Variational Bayesian Multi-channel Robust NMF for Human-voice Enhancement with a Deformable and Partially-occluded Microphone Array

Yoshiaki Bando*, Katsutoshi Itoyama*, Masashi Konyo†, Satoshi Tadokoro†,
Kazuhiro Nakadai‡, Kazuyoshi Yoshii*, and Hiroshi G. Okuno§
*Graduate School of Informatics, Kyoto University †Graduate School of Information Science, Tohoku University
‡Graduate School of Information Science and Engineering, Tokyo Institute of Technology
§Graduate Program for Embodiment Informatics, Waseda University

***Abstract*—This paper presents a human-voice enhancement method for a deformable and partially-occluded microphone array. Although microphone arrays distributed on the long bodies of hose-shaped rescue robots are crucial for finding victims under collapsed buildings, human voices captured by a microphone array are contaminated by non-stationary actuator and friction noise. Standard blind source separation methods cannot be used because the relative microphone positions change over time and some of them are occasionally shaded by rubble. To solve these problems, we develop a Bayesian model that separates multi-channel amplitude spectrograms into sparse and low-rank components (human voice and noise) without using phase information, which depends on the array layout. The voice level at each microphone is estimated in a time-varying manner for reducing the influence of the shaded microphones. Experiments using a 3-m hose-shaped robot with eight microphones show that our method outperforms conventional methods by the signal-to-noise ratio of 2.7 dB.**

I. Introduction

Among the rescue robots developed for gathering information in places humans or animals cannot go are hose-shaped robots specialized for penetrating into narrow gaps under collapsed buildings [1], [2]. The Active Scope Camera, for example, can move forward by vibrating cilia covering its long, thin, and flexible body [2]. Using a microphone array and a tip camera equipped on the robot (Fig. 1), a robot operator searches for victims. These microphones are distributed along the body to avoid all of them being covered by obstacles [3].

Human voices captured by a hose-shaped robot are contaminated by non-stationary ego-noise (*e.g.*, motor and friction noise). The naïve "stop-and-listen" strategy prevents a robot operator from finding victims from wide areas as quickly as possible. Conventional methods of ego-noise suppression based on pre-trained noise dictionaries [4]–[7] cannot be used because the ego-noise changes over time according to the robot's movements and surrounding materials.

In the aspect of microphone-array processing, human-voice enhancement for a hose-shaped robot faces two problems:

1) Deformable layout of microphones: The relative microphone positions change over time because of the vibration and deformation of the robot body.

2) Partial occlusion of microphones: Some of the microphones often fail to capture human voice when they are shaded by rubble around the robot.

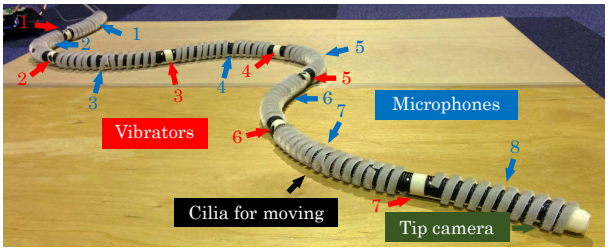


Fig. 1. A hose-shaped rescue robot with an eight-channel microphone array.

One possibility to solve the first problem is to estimate the time-varying body shape [8]. The accuracy of the shape estimation is, however, insufficient for phase-based source separation methods that solve the second problem [9], [10].

In this paper we present a Bayesian human-voice enhancement method for a deformable and partially-occluded microphone array. Our method works on the amplitude domain [11] to avoid using unreliable phase information sensitively affected by the array layout. Multi-channel amplitude spectrograms are separated into sparse and low-rank components (human voice and noise) without prior training [12]–[14]. To reduce the influence of the shaded microphones, the voice level at each microphone is estimated in a time-varying manner.

II. Related Work

This section discusses conventional methods of phase-based and amplitude-based source separation.

