2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Ego-motion Noise Suppression for Robots Based on Semi-Blind Infinite Non-negative Matrix Factorization

Taiki Tezuka, Takami Yoshida, and Kazuhiro Nakadai

***Abstract*— This paper addresses ego-motion noise suppression for a robot. Many methods use motion information such as position, velocity and acceleration of each joint to infer ego-motion noise. However, such inference is not reliable since motion information and ego-motion noise are not at all times correlated. We propose a new framework for ego-motion noise suppression based on single channel processing without using any explicit motion information. In the proposed framework, ego-motion noise features are estimated in advance from an ego-motion noise input with Infinite Non-negative Matrix Factorization (INMF) which is a non-parametric Bayesian model. After that, the proposed Semi-Blind INMF(SB-INMF) is applied to an input signal consisting of both the target and ego-motion noise signals. The ego-motion noise features which are obtained with INMF are used as input to the SB-INMF and treated as the fixed features to extract the target signal. Finally, the target signal is extracted using newly-estimated features with SB-INMF. The proposed framework was applied to ego-motion noise suppression on two types of humanoid robots. Experimental results showed that ego-motion noise was suppressed well compared to a conventional template-based ego-motion noise suppression method using motion information, and thus it worked properly on a robot which does not have an interface to provide the robot's motion information.**

***Index Terms*— robot audition, ego-noise suppression, non-parametric Bayesian**

I. Introduction

Robot audition is one of the most essential capabilities for a robot to interact with people. It has been studied since it was proposed in 2000 [1]. It aims at building auditory functions using a robot's embedded microphones. Many kinds 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, and one common issue for these topics is ego-motion noise suppression. A robot should be able to interact with people simultaneously even when it is performing other 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, there exists several studies on ego-noise suppression. They are classified into two approaches. One is to use additional sensors to obtain ego-motion noise, and the other is the use of joint status which is regarded as correlated with ego-motion noise.

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 is located at the ear positions of a robot cover (external microphones), and the other is located inside the cover (internal microphones). By comparing these two pairs, a robot can distinguish its internal and external worlds. In short, a robot can identify if the sound originates either from the inside or outside of it. When the sound comes from inside the robot, it ignores such noisy time frames by treating them as ego-motion parts. Indeed, that approach was effective for sound source localization, but difficult to apply to sound source separation and speech enhancement because these functions require a continuous signal. Even et al. applied *Frequency-Domain Blind Signal Separation (FD-BSS)* to internal noise estimation by installing additional sensors inside a robot mainly to detect internal noise [16]. A multichannel Wiener filter was performed to enhance a target speech signal. This was applied to a hands-free spoken dialog system for a robot. However, their method requires additional sensors in addition to microphones, and their placement which would be crucial in terms of performance was not discussed. In the end, in this approach, additional microphones or sensors are necessary to obtain ego-motion noise, which makes a system more complicated, and results in expensive computational cost.

The second approach is based on the fact that ego-motion noise is generated by movements of joints, and thus, ego-motion noise is strongly correlated with joint status. For instance, Ito *et al.* introduced *Artificial Neural Network (ANN)* to estimate ego-motion noise model [17]. They first observed joint positions and angles and fed them to ANN to train an ego-motion noise model. After that, they estimated ego-motion noise with the trained ANN 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 with conventional *Spectral Subtraction (SS)* [18]. They showed the effectiveness of their methods although synthesized data was used as input.

Nishimura et al. reported command-based ego-motion noise estimation [3]. They prepared an ego-motion noise template corresponding to a motion command, and ego-

T. Tezuka, and K. Nakadai are with Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, JAPAN. tezuka@cyb.mei.titech.ac.jp

T. Yoshida was with Graduate School of Information Science and Engineering, Tokyo Institute of Technology when this work was done, and he is currently with Toshiba Corp.

K. Nakadai is with Honda Research Insititute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama 351-0114, JAPAN. nakadai@jp.honda-ri.com

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

motion noise was queried from the templates according to a motion command. The estimated ego-motion noise was subtracted with SS [17], and they also applied missing feature theory [19] to deal with a distorted speech signal with SS in *Automatic Speech Recognition (ASR)*. They showed that the proposed method worked well using a real humanoid robot, even though it is difficult to deal with a non-prepared motion and a more complicated motion like a combination of multiple motions with their motion-command-based method.

