https://doi.org/10.20965/jrm.2017.p0198

Paper:

Low Latency and High Quality Two-Stage Human-Voice-Enhancement System for a Hose-Shaped Rescue Robot

Yoshiaki Bando[*1], Hiroshi Saruwatari[*2], Nobutaka Ono[*3], Shoji Makino[*4], Katsutoshi Itoyama[*1], Daichi Kitamura[*5], Masaru Ishimura[*4], Moe Takakusaki[*4], Narumi Mae[*4], Kouei Yamaoka[*4], Yutaro Matsui[*4], Yuichi Ambe[*6], Masashi Konyo[*6], Satoshi Tadokoro[*6], Kazuyoshi Yoshii[*1], and Hiroshi G. Okuno[*7]

[*1]Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
E-mail: yoshiaki@kuis.kyoto-u.ac.jp
[*2]Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[*3]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
[*4]Graduate School of Systems and Information Engineering, Tsukuba University
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
[*5]Department of Informatics, School of Multidisciplinary Sciences, SOKENDAI
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
[*6]Graduate School of Information Science, Tohoku University
6-6-01 Aramaki Aza Aoba, Aoba-ku, Sendai 980-8579, Japan
[*7]Graduate Program for Embodiment Informatics, Waseda University
2-4-12 Okubo, Shinjuku, Tokyo 169-0072, Japan
[Received July 25, 2016; accepted October 18, 2016]

This paper presents the design and implementation of a two-stage human-voice enhancement system for a hose-shaped rescue robot. When a microphone-equipped hose-shaped robot is used to search for a victim under a collapsed building, human-voice enhancement is crucial because the sound captured by a microphone array is contaminated by the ego-noise of the robot. For achieving both low latency and high quality, our system combines online and offline human-voice enhancement, providing *an overview first and then details on demand*. The online enhancement is used for searching for a victim in real time, while the offline one facilitates scrutiny by listening to highly enhanced human voices. Our online enhancement is based on an online robust principal component analysis, and our offline enhancement is based on an independent low-rank matrix analysis. The two enhancement methods are integrated with Robot Operating System (ROS). Experimental results showed that both the online and offline enhancement methods outperformed conventional methods.

Keywords: hose-shaped rescue robot, blind human-voice enhancement, search and rescue, robot audition

1. Introduction

Hose-shaped rescue robots have been developed for gathering information in narrow spaces under collapsed buildings where humans or animals cannot enter [1–3]. They have thin, long, and flexible bodies and have self-locomotion mechanisms. The Active Hose-II robot [2], for example, has small powered wheels enabling it to move forward, and the Active Scope Camera robot [1, 3] can move forward by vibrating the cilia covering its body (see **Fig. 1**). In 2008, the Active Scope Camera robot was used in an actual search-and-rescue mission in Jacksonville, Florida, USA [4].

Rescue robots should keep moving because a rescue activity is a race against time [5]. Owing to the ego-noise of the robot, it is difficult for the remote operator to hear the voice of a victim at an unseen and distant place [6]. Stopping the actuators of the robot periodically so that the operator can hear a human voice is an inefficient use of search time, and does nothing to help the operator hear a voice while the robot is moving. Real-time human-voice enhancement should cope with ego-noise that changes dynamically according to the movements of wheels or vibrators and the friction between the robot's body and surrounding materials. However, conventional enhancement methods [7–10] cannot work effectively because they assume the noise stable or known in advance.

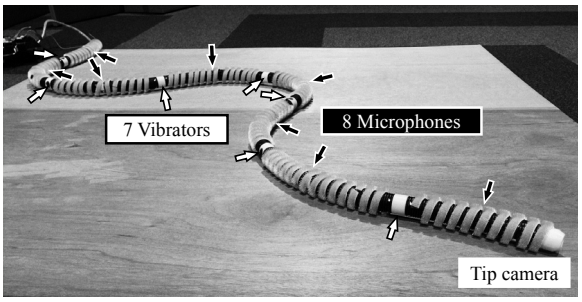


Fig. 1. A hose-shaped rescue robot that has eight-channel microphone array and is covered by cilia for moving.

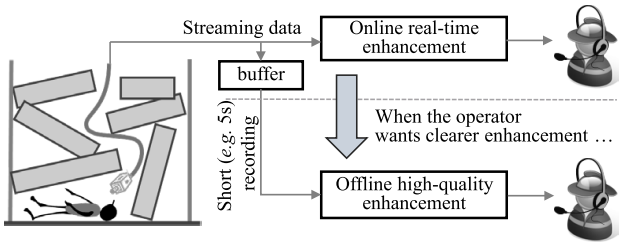


Fig. 2. Overview of two-step human-voice enhancement system.

A human-voice enhancement system for a hose-shaped rescue robot requires low latency for search-and-rescue activities. It also needs to address extremely low signal-to-noise ratio (SNR) conditions because the ego-noise generated from the body of the robot is much louder at robot-mounted microphones and often masks targeted human voices. Conventional blind source separation (BSS) methods also have a trade-off between latency and enhancement quality [11–14].

This paper presents the design and implementation of a two-stage human-voice enhancement system for a hose-shaped rescue robot. The system combines online real-time enhancement and offline high-quality enhancement to attain low latency and high quality. The online enhancement facilitates hearing a trapped victim's voice in real time, while the offline one facilitates scrutiny by listening to highly enhanced human voices.

To enhance a human voice in real-time, we developed an online robust principal component analysis (RPCA)-based enhancement. RPCA can decompose an input amplitude spectrogram into frequency components that appear repeatedly (e.g., the ego-noise of a hose-shaped rescue robot) and other components that occur infrequently (e.g., a human voice) without prior learning [11, 15]. As RPCA is designed for a single-channel input signal, we first apply RPCA to each microphone input of the microphone array, and then we combine the results of the microphones to improve the enhancement performance.

To obtain a high-quality enhancement result, we use an offline independent low-rank matrix analysis (ILRMA)-based enhancement [12, 16]. Although RPCA can work in real-time without prior learning, its enhancement result is distorted and includes artificial noise (called musical noise) caused by its non-linearity. As ILRMA is a linear BSS method, the separation results are not distorted by musical noise [12, 16]. We first apply ILRMA to a multichannel audio input, and the estimated human-voice sound is further refined by postfiltering.

The rest of this paper is organized as follows. Section 2 presents the design and implementation of our human-voice enhancement system. Section 3 reports the experimental results obtained using an actual hose-shaped robot. Section 4 summarizes the key findings and mentions future research. It should be noted that the multi-channel online enhancement in this paper is partially based on an international conference paper [6] written by some of the authors. The contribution of this study is on the design, implementation, and evaluation of a two-stage human-voice enhancement system.

2. Two-Stage Human-Voice Enhancement System for a Hose-Shaped Rescue Robot

In this section, we first discuss the design criteria for our system that combines online and offline enhancements; then, we describe the online and offline enhancement methods used in the system. Finally we explain the implementation of our system based on Robot Operating System (ROS) [17].

2.1. Design Criteria

The proposed system combines online real-time human-voice enhancement and offline high-quality enhancement, providing *an overview first and then details on demand* (**Fig. 2**). The system provides real-time enhanced signals to a remote operator searching for a trapped victim. When the remote operator wants to make the enhanced voice clearer, such as when he detects a very weak human voice in a real-time result, the system provides an offline high-quality enhanced signal for the final several seconds of audio input.

The idea of *an overview first and then details on demand* was originally proposed by Schneiderman as an overview first zoom and filter, then details on demand [18]. The original version was developed as a guideline for designing an interactive visualization system. In the case of a human-voice enhancement system for a rescue robot, the real-time enhanced signal to which the zoom and filter are already applied should be presented to a remote operator by default. Therefore, our system has two steps, namely, 1) providing online enhancement as an overview, and 2) providing offline enhancement as details.

2.2. Hose-Shaped Robot with a Microphone Array

Figure 1 shows the hose-shaped rescue robot used in this study. The body is made with a corrugated tube 38 mm in diameter and 3 m long. This robot has $M = 8$ microphones positioned on its body at intervals of 40 cm and has a USB video camera at the tip. The audio signals of the microphones are captured at 16 kHz and 24-bit sampling by a synchronized multichannel A/D converter

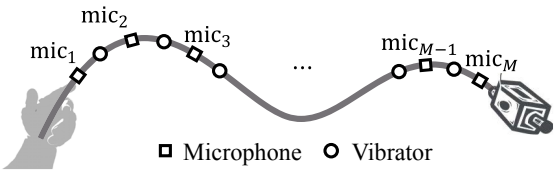


Fig. 3. Configuration of microphones and vibrators on the hose-shaped rescue robot.