A. Phase-based source separation methods

Blind source separation based on the phase differences between the microphones can be used without prior knowledge about microphones and sources [9], [15]–[18]. Multi-channel nonnegative matrix factorization (MNMF) [16]–[18], for example, decomposes given multi-channel complex spectrograms into multiple low-rank source spectrograms and their transfer functions. Kounades-Bastian *et al.* [18] extended MNMF for moving sources by assuming a Markov chain of time-varying transfer functions. Since these methods assume that each source is observed by all the microphones, the separation performance is degraded when each microphone is contaminated by ego-noise specific to that microphone. In addition,

*Demo page: http://sap.ist.i.kyoto-u.ac.jp/members/yoshiaki/demo/eusipco2016

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

these methods are degraded when the transfer functions are changed finely and randomly by the actuator vibration.

B. Amplitude-based source separation methods

One way to avoid estimating the time-varying transfer functions of sound sources is to perform multi-channel source separation in the amplitude domain. Chiba *et al.* [11], for example, proposed a source separation method for a set of asynchronous microphone arrays. The phase differences between asynchronous microphones are gradually changed over time by the clock rate differences between those microphones. Since the phase information is unreliable, NMF is applied to multi-channel amplitude spectrograms under a limited condition that the volume level ratios of each sound source among channels is known in advance. It is, however, difficult to know such information in rubble-existing environments.

Low-rank and sparse decomposition is a popular approach to suppressing non-stationary noise and enhancing human voice without prior training [12]–[14], [19], [20]. Robust principal component analysis (RPCA), for example, can be used for decomposing a single-channel amplitude spectrogram into low-rank and sparse amplitude spectrograms corresponding to noise and human voice [12], [13]. RPCA can be extended in a Bayesian manner to deal with uncertainty of latent low-rank and sparse components [21], [22]. To estimate background and foreground images from video streams, Ding *et al.* [21] proposed a method that imposes a Markovian constraint on sparse components of video images (foreground images) for reducing salt-and-pepper noise. Babacan *et al.* [22] derived a variational Bayesian (VB) algorithm for Bayesian RPCA (VB-RPCA) to reduce the computational cost. Application of RPCA to audio and image data, however, is not physically justified because RPCA allows input, low-rank, and sparse amplitude spectrograms or images to take negative values.

To analyze audio spectrograms or video images, robust NMF (RNMF) has been studied for decomposing an input nonnegative matrix into nonnegative low-rank and sparse matrices [19], [20]. Sun *et al.* [20] proposed a variant of RNMF having a cost function based on the Kullback-Leibler divergence, which has widely been used in NMF-based audio source separation. Like Bayesian RPCA, Bayesian interpretation of RNMF is expected to enhance further extensions for multi-channel audio data.

III. Variational Bayesian Multi-channel RNMF

This section describes the proposed method that decomposes multi-channel audio data into channel-wise low-rank components and sparse components common to all the channels as shown in Fig. 2. The volume level of the common sparse components is estimated at each microphone. We first formulate variational Bayesian RNMF (VB-RNMF) that is a counterpart of VB-RPCA for single-channel audio data. Its multi-channel extension (VB-MRNMF) is then formulated.

A. Problem statement

The hose-shaped robot assumed in this paper has microphones distributed along its body as shown in Fig. 1. The microphones indices range from 1 at the operator's hand position

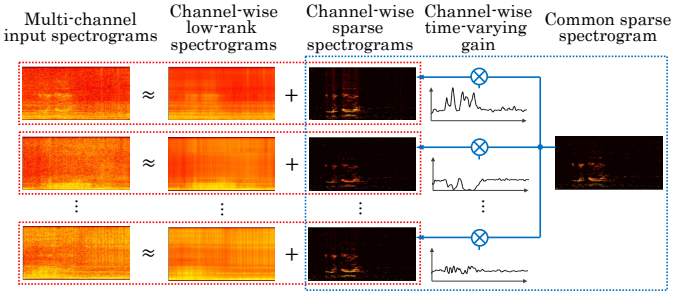


Fig. 2. Overview of the proposed multi-channel robust NMF.