Ince et al. have reported extensive work on ego-motion noise estimation for robots. They proposed to use frame-based templates, and ego-motion noise estimation was performed in 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 in robot audition, *i.e.*, sound source localization, sound source separation and ASR [20] by combining it with microphone array processing. They applied their method to an interactive musical robot, which can dance according to musical beats detected with its own microphones, and at the same time, it can listen to a user's question and answer it under an ego-motion and music noise condition [21]. However, they assume that the same motion generates the same ego-motion noise all the time. In a real situation, this assumption does not hold, and the error of ego-motion noise estimation makes severe distortion after SS, which degrades ASR performance drastically.

The second approach worked properly compared to the conventional noise suppression techniques such as microphone array processing and echo cancellation, since joint status was used as prior information. Ego-motion noise may be correlated to some extent with joint status information, but in most cases, it is not. It means that ego-motion noise inference based on joint status information inevitably has an error, which usually causes performance degradation in post processing such as ASR. Since the second approach is based on unreliable information to some extent, it would be difficult to achieve further improvements of the performance with this approach, and there is a long way to realize ego-noise suppression for active audition which utilizes active motions for better perception.

Therefore, we propose a new ego-motion noise suppression method which requires only one microphone without using any additional sensor and which does not rely on problematic joint status information by introducing *Semi-Blind Infinite Non-negative Matrix Factorization (SB-INMF)* which is a *Non-Parametric Bayesian (NPB)* model. The methods based on joint status information cannot avoid errors caused by spectral and temporal fluctuation of ego-motion noise for one motion as discussed above. The proposed SB-INMF directly models ego-motion noise from an input signal in a non-parametric Bayesian approach to avoid such errors, and it can explicitly separate ego-motion noise and target signals with a linear process by introducing a semi-blind approach, which produces less distortions compared to conventional SS-based methods.

TABLE I
NOTATION OF VARIABLES

Meanings	Notation
Number of microphones	$N_{mic} \in \mathbb{N}$
Number of frequency bins	$N_f \in \mathbb{N}$
Number of latent features	$N_k \in \mathbb{N}$
Truncation Level	$K \in \mathbb{N}$
Sample size	$N_s \in \mathbb{N}$
Power spectrum of a captured audio signal	$Y \in \mathbb{R}_{+0}^{N_f+1}$
Power spectrum of an ego-motion noise	$X \in \mathbb{R}_{+0}^{N_f+1}$
Latent features	$F \in \mathbb{R}_{+0}^{N_f \times N_k}$
Activation of features	$Z \in \mathbb{R}_{+0}^{N_s \times N_f}$
Amplitude of features	$\theta \in \mathbb{R}_{+0}^{N_k}$

$\mathbb{R}_{+0}$ is a non-negative real number, $\mathbb{N}$ is a natural number.

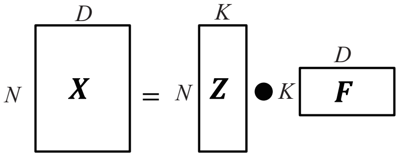


Fig. 1. Latent Feature Model

We also show that the proposed method achieves flexible and accurate ego-motion noise estimation.

The rest of this paper is organized as follows: Section II formulates an ego-motion noise model and proposes its estimation algorithm *Semi-Blind Non-negative Matrix Factorization (SB-INMF)* based on a non-parametric Bayesian model. Section III describes a prototype ego-motion noise suppression system based on SB-INMF. Section IV evaluates the system and gives discussions. The last section concludes this paper with insights of our future work.

II. SEMI-BLIND NON-NEGATIVE MATRIX FACTORIZATION FOR EGO-MOTION NOISE SUPPRESSION

Tab. I describes notations used in this paper. Ego-motion noise is mainly caused by moving joints, and it is natural to consider that ego-motion noise can be represented by a combination of noise generated by all moving joints. In such case, *Latent Feature Model (LFM)* which explains data by adding multiple features selected from finite feature candidates as shown in Fig. 1, where $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{F}$ show $N \times D$ data, $N \times K$ activation, and $K \times D$ feature matrices, respectively.

