Robust Audio Scene Analysis for Rescue Robots

Yoshiaki BANDO

Abstract

Audio scene analysis and its application to robot audition is indispensable for a rescue robot, both to detect victims and to sense the robot itself (i.e., location and posture) in disaster environments where visual and/or GPS sensors cannot be used. This work focuses on hose-shaped rescue robots used for probing narrow gaps under rubble. Arrays of microphones, inertial sensors, and loudspeakers on the robot are used for audio scene analysis.

Two fundamental problems of audio scene analysis are addressed: speech enhancement and posture (shape) estimation. Speech enhancement is crucial for detecting speech sounds in captured noisy signals. Posture estimation is essential for enabling an operator to control the flexible robot and for localizing the speech source by using the deformable microphone array. In addition, the integration of these techniques enables the robot to find and approach a victim autonomously. The major difficulties are that the layout of the microphones changes as the robot moves and that some of them are occasionally occluded by rubble. In this study, Bayesian signal processing is used for speech enhancement. The latent speech signals and parameters, which depend on the surrounding environment, are simultaneously estimated. Multi-modal signal processing is used for posture estimation. Unreliable audio measurements due to occlusion are compensated for by using the measurements of other sensors.

This thesis consists of seven chapters. Chapter 2 overviews audio scene analysis for rescue robots and reviews speech enhancement and posture estimation.

Chapter 3 describes a speech enhancement method called Bayesian robust non-negative tensor factorization. To deal with the dynamic configuration of the microphones, speech and noise signals are separated on the basis of their spectral

pattern difference instead of on the phase difference (which is unreliable). Under the assumption that the speech and noise spectrograms are sparse and low-rank, respectively, they are separated from the input multichannel spectrogram without prior training. To cope with the partial occlusion of microphones, the speech volume gain at each microphone is estimated on the basis of its feasible gain. Experimental results showed that this method outperforms conventional multichannel methods even when a half of the microphones are occluded.

To further improve the enhancement performance, a deep prior distribution on speech signals is introduced in Chapter 4. Instead of using the unrealistic sparse assumption for speech signals, a deep generative model is trained with clean speech signals from a large database. Posterior estimates of clean speech are obtained using the speech model as a prior distribution while adapting a noise model to the observed noisy signals. Experimental results showed that this method outperforms a method based on the low-rank and sparse decomposition. The results also showed that the method outperforms a conventional supervised method with deep learning in unseen noisy environments.

Chapter 5 describes an audio-based posture estimation method that can deal with the dynamic configuration of microphones. The time differences of arrival (TDOAs) of beacon sounds, which depend on the locations of the microphones and loudspeakers, are used to estimate the posture. A state-space model representing the posture dynamics is formulated, and the current posture is tracked by estimating the posture change rate and predicting the current posture.

In Chapter 6, the audio-based posture estimation method is extended to a multi-modal 3D posture estimation method that can work when the microphones are partially occluded. The method excludes TDOA measurements distorted by obstacles and compensates for the missing posture information by using the tilt angles obtained from accelerometers. Experiments using a 3 m hose-shaped rescue robot showed that this method reduces the tip position error of the initial state to about 0.2 m. When the initial error of the initial state is less than 20 %, it can estimate the correct 3D posture in real time.

Chapter 7 concludes this thesis with a brief look at future work.

Acknowledgments

This work was accomplished at Kawahara Laboratory, Graduate School of Informatics, Kyoto University. I express my gratitude to all people who helped me and this work.

I would like to express my special thanks and appreciation to my supervisor Prof. Tatsuya Kawahara. He has allowed me to pursue a research topic that matters to me and advised my research through his broad insights. His comments were essential for organizing the significance of my research from a broad perspective and presenting it as a unified work. Without his continuing encouragement and generous support, this work could not have been completed.

I would also like to express my special thanks and appreciation to Prof. Hiroshi G. Okuno who was my first supervisor in my master course. He taught me the interesting research topic of robot audition when I was the first grade of my undergraduate course. I have always benefited from his brilliant ideas, continuous enthusiasm, rigorous attitude to science, and valuable advice. He also gave many opportunities to work outside of the lab.

Furthermore, I would like to express my deep and sincere appreciation to Dr. Kazuyoshi Yoshii who had supervised my master course after Prof. Okuno and continuously supported this study. I would not have written the dissertation without his insightful comments from his wide and deep knowledge about machine learning and audio signal processing.

I also express my special thanks and appreciation to the members of my dissertation committee, Prof. Toshiyuki Tanaka and Prof. Hisashi Kashima for their time and valuable comments and suggestions.

My thesis cannot be accomplished without Prof. Satoshi Tadokoro, Dr.

Masashi Konyo and the members of Tadokoro Lab. in Tohoku University. The basic structure of a hose-shaped rescue robot used in this study was developed by Prof. Tadokoro and Dr. Konyo. They gave a lot of insightful comments from the viewpoint of rescue robotics. They also give me the great opportunity to receive field research training.

I deeply thank Kawahara Lab. members, who are good colleagues and friends. Dr. Katsutoshi Itoyama told me fundamental skills as a researcher; logically thinking, validating the claim, and so forth. The discussions with Mr. Koji Inoue, Mr. Ryo Nishikimi, Mr. Yuta Ojima, Mr. Kazuki Shimada, and the other students improved my research and knowledge. Dr. Eita Nakamura always gave me insightful advice from his deep knowledge of machine learning and mathematics. I thank Ms. Mayumi Abe, secretary of Kawahara Lasb., who helped me so much. I could not be able to concentrate on my research without their help.

I would like to thank Dr. Kei Shimonishi, Mr. Ryo Kawahara, Mr. Motoharu Sonogashira, and Mr. Tomohiro Sakaguchi who are my colleagues in the same department, Graduate School of Informatics Department of Intelligent Science and Technology.

I also deeply thank the members of Okuno Lab. that I spent my master course. Dr. Takeshi Mizumoto and Dr. Takuma Otsuka gave me a lot of valuable comments from their broad knowledge about robot audition and signal processing. I would like to thank the excellent secretary, Ms. Yumi Okazaki.

I am grateful to the Japan Society for the Promotion and Science (JSPS) for their financial support as a Fellowship for Young Scientists DC1.

Last but not least, I am truly grateful to my parents, Shoji Bando and Yoshiko Bando for their support of my long student life.

Contents

List of Figures

List of Tables

Chapter 1

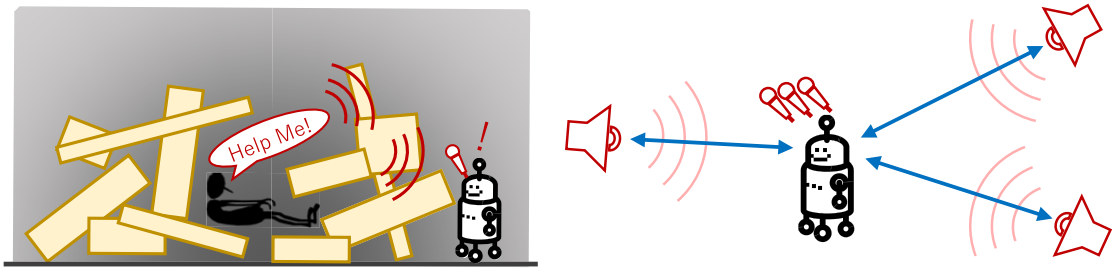
Introduction

This thesis addresses audio scene analysis for rescue robots designed to work in severely adverse environments. This chapter presents the background to this work, describes the problems addressed, and introduces the approaches used.

1.1 Background

The mission of rescue robots is to sense the environment at distant disaster sites where people and animals cannot go and to act accordingly [1–7]. Since a rescue operation is a race against time, the robot should undertake and complete it as soon as possible. This has led to the development of various types of rescue robots that can meet the demands for various disaster situations [3–9]. To widely and quickly monitor a post-disaster environment, aerial rescue robots, such as multi copters [7–9] and micro airplanes [10], have been developed. To search in polluted areas and narrow gaps under rubble, ground rescue robots, such as rover and flexible robots, have been developed [3–5]. Marine rescue robots, such as boat and submarine robots, are helpful for probing a rubble-strewn marine environment [11].

To quickly reach and search a target area, it is critical to collect information about the environment around the robot and the state of the robot. Video cameras and microphones are widely used for helping to manipulate a robot and searching for targets (e.g., victims). Since the raw observations of such sensors are often confusing for a human operator, various sensor systems for helping the



(a) Searching for victims (b) Estimating robot location

Figure 1.1: Applications of audio scene analysis for rescue robot.

operator to interpret the observations have been developed [2]. For example, it is difficult to spot clues about victim locations from video images of a complicated rubble-filled environment. A method that classifies the objects captured in a video image was thus developed [12]. Simultaneous localization and mapping (SLAM) methods have also been developed. They are used along with video cameras and/or laser rangefinders to estimate the robot location and to create a map of the area around the robot [13, 14]. The global positioning system (GPS), magnetometers, and inertial sensors (accelerometers and gyroscopes) are widely used for estimating the location and posture of a robot [1].

Audio scene analysis and its applocation to robot audition [15–17] is indispensable for a rescue robot, both to locate victims and to estimate the state of the robot in disaster environments where visibility is low and/or GPS sensors cannot be used (Figure 1.1). Even if a victim is hidden by obstacles so that the robot cannot see him or her, sounds created by the victim might still reach the robot. Microphone array signal processing can detect and localize audio events around the robot [16–19]. Drones with microphone arrays have actively been studied for finding victims quickly from the sky [20]. The location and posture of a robot can be estimated by transmitting reference (beacon) and then using microphones on the robot to capture them [21, 22]. Submarine robots localize themselves using audio beacons because GPS signals are blocked and the visual range is limited in the sea [21].

(a) Phase 1 (b) Phase 2 (c) Phase 3

Figure 1.2: Typical usage scenario of hose-shaped rescue robot.

1.2 Hose-Shaped Rescue Robot

This thesis focuses on *hose-shaped rescue robots*, which are typically used as ground rescue robots for finding victims buried under collapsed buildings [23–25]. Such robots have a long thin flexible body and can penetrate the narrow gaps in collapsed buildings. A remote operator steers the robot to a target location by using its specialized locomotion mechanism. Active Hose-II [23], for example, has small powered wheels to move forward. Active Scope Camera (ASC) [26] moves forward by vibrating the cilia covering its body. It was used for real search-and-rescue operation in Jacksonville, Florida in 2008 [27].

The typical usage scenario of a hose-shaped rescue robot consists of three phases, as shown in Figure 1.2. First, the operator of the robot inserts it into the target collapsed building and steers the robot to an area where a surviving victim is likely trapped. After reaching the target area, the operator calls to the victim with a loudspeaker and searches for him with a video camera and a microphone on the robot. If a victim is found, the operator determines his or her condition and directs the rescue team to the victim's location.

This study focuses on audio scene analysis for a hose-shaped rescue robot, which works under collapsed buildings where visibility is poor and GPS cannot be used. A microphone array is placed on the robot, as shown in Figure 1.3. The robot has the same self-propelling mechanism as the ASC robot. Eight microphones are distributed along the body so that all the microphones are not obstructed by rubble at the same time. Eight inertial sensors and seven loudspeakers are attached to the robot. These sensors are used to sense the surrounding environments and the robot itself.

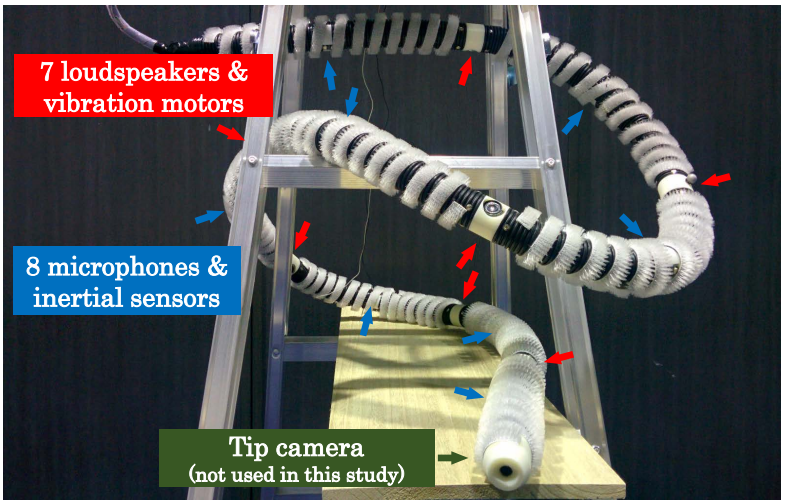


Figure 1.3: Hose-shaped rescue robot with eight-channel microphone array.

1.3 Audio Scene Analysis

Various types of audio scene analysis and its application to robot audition have been studied (Figure 1.4). Acoustic event identification [28–30] and automatic speech recognition (ASR) [31, 32], for example, estimate the content of a sound. Voice activity detection (VAD) detects whether speech sounds exist in an observed signal [33]. Since captured audio signals often include multiple sounds originating at different locations, these recognition methods require sound source separation for extracting each sound source signal and sound source localization for understanding the audio scene spatially.

Sound sources can be extracted from a mixture recording on the basis of the spectral pattern difference across source signals [34, 35] and the spatial information (e.g., relative source locations) [36, 37]. Since many types of robots generate ego-noise from their actuators, ego-noise reduction is essential for understanding the audio scene around the robot [34, 38, 39]. Speech enhancement, which suppresses noise signals and extracts speech signals, is also important for ASR and VAD [40].

The locations of sound sources can be estimated with a microphone array on the basis of power or phase differences between the microphones [41–43]. Like bats and dolphins, echolocation and active sonar systems transmit a beacon

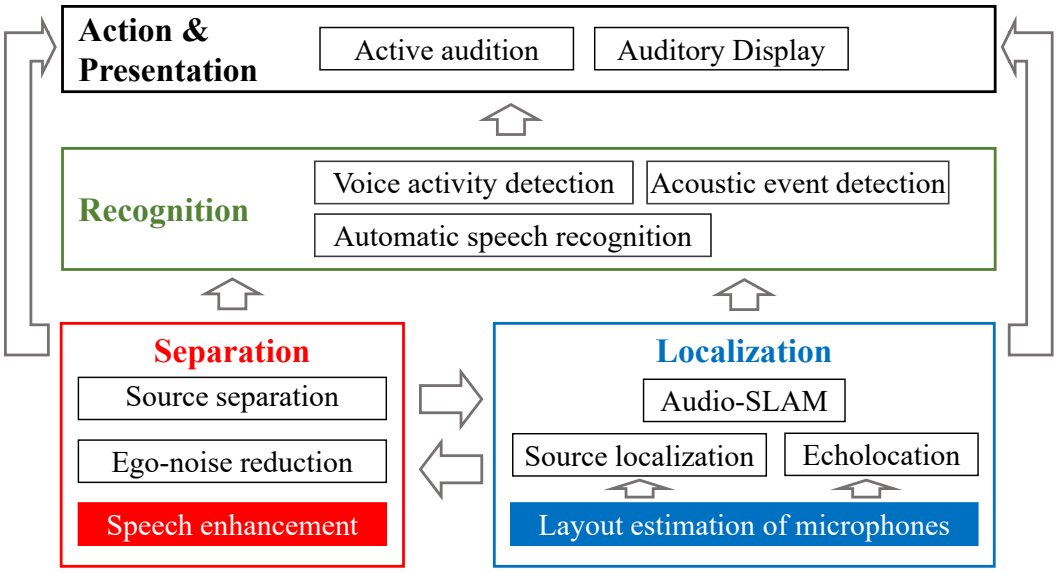


Figure 1.4: Overview of audio scene analysis.

signal, receive it with microphones, and localize objects that reflect the beacon sound [21, 44, 45]. The localization results of sound sources can be used for a SLAM system in a way similar to visual-based SLAM [46]. An underlying assumption for most localization methods is that the layout of microphones is known in advance. Simultaneous localization of microphones and sound sources has been studied for estimating the layout when it is unknown [18, 22].

The results of the separation, localization, and recognition are used for selecting the robot actions automatically and presenting the audio scene to the remote operator [42, 47, 48]. Although localization results with a single microphone array have distance ambiguities, a robot can localize and approach the 3D point of a sound source by moving around the source [42, 49]. Active audition has been studied for planning efficient moving strategies [47, 49]. Since the robot movements generate ego-noise from its actuators, active audition requires not only localizing the sound sources but also separating and distinguishing the sources [50]. Visualizing the audio scenes is an effective way for the distant operator to recognize the scene. A telepresence robot with robot audition displayed sound source directions superposed on a remote display [17]. This system demonstrated the feasibility of an auditory display for telepresence robots.

1.4 Problems

This thesis addresses two fundamental problems of audio scene analysis for a hose-shaped rescue robot: speech enhancement and posture (shape) estimation. Speech enhancement is crucial for detecting speech sounds in captured noisy signals. Posture estimation is essential for enabling the operator to control the flexible robot. Together they are essential for higher-level audio scene analysis (Figure 1.4). Since the microphone layout depends on the robot posture, source localization of a victim's speech sound requires posture estimation. The enhancement results enable automatic detection of speech sound (i.e., VAD). The integration of their results enables active audition and auditory display.

1.4.1 Speech Enhancement

When the robot operator searches for victims using microphones and a tip camera on the robot, speech signals captured with the microphones are contaminated by non-stationary ego-noise (e.g., motor and friction noise). While hose-shaped rescue robots keep moving in order to search a wide area in a limited amount time, conventional robots must stop their actuators and remain silent in order to detect external sounds. This constraint is inconvenient, and speech occurring when the robot is moving is often missed. Speech enhancement helps prevent the operator from failing to detect speech sounds even when the robot is moving.

Speech enhancement is the task of suppressing noise and enhancing speech signals included in a captured noisy signal. It has been studied for a wide variety of applications such as speech telecommunication, speech recognition, and hearing aids [37, 51–53]. In these applications, it is often difficult to anticipate the usage situations. The speech recognition system on a smartphone, for example, is used in various noisy environments such as houses, train stations, and outdoor environments. To cope with unknown noise, blind speech enhancement, which works without using noise signal training data has been studied [37, 53–55]. Speech enhancement for a hose-shaped rescue robot must also deal with environment-dependent noise because the ego-noise of the robot

depends on the robot's movements and surrounding materials.

For microphone array signal processing, speech enhancement must deal with two additional characteristics of a hose-shaped rescue robot:

Dynamic configuration of microphones – The relative positions of the microphones change due to vibration and deformation of the robot body. Most conventional blind enhancement are degraded because they assume the relative layout of microphones is stable [37, 53–55].

Partial occlusion of microphones – Microphones may fail to capture target speech signals when they are obstructed by rubble around the robot. Since such microphones degrade enhancement performance, detection of obstructed microphones is important.

1.4.2 Posture Estimation

It is crucial to estimate the unseen posture of the robot because the unexpected bending of the flexible body makes it difficult for an operator to control the robot as desired. A posture estimation method for a hose-shaped rescue robot was proposed by using gyroscopes distributed on the long body of the robot [56]. Since this method estimates the robot posture by using the angular velocities obtained with the gyroscopes, the estimation error is gradually accumulated as time passes. Shape estimation of a flexible cable has also been conducted with magnetometers [57, 58]. Magnetometers, however, cannot be used for a hose-shaped robot because the magnetic fields are easily distorted in rubble-existing environments.

Both the microphones and loudspeakers on the robot can be simultaneously localized using time differences of arrival (TDOAs) of the beacon sounds transmitted by the loudspeakers [18, 22, 59, 60]. Since the TDOAs depend only on the current relative positions of the microphones and loudspeakers, the accumulative error problem can be avoided. The audio-based approach can be used in any enclosed space allowing sound propagation. This means that sound-based posture estimation is complementary to gyroscope-based and magnetometer-based posture estimation. The simultaneous localization problem is also known

as unsupervised microphone array calibration. Various types of problem settings have been studied such as localization of sub-microphone arrays (pairs of microphones) and asynchronous microphones [18, 22, 59, 60].

The partial occlusion and dynamic configuration of microphones are the main issues in posture estimation:

Dynamic configuration of microphones – As in speech enhancement, most existing localization methods are based on the assumption that the relative locations of the microphones and/or sources are fixed [18, 61].

Partial occlusion of microphones – Since the TDOAs in rubble are much different from those in an open space, it is crucial to detect degraded TDOAs automatically. When many of the microphones are obstructed in a narrow space, posture estimation using only audio measurements is difficult.

1.5 Approaches

This thesis tackles speech enhancement and posture estimation by exploiting two key ideas: using *Bayesian signal processing* and using *multi-modal signal processing*.

1.5.1 Bayesian Speech Enhancement

Bayesian signal processing is used for speech enhancement. The latent speech signals and parameters, which depend on the surrounding environment, are estimated simultaneously. Bayesian inference can be used to stabilize parameter inference by assuming appropriate prior distributions [62]. This is because prior distributions can be regarded as regularization terms from the viewpoint of optimization. In this study, to deal with the partial occlusion of microphones, the speech volume gain at each microphone is estimated in a time-varying manner. This estimation is stably conducted by putting a prior distribution representing feasible gains. Prior distributions represent the statistical characteristics of parameters. To cope with the dynamic configuration of microphones, speech and noise signals are separated on the basis of their spectral pattern difference instead of the phase difference, which is unreliable as it is sensitively affected by

the array layout. The spectral pattern difference is treated as the structure difference of their prior distributions. In addition, Bayesian modeling can use prior distributions trained in advance. Deep generative models, such as variational autoencoders (VAEs), have recently been proposed for learning a probability distribution over complicated data [63–67]. A method that uses a speech prior distribution based on a pre-trained VAE is developed in this study. Utilizing a deep generative model, it enhances speech in unseen noisy signals by estimating the latent speech and noise.

1.5.2 Multi-Modal Posture Estimation

Posture estimation is tackled using multi-modal signal processing. It is important to detect sensor failures and integrate multiple sensors that compensate for each other's weaknesses [1, 68–70]. A 3D posture estimation method is developed in this study by integrating TDOAs obtained from microphones and tilt angles obtained from accelerometers. Unreliable audio measurements due to occlusion are compensated for by using the measurements of accelerometers. Although accelerometers capture only partial information of the robot posture (tilt angles), they are robust against external environments. Multi-modal estimation is done based on a unified state-space model representing the sensor measurements and temporal dynamics of the posture.

1.6 Organization

The organization of this thesis is outlined in Figure 1.5. Chapter 2 reviews audio scene analysis for rescue robots, introduces speech enhancement and sound source separation, and explains the posture estimation of flexible cables and simultaneous localization of microphones and loudspeakers. Chapter 3 presents a blind multi-channel speech enhancement method that can cope with the partial occlusion and dynamic configuration of microphones To improve enhancement performance, Chapter 4 introduces a deep prior distribution on speech signals. Chapter 5 presents an audio-based method that can cope with the dynamic con-

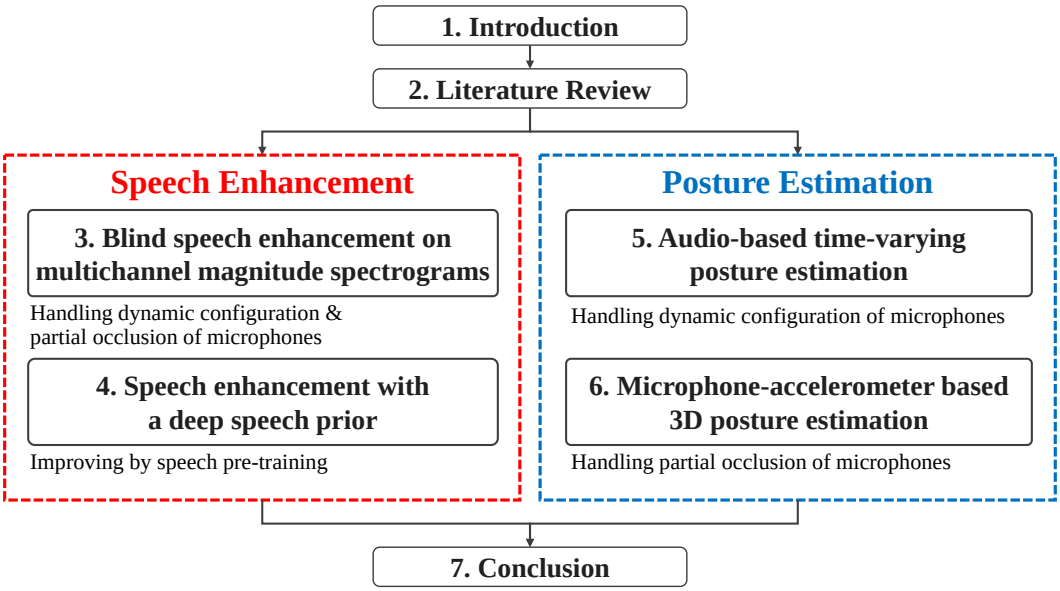


Figure 1.5: Organization of this thesis.

figuration of microphones by estimating the posture change rate and predicting the current posture. In Chapter 6, the audio-based posture estimation method is extended to a multi-modal 3D posture estimation method that can work when there is partial occlusion of microphones. The thesis conclude in Chapter 7 with a brief look at future work.

Chapter 2

Literature Review

This chapter reviews the literature related to audio scene analysis for rescue robots. Section 2.1 summarizes the audio scene analysis for existing rescue robots. Section 2.2 reviews existing methods of speech enhancement and source separation. Section 2.3 describes studies of simultaneous localization of microphones and sources and posture estimation of flexible cables.

2.1 Audio Scene Analysis for Rescue Robots

Rescue robots are categorized into aerial rescue robots, marine rescue robots, and ground rescue robots [1]. This section describes audio scene analysis for each of these categories.

