This paper addresses ego-motion noise suppression for a robot. Many ego-motion noise suppression methods use motion information such as position, velocity, and the acceleration of each joint to infer ego-motion noise. However, such inferences are not reliable, since motion information and ego-motion noise are not always correlated. We propose a new framework for ego-motion noise suppression based on single channel processing using only acoustic signals captured with a microphone. In the proposed framework, ego-motion noise features and their numbers are automatically estimated in advance from an ego-motion noise input using Infinite Non-negative Matrix Factorization (INMF), which is a non-parametric Bayesian model that does not use explicit motion information. After that, the proposed Semi-Blind INMF (SB-INMF) is applied to an input signal that consists of both the target and ego-motion noise signals. Ego-motion noise features, which are obtained with INMF, are used as inputs to the SB-INMF, and are treated as the fixed features for extracting the target signal. Finally, the target signal is extracted with SB-INMF using these newly-estimated features. The proposed framework was applied to ego-motion noise suppression on two types of humanoid robots. Experimental results showed that ego-motion noise was effectively and efficiently suppressed in terms of both signal-to-noise ratio and performance of automatic speech recognition compared to a conventional template-based ego-motion noise suppression method using motion information. Thus, the proposed method worked properly on a robot without a motion information interface.1

1. Introduction

Robot audition is one of the most essential capabilities for a robot interacting with people, and it has been studied since it was proposed in 2000 [1]. Robot audition aims to build auditory functions using a robot’s embedded microphones. Many types of robot audition systems have been reported, such as binaural approaches [2–7], microphone array based approaches [8–11], multi-modal integration [2, 12], and the use of ubiquitous sensors [13–15]. Sound source localization, separation and Automatic Speech Recognition (ASR) are primary research topics in robot audition. A common issue with these functions is ego-motion noise suppression. A robot should be able to interact with people even when it is performing tasks such as manipulating objects, dancing, etc. In such cases, however, ego-motion noise is inevitably generated, and a robot’s auditory functions easily deteriorate.

Since ego-motion noise is an essential problem in auditory processing for robots, several studies have been performed on ego-noise suppression. These studies can be classified into three approaches. The first approach uses additional sensors to obtain ego-motion noise, and the second uses joint status, which is considered to be correlated with ego-motion noise. The last approach estimates ego-noise using only recorded acoustic signals captured by microphones, which are simply attached to arbitrary positions on a robot’s body (e.g., head).

For the first approach, Nakadai et al. proposed ego-noise suppression by introducing the concepts of auditory embodiment [1]. They used two pairs of microphones; one pair was located at the ear positions of a robot cover (external microphones), and the other was located inside the cover (internal microphones). By comparing the noise recorded by these two pairs of microphones, a robot can distinguish its internal and external worlds. In short, a robot can identify if a sound originates inside or outside itself. When a sound originates inside the robot, the robot can ignore such a noisy period by treating them as ego-motion noise. Indeed, that approach was effective for sound source localization, but was difficult to apply to problems of sound source separation and speech enhancement, because these functions require a continuous signal.

Keywords: robot audition, ego-noise suppression, non-parametric Bayesian

Even et al. applied Frequency-Domain Blind Signal Separation (FD-BSS) to internal noise estimation by installing additional sensors inside a robot primarily to detect internal noise [16]. A multichannel Wiener filter was performed to enhance the target speech signal. This was applied to the hands-free spoken dialog system for the robot. However, this method requires the use of sensors in addition to microphones, and the study did not discuss the placement of these sensors, which would be crucial to the method’s performance. In the end, in this approach, additional microphones or sensors are necessary for obtaining ego-motion noise, which makes the overall system more complicated and results in a high computational cost.

The second approach is based on the fact that ego-motion noise is generated by joint movement, and that ego-motion noise is thus strongly correlated with joint status. For instance, Ito et al. introduced an Artificial Neural Network (ANN) estimating ego-motion noise model [17]. They first observed joint positions and angles, and then fed this information to an ANN to train an ego-motion noise model. After that, they used the trained ANN to estimate ego-motion noise using a set of joint positions and angles observed in a frame-by-frame manner. Finally, the estimated ego-motion noise was subtracted from the noisy input signal using conventional Spectral Subtraction (SS) [18]. They demonstrated the effectiveness of their method, although synthesized data was used as an input.