In signal processing, to solve such a decomposition problem, *Blind Source Separation (BSS)* [11], [22] is a common approach. BSS is used for sound source separation, and it is based on statistical information and microphone array processing. Recently, much attention has been drawn to *Non-negative Matrix Factorization (NMF)* since it can separate sound sources even when an input signal is monaural [23]. Since ego-motion noise of each joint has 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 [24].

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

where $x_{nd} \geq 0, z_{nk} \geq 0$, and $f_{kd} \geq 0$ are elements of $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{F}$, respectively.

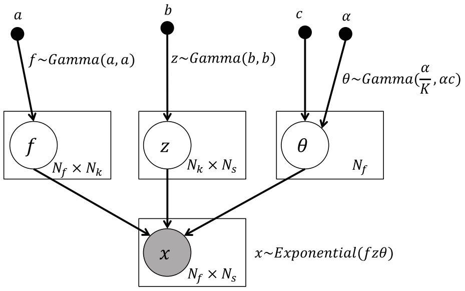


Fig. 2. A Graphical Model for INMF

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 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 *Infinite Latent Feature Model (ILFM)*. It allows LFM to have an unbounded number of feature candidates, and it automatically estimates the most likely number of feature candidates to represent the data. In this paper, ILFM is modeled using a Gamma process, which is a non-parametric stochastic process. NMF is also extended so that it can deal with ILFM, which is called *Infinite Non-negative Matrix Factorization (INMF)* [26]. INMF for ego-motion noise is formulated by

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

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

$$\theta_k \sim \text{Gamma}\,(\alpha/K,\alpha c)\,, \tag{4}$$

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

where a, b, c are parameters for Gamma distributions, and K shows a truncation level.

Since this model assumes a prior distribution based on a Gamma process for each element of a non-negative matrix, it is also called *Gamma Process Non-negative Matrix Factorization (GaP-NMF)*. Fig. 2 illustrates a graphical model of INMF. White and dark circles show latent and observed variables. A black dot shows manually-specified parameters, and multiple nodes can be bundled using a plate. This model simultaneously estimates f, z, and θ 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, 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 features correspond to ego-motion noise among the obtained features. To solve this problem, we propose *Semi-Blind INMF (SB-INMF)*. SB-INMF takes a mixture of ego-motion noise and a target signal as an input.

$$x_{nd} \sim \text{Exponential}\left(\sum_k \theta_k z_{nk} f_{kd} + \sum_l \tilde{\theta}_l \tilde{z}_{nl} \tilde{f}_{ld}\right), \tag{6}$$

where θ_k, z_{nk}, and f_{kd} correspond to ego-motion noise, and $\tilde{\theta}_l$, $\tilde{z}_{nl}$, and $\tilde{f}_{ld}$ correspond to a target signal.

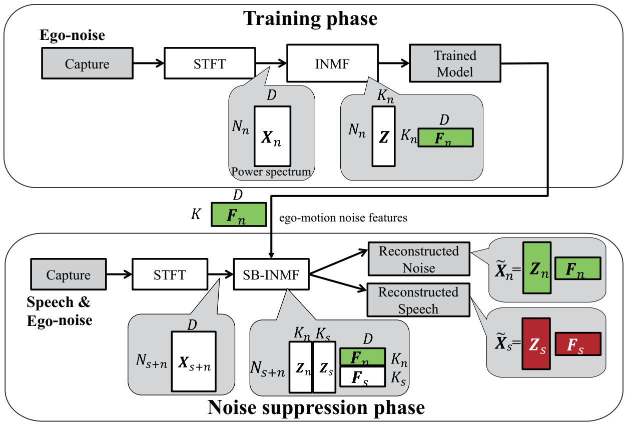


Fig. 3. System Architecture for Ego-motion Noise Suppression

The ego-motion noise feature f_{kd} obtained with INMF is used as given features, and target signal features ($\tilde{\theta}_l$, $\tilde{z}_{nl}$, $\tilde{f}_{ld}$) as well as other ego-motion noise features (θ_k, z_{nk}) are estimated. The noise-suppressed target signal is obtained by multiplying $\tilde{z}_{nl}$ and $\tilde{f}_{ld}$. Because ego-motion noise is suppressed without using non-linear processing such as spectral subtraction [18], it produces less distortions. For estimation with INMF, we use a variational Bayesian method described in [27].