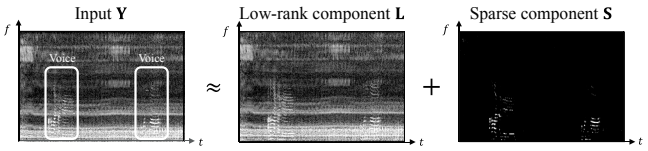


Fig. 4. RPCA separates ego-noise and a human voice as low-rank and sparse components, respectively. Input is a mixture of a human voice and ego-noise of the hose-shaped robot.

called RASP-ZX (System in Frontier Corp.). The body was rotated at 90° after installing each microphone in order to avoid having all microphones obstructed by the ground. This robot moves forward by a mechanism in the same way as that of the Active Scope Camera robot [1]: the entire surface of the robot is covered by cilia, and the robot moves forward by vibrating them. This vibration is generated by seven vibration motors installed in the robot (**Fig. 3**).

2.3. Online Human-Voice Enhancement

The online human-voice enhancement is conducted based on an online RPCA [6, 11].

2.3.1. Motivation

The online enhancement is required to work in real time and to address the deformation of the microphone array, the layout of which changes as the robot moves. The conventional online BSS methods, which separate sound sources based on the phase differences among microphones, assume that the array layout is stable or known in advance [14, 19, 20]. Although there are several offline BSS methods that track the time-varying phase differences or array layout [21], it is difficult to conduct such tracking in real time and in an online manner.

To avoid using phase information that is sensitive to the array layout, the proposed online enhancement method is based on an online RPCA that works on an audio amplitude spectrogram [11, 15, 22]. RPCA can separate the ego-noise and human voice based on the low-rankness and sparseness of their amplitude spectrograms instead of their phase information (**Fig. 4**). As the ego-noise of our robot mainly consists of the periodic sounds generated by vibrators and friction between the robot body and surrounding materials, it has a low-rank tendency. The human voice, on the other hand, has sparse tendency because it is non-stationary and infrequently appears. These tendencies enable an online RPCA to separate the ego-noise and human voice without any prior training.

In this study, we improve the enhancement performance of online RPCA by combining the single-channel online RPCA results of multiple microphones. Because the microphones and vibrators are alternately installed on the long body, we can assume that each microphone captures the different ego-noise generated by different vibrators. The target voice, on the other hand, is assumed to be similarly recorded by all the microphones because the sound source is single and it propagates in the air. Based on these assumptions, we first apply online RPCA to each microphone recording and then extract the components common among the single-channel results.

2.3.2. Problem Statement

The online human-voice enhancement problem is defined as follows:

Input: M-channel synchronized amplitude spectra $\boldsymbol{y}_{1j},\ldots,\boldsymbol{y}_{Mj} \in \mathbb{R}^I$
Output: denoised amplitude spectrum $\boldsymbol{s}_j \in \mathbb{R}^I$

where I and j are the number of frequency bins and the time frame index, respectively. The input amplitude spectra are obtained by taking the absolute values of the short-time Fourier transform (STFT) of captured signals.

2.3.3. Overview of Online RPCA

The proposed method uses an online extension of batch RPCA [11]. The input amplitude spectrum of each channel $\boldsymbol{y}_{mj}$ is decomposed to a low-rank component $\boldsymbol{l}_{mj}$ and sparse component $\boldsymbol{s}_{mj}$ by conducting the online RPCA:

$$\boldsymbol{y}_{mj} \approx \boldsymbol{l}_{mj} + \boldsymbol{s}_{mj}. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

The ego-noise that changes periodically is separated into the low-rank component, and the voice signal and other sparse noise are separated into the sparse component [15].

To explain online RPCA that is independent of the microphone index m, in the rest of this section we leave it out. Let $I \times j$ matrices of input, low-rank, and sparse spectrograms be $\boldsymbol{Y} = [\boldsymbol{y}_1,\ldots,\boldsymbol{y}_j]$, $\boldsymbol{L} = [\boldsymbol{l}_1,\ldots,\boldsymbol{l}_j]$, and $\boldsymbol{S} = [\boldsymbol{s}_1,\ldots,\boldsymbol{s}_j]$, respectively. The original batch RPCA [23] decomposes the input matrix into low-rank and sparse matrices by solving the following problem:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \left\{ \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{L}-\boldsymbol{S}\|_F^2 + \lambda_1\|\boldsymbol{L}\|_* + \lambda_2\|\boldsymbol{S}\|_1 \right\} \quad . \quad (2)$$

where $\|\cdot\|_F$, $\|\cdot\|_*$, and $\|\cdot\|_1$ represent the Frobenius, nuclear, and L1 norms, respectively, and $\min_x f(x)$ is the minimum point x of $f(x)$. The parameter $\lambda_1 > 0$ controls the low-rankness of the low-rank matrix $\boldsymbol{L}$, and the parameter λ_2 controls the sparseness of the sparse matrix $\boldsymbol{S}$. As this optimization, particularly the second term of the nuclear norm, accesses all samples of the input matrix, it is difficult to solve this RPCA problem in an online manner [23].

To overcome this difficulty, the online RPCA [11]

solves the following alternative problem:

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{S}} \left\{ \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{H}^T-\boldsymbol{S}\|_F^2 + \frac{\lambda_1}{2}(\|\boldsymbol{W}\|_F^2+\|\boldsymbol{H}\|_F^2)+\lambda_2\|\boldsymbol{S}\|_1 \right\} \quad (3)$$

where $\boldsymbol{W} \in \mathbb{R}^{I\times K}$ and $\boldsymbol{H}=[\boldsymbol{h}_1,\ldots,\boldsymbol{h}_j]^T \in \mathbb{R}^{j\times K}$ $(K<I,j)$ represent the basis vectors of the low-rank spectrogram and its coefficient vectors ($\boldsymbol{L}=\boldsymbol{W}\boldsymbol{H}^T$). This alternative problem is derived by using the upper bound of the nuclear norm $\|\boldsymbol{L}\|_*$ as follows [11, 24]:

$$\|\boldsymbol{L}\|_* = \|\boldsymbol{W}\boldsymbol{H}\|_* \leq \inf_{\boldsymbol{W},\boldsymbol{H}} \left\{ \frac{1}{2}\|\boldsymbol{W}\|_F^2+\frac{1}{2}\|\boldsymbol{H}\|_F^2 \right\}. \quad (4)$$

where $\inf_x f(x)$ is the infimum point x of $f(x)$. The online RPCA problem (Eq. (3)) can be solved in an online manner by minimizing its following transformations:

$$f_j(\boldsymbol{W}) = \frac{1}{s}\sum_{j'=1}^{j} l(\boldsymbol{y}_{j'},\boldsymbol{L}) + \frac{\lambda_1}{2j}\|\boldsymbol{W}\|_F^2 \quad \ldots \quad (5)$$

$$l(\boldsymbol{y}_j,\boldsymbol{W}) = \min_{\boldsymbol{h}_j,\boldsymbol{s}_j} \left\{ \frac{1}{2}\|\boldsymbol{y}_j-\boldsymbol{W}\boldsymbol{h}_j-\boldsymbol{s}_j\|_F^2 + \frac{\lambda_1}{2}\|\boldsymbol{h}_j\|_F^2+\lambda_2\|\boldsymbol{s}_j\|_1 \right\}. \quad (6)$$

This cost function is minimized using an off-the-shelf solver and block-coordinate descent with warm restarts [11].

2.3.4. Online Normalization of Input Spectrum

The ego-noise of our hose-shaped rescue robot has large powers at low frequency bins. As the online RPCA estimates the low-rank components with the same weight for all the frequency bins (owing to the first term of Eq. (3)), it over-fits to the low frequency bins. We therefore apply a normalization coefficient $\boldsymbol{g}_{mj} = [g_{m1j},\ldots,g_{mIj}]^T \in \mathbb{R}^I$ to the input $\boldsymbol{y}_{mj}$:

$$y'_{mij} = \frac{1}{g_{mij}} y_{mij}. \quad \ldots \quad (7)$$

As the peaks of the ego-noise changes depending on the environment around the robot, the proposed method learns the normalization coefficient in an online manner. We assume that the average ego-noise does not change frequently and drastically; therefore, the normalization coefficient is updated as follows:

$$\boldsymbol{g}_{mj} = (1-\alpha)\boldsymbol{g}_{m(j-1)} + \alpha \boldsymbol{y}_{mj} \quad \ldots \quad (8)$$

where α is a learning weight parameter that is set to a small value (e.g., 1.0×10^{-2}). The flowchart of the channel-wise human-voice enhancement based on the online RPCA is summarized in **Fig. 5**.