Nishimura et al. reported command-based ego-motion noise estimation [3]. They prepared an ego-motion noise template corresponding to a motion command. Ego-motion noise was then queried from that template according to a motion command, and the estimated ego-motion noise was subtracted from SS [17]. They also applied missing feature theory [19] to deal with a distorted speech signal using SS in ASR. They demonstrated that the proposed method worked well on a real humanoid robot, even though it is difficult to handle non-prepared motion and more complicated motion (like a combination of multiple motions) using their motion-command-based method.

Ince et al. have reported extensive work on ego-motion noise estimation for robots. They proposed using frame-based templates, and ego-motion noise estimation was performed on a frame-by-frame basis by combining templates according to the observed joint status such as angle, angular velocity and angular acceleration. They confirmed that their method improves the three main functions of robot audition, i.e., sound source localization, sound source separation and ASR [20] by combining the proposed method with microphone array processing. They applied their method to an interactive musical robot, which can simultaneously dance according to musical beats detected with its own microphones and answer a user’s spoken questions under ego-motion and musical noise conditions [21]. However, they assumed that the same motion generates the same ego-motion noise at all times. In a real situation, this assumption does not hold, and the error associated with ego-motion noise estimation causes severe distortion after SS, which drastically degrades ASR performance.

The second approach worked properly compared to conventional noise suppression techniques such as microphone array processing and echo cancellation, because joint status was used as prior information. Ego-motion noise may be correlated with joint status information to some extent, but in most cases, it is not. This means that ego-motion noise inferences based on joint status information inevitably involve error, which usually causes performance degradation in post processing such as ASR. Since the second approach is to some extent based on unreliable information, it would be difficult to further improve the performance of this approach. Significant further work would be required to realize ego-noise suppression for active audition, which utilizes active motion to achieve better perception.

The third approach tries to solve the problems of the second approach. Deleforge et al. have reported ego-noise cancellation based on a similar concept using K-Singular Value Decomposition (K-SVD), which is a clustering algorithm known as generalized K-means clustering [22]. This method estimates ego-noise without using preprocessed multi-channel acoustic signals, by taking phase information into account. However, the method requires information about the number of clusters, $K$, which cannot be obtained in advance.

In this study, we propose an ego-motion noise suppression method that requires a single microphone and no additional sensors, that does not rely on problematic joint status information, and that automatically estimates the number of clusters, using Semi-Blind Infinite Non-negative Matrix Factorization (SB-INMF) [23]. This method has basically three advantages: 1) The original SB-INMF algorithm, NMF, is known as a single channel sound source separation method, and our method can thus be applied to a robot with only a single microphone, 2) Since INMF is an extended version of NMF based on a Non-Parametric Bayesian (NPB) model, it can automatically and directly estimate the number of latent features included in ego-noise from the captured ego-noise, and 3) SB-INMF uses INMF in a semi-blind way to separate ego-motion noise and target signals with a linear process, and it produces fewer distortions compared to conventional SS-based methods. We also show that the proposed method achieves flexible and accurate ego-motion noise estimation using two types of robots with sound source separation and ASR performance.

The rest of this paper is organized as follows: Section 2 formulates an ego-motion noise model and proposes its estimation algorithm SB-INMF based on a non-parametric Bayesian model. Section 3 describes a prototype ego-motion noise suppression system based on SB-INMF. Section 4 evaluates and discusses the system. The last section concludes this paper with insights into our future work.
2. Semi-Blind Non-Negative Matrix Factorization for Ego-Motion Noise Suppression

Table 1 describes the notations used in this paper. Ego-motion noise is mainly caused by moving joints, and it is natural to consider ego-motion noise to be represented by a combination of the noise generated by all moving joints. In such cases, the Latent Feature Model (LFM) explains data by adding multiple features selected from finite feature candidates as shown in Fig. 1, where $X$, $Z$, and $F$ are $N_s \times N_f$, $N_k \times N_f$, and $N_k \times N_f$ data, $N_s \times N_k$ activation, and $N_k \times N_f$ feature matrices, respectively.

In signal processing, Blind Source Separation (BSS) [11, 24] is a common approach for solving this type of decomposition problem. BSS is used for sound source separation, and is based on statistical information and microphone array processing. Recently, much attention has been drawn to Non-negative Matrix Factorization (NMF), which can separate sound sources even when the input signal is monaural [25]. Since the ego-motion noise of each joint has a non-negative power, it is natural to use NMF to represent the whole ego-motion noise.