III. Ego-Motion Suppression System

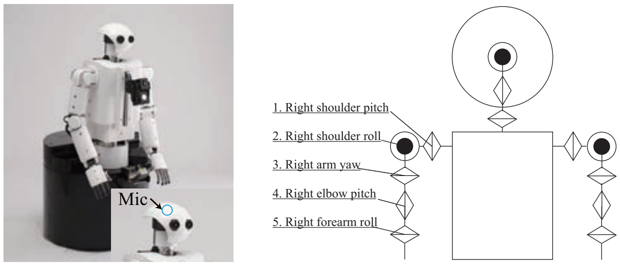
Fig. 3 depicts the system architecture for ego-motion noise estimation and suppression based on SB-INMF. It consists of two phases, that is, training and noise suppression. The upper panel of Fig. 3 shows the training phase to obtain ego-motion noise features. Ego-motion noise is captured with a microphone at a 16 kHz sampling rate. After *Short-Time Fourier Transform (STFT)* with a 256 pt window and a 160 pt window shift is performed, $\boldsymbol{X}_n$ with D frequency bins and N_n time frames is obtained. INMF is performed for $\boldsymbol{X}_n$, and K_n ego-motion noise features are obtained as the trained model $\boldsymbol{F}_n$. The lower panel shows the noise suppression phase. The input speech is contaminated with ego-motion noise. It is transformed into $\boldsymbol{X}_{s+n}$ with STFT. SB-INMF is performed for inference. In SB-INMF, the trained ego-motion noise model $\boldsymbol{F}_n$ is used as a set of given features. Since $\boldsymbol{X}_{s+n}$ also includes a speech signal, an additional feature set $\boldsymbol{F}_s$ which corresponds to speech is obtained. The noise-suppressed speech is reconstructed by multiplying the obtained $\boldsymbol{F}_s$ and its activation matrix $\boldsymbol{Z}_s$. Ego-motion noise is estimated at the same time using the given feature set $\boldsymbol{F}_n$ and the obtained activation matrix $\boldsymbol{Z}_n$.

IV. Evaluation

We evaluated the proposed method as follows:

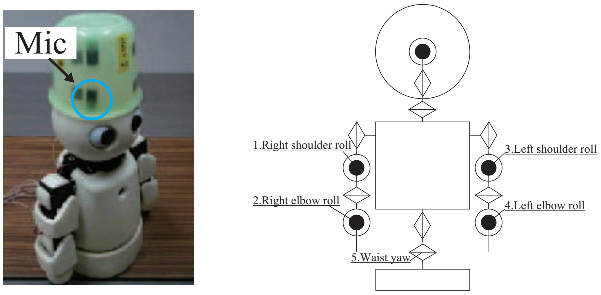
1) ego-motion noise estimation
2) ego-motion noise suppression for synthesized data
3) ego-motion noise suppression for recorded data

The purpose of the first experiment is to validate that ego-motion noise can be modeled in the training phase. Thus, we



a) photo b) joint and link structure

Fig. 4. Hearbo



a) photo b) joint and link structure

Fig. 5. Hearbo and Robovie W

did not add any other sound sources to an input besides ego-motion noise. The other two experiments are to validate that the proposed ego-motion noise suppression works properly for simulated and real data.

A. Humanoid Robots

We used two types of humanoid robots Hearbo and Robovie-W. Hearbo includes 34 Degrees-of-Freedom (DoFs), shown in Fig. 4. An 8 ch microphone array is mounted on the top of the robot's head, and we used the microphone on the forehead for recording. In this robot, each joint status can be obtained synchronously with audio information using ROS. Robovie-W is a commercially-available small robot with 17 DoFs. We put a hat with a microphone on the head for audio recording. We can send commands to control this robot, but it is not possible to obtain joint status. This means that motion information cannot be used for ego-motion noise suppression in this case.