2.3.5. Combining Online RPCA Results

The sparse components of the microphones, $\boldsymbol{s}_{mj} = [s_{m1j},\ldots,s_{mIj}]^T$, are integrated to extract the common component, $\boldsymbol{s}_j = [s_{1j},\ldots,s_{Ij}]^T$ (**Fig. 6**). Because the ego-noise is generated from the whole body of a robot and the human voice is propagated in the air, this integration is based on the assumption that the target human-voice is similar at each microphone whereas the ego-noise differs at different microphones. Each sparse component includes musical noise and sparse noise measured when the corresponding microphone touches the environment. The integration is conducted by taking a median at each frequency bin as follows:

$$s_{ij} = \text{Median}(s_{1ij},\ldots,s_{Mij}) \text{ for all } i=1,\ldots,I \quad (9)$$

where $\text{Median}(\cdots)$ represents the median of the arguments.

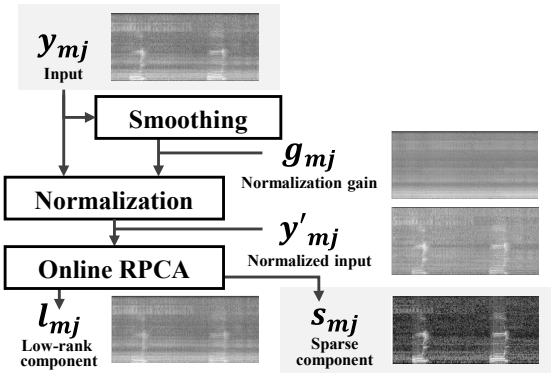


Fig. 5. Human-voice enhancement using channel-wise online RPCA for each microphone input signal.

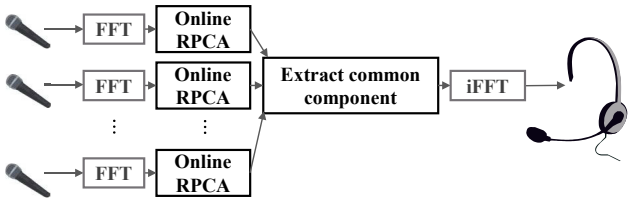


Fig. 6. Overview of online human-voice enhancement.

2.4. Offline Human-Voice Enhancement

The offline human-voice enhancement is conducted by applying ILRMA to a multichannel audio input, and the estimated human-voice signal is further refined by postfiltering.

2.4.1. Motivation

BSS is a technique taken to separately estimate the sources without knowing any prior information, namely, the sensor positions and source locations. It is well known that BSS, which uses multichannel signals, is one of the effective algorithms for human-voice enhancement because it utilizes the spatial information of sound sources, e.g., the difference in directions of arrival of sources, as well as the spectral characteristics of sources. This property is a strong motivation to apply the BSS technique into the hose-shaped rescue robot, where multiple microphones with unknown locations are attached on the flexible robot body.

In order to solve the BSS problem, many methods, e.g., independent component analysis (ICA) [25, 26] have been proposed so far. Therefore, we should carefully select the most appropriate BSS algorithm based on the specific signal and acoustic characteristics assumed in the rescue robot. The dominant factors for ego-noise are the vibration sound generated by vibrating motors and fricative sounds. Thus, we assume that ego-noise can be efficiently expressed by nonnegative matrix factorization (NMF) because the time-frequency structure is obtained by repeating several types of similar spectra. In addition, because the hose-shaped rescue robot moves very slowly and the input signal is short enough for our use, the source separation using the inverse of the linear time-invariant mixing system can be available, owing to the fact that the positional relationship between the ego-noise sources and the microphones barely changes. These assumptions greatly motivate us to introduce the ILRMA [12, 16] proposed by some of the authors for the human-voice enhancement of the rescue robot. Moreover, as the separation performance is often insufficient, particularly for the purpose of actual acoustic sound separation, we propose an extended system that combines ILRMA with the statistical postfiltering technique.

2.4.2. Overview of ILRMA

First, several preliminaries and definitions for signals and system, which are different from those of the online enhancement, are provided. The number of sources and the number of microphones are assumed to be M. We describe multichannel sound source signals, observed signals, and separated signals in each time-frequency slot as follows:

$$\boldsymbol{s}_{ij} = (s_{ij,1} \cdots s_{ij,M})^T, \quad \ldots\ldots\ldots \quad (10)$$

$$\boldsymbol{x}_{ij} = (x_{ij,1} \cdots x_{ij,M})^T, \quad \ldots\ldots\ldots \quad (11)$$

$$\boldsymbol{y}_{ij} = (y_{ij,1} \cdots y_{ij,M})^T, \quad \ldots\ldots\ldots \quad (12)$$

where $1 \leq i \leq I$ $(i \in \mathbb{N})$ describes the frequency index, $1 \leq j \leq J$ $(j \in \mathbb{N})$ describes the time index, ^{T} denotes the vector transpose, and all entries for these vectors are complex values. We can approximately represent the observed signals as

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \quad \ldots\ldots\ldots\ldots\ldots \quad (13)$$

where $\boldsymbol{A}_i = [\boldsymbol{a}_{i,1} \cdots \boldsymbol{a}_{i,M}]$ expresses the mixing matrix of the observed signals ($\boldsymbol{a}_{i,m}$ is often called the steering vector). When $\boldsymbol{W}_i = [\boldsymbol{w}_{i,1} \cdots \boldsymbol{w}_{i,M}]^H$ refers to the demixing matrix, the separated signal $\boldsymbol{y}_{ij}$ is represented as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \quad \ldots\ldots\ldots\ldots\ldots \quad (14)$$

where $\boldsymbol{w}_{i,m}$ is the demixing filter, and ^{H} is the Hermitian transpose. The optimization of the demixing matrix $\boldsymbol{W}_i$ can be performed so that each component of $\boldsymbol{y}_{ij}$ becomes mutually independent.

Next, the formulation of ILRMA is derived as indicated below. In ILRMA, the observed signal is represented by the correlation matrix between the channels, X_{ij}, as

$$\mathsf{X}_{ij} = \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^H. \quad \ldots\ldots\ldots\ldots\ldots \quad (15)$$

The separation model, $\hat{\mathsf{X}}_{ij}$, that approximates X_{ij} is represented as

$$\mathsf{X}_{ij} \approx \hat{\mathsf{X}}_{ij} = \Sigma_k (\Sigma_m \mathsf{H}_{i,m} z_{mk}) t_{ik} v_{kj}, \quad \ldots\ldots \quad (16)$$

where $m = 1, \ldots, M$ is the index of sound sources, and $k = 1, \ldots, K$ is the index of the spectral bases for NMF. $\mathsf{H}_{i,m}$ is an $M \times M$ spatial covariance matrix for each frequency i and source m, and $\mathsf{H}_{i,m} = \boldsymbol{a}_{i,m} \boldsymbol{a}_{i,m}^H$ is limited to a rank-1 matrix. The parameter, $z_{mk} \in \mathbb{R}_{[0,1]}$, is a weight for distributing K NMF bases (frequently appearing spectra) to each sound source. It shows that the k-th basis contributes to only the m-th source. In addition, $t_{ik} \in \mathbb{R}_+$ and $v_{kj} \in \mathbb{R}_+$ are the elements of the basis matrix $\boldsymbol{T}$ and the activation matrix $\boldsymbol{V}$; thus $\boldsymbol{TV}$ is the modeled spectrogram via NMF representation.

ILRMA models each sound source spectrogram as a low-rank nonnegative matrix and decomposes the sources based on their independent nature. This results in the minimization problem of the following Q function:

$$Q = \sum_{i,j} \left[\sum_m \frac{|y_{ij,m}|^2}{\sum_k z_{mk} t_{ik} v_{kj}} - 2 \log |\det \boldsymbol{W}_i| + \sum_m \log \sum_k z_{mk} t_{ik} v_{kj} \right], \quad \ldots\ldots\ldots \quad (17)$$

where the first and second terms in the right side are related to the independence of sources, and the first and third terms are related to the low-rank modeling of the sources. To minimize the Q function while keeping nonnegativity of t_{ik} and v_{kj}, the auxiliary function method (also known as majorization-minimization method) can be applied. The update rules of the demixing matrix $\boldsymbol{W}_i$ to obtain the separated signal $\boldsymbol{y}_{ij}$ are as follows [12]:

$$r_{ij,m} = \Sigma_k z_{mk} t_{ik} v_{kj}, \quad \ldots\ldots\ldots\ldots \quad (18)$$

$$V_{i,m} = \frac{1}{J} \Sigma_j \frac{1}{r_{ij,m}} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^H, \quad \ldots\ldots\ldots \quad (19)$$

$$\boldsymbol{w}_{i,m} \leftarrow (\boldsymbol{W}_i V_{i,m})^{-1} \boldsymbol{e}_m, \quad \ldots\ldots\ldots \quad (20)$$

where $\boldsymbol{e}_m$ is the unit vector and the only m-th element equals 1. The partition function z_{mk}, the elements of the basis matrix, t_{ik}, and those of the activation matrix, v_{kj}, are updated as follows.

$$z_{mk} \leftarrow z_{mk} \sqrt{\frac{\Sigma_{i,j} |y_{ij,m}|^2 t_{ik} v_{kj} \left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\Sigma_{i,j} t_{ik} v_{kj} \left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}}. \quad (21)$$

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\Sigma_{j,m} |y_{ij,m}|^2 z_{mk} v_{kj} \left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\Sigma_{j,m} z_{mk} v_{kj} \left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}}. \quad (22)$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\Sigma_{i,m} |y_{ij,m}|^2 z_{mk} t_{ik} \left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\Sigma_{i,m} z_{mk} t_{ik} \left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}}. \quad (23)$$

From the equations above, we find the separated signals

by updating $\boldsymbol{W}_i$, z_{mk}, t_{ik}, and v_{kj} alternately and repeatedly. Finally, we restore the signal scale by applying a projection-back technique.