In particular, exponential distributions are used to separate an acoustic signal with NMF [26].

$$x_{nd} \sim \text{Exponential} \left( \sum_k z_{nk} f_{kd} \right), \quad \ldots \ldots \quad (1)$$

where $x_{nd} \geq 0$, $z_{nk} \geq 0$, and $f_{kd} \geq 0$ are elements of $X$, $Z$, and $F$, respectively. Indices $n$, $d$, and $k$ represent the sample, frequency bin, and latent feature shown in Table 1, respectively.

When we use NMF to estimate LFM for ego-motion noise estimation, we have to consider two issues; ego-motion noise and joint status information are not fully correlated, and the number of features obtained by NMF for ego-motion noise may not correspond to that of the joints, i.e., the appropriate number of features is unknown. To deal with these issues, we introduce an extension of LFM using an NPB approach called the Infinite Latent Feature Model (ILFM). This approach allows LFM to have an unbounded number of feature candidates, and automatically estimates the most likely number of feature candidates needed to represent the data. In this paper, ILFM is modeled using a Gamma process, which is a non-parametric stochastic process. ILFM is estimated using an Indian Buffet Process (IBP), which is a non-parametric stochastic process based on a beta process that defines a probability distribution over multiple classes of sparse binary matrices with a finite number of rows $N$ and an unbounded number of columns $K$ [27]. NMF is also extended so that it can manage ILFM, in a process referred to as Infinite Non-negative Matrix Factorization (INMF) [28]. INMF for ego-motion noise is formulated by

$$f_{kd} \sim \text{Gamma}(a, a), \quad \ldots \ldots \quad (2)$$

$$z_{nk} \sim \text{Gamma}(b, b), \quad \ldots \ldots \quad (3)$$

$$\theta_k \sim \text{Gamma} \left( \frac{\alpha}{K}, ac \right), \quad \ldots \ldots \quad (4)$$

$$x_{nd} \sim \text{Exponential} \left( \sum_k \theta_k z_{nk} f_{kd} \right), \quad \ldots \ldots \quad (5)$$

where $a$, $b$, and $c$ are parameters for Gamma distributions, $\theta_k$ is an element of $\theta$, and $K$ indicates the truncation level.

Since this model assumes a prior distribution based on a Gamma process for each element in a non-negative matrix, it is also called Gamma Process Non-negative Matrix Factorization (GaP-NMF). Fig. 2 illustrates a graphical model of INMF. The white and dark circles show latent and observed variables. A black dot indicates manually-specified parameters; multiple nodes can be bundled using a plate. This model simultaneously estimates $f$, $z$, and $\theta$ by assuming Eqs. (2)–(5) when ego-motion noise $x$ is observed.

Ego-motion noise can be modeled with INMF as long as an input signal consists only of ego-motion noise.
However, in reality, the input signal may contain a target signal such as speech. When such a signal is included, the features obtained by INMF are affected – features including both ego-motion noise and the target signal are obtained, and/or it is difficult to know which of the obtained features correspond to ego-motion noise. To solve this problem, we propose Semi-Blind INMF (SB-INMF). SB-INMF uses a mixture of ego-motion noise and target signals as an input.

\[
\begin{align*}
x \sim & \text{ Exponential} \left( \sum_k \theta_k z_{ak} f_{kd} + \sum_l \tilde{\theta}_l z_{nl} \tilde{f}_{ld} \right),
\end{align*}
\]

(6)

where \( \theta_k, z_{ak}, \) and \( f_{kd} \) correspond to ego-motion noise, and \( \tilde{\theta}_l, z_{nl}, \) and \( \tilde{f}_{ld} \) correspond to the target signal.

The ego-motion noise feature \( f_{kd} \) obtained with INMF is used as a given feature, and target signal features (\( \tilde{\theta}_l, \tilde{z}_{nl}, \) and \( \tilde{f}_{ld} \)), as well as other ego-motion noise features (\( \theta_k \) and \( z_{ak} \)) are estimated. The noise-suppressed target signal is obtained by multiplying \( z_{nl} \) and \( \tilde{f}_{ld} \). Because ego-motion noise is suppressed without using non-linear processing such as spectral subtraction [18], it produces fewer distortions. For estimation with INMF, we use the variational Bayesian method described in [29].