2.1.1 Aerial Rescue Robots

Aerial rescue robots with microphone arrays have been developed for quickly finding a victim from the sky. Since small microphone arrays on these robots can only estimate the direction of arrival (DOA) of a target sound, they move fast in a large search area and localize the source by integrating the DOAs observed at multiple locations [10, 20]. The relative locations of multiple robots can also be estimated by submitting beacon signals and localizing them [71, 72]. Multi-copters (or drones) have recently gained attention because of their high maneuverability [19, 20, 39]. The main problem of these robots is continuous and large ego-noise caused by their propellers [39]. Ohata et al. developed a

localization method robust against ego-noise based on the multiple signal classification (MUSIC) algorithm [19]. A source classification method was proposed based on deep learning for detecting sound events directly from a captured noisy signal [28].

2.1.2 Marine Rescue Robots

The sound navigation and ranging (sonar) has long been studied for marine vehicles including the rescue robots [11,73,74]. Sonar systems on marine rescue robots are used for localizing objects around the robot and imaging the surface of the sea floor [11]. Since radio waves including the GPS signals tend to be blocked in the sea, acoustic self-localization is used by the submarine robots [21]. Beacon buoys that can observe GPS signals submit beacon sounds into the sea and the submarine robots localize themselves from the beacon signals.

2.1.3 Ground Rescue Robots

Many ground rescue robots have microphones for finding a victim and detecting relevant auditory events [75–79]. The microphones are also used by an operator for feeding back on the behavior of a robot (e.g., slipping or not) [80]. General methods for audio scene analysis, such as ego-noise reduction, localization, and separation, can be applied to the rescue robots that move on the ground. Several ground rescue robots have microphone arrays and localize sound events around the robot [81, 82]. Audio scene analysis for flexible or snake-like robots that penetrate into gaps under rubble, however, has not been well investigated. Most of these robots only have a single microphone and the operators have to carefully listen to the captured signal and check whether the victim exists [25].

2.2 Speech Enhancement

This section first introduces conventional single-channel and multichannel methods for speech enhancement. Then, enhancement methods with deep learning are introduced, which recently gained a lot of attention.

2.2.1 Single-channel Speech Enhancement

Single-channel speech enhancement or source separation works based on the spectral pattern difference between speech and noise signals. For example, conventional Wiener filters and spectral subtraction methods assume that the noise spectrogram is stable and extract the non-stable speech signal from a noisy audio signal. Their performances easily deteriorate in actual recordings because real noise signals are often unstable (e.g., engine sounds of automobiles). An approach for dealing with non-stationary noise signals is to model speech and noise spectrograms with hidden Markov models representing the dynamics of spectral changes [83]. This approach, however, requires a lot of training data for both speech and noise signals in advance. Another approach for non-stationary noise signals is to use non-negative matrix factorization (NMF). NMF assumes that an observed mixture signal is represented by spectral basis (template) vectors and their activation vectors [84, 85]. By preparing the basis vectors for speech signals in advance, NMF can estimate those for noise from an observed signal and separate speech and noise signals [85, 86]. Since the spectrogram of a speech signal has large time variation and cannot be represented by a small number of basis vectors, its performance for speech enhancement is limited.

Low-rank and sparse decomposition can suppress non-stationary noise and enhance speech without prior training [87–91]. Robust principal component analysis (RPCA), for example, decomposes an amplitude spectrogram into low-rank and sparse spectrograms corresponding to noise and speech [87,88]. RPCA can be extended in a Bayesian manner to deal with the uncertainty of latent low-rank and sparse components [92,93]. Ding et al. [92] proposed a Bayesian RPCA whose prior distribution of sparse components has a Markovian constraint. This model was used for separating background and foreground images from video streams and reduced salt-and-pepper noise of estimated foreground images. Application of RPCA to audio data, however, is not physically justified because RPCA allows input, low-rank, and sparse amplitude spectrograms to take negative values. Robust NMF (RNMF) has been studied for decomposing an input non-negative matrix into non-negative low-rank and sparse matrices [90, 91].

2.2.2 Multichannel Speech Enhancement

Multichannel speech enhancement or source separation is conducted based on not only the spectral pattern information but also the power and time differences across microphones. Since these differences depend only on the geometric configurations of microphones and sound sources, multichannel signal separation can be robust against the spectral characteristics of source signals. In addition, since the microphones on a hose-shaped rescue robot are distributed on the whole body, multichannel speech enhancement can deal with microphone occlusion problem by using only the available microphones.

The most basic method of multichannel speech enhancement is beamforming. By using steering vectors that represent the spatial relationship of the sound sources and microphones, beamformers can extract each of source signals from a multichannel mixture recording [36,94,95]. The steering vectors are estimated from the relative locations of sources and microphones. In the case of a hose-shaped rescue robot, obtaining such information precise enough for beamforming is difficult. The steering vectors can also be estimated from the power spectrograms of speech signals obtained by the single-channel enhancement [96]. This approach can deal with microphone failures (and occlusions) by using the speech power at each microphone. Since the single-channel and multichannel enhancements are serially cascaded, the performance will be improved by feeding back the multichannel results to the single-channel results.

Blind source separation based on the phase differences between the microphones can be used without prior knowledge about microphones and sources [37, 53–55,97,98]. Frequency-domain independent component analysis (FD-ICA), for example, separates sound sources by maximizing the statistical independence between the separated sources. Since FD-ICA is independently conducted at frequency bins, it has signal permutation ambiguities over frequency bins. To solve the permutation problem, independent vector analysis (IVA) was proposed by modeling the source spectra as multivariate random vectors. Since this method assumes that source signals are stationary, its performance is degraded by the mixture recordings of non-stationary source signals.

Multichannel non-negative matrix factorization (MNMF) [54, 55, 99] decomposes given multichannel complex spectrograms into multiple low-rank source spectrograms and their transfer functions. As in single-channel NMF, each of source spectrograms is represented as a product of spectral basis vectors and their temporal activation vectors. Kitamura et al. [55] proposed independent low-rank matrix analysis (ILRMA), which is a variant MNMF, by integrating IVA and NMF. Kounades-Bastian et al. [99] extended MNMF for moving sources by assuming a Markov chain of time-varying transfer functions. Its performance may, however, be degraded by the unexpected moving of sources.

One way to avoid estimating the time-varying transfer functions of sound sources is to perform source separation over multichannel magnitude (or power) spectrograms, which are insensitive to relatively small motions. Non-negative tensor factorization (NTF) [100–103] has been used for decomposing the multichannel spectrogram into source spectrograms and their magnitude transfer functions. Murata et al. [101] proposed an NTF by marginalizing out the phase term of MNMF. This method, however, requires the basis vectors for each source in advance. Although another NTF that does not need information about the basis vectors was proposed [103], it requires the volume level ratio of each source in the channels in advance. To make NTF completely blind, it is necessary to import other separation criteria that can remove the constraints.

2.2.3 Speech Enhancement with Deep Learning

With the high non-linearity and model flexibility, deep neural network (DNN)-based speech enhancement demonstrates excellent performance. Various network architectures and cost functions for enhancing speech signals have been reported [51, 52, 104–107]. The popular approach of DNN-based speech enhancement is to train a DNN to represent clean speech directly [107]. The DNN is trained using simulated noisy data constructed by adding noise to speech as input and clean speech as the target. Several methods combine a supervised NMF and a DNN [108, 109]. A DNN is trained to estimate activation vectors of the pre-trained basis vectors corresponding to speech and noise. Bayesian

WaveNet [110] uses two networks: one, called a prior network, represents how likely a signal is speech and the other, called a likelihood network, represents how likely a signal is included in the observation. These two networks enhance the noisy speech signal with a maximum a posteriori (MAP) estimation. Another reported method uses two networks that are trained to represent how likely the input signal is speech or noise, respectively [111]. The speech signal is enhanced by optimizing a cost function so that the estimated speech maximizes the speech-likelihood network and minimizes the noise-likelihood network. All the methods mentioned above are trained with sufficient amount of datasets of both speech and noise signals.

A DNN-based method using only training data of speech signals was reported [112]. This method represents speech and noise spectra with two autoencoders (AEs). The AE for speech is pre-trained, whereas that for noise is trained at the inference for adapting to the observed noise signal. Since the inference in this framework is under-determined, the estimated speech is constrained to be represented by a pre-trained NMF model. It, thus, might have the same problem as the semi-supervised NMF.

DNN-based models have a potential to be a source model of blind speech enhancement described in the above sections. One way to represent the complex distribution of speech signals is to use a deep generative model. Deep generative models [63–65] are deep-learning frameworks for representing a probabilistic distribution of a training data set. Instead of formulating a parametric prior distribution of speech signals, such models trained with a large number of clean speech signals can be used [66,67,113,114]. The recently proposed frameworks of generative adversarial networks (GANs) and variational autoencoders (VAEs) can represent the probabilistic distribution of speech signals [66, 67, 113] and are used to synthesize speech signals. A VAE is formulated as a Bayesian probabilistic model that can be integrated into other probabilistic frameworks. A speech signal can be separated from a noisy input signal by estimating the posterior distribution of the integrated probabilistic generative process of the noisy speech signal.

2.3 Posture Estimation

This section first introduces existing methods for estimating the posture of a flexible cable, and then reviews methods for simultaneous localization of microphones and sound sources.

2.3.1 Posture Estimation of Flexible Cables

Posture estimation for a flexible cable has been studied for medical robots [57, 115]. Lee et al. [57] estimated the posture of a medical robot that works in the colon by using a set of 3-axis accelerometers and 3-axis magnetometers. The robot posture is modeled as a link model and estimated by using orientation information obtained by the sensors. Tully et al. [115] reported a posture estimation method for a surgical robot. This method is also based on magnetometers. Because of sensor noise or poor estimation models, conventional methods often output unrealistic locations (e.g., inside of an organ when the robot is outside of it). The authors proposed a Kalman filter that truncates the probabilistic density function of latent variables. This truncation constrains the robot posture to be a possible location. These methods, however, cannot be used for hose-shaped rescue robots used in collapsed buildings where magnetic fields are easily distorted.

Posture estimation has also been developed for towed array sonars, which are one of the passive sonars [73,74,116,117]. The towed array sonar is a hydrophone array used for localizing other ships or submarines. The posture is estimated by using orientation sensors that consist of depth sensors and magnetometers. To deal with the situation when these sensors are not available, audio-based posture estimation have long been studied [74, 116, 117]. The posture is estimated from the sounds caused by other ships around the array. Since the sound source is assumed to be far from the towed array, such a method uses a plane-wave source model, which is parameterized only with the direction of a sound source instead of its location. This assumption does not hold on the microphones and loudspeakers on a hose-shaped rescue robot.

Ishikura et al. [56] estimated the posture of a hose-shaped rescue robot by using gyroscopes, which measures 3-axis angular velocities. By integrating the observed angular velocities, current sensor orientations can be estimated. They formulated a flexible posture model called absolute nodal coordinate formulation, which is based on the Euler-Bernoulli beam theory. The time-varying posture is recursively estimated based on this posture model and an observation model for the gyroscopes by using an unscented Kalman filter. Since this method uses only inertial sensors, the estimation accuracy is not affected by the environment around a robot. The estimation performance, however, deteriorates as time passes because gyroscopes have the accumulative error problem. In addition, the posture model requires high computational costs for simulating the curve of the flexible robot.

2.3.2 Simultaneous Localization of Microphones and Sources

Simultaneous localization of microphones and sound sources have been studied in various problem settings. A simple way is to use time of arrivals (ToAs), which are flight times of sounds from sources to microphones. By multiplying a ToA by the speed of sound (e.g., 340 m/s), the distance between a microphone and a sound source can be obtained. Based on the distances obtained from ToAs, the location of the microphones and sources can be determined with a closed form solution [118]. Since ToA-based localization needs synchronization of both the microphones and sound sources (loudspeakers), alternative methods that do not need synchronization between them have been studied.

Chen et al. [119] reported a method that uses the energies of a sound measured at microphones, each of which depends on the distance between the sound source and the microphone. This method estimates the microphone and loudspeaker positions based on the attenuation model of a sound. Compared to the ToA-based methods, this approach needs rough synchronization among microphones and sources because the energies are measured by taking the averages of recorded signals. The performance of this method is, however, severely degraded by external noise, which easily distorts the energy measurements.

Time differences of arrivals (TDOAs) are used for the simultaneous localization that requires only the synchronization of microphones [18, 22, 61, 120, 121]. A TDOA is the onset time differences between two microphones observing a single source signal. A TDOA provides the distance difference between two microphones and a source. The TDOA estimation is more robust against noise signals than the energy estimation when the noise signals and a target reference signal are uncorrelated [122]. Le et al. [121] derived a closed form solution for this localization problem. This method can be applied even when the microphones are not synchronized because it estimates the recording offset times at microphones. Bando et al. [123] proposed a audio-based posture estimation for a hose-shaped robot by using microphones and loudspeakers put on the robot. This method designed to estimate the stable locations of the microphones and loudspeakers. Miura et al. [22] built a TDOA-based online method that localizes a moving sound source and asynchronous microphones. The latent variables are estimated based on a SLAM frame work with an extended Kalman filter (EKF-SLAM). The method localizes microphones as the map and a moving sound source as the self-location. Several methods allow the microphone moving by assuming that the source locations are fixed or some of microphones and sources are relatively fixed [124, 125]. To estimate the time-varying posture of a hose-shaped rescue robot, the dynamic configuration of both the microphones and loudspeakers on the robot has to be estimated.

2.4 Summary

This chapter reviewed audio scene analysis for rescue robots and summarized the works relevant to speech enhancement and posture estimation. Audio scene analysis has widely been studied for the aerial and marine robots and the robots that move on the ground. That for flexible or snake-like ground robots penetrating into narrow gaps has not been well investigated. The dynamic configuration and partial occlusion of microphones are the remaining problems of both speech enhancement and posture estimation for such rescue robots.

Chapter 3

Blind Speech Enhancement on Multichannel Magnitude Spectrograms

This chapter presents a multichannel blind speech enhancement method robust against the dynamic configuration and partial occlusion of microphones. The method is formulated as a Bayesian generative model and extended to a state space model for real-time enhancement.

3.1 Introduction

Blind speech enhancement has been studied for various applications such as speech recognition and speech telecommunication [37, 40, 51, 52, 126–128]. In these applications, it is often difficult to assume the typical usage situation, such as noise characteristics and the relative layout of sources and microphones. To enhance noisy speech signals with few assumptions on the usage situation, blind speech enhancement has been studied by focusing on some statistical structures of observed signals [37, 40, 54, 91, 128, 129]. Single-channel speech enhancement, for example, focuses on the spectral pattern difference between speech and noise signals [126, 128]. Multichannel speech enhancement focuses on the inter-channel correlation difference between them, which depends on the relative layout of sources and microphones [37, 40, 54].

This study addresses developing a blind multichannel speech enhancement method that is robust against the time-varying layout of sources and microphones. While most of the existing blind speech enhancement (or source sepa-

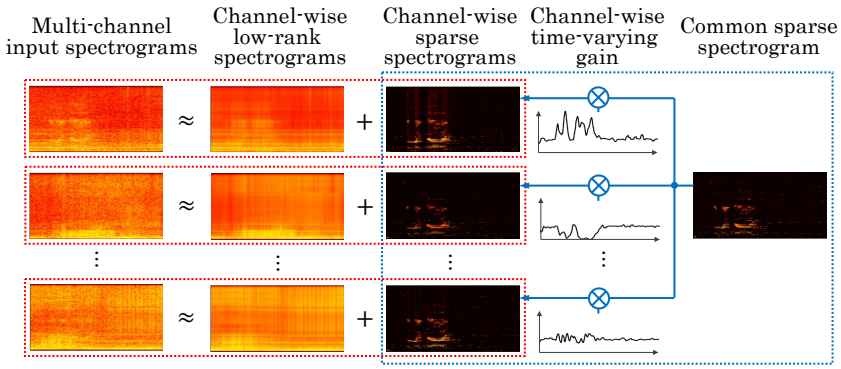


Figure 3.1: Overview of the proposed Bayesian RNTF.

ration) methods assume that the mixing system is time-invariant, this assumption does not always hold [53–55, 97]. A possible way is to enhance speech in magnitude (or power) spectrogram domain, which is insensitive to relatively small changes of the layout. Non-negative tensor factorization (NTF), for example, can separate a multichannel magnitude spectrogram into source spectrograms [100–103]. NTF, however, requires prior information such as the spectral bases (templates) of each source spectrogram, and thus it is not completely blind.

This chapter presents blind multichannel speech enhancement that works in magnitude spectrogram domain based on low-rank and sparse decomposition. Low-rank and sparse decomposition, such as robust non-negative matrix factorization (RNMF), can decompose a magnitude spectrogram into low-rank and sparse spectrograms without any prior training [87–91, 130]. The low-rank spectrogram corresponds to a noise spectrogram that can be represented by a small number of spectral bases (e.g., motor noises). The sparse spectrogram corresponds to a speech spectrogram that has harmonic structures. The method is inspired by NTF and RNMF, and decomposes a multichannel magnitude spectrogram into channel-wise low-rank noise spectrograms and sparse speech spectrogram common to all the channels (Figure 3.1). It is formulated as a Bayesian generative model called Bayesian robust NTF (Bayesian RNTF). Since its mixing system is independently estimated at each time frame, it is robust against the time-varying layout of sources and microphones.

Bayesian RNTF is applied to speech enhancement with a microphone array

on a hose-shaped rescue robot. The following three characteristics make the speech enhancement for the robot difficult:

1. **Environment-dependence of ego-noise**: The ego-noise changes over time depending on the robot's movements and surrounding materials.
2. **Dynamic configuration of microphones**: The relative positions of the microphones change over time because of the vibration and deformation of the robot body.
3. **Partial occlusion of microphones**: Some of the microphones often fail to capture target speech when they are shaded by rubble around the robot.

These problems make it impossible to use conventional supervised methods [34, 35, 51, 52, 127, 131], and degrade the conventional blind methods that assume a time-invariant mixing system [53–55, 97]. On the positive side, since the ego-noise is generated from the vibration motors, the noise spectrogram has repetitive structures and is, thus, considered as low rank. The proposed Bayesian RNTF is based on the low-rank and sparse decomposition and time-varying mixing system, and thus it is robust against the first two problems. In addition, it can deal with the occlusion problem because it estimates the speech level at each microphone.

In actual rescue activities searching for victims, real-time speech enhancement is crucial. Bayesian RNTF is extended to a state-space model called Bayesian streaming RNTF (Bayesian SRNTF) that represents the dynamics of the latent variables. The Bayesian inference of Bayesian SRNTF is conducted in a mini-batch manner with a variational Bayesian (VB) framework [62, 93]. Experimental results show that the method works in real time on a mobile general-purpose graphics processing unit (GPGPU).

3.2 Low-Rank and Sparse Decomposition

Low-rank and sparse decomposition is a popular approach to suppressing non-stationary periodic noise and enhancing target speech without prior training

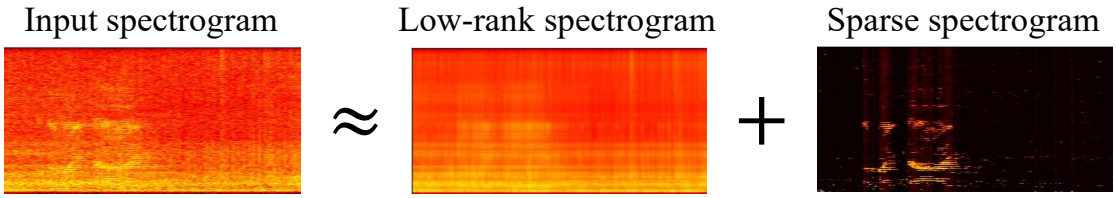


Figure 3.2: Speech enhancement by low-rank and sparse decomposition. Speech signals that have a sparse structure are separated from low-rank noise signals.

(Figure 3.2) [87–91]. Let $\mathbf{Y} \in \mathbb{R}^{F \times T}$, $\mathbf{L} \in \mathbb{R}^{F \times T}$, and $\mathbf{S} \in \mathbb{R}^{F \times T}$ be input, low-rank and sparse matrices (magnitude spectrograms with F frequency bins and T time frames), respectively. This decomposition was originally proposed in robust principal component analysis (RPCA) and is conducted by solving the following minimization problem with the augmented Lagrange multiplier framework [88]:

$$\underset{\mathbf{L},\mathbf{S}}{\operatorname{argmin}} \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \quad \text{s.t.} \quad \mathbf{Y} = \mathbf{L} + \mathbf{S}, \tag{3.1}$$

where $\|\cdot\|_*$ is the nuclear norm representing the low-rankness, $\|\cdot\|_1$ is the L1 norm representing the sparsity, and λ represents a scale parameter controlling the sparseness of $\mathbf{S}$. To reduce the processing time of RPCA, the following relaxed problem of Eq. (3.1) is proposed by replacing the equality constraint with a penalty term [132, 133]:

$$\underset{\mathbf{L},\mathbf{S}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y} - \mathbf{L} - \mathbf{S}\|_F^2 + \lambda_1 \|\mathbf{L}\|_* + \lambda_2 \|\mathbf{S}\|_1, \tag{3.2}$$

where $\|x\|_F$ is the Frobenius norm and λ_1 and λ_2 are the scale parameters. When these scale parameters are small enough, the solutions to Eq. (3.2) approach the solutions to Eq. (3.1).

Eq. (3.2) can be interpreted as a log-likelihood function ($\frac{1}{2}\|\mathbf{Y} - \mathbf{L} - \mathbf{S}\|_F^2$) with priors for the latent variables ($\lambda_1\|\mathbf{L}\|_*$ and $\lambda_2\|\mathbf{S}\|_1$). Bayesian RPCA has been studied for dealing with uncertainty of latent low-rank and sparse components [92, 93]. Babacan et al. [93] derived a VB algorithm for Bayesian RPCA (VB-RPCA) to reduce the computational cost. Bayesian RPCA represents the low-rank matrix $\mathbf{L}$ as the product of K basis vectors $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}^{F \times K}$ and their coefficient vectors $\mathbf{H} \in \mathbb{R}^{K \times T}$ as follows:

$$\mathbf{L} = \mathbf{W}\mathbf{H}. \tag{3.3}$$

Note that the rank of the low-rank matrix $\mathbf{L}$ is constrained to be K or less. Using this low-rank model, the likelihood function is defined with a Gaussian distribution (denoted by $\mathcal{N}$) as follows:

$$p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S}) = \prod_{f,t} \mathcal{N}\left(y_{ft} \left| \sum_k w_{fk}h_{kt} + s_{ft}, \sigma \right.\right) \propto \exp\left(-\frac{1}{\sigma}\|\mathbf{Y}-\mathbf{W}\mathbf{H}-\mathbf{S}\|_F^2\right), \tag{3.4}$$

where σ is a variance parameter and is simultaneously estimated with other parameters. The low-rankness and sparseness of $\mathbf{L}$ and $\mathbf{S}$ are controlled by this structural constraint and their prior distributions.

By constraining the low-rank and sparse matrices to be non-negative, RNMF was proposed for analyzing audio spectrograms or video images [90, 91, 128–130]. Since the Frobenius norm (Euclidean distance) in Eqs. (3.2) and (3.4) often causes over-emphasis of high-energy components in a magnitude spectrogram, Li et al. [128] proposed an RNMF with the Kullback-Leibler (KL) divergence, which has been widely used in audio source separation:

$$\underset{\mathbf{W},\mathbf{H},\mathbf{S}}{\operatorname{argmin}} \ \mathrm{KL}(\mathbf{Y}|\mathbf{W}\mathbf{H}+\mathbf{S}) + \lambda\|\mathbf{S}\|_1, \tag{3.5}$$

where $\mathrm{KL}(\cdot|\cdot)$ represents the KL divergence. Min et al. [129] proposed an RNMF with the Itakura-Saito divergence, which is derived from a statistical generative model of acoustic signals.

3.3 Robust Non-Negative Tensor Factorization

This section describes the proposed RNTF model that represents a multichannel magnitude spectrogram by channel-wise low-rank components and sparse components common to all the channels as shown in Figure 3.1. Since the proposed method does not use the phase information, the phase differences across channels do not affect the result. To derive the proposed multichannel model, this section first formulates a Bayesian reformulation of RNMF (Bayesian RNMF) that is inspired by Bayesian NMF [134] and Bayesian RPCA [93]. Then, its multichannel extension (Bayesian RNTF) is formulated as a statistical generative

model, and finally the mini-batch extension called Bayesian SRNTF is derived by reformulating the batch Bayesian RNTF to a state-space model.

3.3.1 Bayesian RNMF for Single-Channel Enhancement

This sub-section formulates an offline single-channel enhancement model called Bayesian RNMF. The problem of Bayesian RNMF is defined as follows:

Input: Single-channel magnitude spectrogram $\mathbf{Y} \in \mathbb{R}_{+}^{F \times T}$
Output: Denoised magnitude spectrogram $\mathbf{S} \in \mathbb{R}_{+}^{F \times T}$
Assumption: The following values are given in advance:
A) Possible maximum rank of noise spectrogram $K \in \mathbb{N}$
B) Hyperparameters $\alpha^w \in \mathbb{R}_{+}$, $\beta^w \in \mathbb{R}_{+}$, $\alpha^h \in \mathbb{R}_{+}$, $\beta^h \in \mathbb{R}_{+}$, and $\alpha^s \in \mathbb{R}_{+}$

where F and T indicate numbers of frequency bins and time frame bins, respectively. The magnitude spectrogram is defined as the absolute values of the short-time Fourier transform (STFT) of a time-domain signal. Interpretations of the hyperparameters are explained below.