B. Data Preparation

For Hearbo, we selected five joints for the right arm such as 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 room of 4 m × 7 m × 3 m. Hearbo was located at the center of the room, and a loudspeaker was located 1.2 m away in front of the robot.

D1 Each joint was moved for around 20 s separately in turn. In total, 110 s of ego-motion noise was recorded.
D2 J1 was moved for 40 s, and in the last 20 s, speech was also played from the loudspeaker.
D3 Two joints (J1 and J2) were moved simultaneously for 40 s, and in 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 in the last half.
D6 Five joints (J1 – J5) were moved simultaneously for 40 s, and speech was played in the last half.

In every experiment, D1 was used as training data, *i.e.*, ego-motion noise features for D1 were estimated with INMF. For the first experiment, D1 was also used for test data. For the second experiment, we added a clean speech signal to D1. As for the clean speech, we used the same signal as played from the loudspeaker for D2–D6. For the last experiment, D2–D6 were used for test data.

For Robovie-W, five joints like waist yaw, shoulder roll (L and R) and elbow roll (L and R). It was put on a table located at another room of 10 m × 10 m × 3 m.

R1 These five joints were randomly moved for five seconds.
R2 A person spoke to the robot at 2 m distance during performing the same motion as R1.

In every experiment, R1 was used as training data. For test data, R1, a mixed signal of R1 and a clean speech signal, and R2 were used for experiment 1–3, respectively.

C. Metrics

In the first experiment, we used a *Noise Estimation Error (NEE)* defined by

$$\mathrm{NEE} = 20\log_{10}\left(\frac{\sum_f\sum_\omega |N(\omega,f)|^2}{\sum_f\sum_\omega (|N(\omega,f)| - |\hat{N}(\omega,f)|)^2}\right), \qquad (7)$$

where ω and f show frequency and time frame. N and $\hat{N}$ are the original and the estimated ego-motion noise, respectively.

For the second experiment, *Signal-to-Noise Ratio (SNR)*, *Signal-to-Inference Ratio (SIR)*, and *Signal-to-Distortion Ratio (SDR)* were computed. SNR (SNR_1) was defined by

$$\mathrm{SNR}_1 = 20\log_{10}\left(\frac{\sum_f\sum_\omega |X(\omega,f)|^2}{\sum_f\sum_\omega ||Y(\omega,f)| - |X(\omega,f)| - |\hat{N}(\omega,f)||^2}\right) \qquad (8)$$

where X and Y are the clean and the noise-contaminated speech signals, respectively.For SIR and SDR, MATLAB toolbox called "BSS Eval[1]" was used.

In the third experiment, reference signals for speech and noise are unavailable, and thus we defined SNR_2 to compute SNR improvement with the proposed method formulated by,

$$\begin{aligned}\mathrm{SNR}_2 &= 20\log_{10}\left(\frac{\sum_f\sum_\omega \left|\hat{S}(\omega,f)\right|^2}{\sum_f\sum_\omega \left||Y_N(\omega,f)| - |\hat{N}(\omega,f)|\right|^2}\right) \\ &\quad -20\log_{10}\left(\frac{\sum_f\sum_\omega |Y_S(\omega,f)|^2}{\sum_f\sum_\omega |Y_N(\omega,f)|^2}\right), \qquad (9)\end{aligned}$$

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

For comparison, we used a template-based method proposed by Ince et al. [20] for Hearbo, since it is known as one of the best methods to deal with ego-motion noise although it uses joint status information. For template database

[1] http://bass-db.gforge.inria.fr/bss_eval/

TABLE II

EGO-MOTION NOISE ESTIMATION AND SUPPRESSION FOR EXP. 1 & 2

Robot	Hearbo							Robovie-W
Method	Proposed	Template-based						Proposed
# of feat. /templ.	7 (noise) 2 (speech)	31	98	303	1,022	3,115	8,431	5 (noise) 3 (speech)
NEE (dB)	**9.4**	8.0	8.2	8.0	9.9	12.3	24.6	**9.3**
SNR_1 (dB)	**7.2**	6.1	6.3	5.9	7.4	8.7	14.6	**7.6**
SIR (dB)	**13.0**	1.4	2.6	2.6	3.0	2.3	2.2	**17.0**
SDR (dB)	**1.3**	-0.8	1.1	1.1	1.3	0.8	0.8	**1.5**