to M at the tip of the robot. Let F and T be the numbers of frequency bins and time frames, respectively, and let f and t be the indices of them. The human voice enhancement problem in this paper is defined as follows:

Input: M-channel amplitude spectrograms $\mathbf{Y}_m \in \mathbb{R}_+^{F\times T}$
Output: Denoised amplitude spectrogram $\mathbf{S} \in \mathbb{R}_+^{F\times T}$

where $\mathbb{R}_+$ represents the set of nonnegative real values. The amplitude spectrogram is defined as the absolute values of the short-time Fourier transform (STFT) of a time-domain signal.

B. VB-RNMF for a single microphone

We first formulate variational Bayesian RNMF (VB-RNMF) for a single-channel input $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T] \in \mathbb{R}_+^{F\times T}$. VB-RNMF approximates an input spectrogram as the sum of a low-rank spectrogram $\mathbf{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_T] \in \mathbb{R}_+^{F\times T}$ and a sparse spectrogram $\mathbf{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T] \in \mathbb{R}_+^{F\times T}$ as follows:

$$\boldsymbol{y}_t \approx \boldsymbol{l}_t + \boldsymbol{s}_t. \tag{1}$$

In the same way as VB-RPCA [22], the low-rank spectrogram is represented by the product of K spectral basis vectors $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}_+^{F\times K}$ and their temporal activation vectors $\mathbf{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T] \in \mathbb{R}_+^{K\times T}$ as follows:

$$\boldsymbol{y}_t \approx \mathbf{W}\boldsymbol{h}_t + \boldsymbol{s}_t. \tag{2}$$

The low-rankness and sparseness of each term can be controlled in a Bayesian manner stated below.

1) Likelihood function: The proposed method tries to minimize the approximation error for the input spectrogram by using the Kullback-Leibler (KL) divergence. Since the maximization of a Poisson likelihood (denoted by $\mathcal{P}$) corresponds to the minimization of a KL divergence, the likelihood function is defined as follows:

$$p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S}) = \prod_{f,t} \mathcal{P}\left(y_{ft} \middle| \sum_k w_{fk}h_{kt} + s_{ft}\right). \tag{3}$$

2) Prior distributions on low-rank components: Our low-rank modeling is inspired by Bayesian NMF [23] that has been studied for low-rank decomposition of audio spectrograms. Since the gamma distribution (denoted by $\mathcal{G}$) is a conjugate prior for the Poisson distribution, gamma priors are put on the basis and activation matrices of the low-rank components as follows:

$$p(\mathbf{W}|\alpha^{wh},\beta^{wh}) = \prod_{f,k} \mathcal{G}(w_{fk}|\alpha^{wh},\beta^{wh}), \tag{4}$$

$$p(\mathbf{H}|\alpha^{wh},\beta^{wh}) = \prod_{k,t} \mathcal{G}(h_{kt}|\alpha^{wh},\beta^{wh}), \quad (5)$$

where $\alpha^{wh} \in \mathbb{R}_+$ and $\beta^{wh} \in \mathbb{R}_+$ represent the shape and rate parameters of the gamma distribution, respectively. Setting the shape parameter to 1.0 or less forces the basis and activation matrices to be sparse [23], which means that the low-rank component $\mathbf{L}$ is forced to be low-rank.

3) Prior distributions on sparse components: In VB-RPCA, Gaussian priors with the Jeffreys hyperpriors are put on sparse components [22]. To force the sparse components to take non-negative values, gamma priors with rate parameters given the Jeffreys hyperpriors are put on those components as follows:

$$p(\mathbf{S}|\alpha^s,\boldsymbol{\beta}^s) = \prod_{f,t} \mathcal{G}(s_{ft}|\alpha^s,\beta^s_{ft}), \quad (6)$$

$$p(\beta^s_{ft}) \propto (\beta^s_{ft})^{-1}. \quad (7)$$

where $\alpha^s \in \mathbb{R}_+$ represents the hyperparameter of the gamma distribution. In our formulation the sparseness is controlled by this shape parameter α^s.