2.4.3. Combination of Postfiltering

In most cases of BSS- or NMF-based signal separation, a statistical postfilter is applied to attain improvement in human-voice enhancement performance. In this study, we propose to use a generalized minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [27]. The generalized MMSE-STSA estimator calculates the spectrum gain that minimizes the average squared error between the true target signal and the estimated signal given the a priori probability distribution of the signal (see [27] for more detailed algorithm). In the estimator, it is necessary to obtain the power spectrum variance of the nontarget signal, and we can use the separated output from ILRMA, $\sum_{m'} y_{ij,m'}$, where m' denotes the nontarget source components, for this purpose.

2.4.4. Further Extension

ILRMA assumes the existence of an inverse of the mixing system, i.e., the demixing matrix should be determined as a linear time-invariant system. Therefore, the proposed method often suffers from adverse effects caused by the time-variant nature of the mixing system. To address this problem, a noise-canceler-based compensation is developed as our ongoing research, where the time-domain noise component (inverse Fourier transform of $\sum_{m'} y_{ij,m'}$) is optimally subtracted from the noisy target component based on time-variant impulse response estimation (see [28] for more details).

Another possible extension is the introduction of a basis supervision. In the application of robot audition, we can often obtain a prototype of the ego-noise signal that can be used as training data in advance. This property is very suitable for embedding the supervision spectral bases into ILRMA, encouraging the rapid convergence of the algorithm [29].

2.5. Implementation of the Enhancement System

We implemented the two-stage human-voice enhancement system using two laptop computers. One was used for controlling the hose-shaped robot, capturing the video and audio streams, and conducting the online human-voice enhancement. The other was used for conducting the offline human-voice enhancement. The online and offline enhancements were conducted separately on these two computers so that the offline enhancement could use the full resource of the computer. The operating system for these computers was a Linux OS called Ubuntu 14.04.[1]

The entire system was implemented on Robot Operating System (ROS) [17]. ROS provides hardware abstraction and application programming interfaces (APIs) for message passing among multiple modules. A robot system constructed on ROS forms a network of executable programs called *nodes*. Each node communicates with other nodes via *topics*, which are data buses over which the nodes exchange messages. Each topic is named uniquely and the nodes communicate by publishing messages to a topic and subscribing to the topic. A ROS system can be easily extended to a multiple-computer system because this topic-based communication is implemented on (transmission control protocol/internet protocol) TCP/IP and the name resolution of each computer is automatically conducted by ROS with the topic name.

The online enhancement was implemented as a ROS node with a robot audition software called HARK[2] [30]. HARK provides various online signal processing modules, such as those needed for sound source localization, separation, and recognition. Because these modules are implemented by using C++ and connected with each other by function calls, HARK attains a real-time low-overhead processing. By using a graphic user interface (GUI) tool called HARK Designer, users of HARK can easily configure the connections among the modules to make a HARK system suitable for their robots. Although most of the separation and enhancement methods implemented in HARK cannot be used with a hose-shaped robot because they assume that microphone locations are known in advance [7, 19], we used HARK's fundamental functions such as Fourier transforms and ROS communications. We implemented online RPCA and median integration modules as HARK modules using C++ and a linear algebra library called Eigen3.[3] These modules were combined to form a single ROS node that enhances human voice in an online manner.

The offline enhancement, on the other hand, was implemented using MATLAB.[4] Because a longer time can be spent for the offline enhancement than the online one, we gave weight to the maintainability instead of the real-time processing. MATLAB provides functions for linear algebraic operations as its standard functions. Moreover, with a minimal change of the MATLAB source code, we can easily introduce multi-core processing and general-purpose computing on graphics processing units (GPGPU).

Figure 7 shows a diagram of the proposed two-stage human-voice enhancement system. The `audio capture node` captures the eight-channel synchronized audio signal with the microphone array on our robot, and publishes the signal to the `audio signal topic`. The HARK node, which performs the online human-voice enhancement, subscribes to the `audio signal topic` and publishes the enhanced signal to the `enhanced signal topic`. The audio signal published to the `enhanced signal topic` is played back by the `playback node`. To perform the offline enhancement, the `WAV file saver node` stores the audio stream published to the `audio signal topic`,

1. http://www.ubuntu.com/ [Accessed July 24, 2016]

2. Honda Research Institute Japan Audition for Robots with Kyoto Univ.
3. http://eigen.tuxfamily.org/ [Accessed July 24, 2016]
4. http://www.mathworks.com/products/matlab/ [Accessed July 24, 2016]

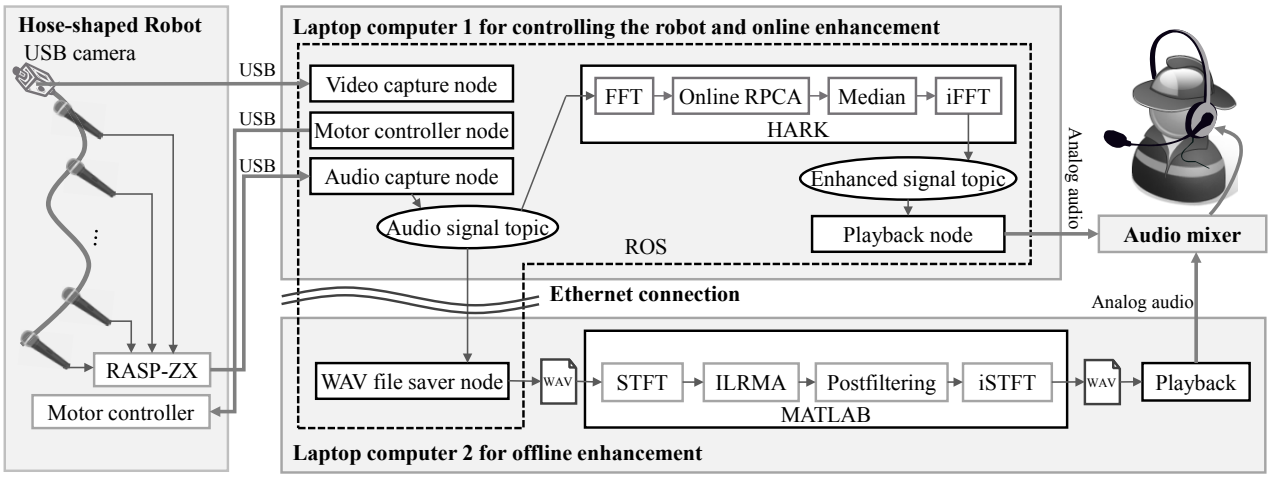


Fig. 7. Configuration of the two-stage human-voice enhancement system.

and writes a WAV file of the last five seconds when this node is required to stop. When the operator wants an offline enhancement signal, it is obtained by quitting the `WAV file saver node`, running the offline enhancement, and playing back the resulting WAV file. Because the online and offline processes are conducted on different computers, their output audio signals are downmixed for presentation to the remote operator.

3. Experimental Evaluation

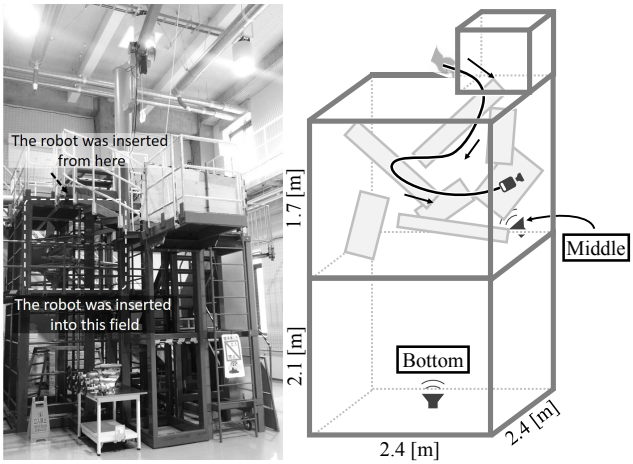
In this evaluation, the performances of online and offline enhancement methods were first evaluated separately with simulated recordings because the inputs of these two methods are different. The online enhancement is used for a streaming input and the offline enhancement is used for a short recording extracted from the streaming input. After independent evaluation of each method, they are compared using an actual recording captured in a simulated collapsed building.