3. Ego-Motion Noise Suppression System

Figure 3 depicts the system architecture for ego-motion noise estimation and suppression based on SB-INMF. The method involves two phases: training and noise suppression. The upper panel of Fig. 3 shows the training phase used to obtain the ego-motion noise features. Ego-motion noise is captured with a microphone. After a Short-Time Fourier Transform (STFT) is performed, \( X_n \) (with \( D \) frequency bins and \( N_n \) time frames) is obtained. INMF is performed for \( X_n \), and \( K_n \) ego-motion noise features are obtained as the trained model \( F_n \). The lower panel shows the noise suppression phase. The input speech is contaminated with ego-motion noise. It is transformed into \( X_{s+n} \) with STFT. SB-INMF is performed for inference of speech in the input. In SB-INMF, the trained ego-motion noise model \( F_n \) is used as a set of given features. Since \( X_{s+n} \) includes a speech signal, an additional feature set \( F_s \) which corresponds to speech is obtained. The noise-suppressed speech is reconstructed by multiplying the obtained \( F_s \) with its activation matrix \( Z_s \). Ego-motion noise is simultaneously estimated using the given feature set \( F_n \) and the obtained activation matrix \( Z_n \).

The training phase uses all training data to estimate ego-motion noise features; that is, it is off-line processing. The noise suppression phase performs SB-INMF for every 100 frames, because the input speech signal may be dynamically changing, which affects the latent features used for the speech signal.

In this study, we used a 16 kHz sampling rate, and performed STFT with a 256-sample hamming window and 160-sample window shift. Generally, a larger STFT window size results in a higher-resolution power spectrogram. This is helpful for obtaining more precise latent features. Additionally, the degree of overlapping frequency components is reduced, because of the increase in spectral sparseness. This means that when a larger window size is used, the performance of the proposed method is expected to improve. There is one side effect: the temporal resolution of the spectrogram decrease for a larger window. Therefore, we need to investigate this parameter carefully, which would be interesting to do in future work.

4. Evaluation

We evaluated the proposed method as follows:

1. ego-motion noise suppression for hyper parameters in SB-INMF
2. ego-motion noise estimation
3. ego-motion noise suppression for synthesized data
4. ego-motion noise suppression for recorded data
5. ASR performance

The purpose of experiment 1 is to select optimal parameters. We performed ego-noise suppression by changing parameters \( a \) and \( b \) in Eq. (2) within \( 10^{-4} \)–\( 10^{0} \) for both Hearbo and Robovie-W. Experiment 2 validates the assumption that ego-motion noise can be modeled in the training phase. Thus, we did not add any sound source to the input other than ego-motion noise. Experiments 3 and 4 validate the ability of the proposed ego-motion noise suppression system to work properly for simulated and real data. The final experiment validates the effectiveness of the proposed method for ASR.

4.1. Humanoid Robots

We used two types of humanoid robots Hearbo and Robovie-W. Hearbo includes 34 Degrees-of-Freedom (DoFs), as shown in Fig. 4. We mounted an 8 ch microphone array on top of the robot’s head and used the forehead microphone for recording. In this robot, the status of each joint can be obtained synchronously with audio information using Robot Operating System (ROS). Robovie-W is a commercially-available small robot with 17 DoFs. For audio recording, we placed a hat with a microphone on the robot’s head. Although eight microphones are attached to the hat, in this study we used just a single microphone, which is marked as a circle in Fig. 5a). We can send commands to control this robot, but it is not possible to obtain information about the robot’s joint status. This means that, in the case of Robovie-W, motion information cannot be used for ego-motion noise suppression.

4.2. Data Preparation

For Hearbo, we selected five joints from the right arm (shoulder pitch (J1), shoulder roll (J2), arm yaw (J3), elbow pitch (J4), and forearm roll (J5)), and recorded the following ego-motion noise data. The recording was conducted in a \( 4 \times 7 \times 3 \) m room with 0.2 s of reverberation time. Hearbo was located at the center of the room, and a loudspeaker was located 1.2 m in front of the robot.

D1: Each joint was moved separately in turn for approximately 20 s (using a back-and-forth motion). In total, 110 s of ego-motion noise was recorded.

D2: J1 was moved for 40 s. During the last 20 s, speech was also played from the loudspeaker.

D3: Two joints (J1 and J2) were moved simultaneously for 40 s. During the last 20 s, speech was also played from the loudspeaker.

D4: Three joints (J1, J2, and J3) were moved simultaneously for 40 s, and speech was played for the last 20 s.

D5: Four joints (J1–J4) were moved simultaneously for 40 s, and speech was played for the last 20 s.