Overview

As in existing low-rank and sparse decomposition methods (Eqs. (3.2), (3.4), and (3.5)), Bayesian RNMF approximates an input spectrogram $\mathbf{Y} \in \mathbb{R}_{+}^{F \times T}$ as the sum of a low-rank spectrogram $\mathbf{L} \in \mathbb{R}_{+}^{F \times T}$ (noise) and a sparse spectrogram $\mathbf{S} \in \mathbb{R}_{+}^{F \times T}$ (target speech) as follows:

$$\mathbf{Y} \approx \mathbf{L} + \mathbf{S}. \tag{3.6}$$

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 $\mathbf{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T] \in \mathbb{R}_{+}^{K \times T}$:

$$\mathbf{Y} \approx \mathbf{W}\mathbf{H} + \mathbf{S}. \tag{3.7}$$

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

Likelihood Function

Bayesian RNMF tries to minimize the approximation error for the input spectrogram by using the KL divergence. Since the maximization of a Poisson likelihood corresponds to the minimization of a KL divergence, the likelihood function is defined as follows:

$$p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S}) = \prod_{ft} \mathcal{P}\left(y_{ft} \middle| \sum_{k} w_{fk}h_{kt} + s_{ft}\right), \tag{3.8}$$

where $\mathcal{P}(x|\lambda) \propto \frac{1}{\Gamma(x+1)}\lambda^x e^{-\lambda}$ denotes a Poisson distribution with a rate parameter $\lambda \in \mathbb{R}_+$. Although the discrete Poisson distribution can be used by quantizing the observation y_{ft}, it has been empirically shown that NMF with the continuous Poisson likelihood performs as well as those of the discrete distribution [135].

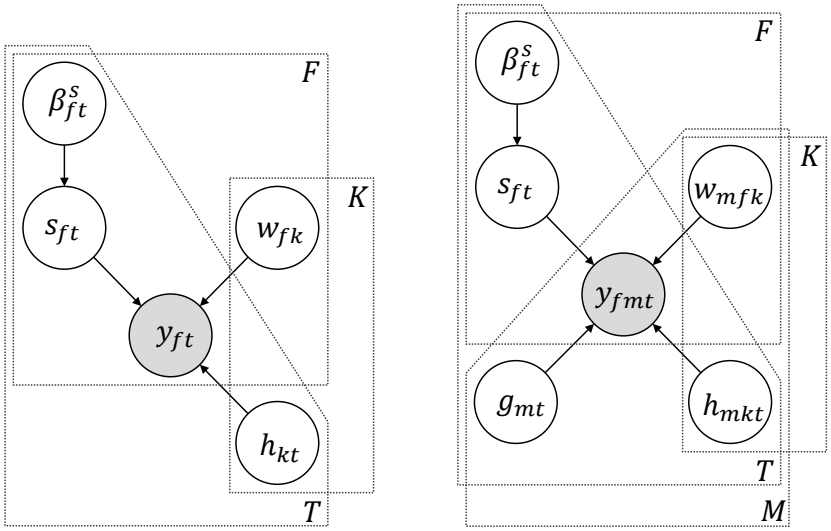
Prior Distributions on Low-Rank Components

The proposed low-rank modeling is inspired by Bayesian NMF [134] that has been studied for low-rank decomposition of audio spectrograms. Since the gamma distribution 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:

$$p(\mathbf{W}|\alpha^w,\beta^w) = \prod_{f,k} \mathcal{G}(w_{fk}|\alpha^w,\beta^w), \tag{3.9}$$

$$p(\mathbf{H}|\alpha^h,\beta^h) = \prod_{k,t} \mathcal{G}(h_{kt}|\alpha^h,\beta^h), \tag{3.10}$$

where $\mathcal{G}(x|\alpha,\beta)$ denotes a gamma distribution with a shape parameter α and a rate parameter β; $\alpha^w \in \mathbb{R}_+$, $\beta^w \in \mathbb{R}_+$, $\alpha^h \in \mathbb{R}_+$, and $\beta^h \in \mathbb{R}_+$ are the hyper-parameters which should be appropriately set in advance. Setting the shape parameters α^w and α^h to 1.0 or less forces the basis and activation matrices to be sparse [134], which means that the low-rank component $\mathbf{L}$ is forced to be low-rank. These prior distributions enhance the low-rankness of this component compared to the original RNMF.



(a) Bayesian RNMF (b) Bayesian RNTF

Figure 3.3: Graphical models for Bayesian RNMF and RNTF.

Prior Distributions on Sparse Components

In Bayesian RPCA, Gaussian priors with the Jeffreys hyperpriors are put on sparse components [93]. To force the sparse components to take non-negative values, gamma priors are put on the sparse components as follows:

$$p(\mathbf{S}|\alpha^s, \boldsymbol{\beta}^s) = \prod_{f,t} \mathcal{G}(s_{ft}|\alpha^s, \beta^s_{ft}), \tag{3.11}$$

where $\alpha^s \in \mathbb{R}_+$ and $\beta^s_{ft} \in \mathbb{R}_+$ represent the shape and rate hyperparameters of the gamma distributions, respectively. To estimate the rate hyperparameters, the Jeffreys hyperpriors are put on them as follows:

$$p(\beta^s_{ft}) \propto (\beta^s_{ft})^{-1}. \tag{3.12}$$

The rate hyperparameters are independently defined at individual time-frequency bins. The significance of each time-frequency bin is automatically estimated by optimizing the rate hyperparameter as in Bayesian RPCA [93]. The shape hyperparameter α^s, on the other hand, controls the sparseness of the sparse component $\mathbf{S}$ and should be set appropriately in advance. The complete graphical model that represents the probabilistic dependency of the latent variables is shown in Figure 3.3-(a).

3.3.2 Bayesian RNTF for Multichannel Enhancement

This sub-section then formulates a multichannel extension of Bayesian RNMF called Bayesian RNTF. The problem in this subsection is defined as follows:

Input: M-channel magnitude spectrograms $\mathbf{Y}_m \in \mathbb{R}_+^{F \times T}$
Output: Denoised magnitude spectrogram $\mathbf{S} \in \mathbb{R}_+^{F \times T}$
Assumption:
The following values are given in advance:
A) Possible maximum rank of noise spectrogram $K \in \mathbb{N}$
B) Hyperparameters $\alpha^w \in \mathbb{R}_+$, $\beta^w \in \mathbb{R}_+$, $\alpha^h \in \mathbb{R}_+$, $\beta^h \in \mathbb{R}_+$, $\alpha^g \in \mathbb{R}_+$, and $\alpha^s \in \mathbb{R}_+$

where m represents the microphone index. Interpretations of the hyperparameters are explained below. Bayesian RNTF is designed for enhancing speech sounds coming from one direction at each time frame. This is considered to be reasonable because multiple speakers located at different directions may not talk simultaneously in disaster situations. Even when a few people speak simultaneously from the same direction, the overlapping speech sounds could be enhanced because those sounds still have sparse harmonic structures and the fine time-frequency fluctuations of speech spectrograms violate the low-rank assumption.

Overview

Bayesian RNTF approximates an input spectrogram at each channel $\mathbf{Y}_m \in \mathbb{R}_+^{F \times T}$ as the sum of channel-wise low-rank spectrogram and channel-wise sparse spectrogram $\mathbf{S}'_m \in \mathbb{R}_+^{F \times T}$:

$$\mathbf{Y}_m \approx \mathbf{W}_m \mathbf{H}_m + \mathbf{S}'_m, \tag{3.13}$$

where $\mathbf{W}_m \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H}_m \in \mathbb{R}_+^{K \times T}$ are channel-wise basis and activation matrices for the low-rank spectrogram, respectively.

The relationship between the target speech signal $\mathbf{S} \in \mathbb{R}_+^{F \times T}$ and its observation at each microphone $\mathbf{S}'_m$ is assumed to be a time-variant and frequency-invariant linear system:

$$s'_{mft} \approx g_{mt} s_{ft}, \tag{3.14}$$

where $g_{mt} \in \mathbb{R}_+$ represents a gain of the target speech signal at microphone m and time t. According to Eqs. (3.13) and (3.14), Bayesian RNTF decomposes the input spectrogram $\mathbf{Y}_m$ into the following four components:

$$y_{mft} \approx \sum_k w_{mfk} h_{mkt} + g_{mt} s_{ft}. \tag{3.15}$$

where $\boldsymbol{g}_m = [g_{m1}, \ldots, g_{mT}]$ is a gain vector. Although magnitude spectrograms are insensitive to relatively small motions [101], the gain g_{mt} depends on the motion of microphones and target speech. The gain g_{mt} is, therefore, independently estimated at each time frame to deal with the movement of microphones and sources.

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 manner as in Bayesian RNMF (Eqs. (3.8) – (3.12)). A gamma prior is put on g_{mt} assuming that its mean is 1:

$$p(g_{mt}|\alpha^g) = \mathcal{G}(g_{mt}|\alpha^g, \alpha^g), \tag{3.16}$$

where $\alpha^g \in \mathbb{R}_+$ is a hyperparameter controlling the variance of the gain parameters. The complete graphical model is shown in Figure 3.3-(b).

3.3.3 Bayesian Streaming RNTF for Real-Time Enhancement

This subsection describes the Bayesian SRNTF. It is formulated as a state-space model representing the latent variable as time-varying latent variables.

Overview

Bayesian SRNTF sequentially enhances target speech for T frames of mini-batch audio inputs (Figure 3.4). The problem in this subsection is defined as follows:

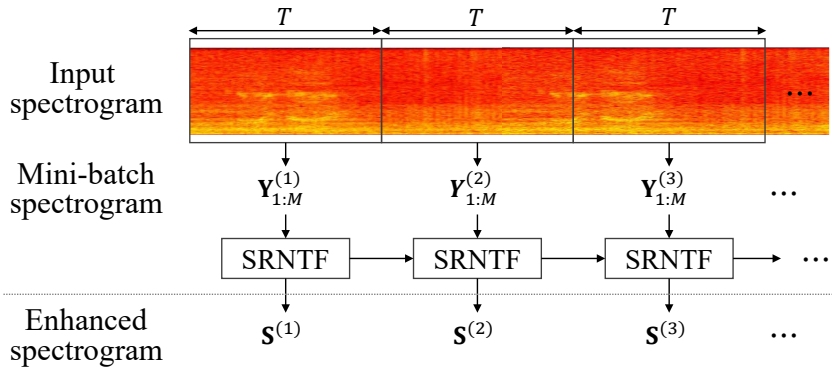


Figure 3.4: Mini-batch processing flow of Bayesian SRNTF.

Input:

1. M-channel magnitude spectrograms $\mathbf{Y}^{(n)}_m \in \mathbb{R}^{F\times T}_+$
2. Posterior distribution at the previous $(n-1)$ mini-batch.

Assumption:

The following values are given in advance:

A) Possible maximum rank of noise spectrogram $K \in \mathbb{N}$

B) Hyperparameters $\alpha^w \in \mathbb{R}_+$, $\beta^w \in \mathbb{R}_+$, $\alpha^h \in \mathbb{R}_+$, $\beta^h \in \mathbb{R}_+$, $\alpha^g \in \mathbb{R}_+$, $\alpha^s \in \mathbb{R}_+$, and $\gamma \in \mathbb{R}_+$

---

where n indicates the mini-batch index ($n = 1, 2, 3, \ldots$). As explained below, the posterior distribution at the prevous mini-batch is used for the prior information of the current latent variables. Interpretations of the hyperparameters are explained below.

Bayesian SRNTF decomposes the mini-batch audio spectrogram $y^{(n)}_{mft}$ into low-rank and sparse components in the same manner as in Bayesian RNTF:

$$y^{(n)}_{mft} \approx \sum_k w^{(n)}_{mfk} h^{(n)}_{mkt} + g^{(n)}_{mt} s^{(n)}_{ft}. \tag{3.17}$$

where $\mathbf{W}^{(n)}_m \in \mathbb{R}^{F\times K}_+$, $\mathbf{H}^{(n)}_m \in \mathbb{R}^{K\times T}_+$, $\boldsymbol{g}^{(n)}_m \in \mathbb{R}^{1\times T}_+$, and $\mathbf{S}^{(n)} \in \mathbb{R}^{F\times T}_+$ are the latent variables for the basis and activation matrices, gain, and sparse matrix notated in the same manner as in Bayesian RNTF, respectively. Let $\boldsymbol{\Theta}^{(n)}$ be a set of all the latent variables at the n-th mini-batch $\{\mathbf{W}^{(n)}_{1:M}, \mathbf{H}^{(n)}_{1:M}, \boldsymbol{g}^{(n)}_{1:M}, \mathbf{S}^{(n)}, \boldsymbol{\beta}^{s(n)}\}$. The proposed state-space model consists of an observation model $p(\mathbf{Y}^{(n)}_{1:M}|\boldsymbol{\Theta}^{(n)})$ and a state update model $p(\boldsymbol{\Theta}^{(n)}|\boldsymbol{\Theta}^{(n-1)})$ that represent the relationship between the observation and latent variables and the dynamics of the latent variables.

Observation Model

The observation model of Bayesian SRNTF $p(\mathbf{Y}_{1:M}^{(n)}|\mathbf{\Theta}^{(n)})$ is formulated with a Poisson distribution in the same manner as in Bayesian RNTF:

$$p\left(\mathbf{Y}_{1:m}^{(n)}\left|\mathbf{\Theta}^{(n)}\right.\right)=\prod_{m,f,t}\mathcal{P}\left(y_{mft}^{(n)}\left|\sum_{k}w_{mfk}^{(n)}h_{mkt}^{(n)}+g_{mt}^{(n)}s_{ft}^{(n)}\right.\right). \tag{3.18}$$

State Update Model

Since the latent variables for the sparse component ($\boldsymbol{g}_{1:M}^{(n)}$, $\mathbf{S}^{(n)}$, and $\boldsymbol{\beta}^{s(n)}$) and the activation matrix for the low-rank component ($\mathbf{H}_{1:M}^{(n)}$) are time-independent, only the basis matrix $\mathbf{W}_{m}^{(n)}$ depends on the previous state $\mathbf{W}_{m}^{(n-1)}$ in the state update model:

$$p\left(\mathbf{\Theta}^{(n)}\left|\mathbf{\Theta}^{(n-1)}\right.\right)=p\left(\mathbf{W}_{1:M}^{(n)}\left|\mathbf{W}_{1:M}^{(n-1)}\right.\right)p\left(\mathbf{H}_{1:M}^{(n)}\right)p\left(\boldsymbol{g}_{1:M}^{(n)}\right)p\left(\mathbf{S}^{(n)}\right)p\left(\boldsymbol{\beta}^{s(n)}\right). \tag{3.19}$$

The priors for $\mathbf{H}_{m}^{(n)}$, $\boldsymbol{g}_{m}^{(n)}$, $\mathbf{S}^{(n)}$, and $\boldsymbol{\beta}^{s(n)}$ are formulated in the same way as in the batch Bayesian RNTF (Section 3.3.2-2).

In this study, the state update model for $\mathbf{W}_{1:M}^{(n)}$ is independently formulated on each of its elements $w_{mfk}^{(n)}$:

$$p\left(\mathbf{W}_{m}^{(n)}\left|\mathbf{W}_{m}^{(n-1)}\right.\right)=\prod_{m,f,k}p\left(w_{mfk}^{(n)}\left|w_{mfk}^{(n-1)}\right.\right). \tag{3.20}$$

The state update model $p(w_{mfk}^{(n)}|w_{mfk}^{(n-1)})$ represents how $w_{mfk}^{(n)}$ varies from the previous state $w_{mfk}^{(n-1)}$. It has the following properties. The mean of $w_{mfk}^{(n)}$ should not be changed from that of $w_{mfk}^{(n-1)}$ because no bias on the update is assumed. The variance of $w_{mfk}^{(n)}$, on the other hand, should be increased from that of $w_{mfk}^{(n-1)}$ because its uncertainty increases over time. As proposed in [136], such an update model can be formulated with a multiplicative process noise $v_{mfk}^{(n)}\in\mathbb{R}_{+}$ as follows:

$$w_{mfk}^{(n)}=v_{mfk}^{(n)}w_{mfk}^{(n-1)}. \tag{3.21}$$

A beta prior distribution is put on $v_{mfk}^{(n)}$ as follows:

$$p(v_{mfk}^{(n)}|\alpha_{mfk}^{(n-1)},\gamma) = \mathcal{B}\left(v_{mfk}^{(n)}\gamma \left| \gamma\alpha_{mfk}^{(n-1)}, (1-\gamma)\alpha_{mfk}^{(n-1)}\right.\right), \tag{3.22}$$

where $\mathcal{B}(\alpha,\beta)$ represents a beta distribution with two shape parameters α and β, and $\gamma \in \mathbb{R}_+$ is a rate parameter controlling the variance of $w_{mfk}^{(n)}$. From Eqs. (3.21) and (3.22), the update model $p(w_{mfk}^{(n)}|w_{mfk}^{(n-1)})$ can be derived as follows:

$$p\left(w_{mfk}^{(n)}\left|w_{mfk}^{(n-1)}\right.\right) = \mathcal{B}\left(\gamma\frac{w_{mfk}^{(n)}}{w_{mfk}^{(n-1)}}\middle|\gamma\hat{\alpha}_{mfk}^{(n-1)}, (1-\gamma)\hat{\beta}_{mfk}^{(n-1)}\right). \tag{3.23}$$

As shown later (Section IV), the posterior $p(w_{mfk}^{(n-1)}|\mathbf{Y}^{(1:n-1)})$ is a gamma distribution $\mathcal{G}(w_{mfk}^{(n-1)}|\hat{\alpha}_{mfk}^{(n-1)},\hat{\beta}_{mfk}^{(n-1)})$ with a shape parameter $\hat{\alpha}_{mfk}^{(n-1)} \in \mathbb{R}_+$ and a rate parameter $\hat{\beta}_{mfk}^{(n-1)} \in \mathbb{R}_+$. As proven in [136], the predictive distribution $p(w_{mfk}^{(n)}|\mathbf{Y}^{(1:n-1)})$ is calculated from $p(w_{mfk}^{(n-1)}|\mathbf{Y}^{(1:n-1)})$ as follows:

$$\begin{aligned} p(w_{mfk}^{(n)}|\mathbf{Y}^{(1:n-1)}) &= \int p(w_{mfk}^{(n)}|w_{mfk}^{(n-1)})p(w_{mfk}^{(n-1)}|\mathbf{Y}^{(1:n-1)})dw_{mfk}^{(n-1)} \\ &= \mathcal{G}(\gamma\hat{\alpha}_{mfk}^{(n-1)}, \gamma\hat{\beta}_{mfk}^{(n-1)}). \end{aligned} \tag{3.24}$$

Note that the mean of this distribution is the same as that of the one in the previous state and its variance is γ^{-1} times larger than that of the one in the previous state.

3.4 Speech Enhancement Based on Bayesian RNTF

This section derives Bayesian inferences of the proposed Bayesian RNMF, RNTF, and SRNTF, and describes the processing flow of speech enhancement based on Bayesian RNTF. The inferences of these models are derived with the VB framework. In summary, the proposed enhancement methods are VB-RNMF, VB-RNTF, and VB-SRNTF collectively.

3.4.1 Variational Inference

The goal is to calculate the full posterior distributions of the proposed models. Since the true posterior is analytically intractable, it is approximated by using a VB algorithm [93, 134].

First, this sub-section shows a brief description of the VB inference framework. Let $\mathbf{Y}$ be an observation variable, $\mathbf{Z}_i$ ($i = 1, \ldots, I$) be parameters whose posterior distributions are estimated, and $\mathbf{\Theta} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_I\}$ be a set of all the parameters. Then, the true posterior distribution $p\left(\mathbf{\Theta}\,|\mathbf{Y}\right)$ is approximated by the product of variational posterior distributions $q\left(\mathbf{Z}_i\right)$ as follows:

$$p\left(\mathbf{\Theta}\,|\mathbf{Y}\right) \approx \prod_i q\left(\mathbf{Z}_i\right). \tag{3.25}$$

VB algorithm estimates the variational distributions $q\left(\mathbf{Z}_i\right)$ by maximizing the following lower bound $\mathcal{L}(q)$ of $\log p\left(\mathbf{Y}\right)$:

$$\begin{aligned}
\log p\left(\mathbf{Y}\right) &= \log \int p\left(\mathbf{Y}, \mathbf{\Theta}\right) d\mathbf{\Theta} && (3.26)\\
&\geq \int q\left(\mathbf{\Theta}\right) \log \frac{p\left(\mathbf{Y}, \mathbf{\Theta}\right)}{q\left(\mathbf{\Theta}\right)} d\mathbf{\Theta} && (3.27)\\
&= \left\langle \log p\left(\mathbf{Y}, \mathbf{\Theta}\right) \right\rangle - \left\langle \log q\left(\mathbf{\Theta}\right) \right\rangle \stackrel{\text{def}}{=} \mathcal{L}(q), && (3.28)
\end{aligned}$$

where $\langle x \rangle$ is the expectation operation. This maximization corresponds to the minimization of the KL divergence between the true and approximated distributions. $\mathcal{L}(q)$ is maximized by alternately and iteratively updating each of $q\left(\mathbf{Z}_i\right)$ as follows:

$$q(\mathbf{Z}_i) = \exp \left\langle \log p(\mathbf{Y}, \mathbf{\Theta}) \right\rangle_{\mathbf{\Theta} \setminus \mathbf{Z}_i}, \tag{3.29}$$

where $\mathbf{\Theta} \setminus \mathbf{Z}_i$ represents a subset of $\mathbf{\Theta}$ obtained by removing $\mathbf{Z}_i$ from $\mathbf{\Theta}$.

VB-RNMF

The target full posterior distribution of Bayesian RNMF is $p(\mathbf{W}, \mathbf{H}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{Y})$. Let $\mathbf{\Theta}$ be a set of all the parameters and $q(x)$ be a variational posterior distribution of x. Then, the true posterior distribution is approximated as follows:

$$p(\mathbf{\Theta}|\mathbf{Y}) \approx q(\mathbf{W})q(\mathbf{H})q(\mathbf{S})q(\boldsymbol{\beta}^s). \tag{3.30}$$

The approximated posterior distributions are estimated by taking a lower bound of $\log p\left(\mathbf{Y}\right)$ and maximizing it.

Since the expectation of $\log p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S})$ includes the following intractable expectations:

$$\langle \log p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S})\rangle = \sum_{f,t} y_{ft}\left\langle \log\left(\sum_k w_{fk}h_{kt} + s_{ft}\right)\right\rangle - \sum_{f,t,k}\langle w_{fk}h_{kt}\rangle - \sum_{f,t}\langle s_{ft}\rangle + \text{const.}, \tag{3.31}$$

this Poisson log-likelihood is lower-bounded by using Jensen's inequality [134]:

$$\begin{aligned}\langle \log p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S})\rangle \geq& \sum_{f,t,k} y_{ft}\phi_{ftk}\left\langle \log\left(\frac{w_{fk}h_{kt}}{\phi_{ftk}}\right)\right\rangle + \sum_{f,t} y_{ft}\psi_{ft}\left\langle \log\left(\frac{s_{ft}}{\psi_{ft}}\right)\right\rangle \\ &- \sum_{f,t,k}\langle w_{fk}h_{kt}\rangle - \sum_{f,t}\langle s_{ft}\rangle + \text{const.} \qquad (3.32)\\ \stackrel{\text{def}}{=}& \langle p'(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S})\rangle \qquad (3.33)\end{aligned}$$

where $\phi_{ftk} \in \mathbb{R}_+$ and $\psi_{ft} \in \mathbb{R}_+$ ($\sum_k \phi_{ftk} + \psi_{ft} = 1$) are the auxiliary variables. Using $\langle p'(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S})\rangle$, $\log p(\mathbf{Y})$ is lower-bounded as follows:

$$\begin{aligned}\log p(\mathbf{Y}) \geq& \langle \log p'(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S})\rangle + \langle \log p(\mathbf{W})\rangle + \langle \log p(\mathbf{H})\rangle \\ &+ \langle \log p(\mathbf{S}|\boldsymbol{\beta}^s)\rangle + \langle \log p(\boldsymbol{\beta}^s)\rangle - \langle \log q\,(\mathbf{W})\rangle - \langle \log q\,(\mathbf{H})\rangle \\ &- \langle \log q\,(\mathbf{S})\rangle - \langle \log q\,(\boldsymbol{\beta}^s)\rangle \stackrel{\text{def}}{=} \mathcal{L}(q). \qquad (3.34)\end{aligned}$$

By maximizing $\mathcal{L}(q)$ with the Lagrange multiplier method, the optimal ϕ_{ftk} and ψ_{ft} are obtained as follows:

$$\phi_{ftk} = \frac{\mathbb{G}[w_{fk}]\mathbb{G}[h_{kt}]}{\sum_k \mathbb{G}[w_{fk}]\mathbb{G}[w_{kt}] + \mathbb{G}[s_{ft}]}, \tag{3.35}$$

$$\psi_{ft} = \frac{\mathbb{G}[s_{ft}]}{\sum_k \mathbb{G}[w_{fk}]\mathbb{G}[h_{kt}] + \mathbb{G}[s_{ft}]}, \tag{3.36}$$

where $\mathbb{G}[x] = \exp(\langle \log x\rangle)$ represents the geometric expectation.