3.1. Evaluation 1: Online Enhancement

This subsection reports the performance of the online human-voice enhancement.

3.1.1. Experimental Settings

This experiment was conducted in a mockup rubble field as shown in **Fig. 8(a)**. Wooden obstacles and several plastic plates were placed in the upper half space of the field. Our robot was inserted into this space from the top of the field. In this evaluation the ego-noise and target voice were recorded separately and then mixed at SNRs from -20 dB to $+5$ dB. The ego-noise was recorded for 60 seconds while inserting the robot into the rubble. The arrangement of the robot and the loudspeaker from which the target voice was emitted is shown in **Fig. 8(b)**. We tested two loudspeaker positions, i.e., middle and bottom. The target voice data consisted of two recordings of male



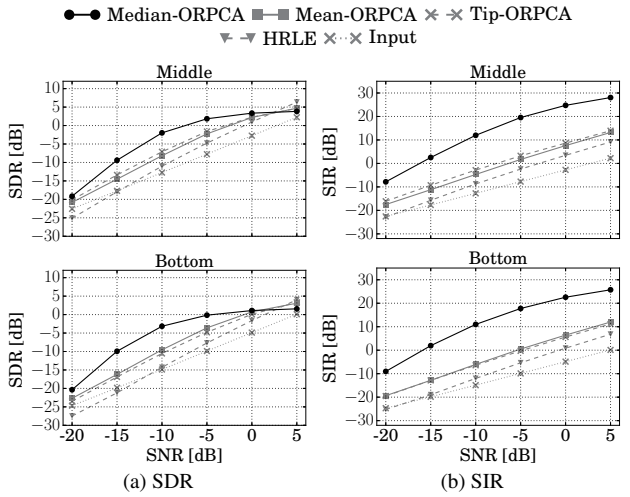
(a) Photograph of the field (b) Configuration of the field

Fig. 8. Mockup rubble field used in Evaluation 1. Two arrangements of the loudspeaker were tested.

voices and two recordings of female voices, each with duration of one minute. Low-noise target voice signals were generated by convoluting the clean voice recorded in an anechoic chamber and the impulse response recorded with the loudspeaker.

The human-voice enhancement performance was evaluated using signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) [31]. SDR measures the overall quality of the retrieved enhancement result, while SIR measures how much the interference due to the ego-noise is suppressed. They were measured using a Python library called MIR-EVAL [32].

The proposed method (Median-ORPCA) was compared with the following three methods: histogram-based recursive level estimation (HRLE) [8], Tip-ORPCA, and Mean-ORPCA. HRLE is one of the conventional spectrum subtraction methods. Because HRLE works with a single-channel input, we evaluated the HRLE result of the tip (8th) microphone. We used HRLE implemented in HARK. Tip-ORPCA was ORPCA applied to the recordings of the tip microphone. The results of Mean-ORPCA



(a) SDR (b) SIR

Fig. 9. SDR and SIR results obtained in Evaluation 1.

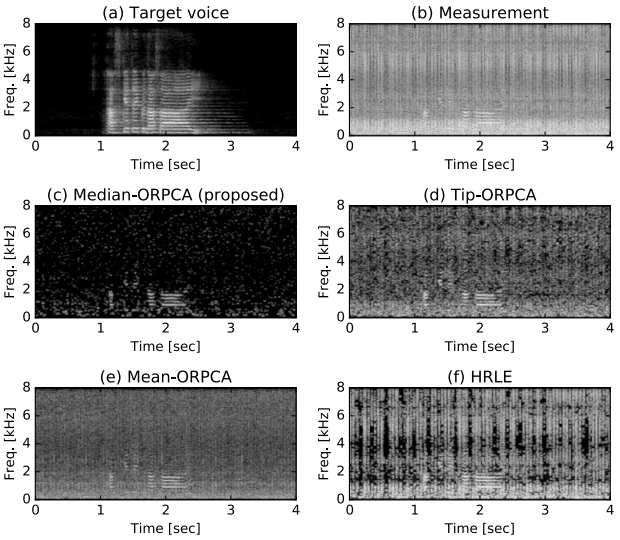


Fig. 10. Examples of online enhancement results obtained when the loudspeaker was at the middle of the field ("Middle") and the SNR was set to -10 dB. A female voice was emitted between 1.0 and 2.5 sec.

were the results obtained by taking mean values of all the microphone results of online RPCA. The frame length of STFT was set to 512 samples, and the frame shift length was 160 samples. The λ_1 and λ_2 of ORPCA were set to $8.0/257 \times 10^{-3}$. The other parameters of the proposed method were decided experimentally, and those of HRLE were set to the default values of the HARK implementation.

3.1.2. Experimental Results

As shown in **Fig. 9(a)**, in both the two loudspeaker conditions, the SDRs of the proposed method were higher compared to those of the other methods at the SNR conditions between -20 dB and 0 dB. Moreover, the SIR of the proposed method was more than 8.2 dB higher than those of the other methods under all the test conditions.

The SIR measures how much the ego-noise is suppressed. As shown in **Fig. 10**, the suppressed spectrogram of the proposed method contains less noise than those of the other suppressed spectrograms. As shown by the vertical-stripe patterns in **Fig. 10(b)**, the ego-noise changes with a frequency of 30 Hz. Because HRLE represents the ego-noise as a single-spectrogram template, it leaves the fluctuation residuals of the ego-noise as vertical-stripe patterns (**Fig. 10(f)**). The result of the proposed method, on the other hand, contains less noise than those of the other methods (**Fig. 10(c)**).

3.2. Evaluation 2: Offline Enhancement

The offline human-voice was evaluated in the same mockup rubble field as Evaluation 1. In this evaluation, the input signals were five-second signals including human-voice signals and ego-noise of the robot.

3.2.1. Experimental Settings

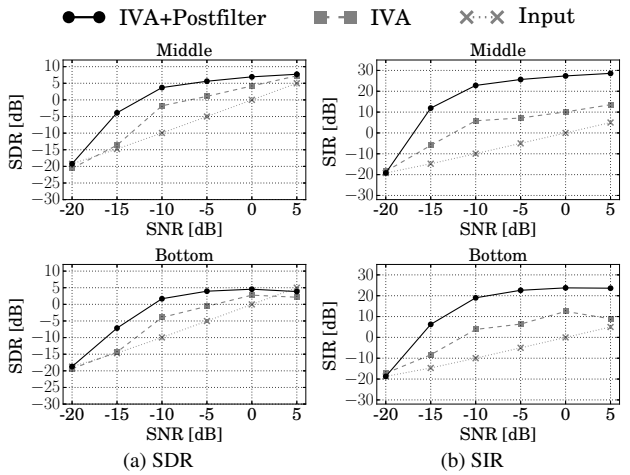
We developed a computationally efficient offline system to consider the real-field robot operation that requires a feasible calculation. To achieve this, in this experiment, the number of bases for each source in ILRMA was set to one which is equal to the case of independent vector analysis (IVA) [33]. Moreover, as for the postfilter, we set specific parameters in the generalized MMSE-STSA to obtain a "spectral-subtraction-type" gain estimator. Then, a smoothing technique [34] was applied to improve the sound quality. The above mentioned simplifications significantly reduced the computational cost, while avoiding serious degradation in the separation performance [35].

The flexible robot had eight microphones with unknown locations, which recorded the observed signals consisting of one target voice signal and ego-noise. The target signal was imitated using clean male and female voice signals with real-recorded impulse responses from the source to each microphone. The multichannel ego-noise signals were independently recorded with the actual dynamics of the robot, and were added into the target voice signals.

The rest of the experimental conditions is as follows. The total length of the observed signals was five seconds. The ego-noise and target voice signals were mixed at SNRs varying from -20 dB to $+5$ dB. The frame length of STFT was set to 2048 samples, and the frame shift length was 512 samples. The number of iterations for parameter updating in ILRMA (IVA) was 100.

3.2.2. Experimental Results

Figure 11 shows the SDR and SIR scores [31] for each condition, where we compare the quality of signals of observation ("Input"), simple IVA ("IVA"), and the proposed method ("IVA+Postfilter"). We can confirm that IVA can increase the SDR and SIR scores to some extent, particularly for the case of low input SNR condition, i.e., -10 dB. In addition, the proposed method significantly outperforms the other methods, resulting in 1.8–9.8 dB improvement in SDR and 11.3–18.5 dB improvement in



(a) SDR (b) SIR

Fig. 11. SDR and SIR results obtained in Evaluation 2.