D6: Five joints (J1–J5) were moved simultaneously for 40 s, and speech was played for the last 20 s.

D7: Speech was played while joints remained still.

In every experiment, D1 was used as training data, i.e., ego-motion noise features for D1 were estimated with INMF. In experiments 1 and 5, we used signals generated by adding D1 to D7 for test data. For experiment 2, D1 was also used for test data. For experiment 3, we added a clean speech signal to D1. For the clean speech, we used the same signal played from the loudspeaker in D2–D6. For experiment 4, D2–D6 were used for test data. For both synthesized and recorded data, the Signal-to-Noise Ratio (SNR) was around 10 dB.

Robovie-W has five joints: waist yaw, shoulder roll (L and R) and elbow roll (L and R). The robot was put on a table located in a room of \( 10 \times 10 \times 3 \) m in size with 0.7 s of the reverberation time.

R1: The five joints were randomly and simultaneously moved for five seconds.

R2: A person spoke to the robot at a 2 m distance while performing the same motion as in R1. SNR was around 4 dB.

---

In each experiment, R1 was used as training data. A signal comprising a mix of R1 and a clean speech signal was used as test data for experiments 1, 3 and 5. R1 and R2 were used for experiments 2 and 4, respectively. Table 2 shows the relationship between the training and test data used in the experiments.

In experiment 5, we used Kaldi for ASR. Kaldi is a hybrid ASR system based on Deep Neural Network (DNN) and Hidden Markov Model (HMM). We trained a base acoustic model using a Corpus of Spontaneous Japanese (CSJ) recipe. However, instead of using CSJ, we used noise-contaminated speech data using the ATR phonetically balanced word corpus (25 persons × 216 words), and the ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS), which includes around 60 h of speech data. We performed additional training for the base acoustic model to allow it to each of four conditions (described as follows). We synthesized noise-contaminated data by adding the ATR phonetically balanced word corpus to D1 for Hearbo and R1 for Robovie-W. The synthesized data were used for additional training to evaluate baseline performance (that is, performance without the proposed method). For the synthesized data, we performed SB-INMF. The noise-suppressed data was used for additional training to evaluate the proposed method. For test data, another 3,240 utterances of the ATR phonetically balanced word corpus (15 persons × 216 words) were selected, which means that the test was word-closed and speaker-open. The language model was trained for isolated word recognition with a 216 word vocabulary.

### 4.3. Metrics

In experiment 1 and 3, we compute SNR, Signal-to-Inference Ratio (SIR), and Signal-to-Distortion Ratio (SDR) which are defined by Eqs. (7), (9), and (10), respectively.

SNR is defined as the SNR improvement calculated by subtracting the SNR of the original signal from the SNR of the noise-refined signal.

$$ SNR = 20 \log_{10} \left( \frac{\sum_{f} \sum_{\omega} |X(\omega, f)|^2}{\sum_{f} \sum_{\omega} |N(\omega, f)|^2} \right) - \left( \frac{\sum_{f} \sum_{\omega} |X(\omega, f)|^2}{\sum_{f} \sum_{\omega} |N(\omega, f)|^2} \right) $$

where $X$ and $N$ are the clean speech signal and pure noise signal, respectively. $f$ and $\omega$ are the frame index and frequency bin, respectively, and $\hat{N}$ is the estimated noise. SIR is a metric for assessing the distortion caused by interference such as ego-motion and environmental noise. SDR is used to assess the total distortion caused by linear, non-linear and interference noise. Consider that the estimated target speech $\hat{x}(t)$ is decomposed, as

$$ \hat{x}(t) = x_{\text{true}}(t) + e_{\text{noise}}(t) + e_{\text{arbg}}(t), $$

where $x_{\text{true}}(t)$ represents speech components, $e_{\text{noise}}(t)$ represents ego-motion noise components, and $e_{\text{arbg}}(t)$ represents other components. SIR, and SDR are defined as Eqs. (9) and (10), respectively.

$$ SIR = 10 \log_{10} \left( \frac{||x_{\text{true}}(t)||^2}{||e_{\text{noise}}(t)||^2} \right), $$

$$ SDR = 10 \log_{10} \left( \frac{||x_{\text{true}}(t)||^2}{||e_{\text{noise}}(t) + e_{\text{arbg}}(t)||^2} \right) .$$

For SIR and SDR, the MATLAB toolbox “BSS Eval” was used.