TABLE III

EGO-MOTION NOISE SUPPRESSION RESULTS FOR EXP. 3

Robot	Dataset	Proposed	Template-based	
		$\text{SNR}_2(dB)$	$\text{SNR}_2(dB)$	# of templates
Hearbo	D2(J1)	5.9	2.5	21
	D3(J1+J2)	5.2	3.1	8
	D4(J1–J3)	6.3	3.2	12
	D5(J1–J4)	3.5	-0.76	244
	D6(J1–J5)	5.3	2.4	45
Robovie-W	R2	3.9	N/A	N/A

generation, D1 was used for the first two experiments, and the first quarter, *i.e.*, around 10 s was used for each of D2 – D6 in the third experiment. However, it cannot be used for Robovie-W due to lack of motion information.

D. Results

Tab. II shows the result for the first and the second experiments. The number of features in the proposed method corresponds to that of templates in template-based ego-motion noise estimation. Even when 7 features for Hearbo and 5 features for Robovie-W were used for ego-motion noise with the proposed method, NEE was better than the template-based method with over 300 templates, which shows that the proposed method models ego-motion noise better. Since five joints were used in this experiment for each robot, it is obvious that joints and features do not have one-to-one correspondence, which indicates that it is difficult to model ego-motion noise when using only joint status information.

In the second experiment, 2 features for Hearbo were additionally estimated as speech features. In total, only 9 features were necessary to model the input signal. In SNR_1, the proposed method demonstrated equivalent performance with the template-based noise suppression with 1,022 templates. In SIR and SDR, the proposed method showed the best performance compared to the template-based method regardless of the number of templates. The larger number of features or templates requires high computational cost, it is a big advantage that only a small number of features is necessary. Figs. 6 and 7 illustrate the spectrograms obtained in the second experiment for Hearbo and Robovie-W, respectively. Fig. 6a) is a mixed signal generated by addition of Fig. 6b) and c). Fig. 6e)-f) is reconstructed signals with the proposed method for a mixed signal, ego-motion noise, and speech, respectively. From these spectrograms, we can see that signals are successfully reconstructed while the reconstructed signals blur to some extent. For Robovie-W, it is impossible to compare with the template based method, but it showed similar performance to the Hearbo's case both in Tab. II and Fig. 7. These results show that ego-motion noise and speech features are estimated properly with the proposed method.

Tab. III shows the results of the third experiment. We selected the best scores for SNR_2 by changing the number of features and templates, and "# of templates" means the number when SNR_2 has the best score. In every case, the proposed method improved SNR_2 and showed better performance compared to the template-based method. This shows the robustness of our method since it is effective even in open tests which means that test data is completely different from the training data. In the case of Robovie-W, the improvement was 3.9 dB. However the recorded sound R2 sounds noisier than Hearbo, which might be caused by a longer distance between the speaker and the robot. For the template-based method, the use of a larger number of templates results in better performance in the closed tests shown in Tab. II. However, in the open test, a small number of templates had the best performance. This supports our claim that the same motion does not always generate the same ego-motion noise, and they are not fully correlated. The result for D5 was different from other cases, and we guess that this is caused by the shortage of training data.

The proposed method also has a problem. Because the search space to feature estimation becomes large due to the introduction of a non-parametric approach, it sometimes results in a local solution. We should find a way to relax this problem, *e.g.*, by using annealing methods for future work.

V. CONCLUSION

We presented a new ego-motion noise suppression method with a single microphone which does not use any joint status information and any additional sensors. The proposed method first estimates features for ego-motion noise using *Infinite Non-negative Matrix Factorization* and then a noise suppressed target signal is obtained with *Semi-Blind INMF (SB-INMF)* by using the estimated ego-motion noise features as given features. Since it is free from non-linear processing, the noise-suppressed signal is less distorted compared to spectral subtraction methods. The effectiveness of the proposed method was shown using two types of humanoid robots. Future work includes the construction of an on-line system, improvement in stability of the proposed method, and validation of the proposed method with automatic speech recognition.