The update rules for the latent variables are obtained such that $\mathcal{L}(q)$ is maximized. Each variational posterior distribution is alternately and iteratively

updated by fixing the other distributions as follows:

$$q(w_{fk}) = \mathcal{G}(\alpha^w + \sum_t y_{ft}\phi_{ftk}, \beta^w + \sum_t \langle h_{kt}\rangle), \tag{3.37}$$

$$q(h_{kt}) = \mathcal{G}(\alpha^h + \sum_f y_{ft}\phi_{ftk}, \beta^h + \sum_f \langle w_{fk}\rangle), \tag{3.38}$$

$$q(s_{ft}) = \mathcal{G}(\alpha^s + y_{ft}\psi_{ft}, \langle \beta^s_{ft}\rangle + 1), \tag{3.39}$$

$$q(\beta^s_{ft}) = \mathcal{G}(\alpha^s, \langle s_{ft}\rangle). \tag{3.40}$$

VB-RNTF

The target full posterior distribution of Bayesian RNTF, $p(\mathbf{W}_{1:m}, \mathbf{H}_{1:m}, \boldsymbol{g}_{1:m}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{Y}_{1:m})$, is approximated in the same manner as that of Bayesian RNMF. The true posterior distribution is approximated as:

$$p(\boldsymbol{\Theta}|\mathbf{Y}_{1:M}) \approx \left\{\prod_m q(\mathbf{W}_m)q(\mathbf{H}_m)q(\boldsymbol{g}_m)\right\} q(\mathbf{S})q(\boldsymbol{\beta}^s). \tag{3.41}$$

The variational posterior distributions are calculated in the same way as in Bayesian RNMF, and each of them is alternately and iteratively updated as follows:

$$q(w_{mfk}) = \mathcal{G}(\alpha^w + \sum_t y_{mft}\phi_{mftk}, \beta^w + \sum_t \langle h_{mkt}\rangle), \tag{3.42}$$

$$q(h_{mkt}) = \mathcal{G}(\alpha^h + \sum_f y_{mft}\phi_{mftk}, \beta^h + \sum_f \langle w_{mfk}\rangle), \tag{3.43}$$

$$q(g_{mt}) = \mathcal{G}(\alpha^g + \sum_f y_{mft}\psi_{mft}, \alpha^g + \sum_f \langle s_{ft}\rangle), \tag{3.44}$$

$$q(s_{ft}) = \mathcal{G}(\alpha^s + \sum_m y_{mft}\psi_{mft}, \langle \beta^s_{ft}\rangle + \sum_m \langle g_{mt}\rangle), \tag{3.45}$$

$$q(\beta^s_{ft}) = \mathcal{G}(\alpha^s, \langle s_{ft}\rangle), \tag{3.46}$$

$$\phi_{mftk} = \frac{\mathbb{G}[w_{mfk}]\mathbb{G}[h_{mkt}]}{\sum_k \mathbb{G}[w_{mfk}]\mathbb{G}[h_{mkt}] + \mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}, \tag{3.47}$$

$$\psi_{mft} = \frac{\mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}{\sum_k \mathbb{G}[h_{mfk}]\mathbb{G}[h_{mkt}] + \mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}. \tag{3.48}$$

where ϕ_{mftk} and ψ_{mft} are auxiliary variables.

VB-SRNTF

VB-SRNTF estimates the current posterior distribution recurrently in prediction and correction steps. The prediction step calculates $p(\boldsymbol{\Theta}^{(n)}|\mathbf{Y}^{(1:n-1)})$ from the previous posterior distribution $p(\boldsymbol{\Theta}^{(n-1)}|\mathbf{Y}^{(1:n-1)})$:

$$p\left(\boldsymbol{\Theta}^{(n)}\left|\mathbf{Y}^{(1:n-1)}\right.\right) = \int p\left(\boldsymbol{\Theta}^{(n)}\left|\boldsymbol{\Theta}^{(n-1)}\right.\right) p\left(\boldsymbol{\Theta}^{(n-1)}\left|\mathbf{Y}^{(1:n-1)}\right.\right) d\boldsymbol{\Theta}^{(n-1)}.$$

According to Eqs. (3.19) and (3.20), the predictive distribution is calculated as follows:

$$p\left(\boldsymbol{\Theta}^{(n)}\left|\mathbf{Y}^{(1:n-1)}\right.\right) = \prod_{m,f,k} \mathcal{G}\left(w_{mfk}^{(n)}\left|\gamma\hat{\alpha}_{mfk}^{(n-1)}, \gamma\hat{\beta}_{mfk}^{(n-1)}\right.\right) \times p\left(\mathbf{H}^{(n)}\right) p\left(\boldsymbol{g}_m^{(n)}\right) p\left(\mathbf{S}^{(n)}\right) p\left(\boldsymbol{\beta}^{s(n)}\right), \tag{3.49}$$

where $\hat{\alpha}_{mfk}^{(n-1)}$ and $\hat{\beta}_{mfk}^{(n-1)}$ are the shape and rate parameters of the gamma distribution $p(w_{mfk}^{(n-1)}|\mathbf{Y}^{(1:n-1)})$, respectively. The correction step estimates the current posterior distribution $p(\boldsymbol{\Theta}^{(n)}|\mathbf{Y}^{(1:n)})$ from the observation $\mathbf{Y}^{(n)}$ and the predictive distribution $p(\boldsymbol{\Theta}^{(n)}|\mathbf{Y}^{(1:n-1)})$ as follows:

$$p\left(\boldsymbol{\Theta}^{(n)}\left|\mathbf{Y}^{(1:n)}\right.\right) \propto p\left(\mathbf{Y}^{(n)}\left|\boldsymbol{\Theta}^{(n)}\right.\right) p\left(\boldsymbol{\Theta}^{(n)}\left|\mathbf{Y}^{(1:n-1)}\right.\right). \tag{3.50}$$

In the correction step, the current posterior distribution is estimated in the same manner as in Eqs. (3.42)–(3.48) by replacing the prior distribution of Bayesian RNTF with the predictive distribution. The initial correction step (n = 1) uses the prior distribution of Bayesian RNTF (Section 3.3.2-2) as the predictive distribution.

3.4.2 Speech Enhancement Based on VB-SRNTF

Figure 3.5 shows the overall processing flow for the speech enhancement using VB-SRNTF. The proposed framework first takes the STFT of each microphone recording and obtains a multichannel magnitude spectrogram. Since each noisy input magnitude spectrogram includes fine fluctuations, the input spectrogram is smoothed for stable low-rank and sparse decomposition. Letting $\mathbf{Y}_m'^{(n)} \in \mathbb{R}_+^{F\times T}$

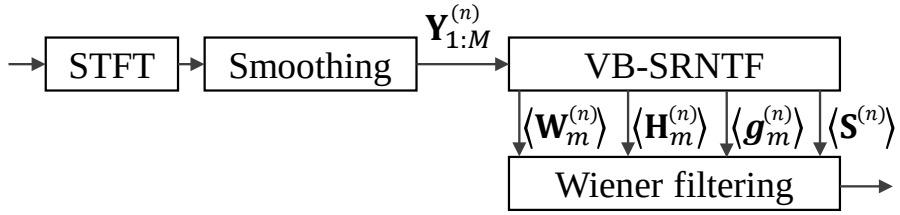


Figure 3.5: Processing flow of the proposed speech enhancement.

be the raw magnitude spectrogram obtained with the STFT, this smoothing preprocessing is conducted as follows:

$$y_{mft}^{(n)} = \frac{1}{9} \sum_{f'=f-1}^{f+1} \sum_{t'=t-1}^{t+1} y_{mf't'}^{\prime(n)}. \tag{3.51}$$

After conducting VB-SRNTF, the framework reconstructs the target speech signal at each microphone with Wiener filtering because VB-SRNTF cannot estimate the absolute scale of the target signal. Letting $\mathsf{Y}_m^{(n)} \in \mathbb{C}^{F \times T}$ be the input complex spectrogram at the m-th microphone, the complex spectrogram of the enhanced speech signal $\mathsf{S}_m^{(n)} \in \mathbb{C}^{F \times T}$ is obtained as follows:

$$\mathsf{s}_{mft}^{(n)} = \frac{\hat{s}_{mft}^{(n)}}{\hat{s}_{mft}^{(n)} + \hat{n}_{mft}^{(n)}} \mathsf{y}_{mft}^{(n)} \tag{3.52}$$

where $\hat{s}_{mft}^{(n)} \in \mathbb{R}_+$ and $\hat{n}_{mft}^{(n)} \in \mathbb{R}_+$ are the estimated power spectrograms of speech and noise signals, respectively. Finally, the time-domain output signal is obtained by taking the inverse STFT of the complex spectrogram.

3.5 Experimental Evaluation with Simulated Data

To analyze the performance of the proposed enhancement methods, and compare them with existing methods, these methods were evaluated using simulated audio signals.

3.5.1 Common Experimental Conditions

As shown in Figure 1.3, the body of the hose-shaped robot used in this evaluation was made from a corrugated tube of 38 mm in diameter and 3 m long. The entire

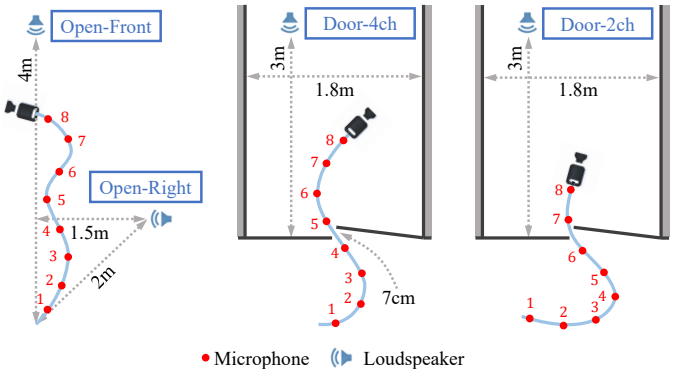


Figure 3.6: Four conditions of robot and loudspeaker in experimental evaluation.

surface of the robot was covered by cilia and seven vibrators used for moving forward by vibrating the cilia. This robot had an 8-ch synchronized microphone array whose microphones were 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 input signals were generated by mixing target speech and ego-noise signals at signal-to-noise ratios (SNRs) varying from -20 dB to $+5$ dB. As shown in Figure 3.6, there were four conditions differing in the relative positions of the robot and the loudspeaker (target speech).

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 loudspeaker was used

for recording the impulse response. Multichannel speech signals were generated by convoluting clean speech signals and the impulse response, and then they were mixed with 20 seconds of the ego-noise recordings. The clean speech data consisted of 24 recordings of three male and three female speech, which were included in the JNAS phonetically balanced Japanese utterances database [137]. In this setting, the location of the target speech did not change as the speech signal was generated with a single impulse response at each condition.

The enhancement performance was evaluated with the source-to-distortion ratio (SDR) [138], speech-to-overall ratio (SOR), and noise reduction ratio (NRR). The SDR measures the power ratio of the target speech and distortion component included in an output signal. Letting $s_t \in \mathbb{R}$ $(t = 1 \ldots, T)$ and $\hat{s}_t \in \mathbb{R}$ be the time-domain signals of reference and estimated speech signals, respectively, the SDR was calculated as follows:

$$\mathrm{SDR} = \frac{\sum_{t=1,\ldots,T} \left(\sum_\tau a_\tau s_{t-\tau}\right)^2}{\sum_{t=1,\ldots,T} (\hat{s}_t - \sum_\tau a_\tau s_{t-\tau})^2}, \tag{3.53}$$

where $a_\tau \in \mathbb{R}$ $(\tau = 0, \ldots, 127)$ is the filter coefficient that compensates the phase and power differences between the estimated and reference speech signals [138]. The SOR measures the average power ratio of the speech section and the whole section in an output signal. Letting $\mathcal{S}$ be the set of time indices where speech exists in the reference signal, the SOR was calculated as follows:

$$\mathrm{SOR} = \frac{\frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \hat{s}_t^2}{\frac{1}{T} \sum_{t=1,\ldots,T} \hat{s}_t^2} \tag{3.54}$$

where $|\mathcal{S}|$ represents the number of elements in $\mathcal{S}$. The SOR represents how prominent the speech is in the output signal. The NRR measures the power ratio of the estimated speech and reference noise signals at the non-speech sections. Let $n_t \in \mathbb{R}$ be the time-domain signal of a reference noise signal, the NRR was calculated as follows:

$$\mathrm{NRR} = \frac{\frac{1}{T-|\mathcal{S}|} \sum_{t \notin \mathcal{S}} \hat{s}_t^2}{\frac{1}{T-|\mathcal{S}|} \sum_{t \notin \mathcal{S}} n_t^2}. \tag{3.55}$$

The NRR represents how suppressed the noise is in an output signal. Since

Table 3.1: Configurations and Results of Bayesian Optimization

Parameters		Group Idx.	α^w	α^h	α^g	α^s	γ	K
Search range	min	–	0.01	0.01	0.01	0.01	0.01	1
	max	–	1.0	1.0	10.0	2.0	1.0	10
VB-RNMF		1	0.87	0.032	–	0.61	–	9
		2	0.56	0.070	–	0.55	–	7
		3	0.77	0.16	–	0.51	–	7
VB-RNTF		1	0.87	0.31	9.0	1.9	–	4
		2	0.36	0.13	5.2	1.9	–	6
		3	0.58	0.073	6.0	1.9	–	6
VB-SRNTF (T=200)		1	0.92	0.86	8.2	1.7	0.71	7
		2	0.98	0.55	8.1	1.7	0.81	5
		3	0.94	0.47	7.7	1.5	0.66	5

VB-RNTF and VB-SRNTF outputs are obtained by applying Wiener filtering to one of the microphones, the results at the tip (8th) microphone were evaluated.

The parameters for VB-RNMF, VB-RNTF, and VB-SRNTF were as follows. The shifting interval and window lengths of the STFT were set to 160 and 1024 samples, respectively. The hyperparameters α^w, α^h, α^g, α^s, γ and the number of bases K, which control the low-rankness and sparseness, were decided by using a Bayesian optimization method [139]. This method regards a target method as a black-box function that takes hyperparameters as input and outputs the value of average SDR. Assuming the function to follow a Gaussian process, the method searches for optimal hyperparameters that maximize the output of the function. The optimization was conducted by using noisy signals with SNRs of -10 dB and -5 dB and with layout conditions of Open-Front and Door-4ch. The noisy signals are separated into three groups at each condition, and the 3-fold cross validation was conducted. The search range and optimization results at each group were summarized in Table 3.1. The rate hyperparameters β^w and β^h were set to α^w and $\alpha^h K$, respectively. VB-RNMF and VB-RNTF were iterated 200 times and VB-SRNTF was iterated 100 times. The latent variables were initialized randomly.

3.5.2 Evaluation of Batch VB-RNTF and VB-RNMF

VB-RNTF and VB-RNMF were compared with existing phase-based blind source separation methods [37, 55] and low-rank and sparse decomposition meth-

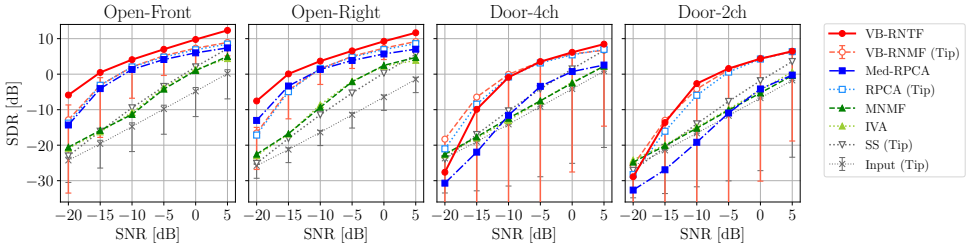


Figure 3.7: Speech enhancement performances in SDR. Each line indicates average SDR at the specified condition. Error bars for VB-RNMF and the input signal span the maximum and minimum SDRs in all the microphones.

ods [87, 93, 128, 140]. The phase-based blind source separation methods were MNMF [55] and independent vector analysis (IVA) [37]. The number of sources was set to eight for MNMF and IVA because seven vibrators generated noise and one target speech existed. This value corresponds to the maximum value tractable in these methods because they cannot perform under-determined source separation. Since these methods cannot distinguish the target speech source and other noise sources, the performance was determined by taking a maximum SDR value from all eight separation results. The low-rank and sparse decomposition methods were conventional RPCA [87, 88], RNMF [128] and VB-RPCA [93]. The results of them were obtained by using the tip (8th) microphone signals. This experiment also evaluated extended RPCA results that were obtained by taking median values of all the microphone results (Med-RPCA) [140]. As a baseline, an adaptive spectral subtraction (SS) method [141] was evaluated by applying to the tip microphone signals.

As shown in Figure 3.7, in the Open-Front and -Right conditions, VB-RNTF performed the best of all the evaluated methods in SDR. The low-rank and sparse decomposition methods (VB-RNTF, VB-RNMF, RPCA, and Med-RPCA) significantly outperformed conventional phase-based methods (MNMF and IVA). In the Door-4ch and Door-2ch conditions where some of the microphones were shaded, Med-RPCA significantly degraded from the Open-Front and Open-Right conditions. VB-RNTF, on the other hand, outperformed other multichannel methods. Although VB-RNTF was also degraded in both of the Door conditions, its performance was comparable to those of single-channel VB-RNMF

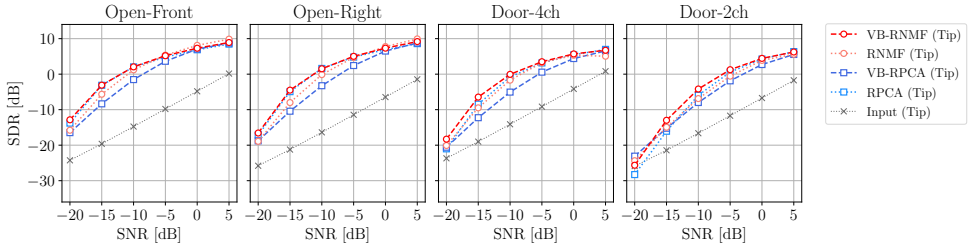


Figure 3.8: Speech enhancement performances of VB-RNMF and existing low-rank and sparse decomposition methods. The SDR of the input signal (gray line) is that of the recordings of the tip microphone.

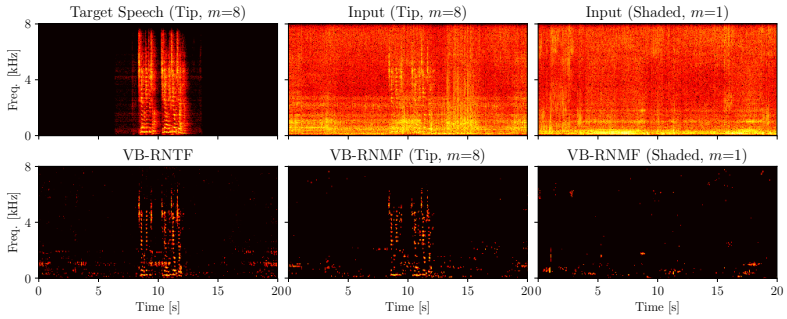


Figure 3.9: Excerpts of enhancement results obtained by VB-RNTF and VB-RNMF when the layout was the Door-4ch condition and the SNR was -5 dB.

and RPCA in these condition except when the SNR was less than -10 dB.

The performances of single-channel VB-RNMF and the existing low-rank and sparse decomposition methods (RPCA, VB-RPCA, and RNMF) are compared in Figure 3.8; where we see that VB-RNMF was comparable to the existing methods. This shows that VB-RNMF provides extensibility of RNMF in a Bayesian manner without performance degradation.

Figure 3.9 illustrates excerpts of enhancement results by VB-RNTF and VB-RNMF in the Door-4ch condition. While the VB-RNMF result applied to the tip (8th) microphone successfully enhanced the target speech, the result on the shaded (1st) microphone failed due to the low-SNR input. On the other hand, VB-RNTF using all the microphones robustly enhanced speech in this condition. Figure 3.10 shows estimated speech magnitudes at microphones $g_{mt} \sum_f s_{ft}$ estimated by VB-RNTF. In the Door-4ch and Door-2ch conditions, the speech magnitudes at the microphones that were separated from the sound source

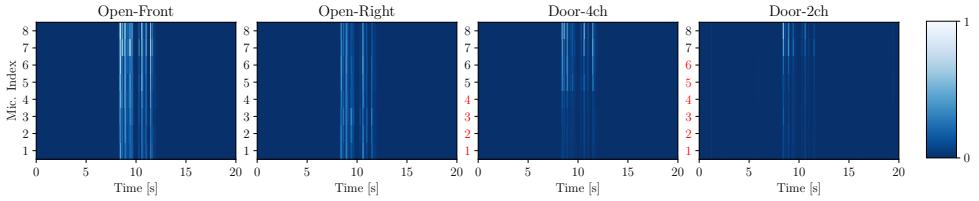


Figure 3.10: Examples of estimated speech magnitudes at microphones $g_{mt} \sum_f s_{ft}$ obtained by VB-RNTF in the condition where SNR was 0 dB. Male speech was emitted between 8 s and 12 s. Microphones shaded in Door-4ch and Door-2ch conditions are highlighted in red.

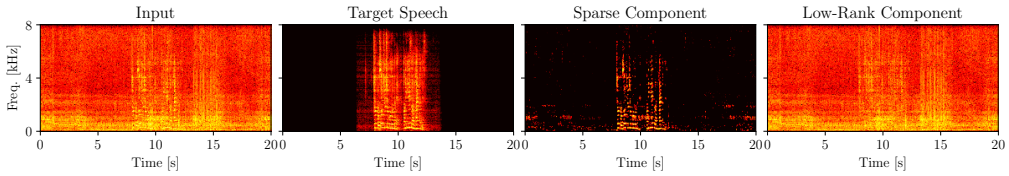


Figure 3.11: Estimated sparse and low-rank components (m=8) obtained by VB-RNTF when the layout was the Door-4ch condition and the SNR was 0 dB.

(highlighted in red) got significantly smaller. This shows that the estimated magnitudes can be used as a reliability of each microphone.

VB-RNTF outperformed the existing methods under high reverberation ($\mathrm{RT}_{60} \geq 750$ ms). The late reverberation can be considered as low-rank because it consists of a large number of reflected sounds. As shown in Figure 3.11, VB-RNTF dealt with reverberations by estimating the most prominent reverberant speech as a speech signal and separating other residuals into the low-rank components.

Figure 3.12 shows the enhancement performances in the SOR and NRR. The SORs of VB-RNTF were almost equivalent to those of the reference speech signals when SNR was more than -15 dB in the Open-Front and -Right conditions, more than -10 dB in the Door-4ch condition, and more than -5 dB in the Door-2ch condition. In the case of the SNR smaller than the above values, the performance of VB-RNTF deteriorated as the SNR decreased. On the other hand, the NRRs of VB-RNTF suppressed noise signals by more than 25 dB in all the conditions. The NRRs of VB-RNTF were almost constant at each layout condition when the SNR was 0 dB or less. These results show that VB-RNTF successfully suppresses noise signals regardless of the SNR and fails to extract speech signals in the

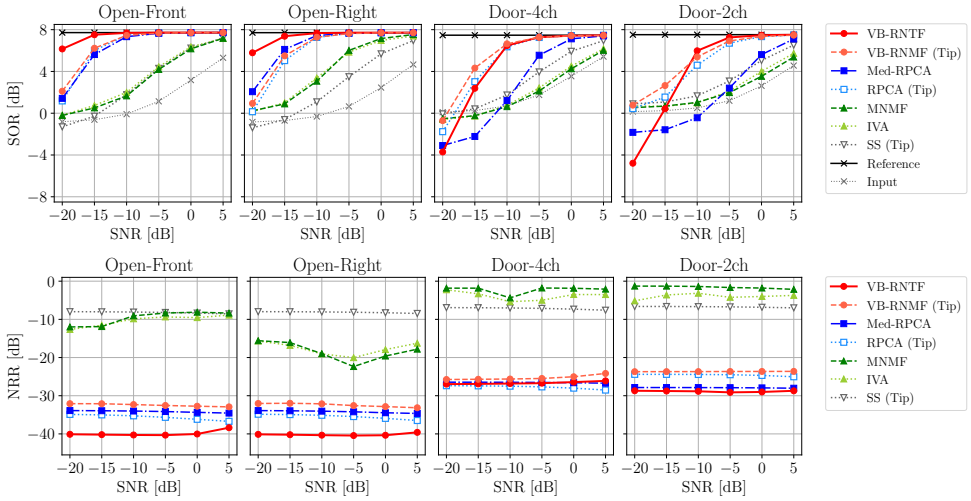


Figure 3.12: Speech enhancement performances of VB-RNTF, VB-RNMF, and existing methods in SOR and NRR.

low-SNR conditions. Especially in the Door-4ch and -2ch conditions, the SORs of VB-RNTF were worth than those of the single-channel VB-RNMF when the SNR was less than -10 dB. One way to improve VB-RNTF in these conditions is selection of valid microphones. The proposed method would be improved by selecting microphones when a few microphones are available. The idea of beta-process NMF [142] will be effective for this extension.

3.5.3 Evaluation of Mini-Batch VB-SRNTF

The performance of VB-SRNTF was evaluated with various mini-batch sizes. The following mini-batch sizes T were tested: 300, 200, 100, 50, and 10 frames. VB-SRNTF was compared with batch VB-RNTF and the following two existing mini-batch inferences: Ind-VB-RNTF and SVI-RNTF. Ind-VB-RNTF simply and independently conducts VB-RNTF at each mini-batch observation. SVI-RNTF is based on the conventional mini-batch VB inference [143] of VB-RNTF. It corresponds to VB-SRNTF whose γ is set to 1.0. VB-SRNTF was also compared with a variant of VB-SRNTF (VB-SRNTF-Raw) that takes a raw magnitude spectrogram without smoothing as input.

As shown in Figure 3.13, the enhancement performance in SDR became

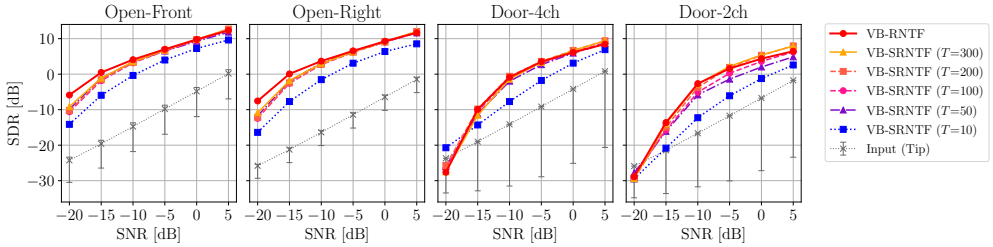


Figure 3.13: SDR performances of VB-SRNTFs with different mini-batch sizes.

