.6$)</td>
<td>0.371</td>
<td>6.55</td>
</tr>
<tr>
<td>GSVD-R ($p = 1.0$)</td>
<td>0.422</td>
<td>5.10</td>
</tr>
<tr>
<td>TSV ($p = 0.4$)</td>
<td><strong>0.411</strong></td>
<td>4.58</td>
</tr>
<tr>
<td>TSV ($p = 0.5$)</td>
<td>0.439</td>
<td>4.66</td>
</tr>
</tbody>
</table>

fered in terms of the directional angle of the loudspeaker. Since the exact block-level reference was difficult to obtain, we simply counted the total number of correct blocks and measured the mean absolute errors (MAE).

The speaker positions were from 0° to 315° with a step size of 45°. There were three different combinations of distance and height, i.e., P1(30,30), P2(90,30), and P3(90,90), which were the same as in the experiments using simulated data. We used signals with two different SNRs, i.e., low and middle, corresponding to a soft voice and a normal-loudness voice, respectively. The estimated SNRs, averaged over four channels, were −5.3 and −14.2 dB, respectively. We turned NAO’s servo off and used the matched-noise correlation matrix.

The methods based on MUSIC were evaluated using criteria based on two block-level correctness ratios, i.e., $R$ and $S$, between the estimated number of directionally correct blocks and the number of active blocks in the reference.

\[
R = \frac{\text{no. of estimated directionally correct blocks}}{\text{no. of active blocks in reference}}, \quad (20)
\]

\[
S = \frac{\text{no. of estimated non-active sound blocks}}{\text{no. of non-active blocks in reference}}. \quad (21)
\]

The point where the two ratios were equal, $R = S$, was used as the criterion for evaluating MUSIC-based SSL. This point was used because $R$ represents localization accuracy and $S$ approximates localization recall, and both values are important. Therefore, we selected the threshold parameters that satisfied $R = S$.

The localization method used in NAOqi was also evaluated, and the event-level correctness ratio $R$ was used instead of the block-level ratio because NAO’s localization API gives a result for every sound period, meaning that block-level information cannot be obtained. Therefore, a direct comparison between the results of MUSIC and NAOqi is not possible, but the obvious differences can be estimated from the results. $R$ was calculated using the reliability (threshold) of the SoundTracker module, which reduced the MAE to less than 15°.

5.2. Results

Table 6 shows the results for real recorded data. NAN means that no results were obtained. TSV-MUSIC outperformed the other methods (except for MAE with an SNR of −5.3 dB), as it did for the simulated data. It achieved an $R$ of 0.375 even under low SNR conditions. Since GSVD-R-MUSIC also outperformed GSVD-L-MUSIC, there is another choice for GSVD-based MUSIC.

The performance of NAOqi was not good owing to the low SNR. Since NAOqi detects a sound event on the basis of the time differences between NAO’s microphones, speech signals masked by internal noise tend not to be detected. This means that a person speaking to NAO should speak loudly to be detected. The localization accuracy of the API cannot outperform that of MUSIC because MUSIC has the advantage of measured impulse responses.

6. Conclusion

We tackled the sound source localization (SSL) problem of soft/weak voices recorded by small humanoid robots, such as NAO. Such robots have a limited number of microphones without sufficient distance between them owing to their small body. MUSIC-based SSL is a promising method that is robust against robot internal noise owing to the whitening of the latter with prior noise information. However, whitening increases the computational cost and the direction-dependent bias of the localization score, which degrades the localization accuracy under low SNR conditions.

We thus developed another MUSIC method based on transformed steering vectors (TSV-MUSIC) as a new implementation of GEVD-MUSIC. Since a transformation equivalent to whitening is applied to the vectors in advance, the real-time computational cost of the proposed method is lower. Moreover, normalization of the transformed vectors cancels the direction-dependent bias and thereby improves the localization accuracy. Experimental results showed that TSV-MUSIC outperformed SEVD-, GSVD-, and GEVD-MUSIC in terms of the F-measures for sound detection and localization. Its processing time was less than that of GEVD-MUSIC.

Future work will focus on optimizing the thresholding of the MUSIC spectrum, reducing the gap in accuracy between detection and localization, improving the ability to adapt to time-varying noise signals, and integrating filtering techniques. Since thresholding and performance are affected by the reverberation time, the microphone arrangement, and the robot’s movement, further investigation of their effects is required. Our ultimate objective is to develop a robot that can detect sound events in any environment.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP15K16051.

References:

Takeda, R. and Komatani, K.


Name: Ryu Takeda
Affiliation: Assistant Professor, The Institute of Scientific and Industrial Research, Osaka University

Address: 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Brief Biographical History: 2011 Received Ph.D. in Informatics from Graduate School of Informatics, Kyoto University 2011-2014 Researcher, Central Research Laboratory, Hitachi, Ltd. 2014- Assistant Professor, The Institute of Scientific and Industrial Research, Osaka University

Main Works:

Membership in Academic Societies:
• The Acoustical Society of Japan (ASJ)
• The Institute of Electrical and Electronics Engineers (IEEE)
• Information Processing Society of Japan (IPSJ)

Name: Kazunori Komatani
Affiliation: Professor, The Institute of Scientific and Industrial Research, Osaka University

Address: 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Brief Biographical History: 2002-2010 Assistant Professor, Graduate School of Informatics, Kyoto University 2010-2014 Associate Professor, Graduate School of Engineering, Nagoya University 2014- Professor, The Institute of Scientific and Industrial Research, Osaka University

Main Works:

Membership in Academic Societies:
• The Japanese Society for Artificial Intelligence (JSAI)
• Information Processing Society of Japan (IPSJ)
• The Association for Natural Language Processing (NLP)
• The Institute of Electronics, Information and Communication Engineers (IEICE)