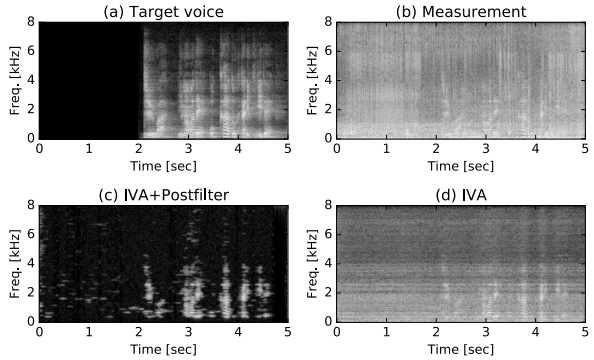


Fig. 12. Examples of offline enhancement results obtained when the loudspeaker was at the middle of the field ("Middle") and the SNR was set to −10 dB. A male voice was emitted between 2.0 and 5.0 sec.

SIR from IVA when the input SNR was between −15 dB and 0 dB.

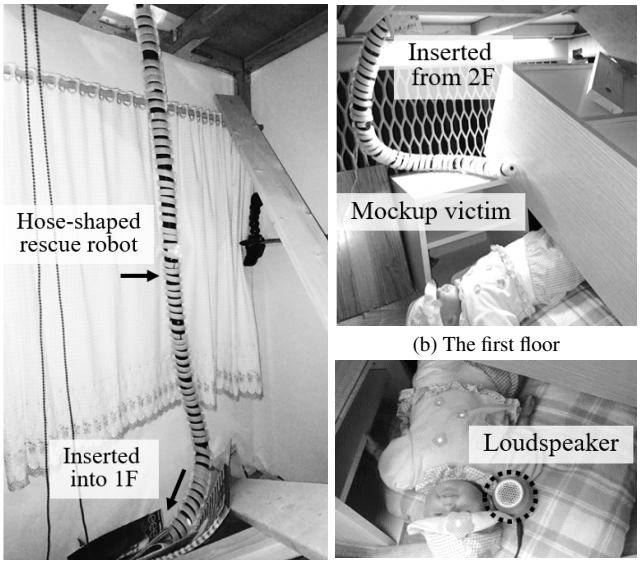
Figure 12 shows an example of spectrograms obtained in this experiment. This clarifies the significant contribution of the postfilter and shows the efficacy of the proposed combination of BSS and postfiltering.

3.3. Evaluation 3: Comparison of Online and Offline Enhancements

We compared the proposed online and offline human-voice enhancement methods using actual data recorded in a simulated collapsed building.

3.3.1. Experimental Settings

This experiment was conducted in a simulated collapsed building at Tohoku University, Miyagi, Japan in 2016. The simulated building consisted of three sections: 1) an attic, 2) a second floor, and 3) a first floor. As shown in **Fig. 13**, the hose-shaped rescue robot was inserted into the second floor from the attic section and penetrated into the first floor. A mannequin mockup victim was placed in the first floor. This building simulated a Japanese wooden house collapsed by an earth quake. In the second floor



(a) The second floor (b) The first floor (c) Mockup victim and loudspeaker

Fig. 13. The robot was placed into a simulated collapsed building with a loudspeaker next to a mannequin.

Table 1. PESQ results in MOS-LQO.

Input signal (tip microphone)	Online enhancement (Sec. 2.3)	Offline enhancement (Sec. 2.4)
1.092	1.126	1.229

were six fallen wooden beams, and in the first floor a wooden shelf had fallen on the mannequin. A loudspeaker was placed next to the mannequin for emitting a target voice signal (**Fig. 13(c)**). The target voice signal was a recording of a female voice that was 30-second long. We recorded the target voice signal while the vibrators of the robot were turned on.

The online and offline systems were compared using perceptual evaluation of speech quality (PESQ) [36]. PESQ is designed for evaluating the speech quality in telephones and telecommunication. As the original PESQ only measures the frequency band from 300 Hz to 3.5 kHz, we used an extension of PESQ called wideband-PESQ which measures the frequency band from 100 Hz to 7.0 kHz. The wideband-PESQ performance is represented in mean opinion scores of listening quality objective (MOS-LQO), which range from 1.02 to 4.56.

3.3.2. Experimental Results

Table 1 shows the PESQ results of the proposed online and offline enhancement methods. The online enhancement improves by 0.034 in MOS-LQO from the raw input recording. Moreover, the offline enhancement improves by 0.137 in MOS-LQO from the raw input. As shown in **Fig. 14**, both the online and offline methods reduce the ego-noise of the robot and enhance the human-voice. The result of the online enhancement has musical noise appearing as salt-and-pepper noise in the spectrogram. The result of the offline enhancement has much less musical noise than the online result.

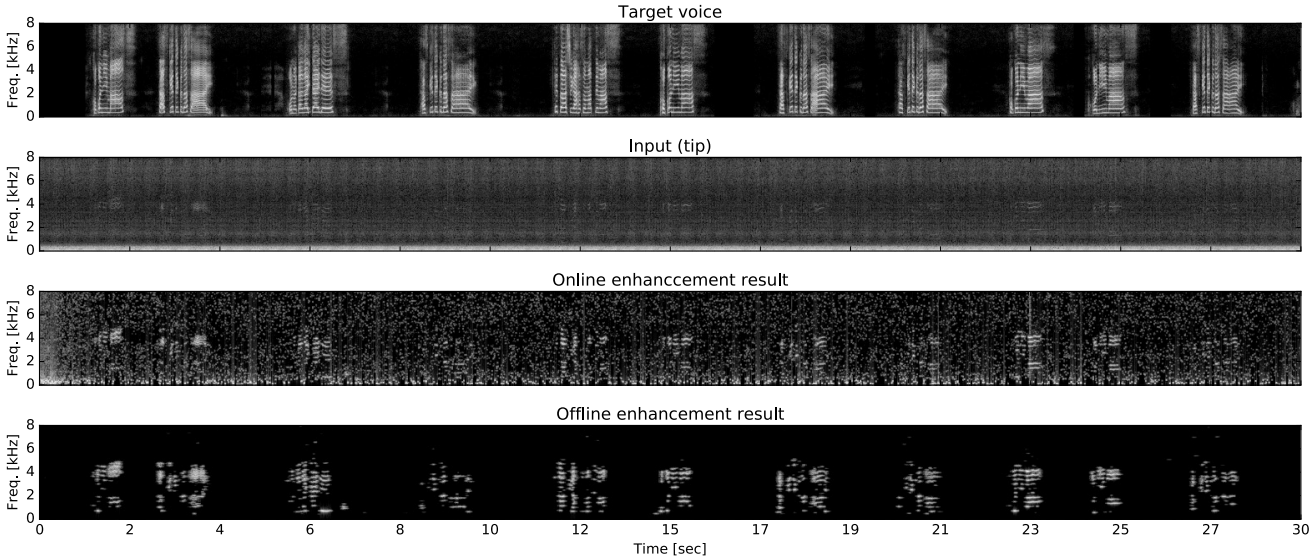


Fig. 14. Results of online and offline enhancement of data recorded in a simulated collapsed building.

Because of the difference in processing time, our system usually outputs online enhancement results, and outputs an offline result of a short-time recording only when the operator wants a clearer result. It should be noted that in the offline enhancement, our target value for the real-time factor is less than four (e.g., five-second data should be processed within 20 seconds), and consequently we should limit the number of iterations in the IVA part within 20. The resultant computational time spent in this signal separation experiment was approximately 17 seconds for the IVA part and 1 second for the postfiltering part or a total of less than 20 seconds using an Intel Core i5-5200U (2-core, 2.20 GHz) laptop computer. Thus, this system achieves the target real-time factor of four. On the other hand, in the online enhancement, our target value for the real-time factor is less than one. The elapsed time for the online enhancement of a 60-second input signal using an Intel Core i7-4500 CPU (2-core, 1.8 GHz) laptop computer was 41 seconds. This value was small enough to allow the online enhancement to work in real time.

4. Conclusion

This paper presented the two-stage human-voice enhancement system for a hose-shaped rescue robot. Our system combines online and offline human-voice enhancement to achieve low latency and high quality. The online enhancement is used for searching for a victim in real-time while the offline one facilitates scrutiny by listening to highly enhanced human voices. The online enhancement is conducted by applying online RPCA to each microphone recording and combining the results. The offline enhancement is conducted by applying ILRMA and postfiltering. These two methods were integrated on ROS for attaining a real-time system, multi-computer processing, and buffering audio recordings. The experimental results showed that both online and offline enhancement methods outperformed the conventional methods.