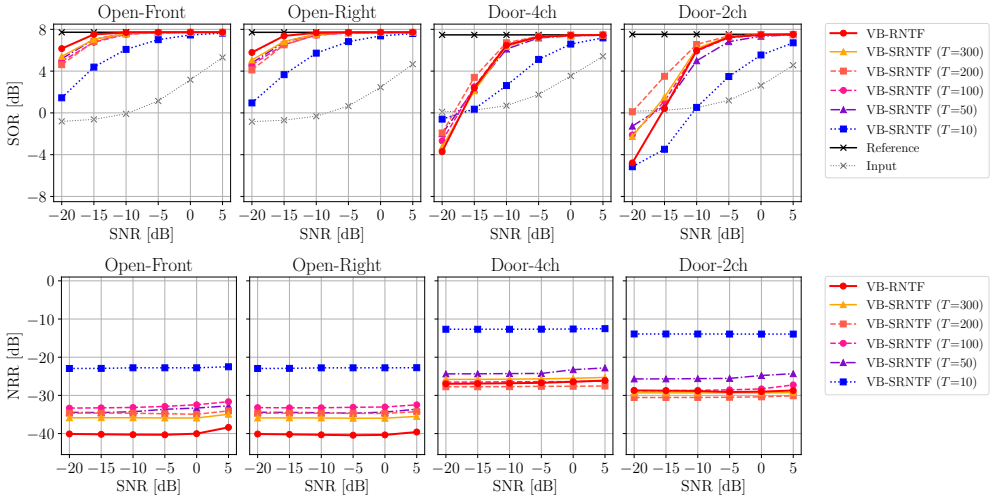


Figure 3.14: Speech enhancement performances of VB-SRNTFs in SOR and NRR.

higher as the mini-batch size was increased. When the mini-batch size was 200 frames or more, the SDR performances tended to be saturated. On the other hand, when the mini-batch size was 50 frames or less, the SDR performances were significantly degraded. Since a large mini-batch size leads to a large latency, there is a trade-off between performance and latency. These results show that a 2.0-second mini-batch (T = 200) was needed for adequate performance.

Figure 3.14 shows the enhancement performances of VB-SRNTFs in SOR and NRR. The SORs of the VB-SRNTFs ($T \geq 100$) were comparable to those of VB-RNTF and the reference speech signals in the conditions where the SORs of VB-RNTF and the reference signals were almost equivalent. The NRRs of the VB-SRNTs ($T \geq 100$) were also less than -25 dB.

Figure 3.15 compares the proposed VB-SRNTF results and other mini-batch inference results. Compared with Ind-VB-RNTF, which did not consider the

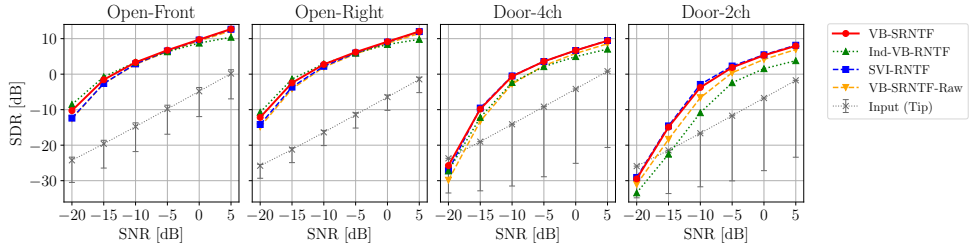


Figure 3.15: Comparison of VB-SRNTF (T=200) and existing mini-batch inferences of Bayesian RNTF.

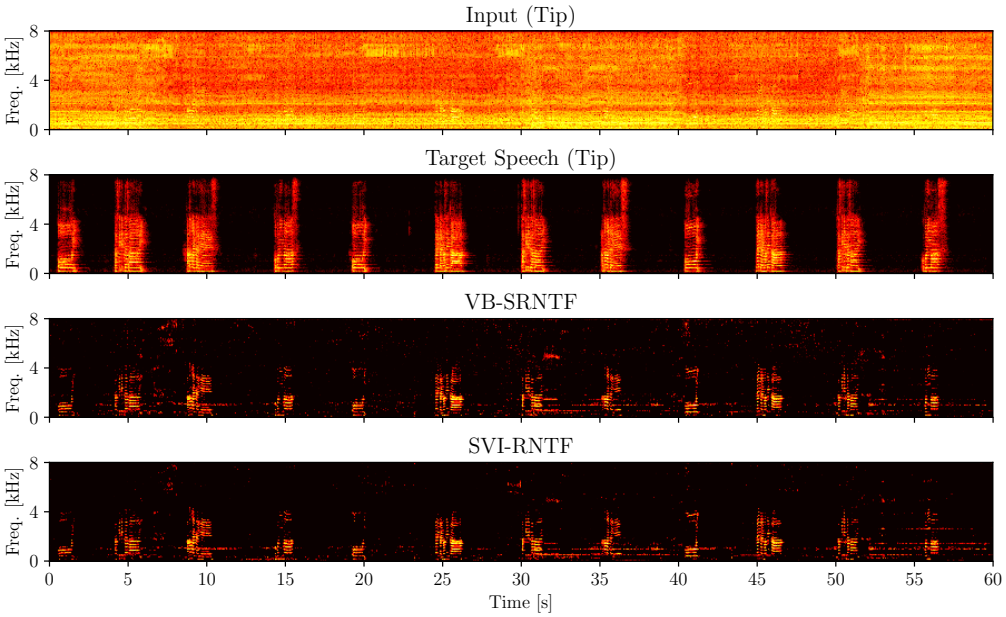


Figure 3.16: Speech enhancement results for 1-minute noisy signal. Female speech and ego-noise were mixed at -5 dB by using the impulse response of the Open-Front condition.

relationship between adjacent mini-batches, VB-SRNTF improved SDRs in the Door-4ch and Door-2ch conditions. Compared with SVI-RNTF, which did not consider the process noise of the basis vectors, the proposed VB-SRNTF slightly improved SDRs when the SNR was less than -5 dB in the Open-Front and -Right conditions. Compared with VB-SRNTF-Raw, which did not smooth the input spectrogram, the proposed VB-SRNTF improved SDRs in all the conditions. Since the VB-SRNTF can deal with the long-term change of ego-noise, the difference between VB-SRNTF and SVI-RNTF can usually be observedwith longer signals. Figure 3.16 shows speech enhancement results of VB-SRNTF

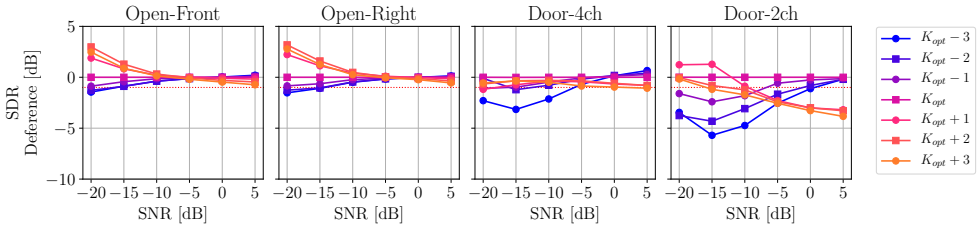


Figure 3.17: SDR differences between VB-SRNTF with different values of K and that with the values in Table 3.1 (K_{opt}).

and SVI-RNTF for a 1-minute noisy signal. The result of SVI-RNTF had more noise residuals than that of VB-SRNTF after 50 s. In actual environments, the ego-noise changes over time depending on the surrounding environments. The robustness against the long-term change of ego-noise is essential for a hose-shaped rescue robot.

Figure 3.17 shows SDR differences between VB-SRNTF with different values of K. Let K_{opt} be the optimized value of K shown in Table 3.1. The SDR degradation in the Open-Front and Open-Right conditions was less than 1.0 dB when K was $K_{opt} - 1$ or more and $K_{opt} + 3$ or less. On the other hand, K has to be set to K_{opt} or $K_{opt} + 2$ in the Door-4ch condition and only K_{opt} in Door-2ch condition. The performance of the method was sensitive to the number of bases K when some of the microphones were shaded. The author confirmed that the SDR degradation was less than 1 dB even when the hyperparameters α^w was changed by 15% from the values of Table 3.1, and α^h and α^g were changed by 20% from the optimal values. The α^s and γ had robustness against 6% and 5% changes, respectively.

3.5.4 Investigation of Gain Parameter Modeling

The gain parameter $g_{mt}^{(n)}$ of VB-SRNTF (Eq. (3.14)) ignores its frequency dependency and temporal continuity. Since the proposed formulation has only one gain parameter shared by all the frequency bins, the frequency characteristics are not considered. It also does not take into account the temporal continuity of the gains as shown in Figure 3.10.

The gain parameter modeling was investigated by evaluating variants of VB-

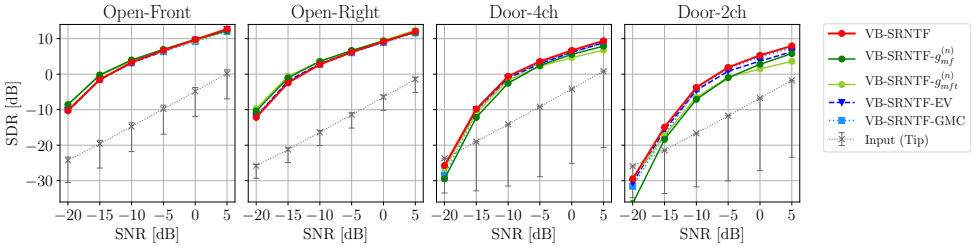


Figure 3.18: Comparison of VB-SRNTF (T=200) and the variants of VB-SRNTF with the frequency-dependent and the temporal-continuous gains.

SRNTF with frequency-dependent gains and temporal-continuous gains. The following two variants with the frequency-dependent gains were evaluated:

1. VB-SRNTF-$g^{(n)}_{mft}$: The gain $g^{(n)}_{mft}$ is both frequency and time dependent.
2. VB-SRNTF-$g^{(n)}_{mf}$: The gain $g^{(n)}_{mf}$ is frequency dependent but time independent.

The following two variants with the temporal-continuous gains were evaluated:

3. VB-SRNTF-EV: The prior distribution of the current state is given with the expected value of the previous posterior distribution $\langle \boldsymbol{g}^{(n-1)}_m \rangle$ as follows:

$$p\left(g^{(n)}_{mt}\middle|\,\alpha^g, \left\langle \boldsymbol{g}^{(n-1)}_m \right\rangle\right) = \mathcal{G}\left(g^{(n)}_{mt}\middle|\alpha^g, \frac{\alpha^g T}{\sum_t \left\langle g^{(n-1)}_{mt}\right\rangle}\right), \tag{3.56}$$

   where $\alpha^g \in \mathbb{R}_+$ is a hyperparameter that controls the strength of the dependencies.

4. VB-SRNTF-GMC: Markov dependencies between adjacent time frames are introduced with a gamma Markov chain prior [144] as follows:

$$p\left(g^{(n)}_{mt}\middle|\,\eta, z^{(n)}_{mt}\right) = \mathcal{G}\left(g^{(n)}_{mt}\middle|\eta, \eta z^{(n)}_{mt}\right), \tag{3.57}$$

$$p\left(z^{(n)}_{mt}\middle|\,\eta, g^{(n)}_{m(t-1)}\right) = \mathcal{G}\left(z^{(n)}_{mt}\middle|\eta, \eta g^{(n)}_{m(t-1)}\right), \tag{3.58}$$

$$p\left(z^{(n)}_{m1}\middle|\,\eta, g^{(n-1)}_{mT}\right) = \mathcal{G}\left(z^{(n)}_{m1}\middle|\eta, \eta g^{(n-1)}_{mT}\right), \tag{3.59}$$

   where $z^{(n)}_{mt}$ is an auxiliary latent variable that makes Markov dependencies between $g^{(n)}_{mt}$ and $g^{(n)}_{m(t-1)}$ in a conjugate manner and $\eta \in \mathbb{R}_+$ is a hyperparameter that controls the strength of the dependencies.

The hyperparameters of these models were determined by using the Bayesian optimization method in the same way as in Section V-A.

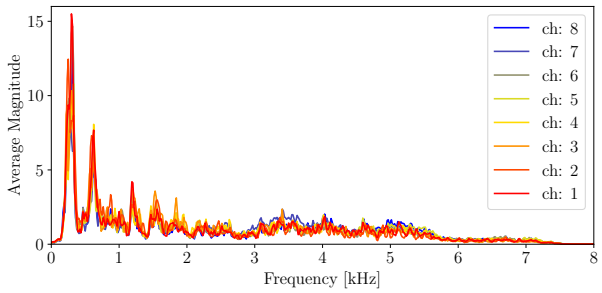


Figure 3.19: Average magnitude spectrum of a speech spectrogram at each microphone. A female speech spectrum with the condition of Open-Front is shown. Note that scale differences across microphones are normalized.

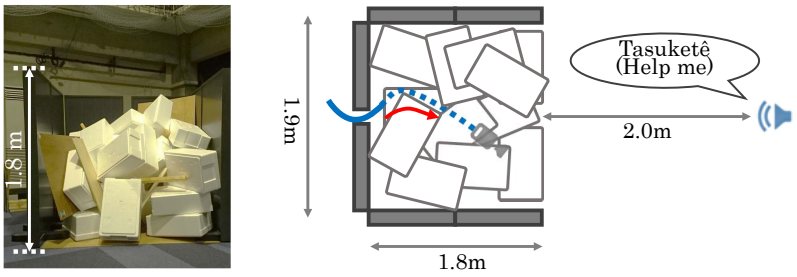
Figure 3.18 compares the SDR performances of the original VB-SRNTF and its variants with frequency-dependant and temporal-continuous gains. Compared with the performances of VB-SRNTF-$g_{mt}^{(n)}$ and -$g_{mft}^{(n)}$, the results show that the performance did not significantly deteriorate even if the frequency differences were ignored. As shown in Figure 3.19, this is because the spectral pattern differences of speech across microphones were small enough to ignore them. Figure 3.18 also shows that the performance of VB-SRNTF was comparable to those of VB-SRNTF-EV and -GMC. Although the temporal continuity of gains is one of the essential clues for blind source separation, only the sparse assumption on speech spectrograms was adequate in this evaluation. The sparseness assumption may cause the method to extract other sparse noise signals (e.g., impact sounds). The gain modeling with the temporal continuity will cope with such situations.

3.6 Experiments with Recorded Data

This section reports experimental results obtained using data recorded in an environment with simulated rubble.

3.6.1 Experimental Conditions

VB-RNTF and VB-SRNTF were evaluated in the condition that the robot moved under simulated rubble. To simulate rubble disturbing sound propagation,



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

Figure 3.20: Condition of rubble and target speech in experiments reported in section 3.6.

styrene foam boxes and wooden plates were piled up (Figure 3.20-(a)). A loudspeaker for playing back target speech signals was put 2 m away from this rubble (Figure 3.20-(b)). The robot was inserted from behind the rubble and captured eight-channel audio signals (mixtures of ego-noise and target speech) for 10 seconds during the insertion. Although the boxes and plates were placed only around the robot, they were enough for disturbing sound propagation around the robot. The target signals were four male and female speech recordings screaming for rescue in Japanese (e.g., "Tasukete kudasai (Help me)" and "Kokoniimasu (I'm here)") and the loudspeaker was calibrated so that its sound pressure level for each utterance was 80 dB. In this experiment, the relative layout of the microphones and target speech source changed over time due to the insertion and vibration. The parameters of the proposed VB-RNMF, VB-RNTF, and VB-SRNTF were set to the values of the optimization results for the first group listed in Table 3.1.

Since it was impossible to obtain clean speech signals captured by the robot microphones, the following SNR was used as an evaluation criterion in this experiment:

$$\mathrm{SNR}(\hat{\mathbf{S}}, \mathbf{S}, a) = 10 \log_{10} \frac{\sum_{f,t} a^2 s_{ft}^2}{\sum_{f,t} (\hat{s}_{ft} - a s_{ft})^2}, \tag{3.60}$$

where $\mathbf{S} \in \mathbb{R}_+^{F \times T}$ and $\hat{\mathbf{S}} \in \mathbb{R}_+^{F \times T}$ represent the magnitude spectrograms of reference and estimated target speech signals, respectively, and a represents a gain parameter compensating for the level difference between $\mathbf{S}$ and $\hat{\mathbf{S}}$. This gain pa-

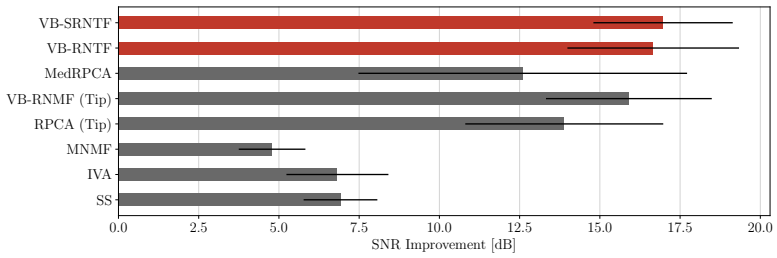


Figure 3.21: Speech enhancement performances in terms of SNR improvement from the input signal (at the tip microphone). Error bars indicate the standard deviation of the results. The average SNR of the input signals was -19.7 dB.

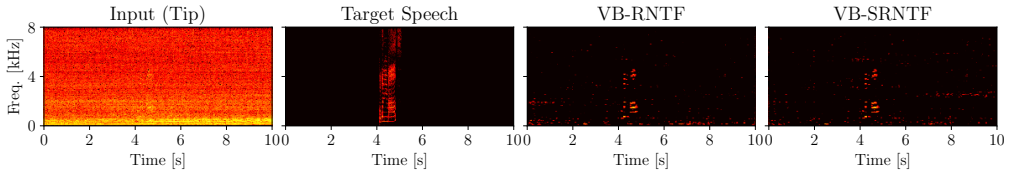


Figure 3.22: Examples of enhancement results obtained in experiments reported in Section 3.6.

rameter was determined with minimum mean-square error estimation (MMSE) between $a\mathbf{S}$ and $\hat{\mathbf{S}}$.

3.6.2 Experimental Results

Figure 3.21 shows that VB-RNTF and VB-SRNTF outperformed all of the other methods. VB-SRNTF improved the SNR by 1.07 dB more than VB-RNMF, which had the second-best performance. Figure 3.22 shows the magnitude spectrogram of an observed signal (at the tip microphone) and the enhanced speech signals obtained by VB-RNTF and VB-SRNTF. These results showed that VB-RNTF and VB-SRNTF suppressed the time-varying ego-noise.

As shown in Figure 3.23, to realize a real-time mobile enhancement system, VB-SRNTF was implemented on an embedded GPGPU board (NVIDIA Jetson TX1) and a standard laptop computer (Dell XPS13). The proposed VB-SRNTF was implemented on the TX1 with GPGPU programming using C++ and CUDA 8.0. The elapsed time for VB-SRNTF with a 20.0 s input signal was 15.2 s when the batch size T was 200 frames. Since this value was small enough compared with the whole signal length, the method could work in real time.

Figure 3.23: Hose-shaped rescue robot system including robot body (back), robot controller (left), laptop PC (middle), and embedded GPGPU board (right).

3.7 Summary

This chapter presented a multichannel blind speech enhancement method based on low-rank and sparse decomposition. The proposed method is formulated as a Bayesian model called Bayesian RNTF. It separates a multichannel magnitude spectrogram into sparse and low-rank spectrograms (target speech and noise) without any prior training. Since Bayesian RNTF works without phase information, it can deal with the time-varying layout of microphones and sound sources. For real-time speech enhancement, Bayesian RNTF is extended to a state-space model called Bayesian SRNTF that represents the dynamics of the latent variables in a mini-batch manner. The Bayesian inferences of these models were derived with a VB framework, so the decomposition methods are abbreviated as VB-RNTF and VB-SRNTF. Experiments using a 3-m hose-shaped rescue robot with eight microphones showed that VB-SRNTF improves the SNR of a speech signal 1.07 dB more than conventional blind methods do. Using an embedded GPGPU board, the proposed VB-SRNTF was fast enough to work in real time.

The proposed methods based on the low-rank noise and sparse speech assumptions have the following limitations. Experimental results showed that the possible maximum rank of the low-rank component K should be given appropriately in advance. The sparseness assumption, on the other hand, may cause the method to enhance not only speech signals but also other sparse noise

signals. For example, if an input signal includes impact noise sounds caused by rubble-removal operations, the method extracts the noise as a speech signal. To relax the low-rank limitation, future work includes the estimation of K based on the non-parametric Bayesian framework [134]. Speech-specific structures can be introduced as prior information of the speech signals because spectrograms of speech signals have dependencies between frequency bins (e.g., harmonic structures) and time frames (e.g., temporal continuity). This extension is addressed in Chapter 4 by using a pre-trained deep generative model.

Chapter 4

Speech Enhancement with a Deep Speech Prior

This chapter presents a single-channel speech enhancement method that combines a DNN-based speech model and a conventional unsupervised noise model. To improve the enhancement performance, this chapter uses a pre-trained deep generative model as a prior distribution of a speech spectrogram instead of the sparse assumption presented in Chapter 3.

4.1 Introduction

Deep neural networks (DNNs) have demonstrated excellent performance in single-channel speech enhancement [51, 52, 104–107]. The denoising autoencoder (DAE), for example, is a typical variant of such networks, which is trained to directly convert a noisy speech spectrogram to a clean speech spectrogram with a supervised training [51]. Alternatively, a DNN can be trained to predict time-frequency (TF) masks called ideal ratio masks (IRMs) that represent ratios of speech to input signals and are used for obtaining a speech spectrogram from a noisy spectrogram [105]. Although it is necessary to prepare as training data a large amount of pairs of clean speech signals and their noisy versions, these supervised methods often deteriorate in unknown noisy environments. This calls for semi-supervised methods that are trained by using only clean speech data in advance and then adapt to unseen noisy environments.

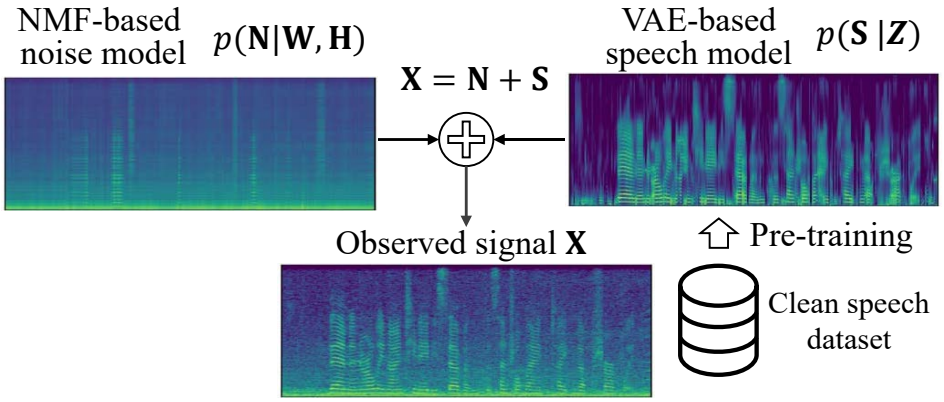


Figure 4.1: Overview of the proposed speech enhancement model.

Statistical source separation methods based on the additivity of speech and noise spectrograms have also been used for speech enhancement [86, 126, 145]. Non-negative matrix factorization (NMF) [84, 85], for example, regards a noisy speech spectrogram as a non-negative matrix and approximates it as the product of two non-negative matrices (a set of basis spectra and a set of the corresponding activations). If a partial set of basis spectra is trained in advance from clean speech spectrograms, the noisy spectrogram is decomposed into the sum of speech and noise spectrograms in a semi-supervised manner. As discussed in Chapter 3, robust principal component analysis (RPCA) [87, 146] is another promising method that can decompose a noisy spectrogram into a sparse speech spectrogram and a low-rank noise spectrogram in an unsupervised manner. These conventional statistical methods, however, have a common problem that the linear representation or the sparseness assumption of speech spectrograms is not satisfied in reality and results in considerable signal distortion.

Recently, deep generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs) have gained a lot of attention for learning a probability distribution over complex data (e.g., images and audio signals) that cannot be represented by conventional linear models [63–67]. GANs and VAEs are both based on two kinds of DNNs having different roles. In GANs [63], a *generator* is trained to synthesize data that fool a *discriminator* from a latent space while the discriminator is trained to detect synthesized data in

a minimax-game fashion. In VAEs [64, 65], on the other hand, an *encoder* that embeds observed data into a latent space and a *decoder* that generates data from the latent space are trained jointly such that the lower bound of the log marginal likelihood for the observed data is maximized. Although in general GANs can generate more realistic data, VAEs provide a principled scheme of inferring the latent representations of both given and new data.

This chapter presents a unified probabilistic generative model of noisy speech spectra by combining a VAE-based generative model of speech spectra with an NMF-based generative model of noise spectra (Figure 4.1). The VAE is trained in advance from a sufficient amount of clean speech spectra and its decoder is used as a prior distribution on clean speech spectra included in noisy speech spectra. Given observed data, the proposed method can estimate both the latent representations of speech spectra as well as the basis spectra and their activations of noise spectra through Bayesian inference based on a Markov chain Monte Carlo (MCMC) algorithm initialized by the encoder of the VAE. The proposed Bayesian approach can adapt to both unseen speech and noise spectra by using prior knowledge of clean speech and the low-rankness assumption on noise instead of fixing all the parameters in advance.