C. VB-MRNMF for multiple microphones

We then formulate variational Bayesian multi-channel RNMF (VB-MRNMF). The relationship between the target voice signal $\boldsymbol{s}_t \in \mathbb{R}_+^F$ and its observation at each microphone $\boldsymbol{y}'_{mt} \in \mathbb{R}_+^F$ is assumed to be represented by a time-variant and frequency-invariant linear system:

$$\boldsymbol{y}'_{mt} \approx g_{mt}\boldsymbol{s}_t, \quad (8)$$

where $g_{mt} \in \mathbb{R}_+$ represents a gain of the target voice signal at microphone m and time t. Using this propagation model, an input spectrogram of each microphone $\mathbf{Y}_m = [\boldsymbol{y}_{m1},\ldots,\boldsymbol{y}_{mT}]$ is decomposed into channel-wise low-rank spectrograms (egonoise) and a sparse spectrogram common to the microphones (target voice) $\mathbf{S} = [\boldsymbol{s}_1,\ldots,\boldsymbol{s}_T] \in \mathbb{R}_+^{F\times T}$ as follows:

$$\boldsymbol{y}_{mt} \approx \mathbf{W}_m\boldsymbol{h}_{mt} + g_{mt}\boldsymbol{s}_t, \quad (9)$$

where $\mathbf{W}_m \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H}_m = [\mathbf{h}_{m1},\ldots,\mathbf{h}_{mT}] \in \mathbb{R}_+^{K\times T}$ represent the basis and activation matrices of the channel-wise low-rank components, respectively.

1) Likelihood function and prior distributions: The likelihood function and prior distributions except for those on the gain parameters g_{mt} are formulated in the same way as VB-RNMF (Eqs. 3–7). A gamma prior is put on g_{mt} as follows:

$$p(g_{mt}|\alpha^g) = \mathcal{G}(g_{mt}|\alpha^g,\alpha^g), \quad (10)$$

where $\alpha^g \in \mathbb{R}_+$ is a hyperparameter controlling the variance of the gain parameters.

D. Variational Bayesian inference

Our goal is to calculate the full posterior distribution of the unknown parameters $p(\mathbf{W}_{1:m},\mathbf{H}_{1:m},\boldsymbol{g}_{1:m},\mathbf{S},\boldsymbol{\beta}|\mathbf{Y}_{1:m})$. Since the true posterior is analytically intractable, we approximate it by using a variational Bayesian (VB) algorithm [22], [23]. Let $\boldsymbol{\Theta}$ be a set of all parameters and $q(x)$ be a variational posterior distribution. The true posterior distribution is approximated as $p(\boldsymbol{\Theta}|\mathbf{Y}_{1:M}) \approx \{\prod_m q(\mathbf{W}_m)q(\mathbf{H}_m)q(\boldsymbol{g}_m)\}\, q(\mathbf{S})q(\boldsymbol{\beta}^s)$, and the parameters of each variational distribution are estimated by minimizing the KL-divergence between the true and approximated distributions.

Since the probability distributions used in VB-MRNMF are in the conjugate exponential family, the form of each posterior approximation can be found by using Jensen's inequality and the Lagrange multiplier framework [23]. Let $\langle x\rangle$ be the mean value of the posterior distribution of x. Each variational posterior distribution is alternately and iteratively updated by fixing the other distributions as follows:

$$q(w_{mfk}) = \mathcal{G}(\alpha^{wh}+\sum_t y_{mft}\lambda^{wh}_{mftk}, \beta^{wh}+\sum_t \langle h_{mtk}\rangle), \quad (11)$$

$$q(h_{mtk}) = \mathcal{G}(\alpha^{wh}+\sum_f y_{mft}\lambda^{wh}_{mftk}, \beta^{wh}+\sum_f \langle w_{mfk}\rangle), \quad (12)$$