In experiment 2, we used a Noise Estimation Error (NEE) defined by Eq. (11).

$$ NEE = 20 \log_{10} \left( \frac{\sum_{f} \sum_{\omega} |N(\omega, f)| - |\hat{N}(\omega, f)|^2}{\sum_{f} \sum_{\omega} |N(\omega, f)|^2} \right) $$
Eq. (12).

\[
\text{SNR}_2 = 20 \log_{10} \left( \frac{\sum_f \sum_\omega |\hat{X}(\omega, f)|^2}{\sum_f \sum_\omega |Y_N(\omega, f)|^2} \right) = -20 \log_{10} \left( \frac{\sum_f \sum_\omega |Y_S(\omega, f)|^2}{\sum_f \sum_\omega |Y_N(\omega, f)|^2} \right), \ldots (12)
\]

where \(Y_S\) and \(Y_N\) are the signal and noise components of the input signal, respectively. \(\hat{X}\) is the estimated signal. Since \(\hat{X}\) is unavailable with a template-based method used for comparison, it is replaced with \(|Y_S(\omega, f)| - |N(\omega, f)|\).

In experiment 5, we simply used a Word Correct Rate (WCR) defined as

\[
WCR = \frac{C}{N} \times 100, \ldots \ldots \ldots \ldots \ldots . \ldots . \ldots (13)
\]

where \(N\) is the total number of words, and \(C\) is the number of successfully recognized words.

For comparison, we used a template-based method proposed by Ince et al. [30] for Hearbo, since such a method is known as one of the best methods for dealing with ego-motion noise (even though it uses joint status information). For template database generation, \(D1\) was used in experiments 2 and 3, and the first quarter (i.e., around 10 s) was used for each of \(D2\)–\(D6\) in experiment 4. However, because of a lack of motion information, this process cannot be used for Robovie-W.

4.4. Results

Figures 6 and 7 show the results. Note that both axes use a log scale. The results from experiments using Robovie-W indicate that \(a\) and \(b\) affect SNR, SIR, and SDR. SIR and SDR have the same tendencies, while SNR has unique characteristics. The results from experiments using Hearbo indicate similar values of SNR and SIR as observed in Robovie-W experiments: SDR, however, had uniform characteristics compared to SNR and SIR. Since \(D7\) includes environmental noise in addition to a speech signal, and because the estimated speech includes less environmental noise (because of suppression), SDR did not improve even for high values of SIR. Therefore, for later experiments, we decided to use \(a = 10^{-2}\) and \(b = 10^{-2}\) for Robovie-W, and \(a = 10^{-1.5}\), and \(b = 10^{-1.5}\) for Hearbo.

Table 3 shows the results for experiments 2 and 3. The number of features in the proposed method corresponds to that in the templates for template-based ego-motion noise estimation. Even when seven features for Hearbo and five features for Robovie-W were used for ego-motion noise with the proposed method, NEE performed better than a template-based method with over 300 templates, which demonstrates that the proposed method models ego-motion noise more effectively. Since five joints were used for each robot in this experiment, it is obvious that joints and features do not have a one-to-one correspondence, which suggests that it is difficult to model ego-motion noise using only joint status information.
Table 3. Ego-motion noise estimation and suppression for experiments 2 and 3.

| Robot | Method | Hearbo | | | Robovie-W |
|-------|--------|--------|---|---|---|---|
|       |        | Proposed | Template-based | Proposed |        |        |
|       | # of feat. | /templ. | 7 (noise) | 2 (speech) | 1,022 | 3,115 | 8,431 | 5 (noise) | 3 (speech) |
|       | NEE [dB] |        | -9.4 | -8.0 | -8.2 | -9.9 | -12.3 | -24.6 | -9.3 |
|       | SNR1 [dB] |        | 13.0 | 1.4 | 2.6 | 3.0 | 2.3 | 2.2 | 17.0 |
|       | SIR [dB] |        | 1.3 | -0.8 | 1.1 | 1.1 | 1.3 | 0.8 | 0.8 |
|       | SDR [dB] |        | 9.3 | 8.0 | 12.3 | 8.0 | 8.0 | 8.0 | 8.0 |

In experiment 3, two Hearbo features were additionally estimated as speech features. In total, only nine features were necessary for modeling the input signal. In SNR1, the proposed method demonstrated equivalent performance with a template-based noise suppression method using 1,022 templates. In terms of SIR and SDR, the proposed method performed better than the template-based method regardless of the number of templates. The larger number of features or templates leads to a high computational cost; a method that requires only a small number of features or templates leads to a high computational cost. Therefore, the proposed method improved SNR2 and showed better performance than the template-based method. These results show that the proposed method properly estimates ego-motion noise and speech features.