ACKNOWLEDGMENTS

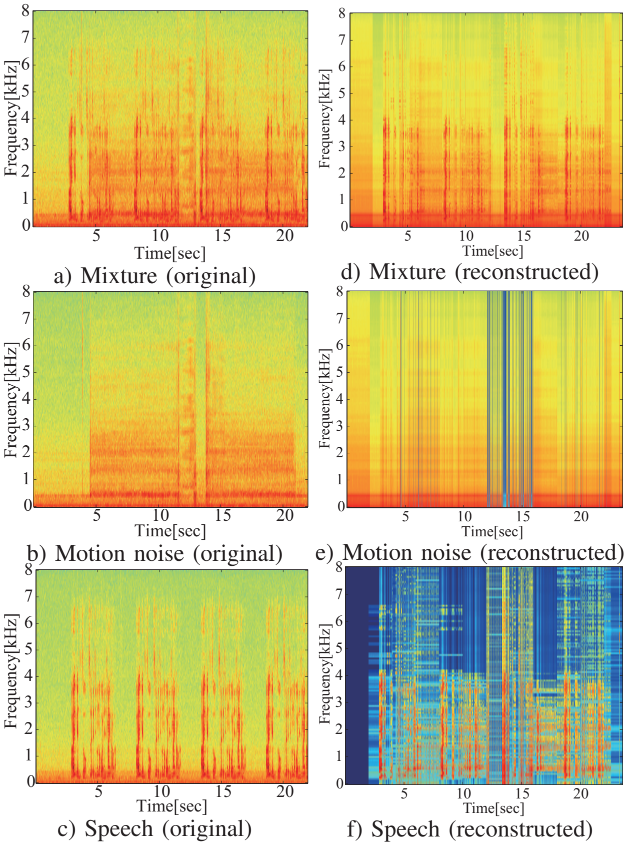
We thank Dr. Randy Gomez, HRI-JP for his great help. This research was partially supported by Grant-in-Aid for Scientific Research No. 24118702, and No. 24220006

REFERENCES

[1] K. Nakadai *et al.*, "Active audition for humanoid," in *Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000)*. AAAI, 2000, pp. 832–839.

[2] K. Nakadai *et al.*, "Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots," *Speech Communication*, vol. 44, pp. 97–112, 2004.

[3] Y. Nishimura *et al.*, "Speech recognition for a humanoid with motor noise utilizing missing feature theory," in *Proc. of Humanoids 2006*. IEEE, pp. 26–33.



a) Mixture (original)

d) Mixture (reconstructed)

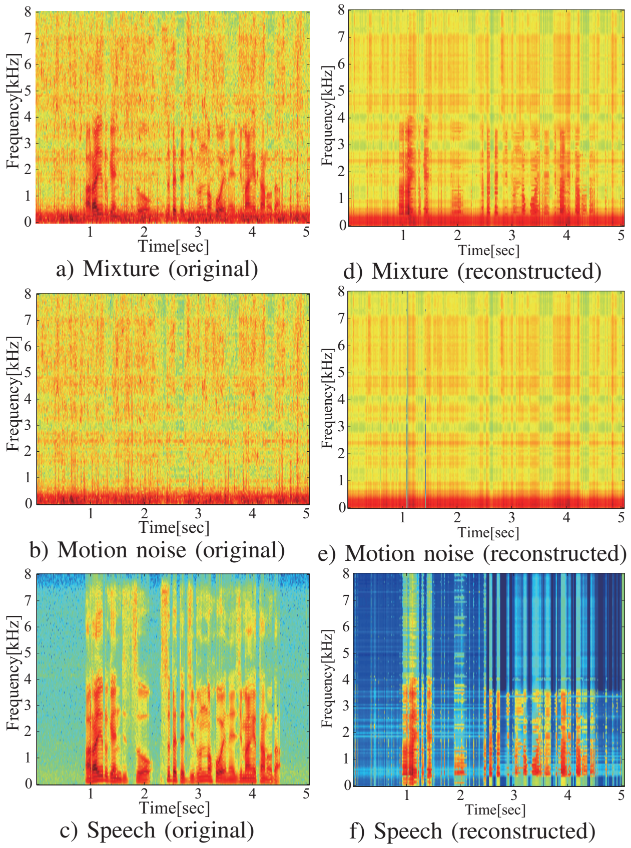
b) Motion noise (original)

e) Motion noise (reconstructed)

c) Speech (original)

f) Speech (reconstructed)