4.2 Variational Autoencoder

A VAE [64] is a framework for learning the probability distribution of a dataset. This subsection denotes by $\mathbf{X}$ a dataset that contains F-dimensional samples $\boldsymbol{x}_t \in \mathbb{R}^F$ ($t = 1, \ldots, T$). The VAE assumes that a D-dimensional latent variable (denoted by $\boldsymbol{z}_t \in \mathbb{R}^D$) follows a standard Gaussian distribution and each sample $\boldsymbol{x}_t$ is stochastically generated from a conditional distribution $p(\boldsymbol{x}_t | \boldsymbol{z}_t)$:

$$\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D), \tag{4.1}$$

$$\boldsymbol{x}_t \sim p(\boldsymbol{x}_t | \boldsymbol{z}_t), \tag{4.2}$$

where $\mathcal{N}(\mu, \sigma)$ represents a Gaussian distribution with mean parameter μ and variance parameter σ. $p(\boldsymbol{x}_t | \boldsymbol{z}_t)$ is called a decoder and parameterized as a well-known probability density function whose parameters are given by nonlinear

functions represented as neural networks. For example, Kingma et al. [64] reported a VAE model that has the following Gaussian likelihood function:

$$\boldsymbol{x}_t \sim p\left(\boldsymbol{x}_t \left| \boldsymbol{z}_t \right.\right) = \prod_f p\left(x_{ft} \left| \boldsymbol{z}_t \right.\right) = \prod_f \mathcal{N}\left(\mu_f^x(\boldsymbol{z}_t), \sigma_f^x(\boldsymbol{z}_t)\right), \tag{4.3}$$

where $\mu_f^x : \mathbb{R}^D \to \mathbb{R}$ and $\sigma_f^x : \mathbb{R}^D \to \mathbb{R}_+$ are neural networks representing the mean and variance parameters, respectively.

The objective of VAE training is to find a likelihood function $p\left(\boldsymbol{x}_t \left| \boldsymbol{z}_t \right.\right)$ that maximizes the log marginal likelihood:

$$\underset{p(\boldsymbol{x}_t|\boldsymbol{z}_t)}{\operatorname{argmax}} \log p\left(\mathbf{X}\right) = \underset{p(\boldsymbol{x}_t|\boldsymbol{z}_t)}{\operatorname{argmax}} \prod_t \int p\left(\boldsymbol{x}_t \left| \boldsymbol{z}_t \right.\right) p\left(\boldsymbol{z}_t\right) d\boldsymbol{z}_t. \tag{4.4}$$

Since calculating this marginal likelihood is intractable, it is approximated with a variational Bayesian (VB) framework. The VAE first approximates the posterior distribution of $\boldsymbol{z}_t$ with the following variational posterior distribution $q\left(\boldsymbol{z}_t\right)$ called an encoder:

$$p\left(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T \left| \mathbf{X} \right.\right) \approx \prod_t q\left(\boldsymbol{z}_t\right) = \prod_{d,t} q\left(z_{dt}\right) \tag{4.5}$$

$$= \prod_{d,t} \mathcal{N}\left(\mu_d^z(\boldsymbol{x}_t), \sigma_d^z(\boldsymbol{x}_t)\right), \tag{4.6}$$

where $\mu_d^z : \mathbb{R}^F \to \mathbb{R}$ and $\sigma_d^z : \mathbb{R}^F \to \mathbb{R}_+$ are nonlinear functions representing the mean and variance parameters, respectively. These functions are formulated with DNNs. By using the variational posterior, the log marginal likelihood is lower-bounded as follows:

$$\log p\left(\mathbf{X}\right) = \sum_t \log \int p\left(\boldsymbol{x}_t \left| \boldsymbol{z}_t \right.\right) p\left(\boldsymbol{z}_t\right) d\boldsymbol{z}_t \tag{4.7}$$

$$\geq \sum_t \int q\left(\boldsymbol{z}_t\right) \log \frac{p\left(\boldsymbol{x}_t \left| \boldsymbol{z}_t \right.\right) p\left(\boldsymbol{z}_t\right)}{q\left(\boldsymbol{z}_t\right)} d\boldsymbol{z}_t \tag{4.8}$$

$$= -\sum_t \mathbb{KL}\left[q\left(\boldsymbol{z}_t\right) \left| p\left(\boldsymbol{z}_t\right) \right.\right] + \sum_k \mathbb{E}_q\left[\log p\left(\boldsymbol{x}_t \left| \boldsymbol{z}_t \right.\right)\right], \tag{4.9}$$

where $\mathbb{KL}\left[\cdot \left| \cdot \right.\right]$ represents the Kullback-Leibler divergence. The VAE is trained so that $p\left(\boldsymbol{x}_t \left| \boldsymbol{z}_t \right.\right)$ and $q\left(\boldsymbol{z}_t\right)$ maximize this variational lower bound. The first term of Eq. (4.9) is analytically tractable and the second term can be approximated with a Monte-Carlo algorithm. The lower bound can be maximized by using a stochastic gradient descent (SGD) [147].

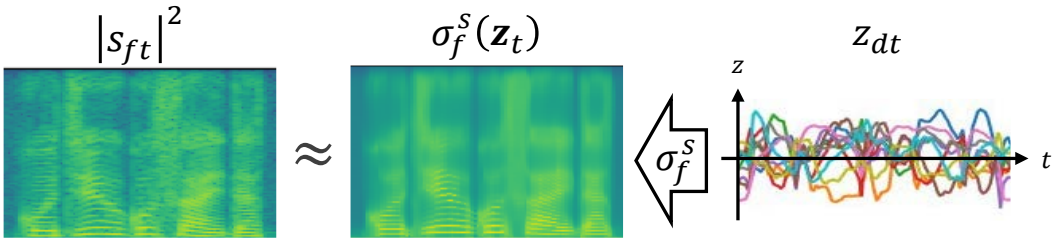


Figure 4.2: VAE representation of a speech spectrogram.

4.3 Probabilistic Combination of VAE and NMF

This section describes the proposed probabilistic generative model called VAE-NMF, that combines a VAE-based speech model and a NMF-based noise model. This section formulates the generative process of an observed complex spectrogram $\mathbf{X} \in \mathbb{C}^{F \times T}$ by formulating the process of a speech spectrogram $\mathbf{S} \in \mathbb{C}^{F \times T}$ and a noise spectrogram $\mathbf{N} \in \mathbb{C}^{F \times T}$. The characteristics of speech and noise signals are represented by their priors based on VAE and NMF, respectively.

4.3.1 VAE-Based Speech Model

The speech model assumes a frame-wise D-dimensional latent variable $\mathbf{Z} \in \mathbb{R}^{D \times T}$. Each time-frame of the latent variable $\boldsymbol{z}_t$ is supposed to represent the characteristics of a speech spectrum such as fundamental frequency, spectral envelope, and type of phoneme. The specific representation of $\boldsymbol{z}_t$ is obtained automatically by conducting the VAE training with a dataset of clean speech spectra. As in the conventional VAEs, the standard Gaussian prior is put on each element of $\mathbf{Z}$:

$$z_{dt} \sim \mathcal{N}(0, 1)\,. \tag{4.10}$$

Since the speech spectra are primarily characterized by its power spectral density (PSD), it follows a zero-mean complex Gaussian distribution whose variance parameter is formulated with $\mathbf{Z}$ (Figure 4.2):

$$s_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sigma_f^s(\boldsymbol{z}_t)\right), \tag{4.11}$$

where $\mathcal{N}_{\mathbb{C}}(\mu, \sigma)$ is a complex Gaussian distribution with mean parameter μ and variance parameter σ. $\sigma_f^s(\boldsymbol{z}_t) : \mathbb{R}^D \rightarrow \mathbb{R}_+$ is a nonlinear function representing the relationship between $\mathbf{Z}$ and the speech signal $\mathbf{S}$. This function is formulated by using a DNN and obtained by the VAE training.

4.3.2 Generative Model of Mixture Signals

In the proposed Bayesian generative model, the input complex spectrogram $\mathbf{X} \in \mathbb{C}^{F \times T}$ is represented as the sum of a speech spectrogram $\mathbf{S}$ and a noise spectrogram $\mathbf{N}$:

$$x_{ft} = s_{ft} + n_{ft}. \tag{4.12}$$

The VAE-based hierarchical prior model (Eqs. (4.10) and (4.11)) is put on the speech spectrogram $\mathbf{S}$. On the other hand, the PSD of the noise spectrogram is assumed to be low-rank by putting NMF-based prior model on it. More specifically, the PSD of a noise spectrogram can be represented as the product of K spectral basis vectors $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}_+^{F \times K}$ and their activation vectors $\mathbf{H} \in \mathbb{R}_+^{K \times T}$. The zero-mean complex Gaussian distribution is put on each TF bin of the noise spectrogram $\mathbf{N}$ as follows:

$$n_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k w_{fk} h_{kt}\right). \tag{4.13}$$

For mathematical convenience, conjugate prior distributions are put on $\mathbf{W}$ and $\mathbf{H}$ as follows:

$$w_{fk} \sim \mathcal{G}(a_0, b_0), \quad h_{kt} \sim \mathcal{G}(a_1, b_1), \tag{4.14}$$

where $\mathcal{G}(\alpha, \beta)$ is a gamma distribution with the shape parameter $\alpha > 0$ and the rate parameter $\beta > 0$; a_0, b_0, a_1, and b_1 are hyperparameters that should be set in advance.

By marginalizing out the speech and noise complex spectrograms $\mathbf{S}$ and $\mathbf{N}$, the following Gaussian likelihood is obtained:

$$x_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k w_{fk} h_{kt} + \sigma_f^s(\boldsymbol{z}_t)\right). \tag{4.15}$$

Since this likelihood function is independent of the phase term of the input spectrogram $\mathbf{X}$, it is equivalent to the following exponential likelihood:

$$\|x_{ft}\|^2 \sim \mathrm{Exp}\left(\sum_k w_{fk}h_{kt} + \sigma_f^s(\boldsymbol{z}_t)\right), \tag{4.16}$$

where $\|x_{ft}\|^2$ is the power spectrogram of $\mathbf{X}$ and $\mathrm{Exp}\,(\lambda)$ is the exponential distribution with a mean parameter λ. Maximization of the exponential likelihood on a power spectrogram corresponds to minimization of Itakura-Saito divergence, which is widely used in audio source separation [84, 134].

4.3.3 Pre-Training of VAE-based Speech Model

The goal of the pre-training of the VAE-based speech model is to find $p\left(\boldsymbol{s}_t\,|\boldsymbol{z}_t\right)$ that maximizes the following marginal likelihood $p\left(\mathbf{S}\right)$ from the dataset of clean speech signal (denoted by $\mathbf{S} \in \mathbb{C}^{F\times T}$ in this subsection):

$$p\left(\mathbf{S}\right) = \prod_t \int p\left(\boldsymbol{s}_t\,|\boldsymbol{z}_t\right) p\left(\boldsymbol{z}_t\right) p\boldsymbol{z}_t. \tag{4.17}$$

As stated in Sec. 4.2, it is difficult to analytically calculate this marginal likelihood. The marginal likelihood is approximated by using the Variational mean-field approximation. Let $q\left(\mathbf{Z}\right)$ be the variational posterior distribution of $\mathbf{Z}$. Since $p\left(\mathbf{S}\,|\mathbf{Z}\right)$ is independent from the phase term of the speech spectrogram $\mathbf{S}$, the variational posterior $q\left(\mathbf{Z}\right)$ is defined by ignoring the phase term as follows:

$$q\left(\mathbf{Z}\right) = \prod_{d,t} q\left(z_{dt}\right) = \prod_{d,t} \mathcal{N}\left(\mu_d^z\left(\|\boldsymbol{s}_t\|^2\right), \sigma_d^z\left(\|\boldsymbol{s}_t\|^2\right)\right), \tag{4.18}$$

where $\|\boldsymbol{s}_t\|^2$ is the power spectrum of $\boldsymbol{s}_t$ and $\mu_d^z : \mathbb{R}_+^F \to \mathbb{R}$ and $\sigma_d^z : \mathbb{R}_+^F \to \mathbb{R}_+$ are nonlinear functions representing the mean and variance parameters of the Gaussian distribution. These two functions are defined with DNNs. The marginal likelihood is approximately calculated as follows:

$$\begin{aligned}
\log p\left(\mathbf{S}\right) &\geq -\mathbb{KL}\left[q\left(\mathbf{Z}\right)|p\left(\mathbf{Z}\right)\right] + \mathbb{E}_q\left[\log p\left(\mathbf{S}\,|\mathbf{Z}\right)\right] &(4.19)\\
&= -\sum_{d,t}\frac{1}{2}\left\{\left(\mu_d^z(\|\boldsymbol{s}_t\|^2)\right)^2 + \sigma_d^z(\|\boldsymbol{s}_t\|^2) - \log\sigma_d^z(\|\boldsymbol{s}_t\|^2)\right\}\\
&\quad + \sum_{f,t}\mathbb{E}_q\left[-\log\sigma_f^s(\boldsymbol{z}_t) - \frac{\|s_{ft}\|^2}{\sigma_f^s(\boldsymbol{z}_t)}\right] + \mathrm{const.} &(4.20)
\end{aligned}$$

The DNNs for σ_f^s, μ_d^z, and σ_d^z are optimized by using SGD so that this variational lower bound is maximized.

4.3.4 Bayesian Inference of VAE-NMF

To enhance the speech signal in a noisy observed signal, VAE-NMF calculates the full posterior distribution: $p\left(\mathbf{W}, \mathbf{H}, \mathbf{Z} \,|\mathbf{X}\right)$. Since the true posterior is analytically intractable, it is approximated with a finite number of random samples by using a Markov chain Monte Carlo (MCMC) algorithm [62]. MCMC alternatively and iteratively samples one of the latent variables ($\mathbf{W}$, $\mathbf{H}$, and $\mathbf{Z}$) according to their conditional posterior distributions.

By fixing the speech parameter $\mathbf{Z}$, the conditional posterior distributions $p\left(\mathbf{W} \,|\mathbf{X}, \mathbf{H}, \mathbf{Z}\right)$ and $p\left(\mathbf{H} \,|\mathbf{X}, \mathbf{W}, \mathbf{Z}\right)$ can be derived with a variational approximation [62, 134] as follows:

$$w_{fk} \mid \mathbf{X}, \mathbf{H}, \mathbf{Z} \sim \mathcal{GIG}\left(a_0, b_0 + \sum_t \frac{h_{kt}}{\lambda_{ft}}, \sum_t \|x_{ft}\|^2 \frac{\phi_{ftk}^2}{h_{kt}}\right), \tag{4.21}$$

$$h_{kt} \mid \mathbf{X}, \mathbf{W}, \mathbf{Z} \sim \mathcal{GIG}\left(a_1, b_1 + \sum_f \frac{w_{fk}}{\lambda_{ft}}, \sum_f \|x_{ft}\|^2 \frac{\phi_{ftk}^2}{w_{fk}}\right), \tag{4.22}$$

$$\lambda_{ft} = \sum_k w_{fk} h_{kt} + \sigma_f^s(\boldsymbol{z}_t), \;\; \phi_{ftk} = \frac{w_{fk} h_{kt}}{\sum_k w_{fk} h_{kt} + \sigma_f^s(\boldsymbol{z}_t)}, \tag{4.23}$$

where $\mathcal{GIG}\left(\gamma, \rho, \tau\right) \propto x^{\gamma-1}\exp(-\rho x - \tau/x)$ is the generalized inverse Gaussian distribution and λ_{ft} and ϕ_{ftk} are auxiliary variables.

The latent variable of speech $\mathbf{Z}$ is updated by using a Metropolis method [62] because it is hard to analytically derive the conditional posterior $p\left(\mathbf{Z} \,|\mathbf{X}, \mathbf{W}, \mathbf{H}\right)$. The latent variable is sampled at each time frame by using the following Gaussian proposal distribution $q\left(\boldsymbol{z}_t^* \,|\boldsymbol{z}_t\right)$ whose mean is the previous sample $\boldsymbol{z}_t$:

$$\boldsymbol{z}_t^* \sim q\left(\boldsymbol{z}_t^* \,|\boldsymbol{z}_t\right) = \mathcal{N}\left(\boldsymbol{z}_t, \sigma \mathbf{I}\right), \tag{4.24}$$

where σ is a variance parameter of the proposal distribution. This candidate $\boldsymbol{z}_t^*$ is randomly accepted with the following probability:

$$a_{\boldsymbol{z}_t^* | \boldsymbol{z}_t} = \min\left(1, \frac{p\left(\boldsymbol{x}_t \,|\mathbf{W}, \mathbf{H}, \boldsymbol{z}_t^*\right) p\left(\boldsymbol{z}_t^*\right)}{p\left(\boldsymbol{x}_t \,|\mathbf{W}, \mathbf{H}, \boldsymbol{z}_t\right) p\left(\boldsymbol{z}_t\right)}\right). \tag{4.25}$$

4.3.5 Reconstruction of Complex Speech Spectrogram

In this chapter the enhanced speech is obtained with Wiener filtering by maximizing the conditional posterior $p(\mathbf{S}|\mathbf{X}, \mathbf{W}, \mathbf{H}, \mathbf{Z})$. Let $\hat{\mathbf{S}} \in \mathbb{C}^{F \times T}$ be the speech spectrogram that maximizes the conditional posterior. It is given by the following equation:

$$\hat{s}_{ft} = \frac{\sigma_f^s(\boldsymbol{z}_t)}{\sum_k w_{fk} h_{kt} + \sigma_f^s(\boldsymbol{z}_t)} x_{ft}. \tag{4.26}$$

The mean values of the sampled latent variables are simply used as $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{Z}$ in Eq. (4.26).

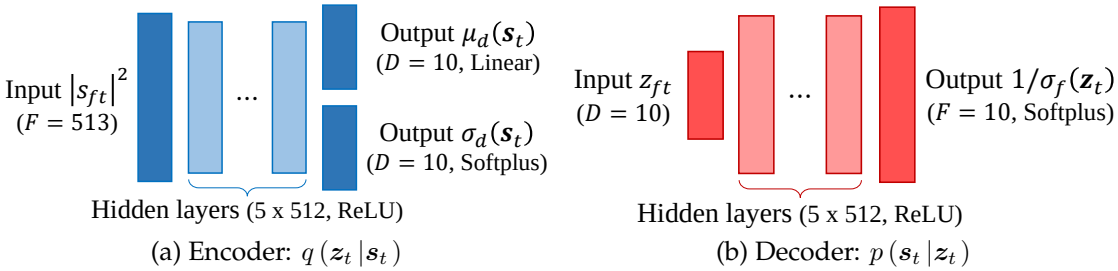
4.4 Evaluation with Datasets of Urban Noise

For evaluating the basic performance of VAE-NMF, this section reports experimental results with noisy speech signals whose noise signals were captured in actual urban environments.

4.4.1 Experimental Settings

To compare VAE-NMF with a DNN-based supervised method, this evaluation used CHiME-3 dataset [96] and DEMAND noise database[1]. The CHiME-3 dataset was used for both the training and evaluation. The DEMAND database was used for constructing another evaluation dataset for unseen noise conditions. The evaluation with the CHiME-3 was conducted by using its development set, which consists of 410 simulated noisy utterances in each of four different noisy environments: on a bus (BUS), in a cafe (CAF), in a pedestrian area (PED) and on a street junction (STR). The average signal-to-noise ratio (SNR) of the noisy speech signals was 5.8 dB. The evaluation with the DEMAND was conducted by using 20 simulated noisy speech signals in each of four different noisy environments: on a subway (SUB), in a cafe (CAF), at a town square (SQU), and in a living room (LIV). These signals were generated by mixing the clean speech signals of the CHiME-3 development set with the noise signals in the

[1] http://parole.loria.fr/DEMAND/



(a) Encoder: $q(\boldsymbol{z}_t|\boldsymbol{s}_t)$ (b) Decoder: $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$

Figure 4.3: Configuration of the VAE used in the Section 4.4.

DEMAND database. The SNR of these noisy speech signals was set to be 5.0 dB. The sampling rate of these signals was 16 kHz. The enhancement performance was evaluated by using the source-to-distortion ratio (SDR) [138].

The prior distribution of speech signals $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$ was obtained by training a VAE that had two networks of $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$ and $q(\boldsymbol{z}_t)$ as shown in Figure 4.3. The dimension of the latent variables D was set to be 10. The training data were about 15 hours of clean speech signals in the WSJ-0 corpus [148]. Their spectrograms were obtained with a short-time Fourier transform (STFT) with a window length of 1024 samples and a shifting interval of 256 samples. To make the prior distribution robust against a scale of the speech power, the average power of the spectrogram was changed between 0.0 and 10.0 at each parameter update.

The parameters for VAE-NMF were as follows. The number of bases K was set to be 5. The hyperparameters a_0, b_0, a_1, b_1, and σ were set to be 1.0, 1.0, and 1.0, $K/scale$, and 0.01, respectively. The $scale$ represents the empirical average power of the input noisy spectrogram. After drawing 100 samples for burn-in, 50 samples were drawn to estimate the latent variables. These parameters had been determined empirically. The latent variables of noise $\mathbf{W}$ and $\mathbf{H}$ were randomly initialized. Since the latent variable of speech $\mathbf{Z}$ depends on the initial state, the initial sample was drawn from $q(\boldsymbol{z}_t|\boldsymbol{s}_t)$ by setting the observation $\boldsymbol{x}_t$ as the speech signal $\boldsymbol{s}_t$.

VAE-NMF was compared with a DNN-based supervised method and the unsupervised RPCA. A DNN that outputs IRMs (DNN-IRM) was implemented.

Table 4.1: Enhancement performance in SDR for CHiME-3 dataset

Method	Average	BUS	CAF	PED	STR
VAE-NMF	10.10	**9.47**	10.62	10.93	9.39
DNN-IRM	**10.93**	8.92	**11.92**	**12.92**	**9.95**
RPCA	7.53	6.13	8.10	9.13	6.77
Input	6.02	3.26	7.21	8.83	4.78

Table 4.2: Enhancement performance in SDR for DEMAND dataset

Method	Average	SUB	CAF	SQU	LIV
VAE-NMF	**11.17**	**10.56**	**9.57**	**12.38**	**12.16**
DNN-IRM	9.85	9.13	9.15	10.69	10.42
RPCA	7.03	6.48	6.37	6.99	8.28
Input	5.21	5.25	5.24	5.19	5.16

It had five hidden layers with ReLU activation functions. It takes as an input 11 frames of noisy 100-channel log-Mel-scale filterbank features and predicts one frame of IRMs[2]. DNN-IRM was trained with the training dataset of CHiME-3, which was generated by using the WSJ-0 speech utterances and noise signals. The noise signals were recorded in the same environments as those in the evaluated data.

4.4.2 Experimental Results

The enhancement performance is shown in Tables 4.1 and 4.2. In the experiments using the CHiME-3 test set (Table 4.1), DNN-IRM, which was trained using the noisy data recorded in the same environments at the test data, yielded the highest average SDR. The proposed VAE-NMF achieved higher SDRs than RPCA in all conditions and even outperformed the supervised DNN-IRM in BUS condition without any prior training of noise signals. From the results obtained using the test set constructed with the DEMAND noise data, we can see that VAE-NMF outperformed the other methods in all the conditions. The noise data in DEMAND is unknown to DNN-IRM trained using the CHiME-3 training

[2]SDRs were evaluated by dropping 2048 samples (5 frames) at both ends.

set, and its enhancement performance deteriorated significantly. These results clearly show the robustness of the proposed VAE-NMF against various types of noise conditions.

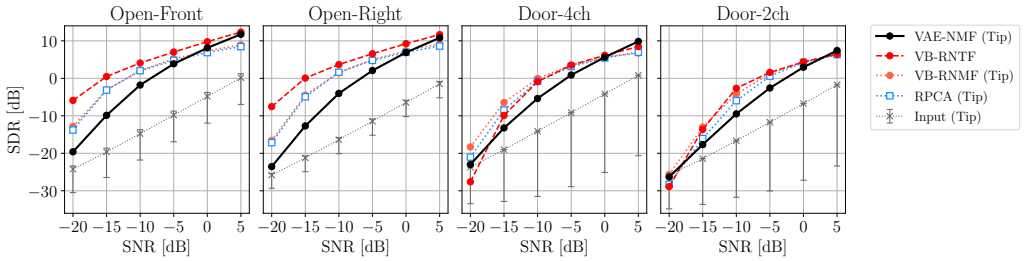
The SDR performance of VAE-NMF for the CAF condition in the DEMAND test set was lower than those for the other conditions. In this condition, the background noise contained conversational speech. Since VAE-NMF estimates speech component independently at each time frame, the background conversations were enhanced at the time frames where the power of the target speech was relatively small. This problem would be solved by making the VAE-based speech model to maintain time dependencies of a speech signal. The variational recurrent neural network [149] would be useful for this extension.

4.5 Evaluation with Hose-Shaped Rescue Robot

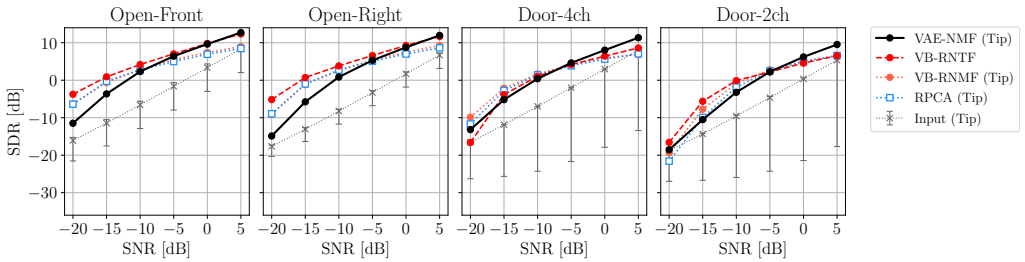
This section reports the enhancement performance of VAE-NMF for a hose-shaped rescue robot.

4.5.1 Experimental Settings

The mixture signals were the same signals as those used in Section 3.5. VAE-NMF was compared with RPCA, VB-RNMF, and VB-RNTF. The enhancement performance was evaluated by using the SDR for the whole enhanced signal. To evaluate the speech quality, the SDR for the speech section of the enhanced signal was also evaluated. The hyperparameters and network parameters of VAE-NMF were set to the same values as those for the previous section. The VAE-NMF was implemented by using python 3.6 and Chainer 3.2 framework. When the inference of VAE-NMF was conducted on a workstation that had an Intel Xeon E5-1650 CPU (6 cores, 3.5 GHz) and NVIDIA GeForce 1080 GPU, the elapsed time with a 20-seconds input signal was 19.0 s.