$$q(g_{mt}) = \mathcal{G}(\alpha^g+\sum_f y_{mft}\lambda^{gs}_{mft}, \alpha^g+\sum_f \langle s_{ft}\rangle), \quad (13)$$

$$q(s_{ft}) = \mathcal{G}(\alpha^s+\sum_m y_{mft}\lambda^{gs}_{mft}, \langle\beta^s_{ft}\rangle+\sum_m \langle g_{mt}\rangle), \quad (14)$$

$$q(\beta^s_{ft}) = \mathcal{G}(\alpha^s, \langle s_{ft}\rangle), \quad (15)$$

$$\lambda^{wh}_{mftk} = \frac{\mathbb{G}[w_{mfk}]\mathbb{G}[h_{mtk}]}{\sum_k \mathbb{G}[w_{mfk}]\mathbb{G}[w_{mtk}] + \mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}, \quad (16)$$

$$\lambda^{gs}_{mft} = \frac{\mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}{\sum_k \mathbb{G}[h_{mfk}]\mathbb{G}[h_{mtk}] + \mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}, \quad (17)$$

where $\mathbb{G}[x] = \exp(\langle\log x\rangle)$ represents the geometric mean, and λ^{wh}_{mftk} and λ^{gs}_{mft} are auxiliary variables.

IV. Experimental Evaluation

We report results of human-voice enhancement using actual recordings and numerically simulated audio signals.

A. Implementation

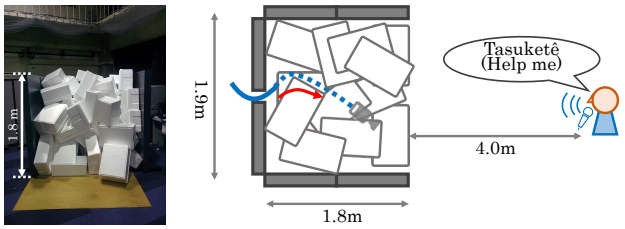
As shown in Fig. 1, the body of a hose shaped robot used in this evaluation was made with a corrugated tube 38 mm in diameter and 3 m long. The entire surface of the robot was covered by cilia and seven vibrators used for moving forward by vibrating the cilia. This robot had $M = 8$ synchronized microphones distributed on its body at 40 cm intervals. The audio signals of these microphones were captured at 16 kHz and with 24-bit sampling.

The parameters for VB-MRNMF were as follows. The shift and window lengths of the STFT were set to 160 and 1024 samples, respectively. The hyperparameters α^{wh}, β^{wh}, α^g, and α^s were set to 1.0, 1.0, 5.0, and 0.7, respectively. The number of bases K was set to 20. These parameters had been determined experimentally. In this paper we iterated VB-MRNMF 200 times with random parameter initialization.

B. Experiment 1: Actual recordings

We evaluated the proposed method in the condition that the robot moved under rubble.

1) Experimental conditions: To simulate rubble disturbing sound propagation, styrene foam boxes are piled up (Fig. 3-(a)). A male subject sit 4 m away from this rubble (Fig. 3-(b)) and called for rescue in Japanese (*e.g.*, "Tasukete," "Ôi," and "Darekâ"). The robot was inserted from behind the rubble and



(a) Pile of rubble (b) Condition of rubble and target voice

Fig. 3. Condition of rubble and target human voice of experiment 1.

TABLE I
SNR IMPROVEMENT (DB) IN EXPERIMENT 1

VB-MRNMF (Sec. III-C)	VB-RNMF (Sec. III-B)	Med-RPCA [3]	RPCA [12]	MNMF [17]	IVA [15]	HRLE [24]
4.29	0.49	1.62	-0.57	-0.18	0.02	-0.89

captured eight-channel audio signals (mixtures of ego-noise and target voice signals) for 60 seconds during the insertion. For reference, the target voice signals were recorded using a microphone close to the subject's mouth.