Table 4 shows the results of experiment 4. We selected the best SNR2 scores by changing the number of features and templates; “# of templates” means the number of templates for the SNR2 with the best score. In every case, the proposed method improved SNR2 and showed better performance than the template-based method. This demonstrates the robustness of our method since it is effective even in open tests (meaning that test data is completely

Fig. 8. Hearbo noise suppression results.

Fig. 9. Robovie-W noise suppression results.
Table 4. Ego-motion noise suppression results for experiment 4.

<table>
<thead>
<tr>
<th>Robot</th>
<th>Dataset</th>
<th>Proposed</th>
<th>Template-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SNR₂ [dB]</td>
<td>SNR₂ [dB]</td>
</tr>
<tr>
<td>Hearbo</td>
<td>D2 (J1)</td>
<td>5.9</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>D3 (J1 + J2)</td>
<td>5.2</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>D4 (J1–J3)</td>
<td>6.3</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>D5 (J1–J4)</td>
<td>3.5</td>
<td>−0.76</td>
</tr>
<tr>
<td></td>
<td>D6 (J1–J5)</td>
<td>5.3</td>
<td>2.4</td>
</tr>
<tr>
<td>Robovie-W</td>
<td>R2</td>
<td>3.9</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Fig. 10. Results of experiment 5.

The proposed method also has a problem. The introduction of a non-parametric approach causes the search space for feature estimation to become large; this sometimes results in a local solution. Future work will investigate methods for minimizing this problem, e.g., by using annealing methods.

5. Conclusion

We presented a new ego-motion noise suppression method that uses a single microphone and does not use any joint status information or additional sensors. The proposed method first estimates ego-motion noise features using INMF and then obtains a noise-suppressed target signal using SB-INMF. The method uses the estimated ego-motion noise features as given features. Since the method does not require non-linear processing, the noise-suppressed signal is less distorted than that produced by SS methods. The effectiveness of the proposed method was shown using two types of humanoid robots. Future work will include improving the stability of the proposed method, and validating the proposed method against SB-NMF, and NMF. Also, online and real-time processing will be important to investigate, as they were not substantially addressed by our MATLAB implementation.

Acknowledgements

This research was partially supported by Grant-in-Aid for Scientific Research No. 24118702, 24220006, 16H02884, 16K00294, and ImPACT Tough Robotics Challenge.

References:


Name: Kazuhiro Nakadai

Affiliation: Honda Research Institute Japan Co., Ltd.
Tokyo Institute of Technology

Address: 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan 2-12-1-W30 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Brief Biographical History:
1995 Received M.E. from The University of Tokyo 1995-1999 Engineer, Nippon Telegraph and Telephone and NTT Comware 1999-2003 Researcher, Kitano Symbiotic Systems Project, ERATO, JST 2003 Received Ph.D. from The University of Tokyo 2003-2009 Senior Researcher, Honda Research Institute Japan Co., Ltd. 2009-2012 Visiting Associate Professor, Tokyo Institute of Technology 2010- Principal Researcher, Honda Research Institute Japan Co., Ltd. 2011- Visiting Professor, Tokyo Institute of Technology 2011- Visiting Professor, Waseda University

Main Works:

Membership in Academic Societies:
• The Robotics Society of Japan (RSJ)
• The Japanese Society for Artificial Intelligence (JSAI)
• The Acoustic Society of Japan (ASA)
• Information Processing Society of Japan (IPSJ)
• Human Interface Society (HIS)
• International Speech and Communication Association (ISCA)
• The Institute of Electrical and Electronics Engineers (IEEE)
Name: Taiki Tezuka

Affiliation: Developer, Office Imaging Products System Technology Development Center, Office Imaging Products Development Group, Office Imaging Products Operations, Canon Inc.

Address: 7-5-1 Hakusan, Toride, Ibaraki 302-8501, Japan

Brief Biographical History: 2014 Received Master of Engineering from Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2014- Joined Canon Inc.


Name: Takami Yoshida

Affiliation: Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation

Address: 1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, Japan

Brief Biographical History: 2010 Received Master of Engineering from Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2013 Received Ph.D from Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2013- Corporate Research & Development Center, Toshiba Corporation