Fig. 6. Noise Suppression Result with Hearbo



a) Mixture (original)

d) Mixture (reconstructed)

b) Motion noise (original)

e) Motion noise (reconstructed)

c) Speech (original)

f) Speech (reconstructed)

Fig. 7. Noise Suppression Result with Robovie-W

[4] T. Rodemann *et al.*, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proc. of IROS 2006*, IEEE/RSJ, pp. 860–865.

[5] J. Hornstein *et al.*, "Sound localization for humanoid robots – building audio-motor maps based on the hrtf," in *Proc. of IROS 2006*, IEEE/RSJ, pp. 1171–1176.

[6] T. Shimoda *et al.*, "Spectral cues for robust sound localization with pinnae," in *Proc. of IROS 2006*, IEEE/RSJ, pp. 386–391.

[7] A. Portello *et al.*, "Active binaural localization of intermittent moving sources in the presence of false measurements," in *Proc. of IROS 2012*, IEEE/RSJ, pp. 3294–3299.

[8] J.-M. Valin *et al.*, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. ICRA 2004*, IEEE-RAS, pp. 1033–1038.

[9] Y. Sasaki *et al.*, "Spherical microphone array for spatial sound localization for a mobile robot," in *Proc. of IROS 2012*, IEEE/RSJ, pp. 713–718.

[10] S. Yamamoto *et al.*, "Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech," in *Proc. of the 2007 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-2007)*. IEEE, 2007, pp. 111–116.

[11] H. Saruwatari *et al.*, "Two-stage blind source separation based on ica and binary masking for real-time robot audition system," in *Proc. of IROS 2005*. IEEE/RSJ, pp. 209–214.

[12] T. Yoshida and K. Nakadai, "Active audio-visual integration for voice activity detection based on a causal bayesian network," in *Proc. of Humanoids 2012*. IEEE, pp. 370–375.

[13] K. Nakadai *et al.*, "Sound source tracking with directivity pattern estimation using a 64ch microphone array," in *Proc. of IROS 2005*. IEEE/RSJ, pp. 196–202.

[14] F. Perrodin *et al.*, "Design and calibration of large microphone arrays for robotic applications," in *Proc. of IROS 2012*, IEEE/RSJ, pp. 4596–4601.

[15] J. Even *et al.*, "Combining laser range finders and local steered response power for audio monitoring," in *Proc. of IROS 2012*, IEEE/RSJ, pp. 986–991.

[16] J. Even *et al.*, "Semi-blind suppression of internal noise for hands-free robot spoken dialog system," in *Proc. of IROS 2009*, IEEE/RSJ, pp. 658–663.

[17] A. Ito *et al.*, "Internal noise suppression for speech recognition by small robots," in *Proc. of European Conference on Speech Communication and Technology (Eurospeech-2005)*, 2005, pp. 2685–2688.

[18] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proc. of 1979 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*. IEEE, 1979, pp. 200–203.

[19] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[20] G. Ince *et al.*, "Incremental learning for ego noise estimation of a robot," in *Proc. of IROS 2011*. IEEE/RSJ, pp. 131–136.

[21] J. L. Oliveira *et al.*, "An active audition framework for auditory-driven hri: Application to interactive robot dancing." in *Proc. of IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2012)*, 2012, pp. 1078–1085.

[22] L. C. Parra and C. V. Alvino, "Geometric source separation: Mergin convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[23] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2006.

[24] S. Abdallah and M. D. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, 2004, pp. 10–14.

[25] T. L. Griffiths and Z. Ghahramani, "The indian buffet process: An introduction and review," *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, 2011.

[26] M. N. Schmidt and M. Mørup, "Infinite non-negative matrix factorization," in *European Signal Processing Conference (EUSIPCO)*, 2010.

[27] M. D. Hoffman *et al.*, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. of the 27th International Conference on Machine Learning (ICML2010)*, 2010, pp. 439–446.