Since it was impossible to obtain pure human-voice signals captured by the robot microphones, we used the signal-to-noise ratio (SNR) as a evaluation criteria of this experiment:

$$\mathrm{SNR}(\hat{\mathbf{S}}, \mathbf{S}, \alpha) = 10 \log_{10} \frac{\sum_{f,t} \alpha^2 s_{ft}^2}{\sum_{f,t} (\hat{s}_{ft} - \alpha s_{ft})^2}, \qquad (18)$$

where $\mathbf{S} \in \mathbb{R}_+^{F \times T}$ and $\hat{\mathbf{S}} \in \mathbb{R}_+^{F \times T}$ represent the amplitude spectrograms of reference and estimated target voice signals, respectively, and α represents a gain parameter compensating for the level difference between $\mathbf{S}$ and $\hat{\mathbf{S}}$. This gain parameter was determined with minimum mean-square error estimation (MMSE) between $\alpha\mathbf{S}$ and $\hat{\mathbf{S}}$. The estimated SNR of the input signal was -14.7 dB.

The proposed VB-MRNMF and VB-RNMF were compared with MNMF [17], independent vector analysis (IVA) [15], RPCA [12], and histogram-based recursive level estimation (HRLE) [24] which is a conventional single-channel spectrum subtraction method. The number of sources was set to eight for MNMF and IVA because seven vibrators generated noise and one target voice existed. Since these methods cannot distinguish the target source and other noise sources, the SNR performance was determined by taking a maximum SNR value in all the eight separation results. The results of VB-RNMF, RPCA, and HRLE were obtained by using the tip (8th) microphone signals. We also evaluated extended RPCA results that were obtained by taking median values of all the microphone results (Med-RPCA) [3].

2) Experimental results: TABLE I shows that VB-MRNMF outperformed any of the other methods. It improved the SNR by 2.7 dB more than Med-RPCA, which had the second performance, did. Comparing VB-MRNMF with VB-RNMF, we see that the proposed multi-channel extension improved the SNR by 3.8 dB. Fig. 4 shows the amplitude spectrogram of an observed signal (at the tip microphone) and a voice-enhanced version obtained by VB-MRNMF. These results showed that the proposed method successfully suppressed the time-varying ego-noise.

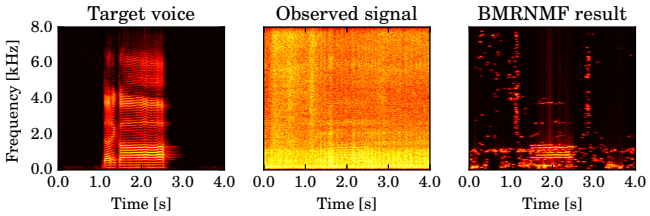


Fig. 4. Examples of enhancement result of experiment 1.

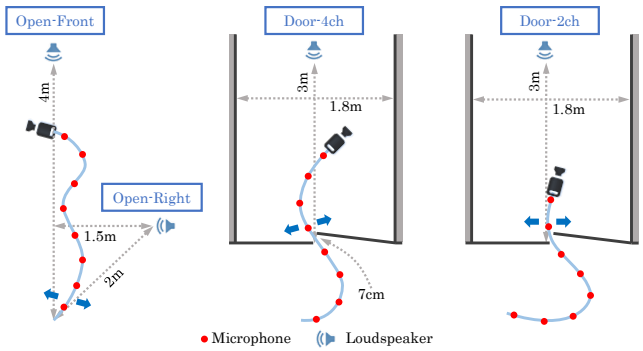


Fig. 5. Four conditions of robot and loudspeaker in experiment 2.

C. Experiment 2: Numerically simulated audio signals

To analyze the performance of the proposed method in detail, we evaluated with numerically simulated audio signals.

1) Experimental conditions: Ego-noise and target voice signals were captured independently and then mixed at SNRs varying from -20 dB to $+5$ dB. As shown in Fig. 5, there were four conditions differing in the relative positions of the robot and a loudspeaker emitting target voice signals.