(a) SDRs for whole parts of enhanced signals



(b) SDRs for speech sections of enhanced signals

Figure 4.4: Speech enhancement performances in SDR. Error bars for the input signal span the maximum and minimum SDRs in all the microphones.

4.5.2 Experimental Results

As shown in Figure 4.4-(a), when the SNR was $+5$ dB, VAE-NMF performed better than RPCA and VB-RNMF, which are the low-rank and sparse decomposition methods. Furthermore, in the Door-4ch and -2ch conditions, VAE-NMF outperformed the multichannel method of VB-RNTF when the SNR was $+5$ dB. As shown in Figure 4.4-(b), when the SNR was more than -5 dB, VAE-NMF performed better than RPCA and VB-RNMF in the SDR for the speech section of the enhanced signal. These results show that the VAE-NMF extracted speech signals more efficiently than RPCA and VB-RNMF did whereas its noise suppression performance was lower than those of the low-rank and sparse decomposition methods. As shown in Figure 4.5, the enhancement results of VAE-NMF includes more residuals of noise signals than the other methods. VAE-NMF would be further improved by introducing a sparse constraint to the VAE speech model.

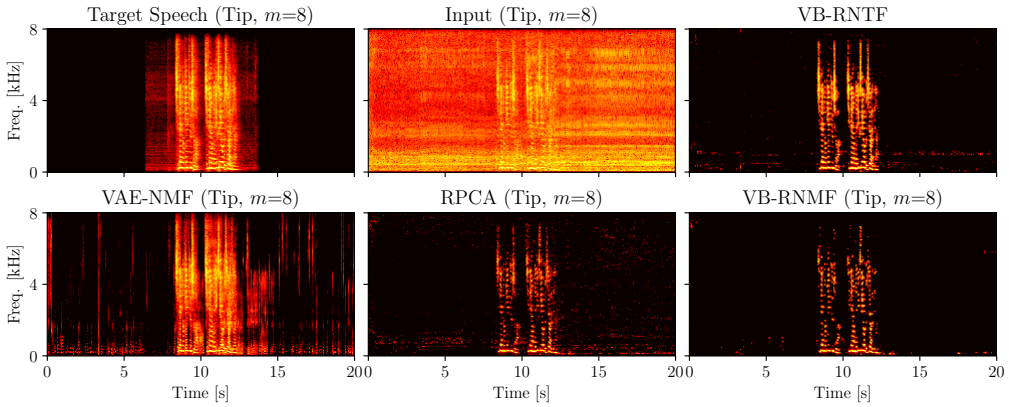


Figure 4.5: Excerpts of enhancement results when the layout was the Front condition and the SNR was $+5$ dB.

4.6 Summary

This chapter presented a semi-supervised speech enhancement method, called VAE-NMF, that involves a probabilistic generative model of speech based on a VAE and that of noise based on NMF. Only the speech model is trained in advance by using a sufficient amount of clean speech. Using the speech model as a prior distribution, posterior estimates of clean speech were obtained by using an MCMC sampler while adapting the noise model to noisy environments. Experimental results showed that VAE-NMF outperformed the conventional method based on low-rank and sparse decomposition. In addition, the results showed that VAE-NMF outperformed the conventional supervised DNN-based method in unseen noisy environments. It was also experimentally confirmed that the enhancement performance of VAE-NMF for a hose-shaped rescue robot was higher than those of VB-RNMF and RPCA when the SNR was relatively high.

Chapter 5

Audio-Based Time-Varying Posture Estimation

This chapter presents an audio-based method that can accurately estimate the time-varying posture of a hose-shaped rescue robot. The estimation is conducted based on a state-space model that represents the posture dynamics.

5.1 Introduction

To control a hose-shaped robot that flexibly changes its posture (shape) over time in an unseen environment, it is necessary to estimate the time-varying posture of the moving robot. Ishikura *et al.* [56], for example, proposed an inertial-sensor-based method that can estimate the posture by integrating the acceleration and angular-velocity information obtained from gyro sensors installed on the robot. Such integral-type methods based on the posture change rate, however, cannot work over a long time because the estimation error is gradually accumulated. Although non-integral-type methods based on information obtained by magnetometers and strain gauges can accurately track the posture independently of the past history [57,150], those methods can neither be used indoors nor be used for a robot with a long body.

A posture of the robot can be estimated by localizing microphone positions with sound generated by itself. This study aims to develop a non-integral-type posture estimation method that robustly works in disaster sites. The proposed method uses a set of microphones and loudspeakers distributed on a hose-

shaped rescue robot. The robot emits a reference signal from a loudspeaker one by one and estimates its posture by measuring the time differences of arrival (TDOAs) at the microphones. Since those TDOAs depend only on the current relative positions of the microphones and loudspeakers, the cumulative error problem can be avoided. The audio-based approach can be used in a closed space allowing sound propagation, whereas the accurate magnetometer-based approach can be used only outdoors for receiving the Earth's magnetic field. This indicates that audio-based posture estimation is complementary to inertial-sensor-based and magnetometer-based posture estimation.

The major requirement of the posture estimation is that the time-varying robot posture should be continuously presented to an operator in real time. Most of the existing audio-based methods are intended for offline use and assume that microphones are stable [18, 61, 120, 121]. Miura et al. [22] proposed a method based on simultaneous localization and mapping (SLAM) framework. This method can localize microphones and a moving sound source in an online manner. However, this method also assumes the microphones to be stable.

This chapter presents an audio-based online method that can accurately estimate the time-varying posture of a moving hose-shaped robot. This is achieved by formulating state-space model that represents the dynamics of not only the posture itself but also its change rate in the state space. The proposed model has two distinct characteristics. First, to use the method in an online manner, the current posture of the robot is predicted from the previous posture by using an unscented Kalman filter (UKF) [151]. Second, the proposed model assumes that the relative positions of the microphones and loudspeakers can change over time under a constraint that the microphones and loudspeakers are serially linked in a specified order.

5.2 Audio-based Posture Estimation

The proposed method estimates the posture of a moving robot by using TDOAs calculated from the recorded signals. The posture of a hose-shaped robot is

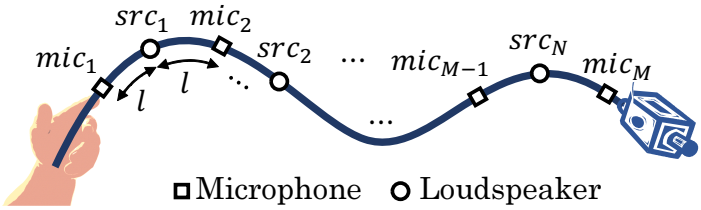


Figure 5.1: Microphone and loudspeaker arrangements.

estimated by repeating the following three steps: 1) generate a reference signal from each loudspeaker, one by one, 2) estimate the TDOAs of the reference signal at the microphones, and 3) estimate the relative positions of the microphones and loudspeakers from the estimated TDOAs.

5.2.1 Problem Specification

Microphones and loudspeakers are installed alternately at a regular interval l on the body of a hose-shaped robot, as shown in Figure 5.1. Note that the bending of the robot body makes the distance between adjacent modules less than l. Let mic_m $(m = 1, \cdots, M)$ and src_n $(s = 1, \cdots, N)$ are the microphones and loudspeakers, respectively, where $N = M - 1$. Let k, $\boldsymbol{x}^{\mathrm{mic}}_{m,k}$, and $\boldsymbol{x}^{\mathrm{src}}_{n,k}$ represent a measurement index, the microphone, and loudspeaker positions, respectively. In this chapter, it is assumed that the microphones and loudspeakers are on a two-dimensional surface.

The problem of the posture estimation is defined for each k as follows:

Input:	Synchronized M-channel audio signals $\boldsymbol{y}_k(t)$ obtained by recording a reference signal $s(t)$ emitted from src_{n_k}.
Output:	The relative positions of microphones $\boldsymbol{x}^{\mathrm{mic}}_{m,k}$ and loudspeakers $\boldsymbol{x}^{\mathrm{src}}_{n,k}$.

The input signals are used for calculating the TDOA of the reference signal at each microphone. Since the TDOA represents the relationship between the microphones and loudspeaker, the output is the *relative* positions of the microphones and loudspeakers. The $\boldsymbol{x}^{\mathrm{mic}}_{1,k}$ and $\boldsymbol{x}^{\mathrm{src}}_{1,k}$ are therefore assumed to be known without loss of generality.

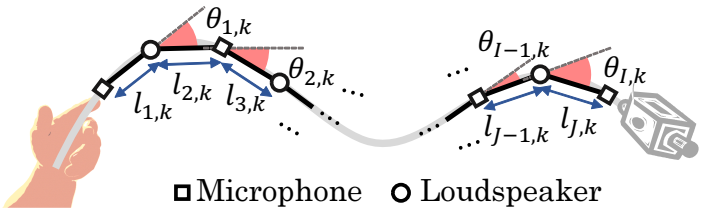


Figure 5.2: Serially-connected link model of robot posture.

5.2.2 State-Space Model of Robot Posture

The method estimates the posture of a moving robot by using the TDOAs calculated from the input data. More specifically, we formulate a nonlinear state-space model that associates a state space representing the posture dynamics with an observation space representing the TDOA. The UKF approximates the posterior distribution $p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k})$ from the likelihood $p(\boldsymbol{y}_k|\boldsymbol{\zeta}_k)$ and prior $p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k-1})$ using unscented transform.

The robot posture is modeled as a serially-connected link model, as shown in Figure 5.2. The posture at the k-th measurement, $\boldsymbol{z}_k$, is defined as

$$\boldsymbol{z}_k = [\theta_{1,k}, \cdots, \theta_{M+N-2,k}, l_{1,k}, \cdots, l_{M+N-1,k}], \tag{5.1}$$

where $\theta_{a,k}$ $(1 \leq a \leq M+N-2)$ is a link angle and $l_{b,k}$ $(1 \leq b \leq M+N-1)$ is a link length. To deal with a moving robot, the proposed method estimates not only posture, $\boldsymbol{z}_k$, but also its change rate, $\dot{\boldsymbol{z}}_k$. The state-space vector, $\boldsymbol{\zeta}_k$, is given by

$$\boldsymbol{\zeta}_k = [\boldsymbol{z}_k, \dot{\boldsymbol{z}}_k]^{\mathrm{T}} \in \mathbb{R}^L, \tag{5.2}$$

where $L = 4M + 4N - 6$ is the dimension of the state space.

The relative positions of the microphones and loudspeakers on the robot, $\boldsymbol{x}^{\mathrm{mic}}_{m,k}$ and $\boldsymbol{x}^{\mathrm{src}}_{n,k}$, can be calculated recursively from the known positions $\boldsymbol{x}^{\mathrm{mic}}_{1,k}$ and $\boldsymbol{x}^{\mathrm{src}}_{1,k}$. Suppose that $\boldsymbol{x}^{*}_{i,k}$ is the i-th member of $[\boldsymbol{x}^{\mathrm{mic}}_{1,k}, \boldsymbol{x}^{\mathrm{src}}_{1,k}, \cdots, \boldsymbol{x}^{\mathrm{mic}}_{M-1,k}, \boldsymbol{x}^{\mathrm{src}}_{N,k}, \boldsymbol{x}^{\mathrm{mic}}_{M,k}]$, each position is given by

$$\boldsymbol{x}^{*}_{i,k} = \boldsymbol{x}^{*}_{i-1,k} + l_{i,k} \times \left[\cos\left(\sum_{a=1}^{i-1}\theta_{a,k}\right), \sin\left(\sum_{a=1}^{i-1}\theta_{a,k}\right)\right]^{\mathrm{T}}.$$

Measurement Model

The TDOA measurement model $p(\boldsymbol{\tau}_k|\boldsymbol{\zeta}_k)$ is defined using a set of TDOAs $\tau^{n_k}_{n_k \to m_k}$ where the m_k is the one of the filtered microphone indices $\mathcal{M}_k$:

$$p(\boldsymbol{\tau}_k|\boldsymbol{\zeta}_k) = \mathcal{N}(\boldsymbol{\tau}_k|[\tau^{n_k}_{n_k \to m}(\boldsymbol{z}_k)|m \in \mathcal{M}_k]^{\mathrm{T}}, \sigma^{\tau}\mathbf{I}), \tag{5.3}$$

where $\sigma^{\tau} \in \mathbb{R}_+$ represents the variance of the measurement noise and TDOA $\tau^{n}_{m_1 \to m_2}(\boldsymbol{\zeta}_k)$ is calculated by using the distances between the two microphones and the loudspeaker as follows:

$$\tau^{n}_{m_1 \to m_2}(\boldsymbol{\zeta}_k) = \frac{|\boldsymbol{x}^{\mathrm{mic}}_{m_2,k} - \boldsymbol{x}^{\mathrm{src}}_{n,k}| - |\boldsymbol{x}^{\mathrm{mic}}_{m_1,k} - \boldsymbol{x}^{\mathrm{src}}_{n,k}|}{c}, \tag{5.4}$$

where c represents the speed of sound. In this thesis, C is assumed to be 340 m/s.

State Update Model

A state update model $p(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1})$ is based on two concepts: a) posture dynamics and b) posture constraint. The posture dynamics $q(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1})$ represents how likely the previous posture $\boldsymbol{z}_{k-1}$ is to change to the current posture $\boldsymbol{z}_k$ with a change rate $\dot{\boldsymbol{z}}_{k-1}$ as follows:

$$q(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1}) = \mathcal{N}\left(\boldsymbol{\zeta}_k|[\boldsymbol{z}_{k-1} + \dot{\boldsymbol{z}}_{k-1}, \dot{\boldsymbol{z}}_{k-1}]^{\mathrm{T}}, \mathrm{diag}\left(\boldsymbol{\sigma}^{\boldsymbol{\zeta}}\right)\right), \tag{5.5}$$

where $\boldsymbol{\sigma}^{\boldsymbol{\zeta}} \in \mathbb{R}^L_+$ is the variance vector of the process noise. The posture constraint $r(\boldsymbol{\zeta}_k)$, on the other hand, is modeled as a Gaussian distribution:

$$r(\boldsymbol{\zeta}_k) = \mathcal{N}(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}, \mathbf{P}), \tag{5.6}$$

where $\boldsymbol{\zeta} \in \mathbb{R}^L$ and $\mathbf{P} \in \mathbb{R}^{L \times L}$ are the mean vector and covariance matrix of the feasible posture. More specifically, the following gaussian prior distribution is put on $\boldsymbol{\zeta}_k$:

$$r(\boldsymbol{\zeta}_k) = \prod_{a=1}^{M+N-2} \left\{\mathcal{N}\left(\theta_{a,k}|0, \sigma^{\theta}\right)\mathcal{N}\left(\dot{\theta}_{a,k}|0, \sigma^{\dot{\theta}}\right)\right\} \prod_{b=1}^{M+N-1} \left\{\mathcal{N}\left(l_{b,k}|l, \sigma^{l}\right)\mathcal{N}\left(\dot{l}_{b,k}|0, \sigma^{\dot{l}}\right)\right\}, \tag{5.7}$$

where σ^{θ}, $\sigma^{\dot{\theta}}$, σ^{l}, and $\sigma^{\dot{l}}$ are the variance parameters of the gaussian distribution.

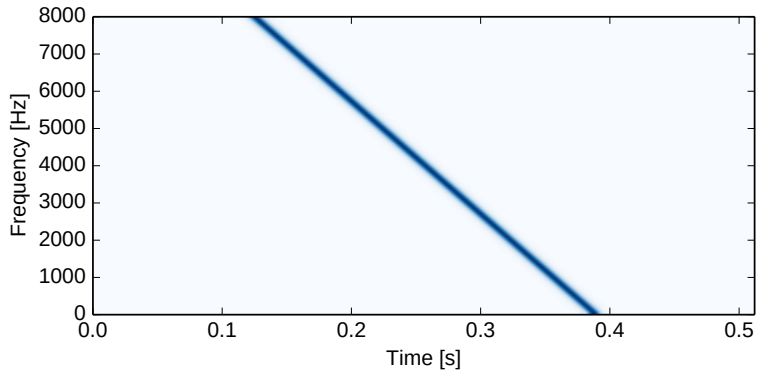


Figure 5.3: TSP signal with length of 8192 samples at 16 kHz.

These two distributions are integrated into the state update model $p(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1})$ on the basis of the product of experts [152]:

$$p(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1}) = \frac{1}{A} q(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1}) r(\boldsymbol{\zeta}_k), \tag{5.8}$$

where $A = \int q(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1}) r(\boldsymbol{\zeta}_k) d\boldsymbol{\zeta}_k$ is a normalization factor.

Estimation Algorithm

The robot posture $\boldsymbol{z}_k$ is estimated from $\boldsymbol{y}_{1:k}$ in an online manner by using an UKF [151] assuming that the posterior distribution of the state variable $\boldsymbol{\zeta}_k$ follows a Gaussian distribution. The UKF approximates the posterior distribution $p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k})$ from the likelihood $p(\boldsymbol{y}_k|\boldsymbol{\zeta}_k)$ and prior $p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k-1})$ using unscented transform. The prior distribution $p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k-1})$ is obtained by calculating the following equation using unscented transform:

$$p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k-1}) = \int p(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1}) p(\boldsymbol{\zeta}_{k-1}|\boldsymbol{y}_{1:k-1}) d\boldsymbol{\zeta}_{k-1}. \tag{5.9}$$

The calculation of the prior distribution $p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k-1})$ can be simplified as follows. Since the $q(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_{k-1})$ is a linear transformation of $\boldsymbol{\zeta}_{k-1}$ (Equation (5.5)) and the $r(\boldsymbol{\zeta}_k)$ is defined as a Gaussian distribution (Equation (5.6)), the state update model can be written as a linear model. The prior distribution $p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k-1})$

is therefore calculated without unscented transform as follows:

$$p(\boldsymbol{\zeta}_k|\boldsymbol{y}_{1:k-1}) = \mathcal{N}(\boldsymbol{\zeta}_k|\boldsymbol{\zeta}_k^-, \mathbf{P}_k^-), \tag{5.10}$$

$$\boldsymbol{\zeta}_k^- = \mathbf{P}_k^-((\mathbf{P}_k^*)^{-1}\hat{\boldsymbol{z}}_{k-1} + \mathbf{P}^{-1}\boldsymbol{\zeta}), \tag{5.11}$$

$$\mathbf{P}_k^- = ((\mathbf{P}_k^*)^{-1} + \mathbf{P}^{-1})^{-1}, \tag{5.12}$$

$$\mathbf{P}_k^* = \mathbf{F}^{\mathrm{T}}\hat{\mathbf{P}}_{k-1}\mathbf{F} + \mathbf{P}_k, \tag{5.13}$$

where $\hat{\boldsymbol{z}}_{k-1}$ and $\hat{\mathbf{P}}_{k-1}$ are the mean vector and covariance matrix of the last posterior distribution $p(\boldsymbol{\zeta}_{k-1}|\boldsymbol{y}_{1:k-1})$. $\mathbf{F} \in \mathbb{R}^{L\times L}$ is the transition matrix representing the linear update model. This calculation is recursively performed over time.

5.2.3 Robust TDOA Estimation

To make TDOA estimation robust against motor noise, we use a time stretched pulse (TSP) [153] as a reference signal (Fig. 5.3). A TSP has a high time resolution because the auto-correlation of the TSP signal become an impulse. In addition, the TSP can be sent with large energy from a loudspeaker. Therefore, the reference signal can be easily distinguished from the motor noise. A TSP signal with a length of W samples is defined in the frequency domain as follows:

$$S(\omega) = \begin{cases} \exp(j2\pi\omega^2/W^2) & 0 \le \omega \le W/2 \\ S(W-\omega) & W/2 \le \omega \le W \end{cases}, \tag{5.14}$$

where $S(\omega)$ is the frequency spectrum of the reference signal $s(t)$ in the frequency domain and ω indicates a frequency. The reference signal $s(t)$ is obtained by the inverse discrete Fourier transform of $S(\omega)$.

As shown in Fig. 5.4, TDOA $\tau^n_{m_1\to m_2}$ is estimated from the recorded signal $\boldsymbol{y}_k(t)$ as follows:

1. Calculate the cross correlation coefficient $G_{m,k}(\tau)$ between each recorded signal $z_{m,k}(t)$ and the reference signal $s(t)$.
2. Calculate the onset times of the input signals, $t_{m_1,k}$ and $t_{m_2,k}$ by detecting the first peak of the correlation coefficient $G_{m_1,k}(\tau)$ and $G_{m_2,k}(\tau)$, respectively.
3. Calculate the TDOA $\tau^n_{m_1\to m_2}$ by subtracting $t_{m_1,k}$ from $t_{m_2,k}$.

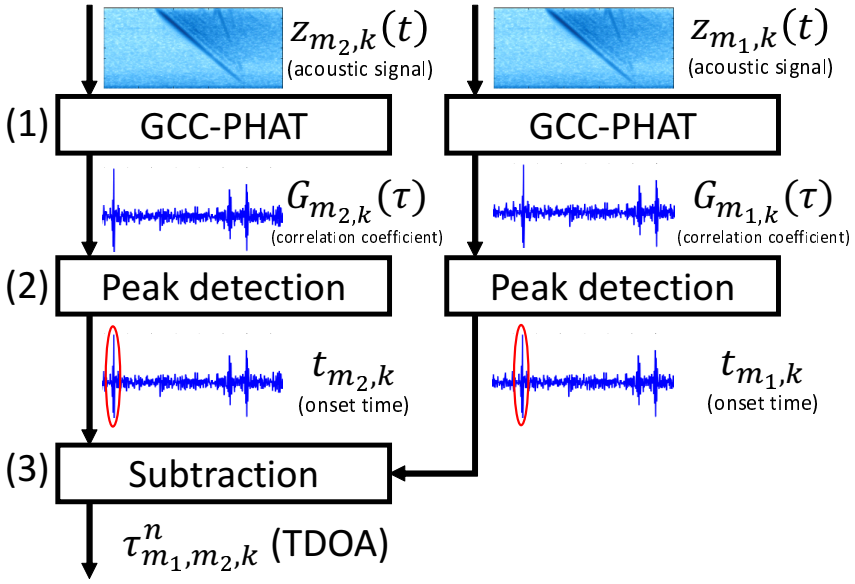


Figure 5.4: Overview of TDOA estimation.

The cross correlation is calculated using the generalized cross correlation with phase transform (GCC-PHAT), which is robust against reverberation [122, 154].

5.3 Experimental Evaluation

This section reports the experiments that were conducted for evaluating the proposed method of online posture estimation using a prototype hose-shaped robot as shown in Figure 5.5.

5.3.1 Experimental Settings

The proposed method was compared with a conventional method that does not consider the posture change rate. The initial shape of the robot was set to one of three postures: C-shape, S-shape, and straight. The TSP reference signal had a length of 8192 samples (512 ms) generated at 16 kHz. The TSP is played by each loudspeaker in order (1, 2, 3, . . ., 7, 1, . . .). To use a UKF, the initial state $\boldsymbol{\zeta}_0 = [\boldsymbol{z}_0, \dot{\boldsymbol{z}}_0]$ was determined in the following manner. The initial posture $\boldsymbol{z}_0$ was sampled from a Gaussian distribution whose mean corresponds to the correct posture and standard deviation was 15°. The initial change rate $\dot{\boldsymbol{z}}_0$ was set to

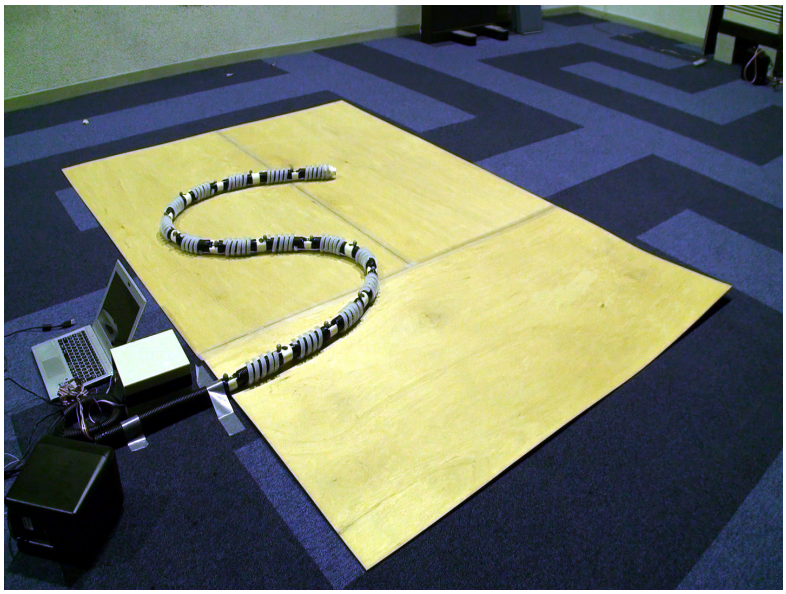
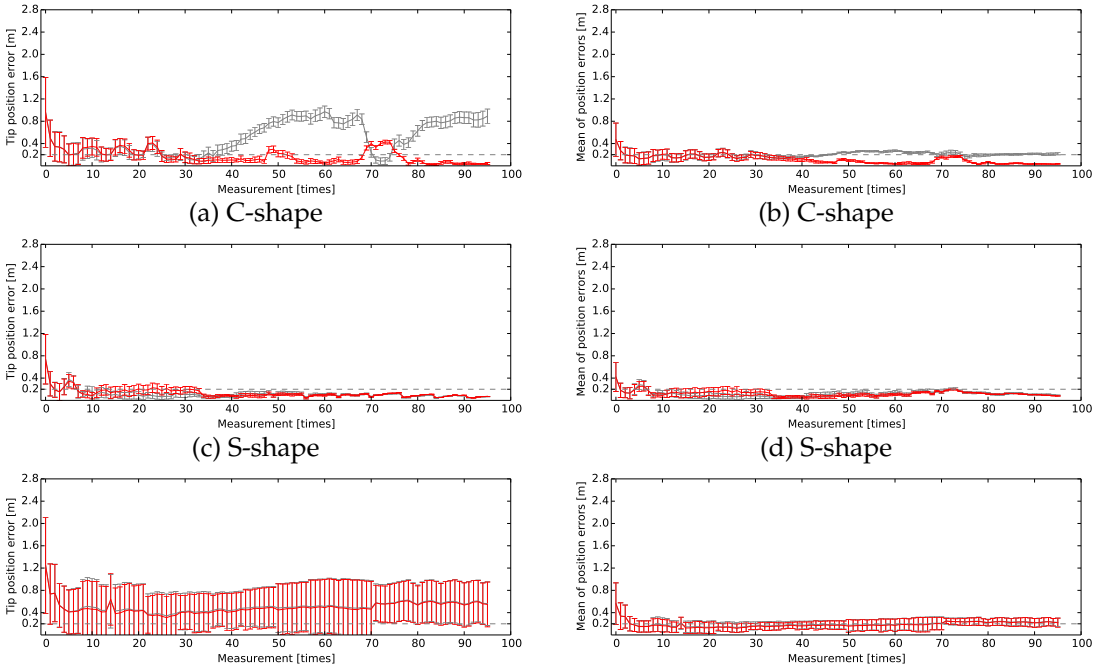


Figure 5.5: Prototype hose-shaped robot placed on experimental room.

zero. The other parameters were determined experimentally.