We have two directions for future research, namely, 1) developing an efficient GUI, and 2) developing an automatic voice activity detection (VAD). As the current system provides only a command line interface and enhanced audio signal, the usability of our system could be improved by implementing a GUI that can switch the online and offline enhancement and visualize the enhancement results. Furthermore, by conducting VAD to the online enhancement results, our system could be able to indicate the existence of a victim to the operator, and speculatively conduct the offline enhancement before the operator manually switches to it. Our system will be further improved by visualizing the VAD results and the speculative offline results.

Acknowledgements

This study was partially supported by ImPACT Tough Robotics Challenge and by JSPS KAKENHI No. 24220006 and No. 15J08765.

References:

[1] J. Fukuda, M. Konyo, E. Takeuchi, and S. Tadokoro, "Remote vertical exploration by Active Scope Camera into collapsed buildings," Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 1882-1888, 2014.

[2] A. Kitagawa, H. Tsukagoshi, and M. Igarashi, "Development of Small Diameter Active Hose-II for Search and Life-prolongation of Victims under Debris," J. of Robotics and Mechatronics, Vol.15, No.5, pp. 474-481, 2003.

[3] H. Namari, K. Wakana, M. Ishikura, M. Konyo, and S. Tadokoro, "Tube-type active scope camera with high mobility and practical functionality," Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 3679-3686, 2012.

[4] S. Tadokoro, R. Murphy, S. Stover, W. Brack, M. Konyo, T. Nishimura, and O. Tanimoto, "Application of active scope camera to forensic investigation of construction accident," Proc. of IEEE Workshop on Advanced Robotics and its Social Impacts, pp. 47-50, 2009.

[5] R. R. Murphy, "Disaster Robotics," MIT Press, 2014.

[6] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, and H. G. Okuno, "Human-Voice Enhancement based on Online RPCA for a Hose-shaped Rescue Robot with a Microphone Array," Proc. of IEEE Int. Symposium on Safety, Security, and Rescue Robotics, pp. 1-6, 2015.
[7] G. Ince, K. Nakadai, T. Rodemann, J. Imura, K. Nakamura, and H. Nakajima, "Incremental learning for ego noise estimation of a robot," Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 131-136, 2011.
[8] H. Nakajima, G. Ince, K. Nakadai, and Y. Hasegawa, "An easily-configurable robot audition system using histogram-based recursive level estimation," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 958-963, 2010.
[9] T. Tezuka, T. Yoshida, and K. Nakadai, "Ego-motion Noise Suppression for Robots Based on Semi-Blind Infinite Non-negative Matrix Factorization," Proc. of IEEE Int. Conf. on Robotics and Automation, pp. 6293-6298, 2014.
[10] B. Cauchi, S. Goetze, and S. Doclo, "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization," Proc. of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments, pp. 28-33, 2012.
[11] J. Feng, H. Xu, and S. Yan, "Online robust PCA via stochastic optimization," Proc. of Advances in Neural Information Processing Systems, pp. 404-412, 2013.
[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization," IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol.24, No.9, pp. 1626-1641, 2016.
[13] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Trans. on Audio, Speech, and Language Processing, Vol.18, No.3, pp. 550-563, 2010.
[14] T. Kim, "Real-Time Independent Vector Analysis for Convolutive Blind Source Separation," IEEE Trans. on Circuits and Systems I: Regular Papers, Vol.57, No.7, pp. 1431-1438, 2010.
[15] C. Sun, Q. Zhang, J. Wang, and J. Xie, "Noise Reduction Based on Robust Principal Component Analysis," J. of Computational Information Systems, Vol.10, No.10, pp. 4403-4410, 2014.
[16] H. Kameoka, T. Yoshioka, M. Hamamura, J. L. Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," Proc. of 9th Int. Conf. on Latent Variable Analysis and Signal Separation, pp. 245-253, 2010.
[17] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: an open-source Robot Operating System," Proc. of IEEE ICRA workshop on open source software, pp. 1-5, 2009.
[18] B. Shneiderman and C. Plaisant, "Designing the user interface," Addison Wesley, 1998.
[19] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind source separation with parameter-free adaptive step-size method for robot audition," IEEE Trans. on audio, speech, and language processing, Vol.18, No.6, pp. 1476-1485, 2010.
[20] J. C. Murray, H. R. Erwin, and S. Wermter, "Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks," Neural Networks, Vol.22, No.2, pp. 173-189, 2009.
[21] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of moving sound sources," Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1-5, 2015.
[22] Z. Chen and D. P. W. Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1-4, 2013.
[23] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," J. of the ACM, Vol.58, No.3, p. 11, 2011.
[24] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," SIAM review, Vol.52, No.3, pp. 471-501, 2010.
[25] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," IEEE Trans. on Speech and Audio Processing, Vol.11, No.2, pp. 109-116, 2003.
[26] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, Vol.22, pp. 21-34, 1998.
[27] C. Breihaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 4037-4040, 2008.
[28] M. Ishimura, S. Makino, T. Yamada, N. Ono, and H. Saruwatari, "Noise reduction using independent vector analysis and noise cancellation for a hose-shaped rescue robot," Proc. of Int. Workshop on Acoustic Signal Enhancement, 2016.
[29] H. Saruwatari, K. Takata, N. Ono, and S. Makino, "Flexible microphone array based on multichannel nonnegative matrix factorization and statistical signal estimation," Proc. of Int. Congress on Acoustics, number ICA2016-312, 2016.
[30] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System HARK – Open Source Software for Listening to Three Simultaneous Speakers," Advanced Robotics, Vol.24, No.5-6, pp. 739-761, 2010.
[31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE Trans. on audio, speech, and language processing, Vol.14, No.4, pp. 1462-1469, 2006.
[32] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: a transparent implementation of common MIR metrics," Proc. of the 15th Int. Society for Music Information Retrieval Conf., pp. 367-372, 2014.
[33] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 189-192, 2011.
[34] T. Esch and P. Vary, "Efficient musical noise suppression for speech Enhancement Systems," Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 4409-4412, 2009.
[35] M. Takakusaki, D. Kitamura, N. Ono, T. Yamada, S. Makino, and H. Saruwatari, "Ego-noise reduction for a hose-shaped rescue robot using determined rank-1 multichannel nonnegative matrix factorization," Proc. of Int. Workshop on Acoustic Signal Enhancement, 2016.
[36] "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," ITU-T Recommendation P.862.2, 2005.

Name:
Yoshiaki Bando

Affiliation:
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
JSPS Research Fellow DC1

Address:
Room 417, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:
2014 Received M.Inf. degree from Graduate School of Informatics, Kyoto University
2015- Ph.D. Candidate, Graduate School of Informatics, Kyoto University

Main Works:
• "Posture estimation of hose-shaped robot by using active microphone array," Advanced Robotics, Vol.29, No.1, pp. 35-49, 2015 (Advanced Robotics Best Paper Award).
• "Variational Bayesian Multi-channel Robust NMF for Human-voice Enhancement with a Deformable and Partially-occluded Microphone Array," European Signal Processing Conf. (EUSIPCO), pp. 1018-1022, 2016.
• "Microphone-accelerometer based 3D posture estimation for a hose-shaped rescue robot," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 5580-5586, 2015.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE)
• The Robotics Society of Japan (RSJ)
• Information Processing Society of Japan (IPSJ)

Name:
Hiroshi Saruwatari

Affiliation:
Professor, The University of Tokyo

Address:
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Brief Biographical History:
1993- Scientist, SECOM IS Lab.
2000- Associate Professor, Nara Institute of Science and Technology
2014- Professor, The University of Tokyo

Main Works:
• "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," IEEE Trans. on Audio, Speech, and Language Processing, Vol.14, No.2, pp. 666-678, 2006.

Membership in Academic Societies:
• The Institute of Electrical and Electronics Engineers (IEEE)
• The Institute of Electronics, Information and Communication Engineers (IEICE)
• Acoustic Society of Japan (ASJ)

Name:
Shoji Makino

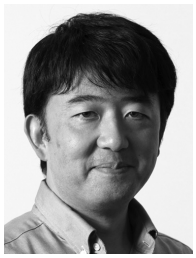
Affiliation:
Professor, University of Tsukuba

Address:
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan

Brief Biographical History:
1981- Research Engineer, NTT Electrical Communication Laboratory
2003- Executive Manager, NTT Communication Science Laboratories
2009- Professor, University of Tsukuba

Main Works:
• "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," IEEE Trans. Speech Audio Processing, Vol.11, No.2, pp. 109-116, Mar. 2003.

Membership in Academic Societies:
• The Institute of Electrical and Electronics Engineers (IEEE), Fellow
• European Association for Signal Processing (EURASIP)
• The Institute of Electronics, Information, and Communication Engineers (IEICE)



Name:
Nobutaka Ono

Affiliation:
National Institute of Informatics

Address:
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Brief Biographical History:
2001- Research Associate, The University of Tokyo
2005- Lecturer, The University of Tokyo
2011- Associate Professor, National Institute of Informatics

Main Works:
• S. Miyabe, N. Ono, and S. Makino, "Blind Compensation of Interchannel Sampling Frequency Mismatch for Ad hoc Microphone Array Based on Maximum Likelihood Estimation," Elsevier Signal Processing, Vol.107, pp. 185-196, Feb. 2015.