1) **Open-Front**: The robot was in an experimental room with no obstacles. The loudspeaker was in front of the robot. The reverberation time (RT_{60}) of the room was 750 ms.
2) **Open-Right**: Same as Open-Front except that the loudspeaker was to the right of the robot.
3) **Door-4ch**: The robot was caught by a door, the loudspeaker was in front of the robot, and four of the microphones were behind the door. The reverberation time was 990 ms.
4) **Door-2ch**: Same as Door-4ch except that six microphones were behind the door.

The ego-noise was recorded for 60 seconds under each condition while sliding the robot left and right by using vibrators and a hand. The target-voice data consisted of two recordings of male voices and two recordings of female voices, each of which was one-minute long. It should be noted that the target source did not move in this experiment because it was recorded when the robot was stationary. In this experiment, the enhancement performance was evaluated by using the signal-to-distortion ratio (SDR) [25], [26].

2) Experimental results: As shown in Fig. 6, in the Open-Front and -Right conditions, VB-MRNMF performed better than any of the other methods. In the Door-4ch and -2ch conditions where some microphones were shaded, the performances of conventional multi-channel methods (MNMF, IVA, and Med-RPCA) were worse than those of the single-channel methods. Although VB-MRNMF was also degraded in these conditions, its performance was comparable to the results of

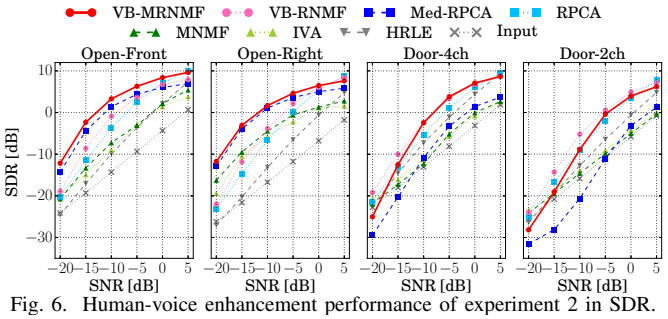


Fig. 6. Human-voice enhancement performance of experiment 2 in SDR.

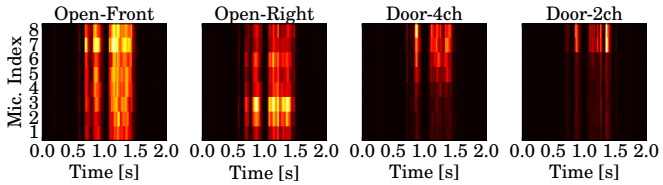


Fig. 7. Examples of time-varying powers of sparse spectrogram obtained by VB-MRNMF at each microphone when SNR was -5 dB. Male voice was emitted between 0.5 sec and 1.5 sec.

VB-RNMF in the Door-4ch condition except for the SNR condition of -20 dB.

Fig. 7 shows time-varying powers of the sparse spectrogram obtained by VB-MRNMF at each microphone. In the Door-4ch and -2ch conditions, the power of channel-wise sparse components that were separated from the sound source (in the Door-4ch condition the 1st to 4th microphones, in the Door-2ch condition the 1st to 6th microphones) got significantly smaller. This shows that the gain estimation can be used for estimating the reliability of each microphone.

One way to improve VB-MRNMF is selection of valid microphones. The SDR performance of VB-RNMF was better than that of VB-MRNMF when only two microphones were available (in the Door-2ch condition). The proposed method would be improved by selecting microphones when a few microphones are available. The idea of beta-process NMF [27] will be effective for this extension.

V. Conclusion

This paper presented a multi-channel blind human-voice enhancement method based on variational Bayesian multi-channel RNMF (VB-MRNMF). Human-voice enhancement for a hose-shaped robot needs to address two main problems: deformable layout of microphones and partial occlusion of microphones. To solve these problems, we developed a Bayesian model that separates multi-channel amplitude spectrograms into sparse and low-rank components (human voice and noise) without using phase information depending on the array layout. The voice level at each microphone is estimated in a time-varying manner for reducing the influence of the shaded microphones. Experiments using a 3-m hose-shaped rescue robot with eight microphones showed that the proposed method improves the SNR of a human voice 2.7 dB more than conventional blind source separation methods do. Future work includes the derivation of online updating for real-time processing that is mandatory for rescue tasks. Furthermore, we will conduct more practical experiments for evaluating the effectiveness of the proposed method in search-and-rescue tasks.

Acknowledgments

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

References

[1] R. R. Murphy, *Disaster Robotics*. MIT Press, 2014.
[2] J. Fukuda *et al.*, "Remote vertical exploration by active scope camera into collapsed buildings," in *IEEE/RSJ IROS*, 2014, pp. 1882–1888.
[3] Y. Bando *et al.*, "Human-voice enhancement based on online RPCA for a hose-shaped rescue robot with a microphone array," in *IEEE SSRR*, 2015, pp. 1–6.
[4] A. Deleforge *et al.*, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *IEEE ICASSP*, 2015, pp. 355–359.
[5] B. Cauchi *et al.*, "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization," in *SMIAE*, 2012, pp. 28–33.
[6] K. Furukawa *et al.*, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *IEEE/RSJ IROS*, 2013, pp. 3943–3948.
[7] G. Ince *et al.*, "Assessment of general applicability of ego noise estimation," in *IEEE ICRA*, 2011, pp. 3517–3522.
[8] Y. Bando *et al.*, "Microphone-accelerometer based 3D posture estimation for a hose-shaped rescue robot," in *IEEE/RSJ IROS*, 2015, pp. 5580–5586.
[9] J. Nikunen *et al.*, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM TASLP*, vol. 22, no. 3, pp. 727–739, 2014.
[10] Y. Tatekura *et al.*, "Sound source separation with shaded microphone array," *JARP*, vol. 3, no. 2, 2013.
[11] H. Chiba *et al.*, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *IWAENC*, 2014, pp. 203–207.
[12] C. Sun *et al.*, "Noise reduction based on robust principal component analysis," *JCIS*, vol. 10, no. 10, pp. 4403–4410, 2014.
[13] E. J. Candès *et al.*, "Robust principal component analysis?" *JACM*, vol. 58, no. 3, p. 11, 2011.
[14] Z. Chen *et al.*, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," in *IEEE WASPAA*, 2013, pp. 1–4.
[15] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE WASPAA*, 2011, pp. 189–192.
[16] A. Ozerov *et al.*, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2010.
[17] D. Kitamura *et al.*, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *IEEE ICASSP*, 2015, pp. 276–280.
[18] D. Kounades-Bastian *et al.*, "A variational EM algorithm for the separation of moving sound sources," in *IEEE WASPAA*, 2015, pp. 1–5.
[19] N. Dobigeon *et al.*, "Robust nonnegative matrix factorization for nonlinear unmixing of hyperspectral images," in *WHISPERS*, 2013, pp. 1–4.
[20] M. Sun *et al.*, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM TASLP*, vol. 23, no. 7, pp. 1233–1242, 2015.
[21] X. Ding *et al.*, "Bayesian robust principal component analysis," *IEEE TIP*, vol. 20, no. 12, pp. 3419–3430, 2011.
[22] S. D. Babacan *et al.*, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE TSP*, vol. 60, no. 8, pp. 3964–3977, 2012.
[23] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *CIN*, vol. 2009, no. 785152, pp. 1–17, 2009.
[24] H. Nakajima *et al.*, "An easily-configurable robot audition system using histogram-based recursive level estimation," in *IEEE/RSJ IROS*, 2010, pp. 958–963.
[25] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.
[26] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. Ellis, "mir eval: a transparent implementation of common MIR metrics," in *ISMIR*, 2014, pp. 367–372.
[27] D. Liang *et al.*, "Beta process non-negative matrix factorization with stochastic structured mean-field variational inference," *arXiv preprint arXiv:1411.1804*, 2014.