Membership in Academic Societies:
• The Institute of Electrical and Electronics Engineers (IEEE)
• Acoustic Society of Japan (ASJ)
• The Institute of Electronics, Information and Communication Engineers (IEICE)

Name:
Katsutoshi Itoyama

Affiliation:
Assistant Professor, Speech and Audio Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Address:
Room 417, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:
2011- Received Ph.D. degree from Graduate School of Informatics, Kyoto University
2011- Assistant Professor, Graduate School of Informatics, Kyoto University

Main Works:
• "Query-by-Example Music Information Retrieval by Score-Informed Source Separation and Remixing Technologies," EURASIP J. on Advances in Signal Processing, Vol.2010, No.1 pp. 1-14, January 17, 2011.

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

Name:
Daichi Kitamura

Affiliation:
Ph.D. Candidate, SOKENDAI (The Graduate University for Advanced Studies)

Address:
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan

Brief Biographical History:
2014 Received M.E. degree in Engineering from Nara Institute of Science and Technology
2014- Ph.D. Course, Department of Informatics, School of Multidisciplinary Sciences, SOKENDAI

Main Works:
• "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol.24, No.9, pp. 1626-1641, Sep. 2016.
• "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol.23, No.4, pp. 654-669, Apr. 2015.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE) Signal Processing Society (SPS)
• The Institute of Electronics, Information and Communication Engineers (IEICE)
• Acoustical Society of Japan (ASJ)

Name:
Masaru Ishimura

Affiliation:
Graduate School of Systems and Information Engineering, University of Tsukuba

Address:
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan

Brief Biographical History:
2015- Graduate School of Systems and Information Engineering, University of Tsukuba

Main Works:
• "Noise reduction using independent vector analysis and noise cancellation for a hose-shaped rescue robot," Proc. IWAENC2016, pp. 1-5, Sept. 2016.

Name:
Moe Takakusaki

Affiliation:
Graduate School of Systems and Information Engineering, University of Tsukuba

Address:
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan

Brief Biographical History:
2016- Graduate School of Systems and Information Engineering, University of Tsukuba

Main Works:
• "Ego-noise reduction for a hose-shaped rescue robot using determined rank-1 multichannel nonnegative matrix factorization," Proc. IWAENC2016, pp. 1-4, Sept. 2016.

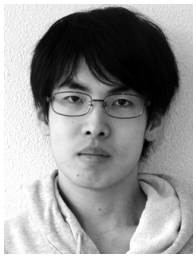
Name:
Narumi Mae

Affiliation:
Graduate School of Systems and Information Engineering, University of Tsukuba

Address:
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan

Brief Biographical History:
2016- Graduate School of Systems and Information Engineering, University of Tsukuba

Main Works:
• "Ego noise reduction for hose-shaped rescue robot combining independent low-rank matrix analysis and noise cancellation," Proc. APSIPA2016, pp. 1-6, Dec. 2016.



Name:
Kouei Yamaoka

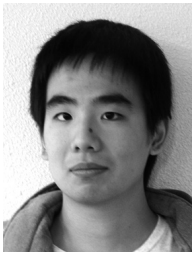
Affiliation:
School of Informatics, University of Tsukuba

Address:
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan

Brief Biographical History:
2013- School of Informatics, University of Tsukuba

Main Works:
• "Performance of maximum SNR beamformer based on virtual increase of channels in reverberant environments," Autumn Meeting of the Acoustical Society of Japan, pp. 379-382, Sept. 2016 (in Japanese).

Membership in Academic Societies:
• Acoustical Society of Japan (ASJ)



Name:
Yutaro Matsui

Affiliation:
School of Informatics, University of Tsukuba

Address:
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan

Brief Biographical History:
2013- School of Informatics, University of Tsukuba

Main Works:
• "Multiple far noises suppression under real environments with transfer-function-gain NMF," Proc. Signal Processing Symposium, pp. 231-235, Nov. 2016 (in Japanese).

Name:
Yuichi Ambe

Affiliation:
Researcher, Graduate School of Information Sciences, Tohoku University

Address:
6-6-01 Aramaki Aza Aoba, Aoba-ku, Sendai 980-8579, Japan

Brief Biographical History:
2016- Researcher, Graduate School of Information Sciences, Tohoku University

Main Works:
• Y. Ambe, T. Nachstedt, P. Manoonpong, F. Wörgötter, S. Aoi, and F. Matsuno, "Stability analysis of a hexapod robot driven by distributed nonlinear oscillators with a phase modulation mechanism," Proc. of the 2013 Int. Conf. on Robotic System (IROS), pp. 5087-5092, 2013.
• Y. Ambe and F. Matsuno, ""Leg-grope walk"-strategy for walking on fragile irregular slopes as a quadruped robot by force distribution," ROBOMECH J., 2016.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE)
• The Robotics Society of Japan (RSJ)
• The society of Instrument and Control Engineers (SICE)

Name:
Masashi Konyo

Affiliation:
Associate Professor, Graduate School of Information Sciences, Tohoku University

Address:
6-6-01 Aramaki Aza Aoba, Aoba-ku, Sendai 980-8579, Japan

Brief Biographical History:
2004- Research Associate, The 21st Century COE Program, Keio University
2005- Assistant Professor, Graduate School of Information Sciences, Tohoku University
2009- Associate Professor, Graduate School of Information Sciences, Tohoku University

Main Works:
• "Vibrotactile Stimuli Applied to Finger Pads as Biases for Perceived Inertial and Viscous Loads," IEEE Trans. on Haptics, Vol.4, No.4, pp. 307-315, 2011.
• "Ciliary Vibration Drive Mechanism for Active Scope Cameras," J. of Robotics and Mechatronics, Vol.20, No.3, pp. 490-499, 2008.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE)
• The Japan Society of Mechanical Engineers (JSME)
• The Robotics Society of Japan (RSJ)

Name:
Satoshi Tadokoro

Affiliation:
Professor, Tohoku University
Program Manager, Japan Cabinet Office ImPACT Tough Robotics Challenge

Address:
6-6-01 Aramaki Aza Aoba, Aoba-ku, Sendai 980-8579, Japan

Brief Biographical History:
2002- President, International Rescue System Institute
2005- Professor, Tohoku University
2014- Program Manager, JCO ImPACT-TRC
2016- President, IEEE Robotics and Automation Society

Main Works:
• "On robotic rescue facilities for disastrous earthquakes – from the Great Hanshin-Awaji (Kobe) Earthquake –," J. of Robotics and Mechatronics, Vol.9, No.1, pp. 46-56, 1997.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE), Fellow
• The Japan Society of Mechanical Engineers (JSME), Fellow
• The Robotics Society of Japan (RSJ), Fellow

Name:
Kazuyoshi Yoshii

Affiliation:
Senior Lecturer, Speech and Audio Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Address:
Room 412, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:
2008- Received Ph.D. degree from Graduate School of Informatics, Kyoto University
2008- Research Scientist, Information Technology Research Institute (ITRI), National Institute of Advanced Industrial Science and Technology (AIST)
2013- Senior Researcher, AIST
2014- Senior Lecturer, Graduate School of Informatics, Kyoto University

Main Works:
• "A Nonparametric Bayesian Multipitch Analyzer Based on Infinite Latent Harmonic Allocation," IEEE Trans. on Audio, Speech, and Language Processing, Vol.20, No.3, pp. 717-730, 2012.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE)
• Information Processing Society of Japan (IPSJ)
• The Institute of Electronics, Information, and Communication Engineers (IEICE)

Name:
Hiroshi G. Okuno

Affiliation:
Professor, Graduate School of Science and Engineering, Waseda University
Professor Emeritus, Kyoto University

Address:
Lambdax Bldg 3F, 2-4-12 Okubo, Shinjuku, Tokyo 169-0072, Japan

Brief Biographical History:
1996 Received Ph.D. of Engineering from Graduate School of Engineering, The University of Tokyo
2001-2014 Professor, Graduate School of Informatics, Kyoto University
2014- Professor, Graduate School of Science and Engineering, Waseda University

Main Works:
• "Design and Implementation of Robot Audition System "HARK"," Advanced Robotics, Vol.24, No.5-6, pp. 739-761, 2010.
• "Computational Auditory Scene Analysis," Lawrence Erlbaum Associates, Mahmoh, NJ, 1998.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE), Fellow
• The Japanese Society for Artificial Intelligence (JSAI), Fellow
• Information Processing Society Japan (IPSJ), Fellow
• The Robotics Society of Japan (RSJ), Fellow