The estimation algorithm was implemented using Python without multiprocessing. A standard laptop computer with an Intel Core i7-3517U CPU (2 cores, 1.9 GHz) and 4.0 GB of memory was used to estimate the TDOAs of the reference signal and the posture of the robot. The CPU time and elapsed time for 50 TDOA estimations (25.6 s) were 8.759 s and 8.843 s, respectively. Those for posture estimation were 2.679 s and 2.697 s, respectively. Therefore, the total computation time for an input signal of 25.6 s was 11.456 s.

The tip position error was the distance between the ground-truth and estimated positions of the tip microphone. The average estimation error was the average distance between the ground-truth and estimated positions of all the microphones. The ground-truth position of each microphone was measured using a motion capture system (OptiTrack, NaturalPoint Inc.). Ishikura et al. [56] achieved a tip position error of under about 0.2 m at 35 sec estimation with a 3.0 m mock-up robot by using the inertial sensor approach. To overcome this method, the objective accuracy was set to 0.2 m of the position error. All experiments were conducted in an experimental room whose reverberation time (RT_{60}) was 800 ms.



(a) C-shape (b) C-shape

(c) S-shape (d) S-shape

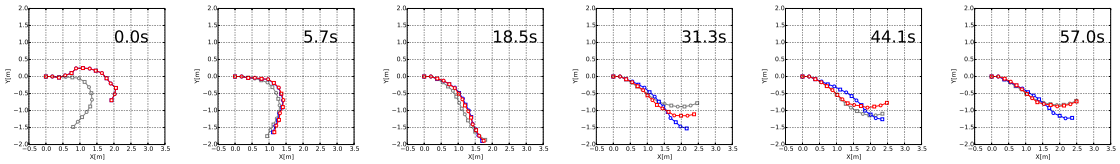
(e) Straight (f) Straight

Figure 5.6: The tip and average position errors obtained by the proposed and baseline methods. The red line represents the proposed method, and the gray line represents the baseline method. The polyline and error bar indicate the mean and standard deviation, respectively.

5.3.2 Experimental Results

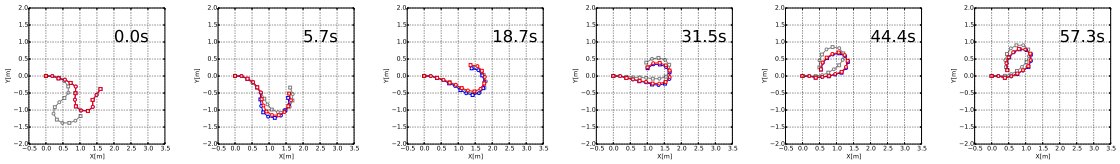
When the initial posture was set to the C-shape or S-shape, as shown in Figures 5.6-(a), -(b), -(c), and -(d), the estimation errors were decreased over time and, as shown in Figures 5.7 and 5.8, the estimated postures followed the moving robot postures accurately. Moreover, when the initial posture was set to the C-shape, the baseline method failed to follow the moving posture and the estimation error increased after the 30-th measurement. On the other hand, the proposed method successfully tracked the moving posture in real time. The estimation errors, when the initial posture was set to the C-shape or S-shape, were almost under 0.2 m after the 40-th measurement.

When the initial posture was straight, as shown in Figures 5.6-(e) and -(f), on the other hand, the estimation error was larger than those obtained in the cases of the other initial postures. This is because of the mirror-symmetrical



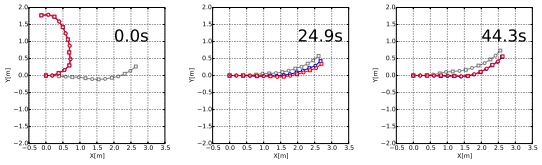
(a) Initial (b) 10-th (c) 30-th (d) 50-th (e) 70-th (f) 90-th

Figure 5.7: Estimation results when the initial posture was set to the C-shape. The red and blue lines indicate the postures estimated by the proposed and baseline methods, respectively. The gray line illustrates the correct posture.



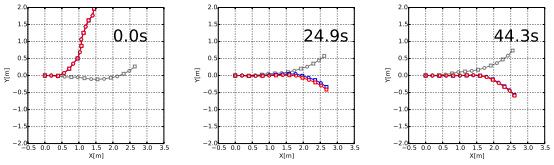
(a) Initial (b) 10-th (c) 30-th (d) 50-th (e) 70-th (f) 90-th

Figure 5.8: Estimation results when the initial posture was set to the S-shape.



(a) Initial (b) 40-th (c) 70-th

Figure 5.9: Estimation results when the initial posture was set to straight.



(a) Initial (b) 40-th (c) 70-th

Figure 5.10: Another set of results when the initial posture was set to straight.

problem. Since the microphones and loudspeakers were installed in a row on the robot, it is difficult to distinguish between two postures which were mirror-symmetrical with respect to mic_1 and src_1. The correct posture was estimated with an initial state as shown in Figure 5.9 whereas the mirror-symmetrical posture was estimated with another initial state as shown in 5.10.

The possible reason why the baseline method failed in the C-shape condition is also due to the mirror-symmetrical problem. As shown in Figure 5.7-(c), more than half part of the posture became straight at the 30-th measurement. The baseline method failed to estimate the correct posture because the method failed to estimate the direction of the robot movement after the 30-th measurement. The experimental results show that the estimation of the posture change rate reduces the mirror-symmetrical problem of a moving robot. However,

additional information is required to solve the problem completely.

One way to solve the mirror symmetrical problem is to use multi-modal information, *i.e.*, integrate various types of information obtained from microphones, accelerometers, and gyro sensors. If a robot has those modalities, mirror-symmetrical postures can be distinguished by considering the posture change history and the robot can work in a closed and narrow space in which some modalities do not work. The mirror-symmetrical ambiguity could be handled with an unscented particle filter [155] that can maintain multiple possibilities about the posture of the robot at the same time.

5.4 Summary

This chapter presented an online method that can accurately estimate the time-varying posture of a moving hose-shaped rescue robot having multiple microphones and loudspeakers. The experiments using a 3 m moving hose-shaped robot showed that the method successfully suppressed the estimation error under 20 cm at the tip position even after the robot moved over a long time. The results also revealed that the purely audio-based method often confuses mirror-symmetrical postures, depending on the initial value of the estimation. The mirror-symmetrical problem will be solved by integrating with other sensors such as gyroscopes. The probabilistic state-space modeling enables us to integrate various types of information obtained from multi-modal sensors in a principled way. The 3D posture estimation is also an important future work. This extension is addressed in Chapter 6.

Chapter 6

Microphone-Accelerometer Based 3D Posture Estimation

This chapter presents 3D posture estimation that can deal with the partial occlusion of microphones. The unreliable audio measurements due to the occlusion is compensated for by using the tilt information obtained from accelerometers.

6.1 Introduction

Sensor systems on rescue robots including the hose-shaped robots typically do not work well in the extreme environments where such robots are intended to be used [1, 68–70]. The accuracy of the GPS, for example, is degraded because the rubble in collapsed buildings blocks signals from the satellites [1], and a video camera inserted into narrow gaps often fails to capture the views there because the lighting causes whiteout or blackout conditions [68]. To develop robust sensor systems, it is thus essential to integrate various modalities compensating each other's weaknesses [2, 156–158].

Although there are many posture estimation methods using various types of sensors [56, 57, 159], these methods face some problems in the disaster environments. The performances of magnetometer-accelerometer based method, for example, are degraded in the disaster environments because magnetic fields are easily affected by the steel frames of collapsed buildings [57]. As presented in Chapter 5, the audio-based posture estimation can be used in a closed space allowing sound propagation among microphones and loudspeakers installed

on the robot. This method uses a set of microphones and loudspeakers installed on the robot, and estimates the posture from the time differences of arrival (TDOAs) using recorded acoustic signals. Nevertheless, an audio-based method often fails to estimate the robot posture accurately because the obstacles around the robot block sound propagation.

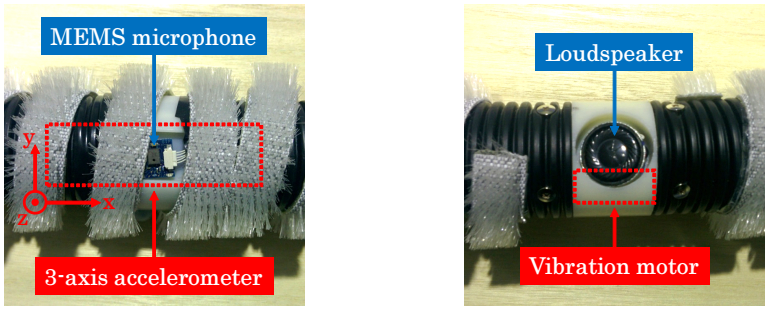
This chapter presents a microphone-accelerometer based 3D posture estimation method for a hose-shaped robot equipped with a set of microphones, loudspeakers, and accelerometers. The microphones and loudspeakers allow to estimate their relative positions using the time differences of arrival (TDOAs) of a reference signal emitted from the loudspeakers, and the accelerometers are used for estimating their tilts by measuring the acceleration of gravity. Since the TDOA-based method is degraded in a rubble-containing environment, the proposed method excludes TDOAs distorted by rubble and fills up the lack of posture information with the tilt information. To do this, it detect TDOAs of direct sound by excluding outliers, and estimate the robot posture based on a nonlinear state-space model integrating TDOA and tilt information by using the unscented Kalman filter (UKF) [151].

6.2 3D Posture Estimation Based on Microphones and Accelerometers

In the proposed method of microphone-accelerometer based 3D posture estimation, the posture of a hose-shaped robot is estimated by repeating the following four steps: 1) generate a reference signal from each loudspeaker, one by one, 2) estimate the reference signal's TDOAs at the microphones, 3) estimate the tilts at 3-axis accelerometers, and 4) estimate the robot posture from the estimated TDOAs and tilts by using the UKF.

6.2.1 Prototype Hose-shaped Robot

As shown in Figure 1.3, a prototype hose-shaped robot is used in this study. This robot has two types of modules, one with a microphone (mic) and 3-axis accelerometer (acc) (Figure 6.1-(a)) and the other with a small loudspeaker (src)



(a) Microphone and accelerometer (b) Loudspeaker and vibrator

Figure 6.1: Modules with a microphone and accelerometer or a loudspeaker and vibrator placed on the robot.

and vibration motor (vib) (Figure 6.1-(b)). As shown in Figure 6.2, $M = 8$ mic-acc modules and $N = 7$ src-vib modules are positioned on the robot at a regular interval $l = 20$ cm. The distance between the modules at the ends is 2.8 m.

6.2.2 Problem Specification

The microphones, accelerometers, and loudspeakers were denoted by mic_m, acc_m ($m = 1, \cdots, M$), and src_n ($n = 1, \cdots, N$), respectively, where $N = M - 1$. Let k be the measurement index and the mic-acc module and src-vib module positions be $\boldsymbol{x}^{\mathrm{mic}}_{m,k}$ and $\boldsymbol{x}^{\mathrm{src}}_{n,k}$, respectively.

The problem of the microphone-accelerometer based posture estimation is defined as follows:

Input:	1) TDOAs $\tau^n_{m_1 \to m_2} \in \mathbb{R}$ $(m_1, m_2 \in \mathcal{M}_k)$ when src_n omits a reference signal, and 2) tilt angles at the accelerometers $\psi_{1,k}, \cdots, \psi_{M,k} \in \mathbb{R}$.
Output:	The positions of each mic-acc module $\boldsymbol{x}^{\mathrm{mic}}_{m,k} \in \mathbb{R}^3$ and each src-vib module $\boldsymbol{x}^{\mathrm{src}}_{n,k} \in \mathbb{R}^3$.

where $\tau^n_{m_1 \to m_2}$ represents a TDOA between mic_{m_1} and mic_{m_2} and $\mathcal{M}_k$ represents a set of indices for microphones that record the direct sound of the reference signal. The TDOAs $\tau^n_{m_1 \to m_2}$ and microphone indices $\mathcal{M}_k$ are estimated from synchronized M-channel audio signals $\boldsymbol{y}_k(t) \in \mathbb{R}^M$ obtained by recording a reference signal $s(t) \in \mathbb{R}$ (Sec. 6.2.3). The tilts $\psi_{1,k}, \cdots, \psi_{M,k}$ are estimated from M-channel 3-axis accelerometer measurements $\boldsymbol{a}_{1,k}, \cdots, \boldsymbol{a}_{M,k} \in \mathbb{R}^3$ (Sec. 6.2.3).

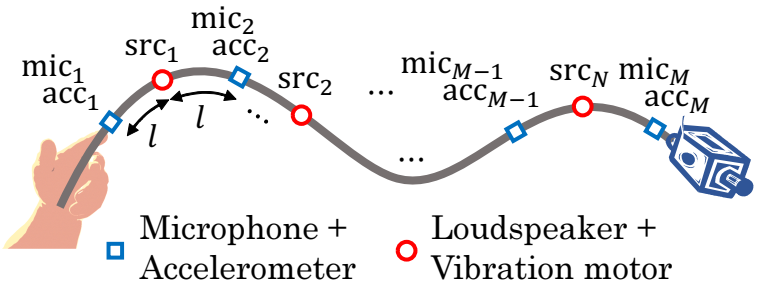


Figure 6.2: Arrangements of microphones, accelerometers, and loudspeakers.

6.2.3 Feature Extraction

The robot posture is estimated by using TDOAs and tilts at the mic-acc modules calculated from the M-ch audio signal $\boldsymbol{y}_k(t)$ and the accelerometer measurements $\boldsymbol{a}_{1,k}, \cdots, \boldsymbol{a}_{M,k}$.

TDOA Estimation

The output of TDOA estimation is a set of TDOAs $\tau^n_{m_1 \to m_2}(m_1, m_2 \in \mathcal{M}_k)$ where $\mathcal{M}_k$ represents a set of microphone indices for microphones that record the direct sound of the reference signal. Since the TDOA between two adjacent microphones cannot be longer than the sound propagation time for the interval length on the robot ($2l$) in an open space, the proposed method excludes the TDOA that does not satisfy this theorem. We formulate the set of indices $\mathcal{M}_k$ for microphones that record the direct sound of the reference signal as follows:

$$\mathcal{M}_k = \{m | m \text{ satisfies } \mathrm{valid}(m)\} \tag{6.1}$$

$$\mathrm{valid}(m) = \begin{cases} \mathrm{valid}(m-1) \wedge |\tau^n_{m-1 \to m}| < \frac{2l}{c} & \text{if } m > n \\ |\tau^n_{n \to n+1}| < \epsilon & \text{if } m = n \\ \mathrm{valid}(m+1) \wedge |\tau^n_{m+1 \to m}| < \frac{2l}{c} & \text{if } m < n \end{cases} \tag{6.2}$$

where c and ϵ represent the speed of sound in an open space and a threshold parameter for regarding the TDOA $\tau^n_{n+1,n} n$ as small enough, respectively. A TDOA $\tau^n_{m_1 \to m_2}$ between mic_{m_1} and mic_{m_2} when src_n omits a reference signal is estimated from the difference of onset times at the microphones $t^n_{m_1,k}$ and $t^n_{m_2,k}$:

$$\tau^n_{m_1 \to m_2} = t^n_{m_2,k} - t^n_{m_1,k}. \tag{6.3}$$

The TDOA is calculated in the same way as in Chapter 5.

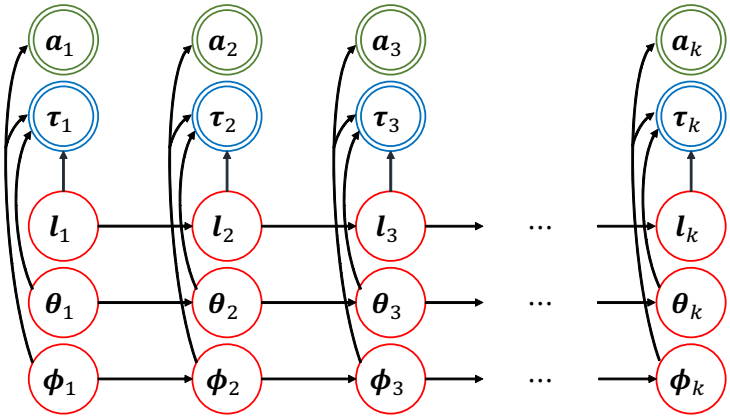


Figure 6.3: Graphical representation of the proposed state-space model.

Tilt Estimation

The tilts at the accelerometers are estimated by measuring the direction of gravitational acceleration. The output of tilt estimation is a set of tilts $\psi_{m,k}(m = 1, \cdots, M)$ at the accelerometers acc_m. The tilt $\psi_{m,k}$ is estimated from the accelerometer measurements $\boldsymbol{a}_{m,k}$ as follows:

$$\psi_{m,k} = \arctan\left(-a^x_{m,k} \Big/ \sqrt{(a^y_{m,k})^2 + (a^z_{m,k})^2}\right) \tag{6.4}$$

where $a^x_{m,k}$, $a^y_{m,k}$, and $a^z_{m,k}$ represent the elements of the input acceleration $\boldsymbol{a}_{m,k}$, respectively.

6.2.4 State-Space Model of Robot Posture

The proposed state-space model associates a state space representing the 3D robot posture with an observation space representing the TDOA and tilt of each mic-acc module (Figure 6.3). The current posture is estimated by using the UKF.

As shown in Figure 6.4, the robot posture is modeled as a serially-connected link model. A posture at the k-th measurement, $\boldsymbol{z}_k$, is defined as follows:

$$\boldsymbol{z}_k = [\theta_{1,k}, \cdots, \theta_{M+N-2,k}, \phi_{1,k}, \cdots, \phi_{M+N-1,k}, l_{1,k}, \cdots, l_{M+N-1,k}]^{\mathrm{T}}, \tag{6.5}$$

where $\theta_{i,k}$, $\phi_{i,k}$, and $l_{i,k}$ are a horizontal link angle, a vertical link angle, and a link length, respectively.

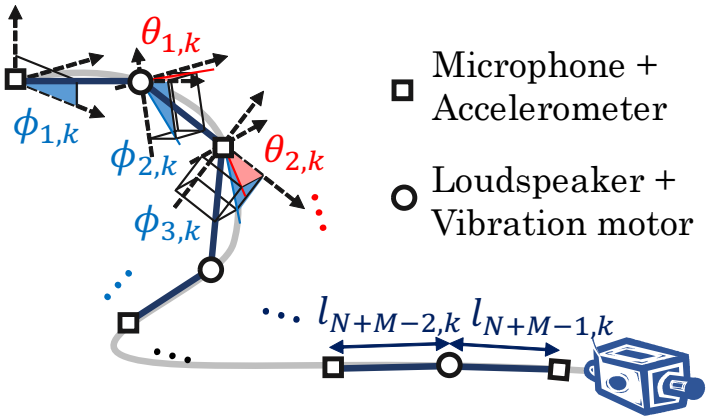


Figure 6.4: 3D serially-connected link model of robot posture.

The relative positions of the microphones and loudspeakers on the robot, $\boldsymbol{x}^{\mathrm{mic}}_{m,k}$ and $\boldsymbol{x}^{\mathrm{src}}_{n,k}$, are calculated recursively from the first position $\boldsymbol{x}^{\mathrm{mic}}_{1,k}$. Suppose that $\boldsymbol{x}^{*}_{i,k}$ is the i-th member of $[\boldsymbol{x}^{\mathrm{mic}}_{1,k}, \boldsymbol{x}^{\mathrm{src}}_{1,k}, \cdots, \boldsymbol{x}^{\mathrm{mic}}_{M-1,k}, \boldsymbol{x}^{\mathrm{src}}_{N,k}, \boldsymbol{x}^{\mathrm{mic}}_{M,k}]$. Then each position is given by

$$\boldsymbol{x}^{*}_{i,k} = \boldsymbol{x}^{*}_{i-1,k} + l_{i-1,k} \begin{bmatrix} \cos(\phi^{*}_{i,k})\cos(\theta^{*}_{i,k}) \\ \cos(\phi^{*}_{i,k})\sin(\theta^{*}_{i,k}) \\ \sin(\phi^{*}_{i,k}) \end{bmatrix}, \quad \phi^{*}_{i,k} = \sum_{j=1}^{i-1} \phi_{j,k}, \quad \theta^{*}_{i,k} = \sum_{j=1}^{i-2} \theta_{j,k}. \tag{6.6}$$

State Update Model

As in Chapter 5, the state update model $p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1})$ is based on two concepts: a) posture dynamics and b) posture constraint. The posture dynamics $q(\boldsymbol{z}_k|\boldsymbol{z}_{k-1})$ is represented as random walk:

$$q(\boldsymbol{z}_k|\boldsymbol{z}_{k-1}) = \mathcal{N}\left(\boldsymbol{z}_k|\boldsymbol{z}_{k-1}, \mathrm{diag}\left(\boldsymbol{\sigma}^{\boldsymbol{z}}\right)\right), \tag{6.7}$$

where $\boldsymbol{\sigma}^{\boldsymbol{z}} \in \mathbb{R}^{L}_{+}$ is the variance vector of the process noise. Note that in this chapter only the posture $\boldsymbol{z}_k$ is estimated for evaluating the effectiveness of integration of the audio and accelerometer measurements. The posture constraint $r(\boldsymbol{z}_k)$, on the other hand, is modeled as a Gaussian distribution:

$$r(\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{z}_k|\boldsymbol{z}, \mathbf{P}), \tag{6.8}$$

where $\boldsymbol{z} \in \mathbb{R}^{L}$ and $\mathbf{P} \in \mathbb{R}^{L\times L}$ are the mean and covariance matrix of the feasible posture. The $\boldsymbol{z}$ and $\mathbf{P}$ are parameterized as in Eq. 5.7. These two distributions

are integrated for the state update model $p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1})$ on the basis of the product of experts [152]:

$$p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1}) = \frac{1}{A} q(\boldsymbol{z}_k|\boldsymbol{z}_{k-1}) r(\boldsymbol{z}_k), \tag{6.9}$$

where $A = \int q(\boldsymbol{z}_k|\boldsymbol{z}_{k-1}) r(\boldsymbol{z}_k) d\boldsymbol{z}_k$ is a normalization factor.

Measurement Model

The measurement model $p(\boldsymbol{\tau}_k, \boldsymbol{\psi}_k|\boldsymbol{z}_k)$ is formulated with two sub models: a) a TDOA measurement model $p(\boldsymbol{\tau}_k|\boldsymbol{z}_k)$ and b) a tilt measurement model $p(\boldsymbol{\psi}_k|\boldsymbol{z}_k)$ as follows:

$$p(\boldsymbol{\tau}_k, \boldsymbol{\psi}_k|\boldsymbol{z}_k) = p(\boldsymbol{\tau}_k|\boldsymbol{z}_k) p(\boldsymbol{\psi}_k|\boldsymbol{z}_k) \tag{6.10}$$

The TDOA measurement model $p(\boldsymbol{\tau}_k|\boldsymbol{z}_k)$ is defined using a set of TDOAs $\tau^{n_k}_{n_k \to m_k}$ where the m_k is the one of the filtered microphone indices $\mathcal{M}_k$:

$$p(\boldsymbol{\tau}_k|\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{\tau}_k|[\tau^{n_k}_{n_k \to m_1}\boldsymbol{z}_k)|m_1 \in \mathcal{M}_k]^{\mathrm{T}}, \mathbf{R}^{\tau}_k), \tag{6.11}$$

where $\mathbf{R}^{\tau}_k$ represents the covariance matrix of the measurement noise and TDOA $\tau^{n}_{m_1 \to m_2}(\boldsymbol{z}_k)$ is calculated by using the distances between the two microphones and the loudspeaker as follows:

$$\tau^{n}_{m_1 \to m_2}(\boldsymbol{z}_k) = \frac{|\boldsymbol{x}^{\mathrm{mic}}_{m_2,k} - \boldsymbol{x}^{\mathrm{src}}_{n,k}| - |\boldsymbol{x}^{\mathrm{mic}}_{m_1,k} - \boldsymbol{x}^{\mathrm{src}}_{n,k}|}{c}, \tag{6.12}$$

where c represents the speed of sound.

The tilt measurement $\boldsymbol{\psi}_k$ is a set of tilts angles $\psi_{m,k}$ at mic-acc modules:

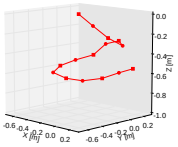
$$q(\boldsymbol{\psi}_k|\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{\psi}_k|[\psi_1(\boldsymbol{z}_k), \cdots, \psi_M(\boldsymbol{z}_k)]^{\mathrm{T}}, \mathbf{R}^{\psi}_k) \tag{6.13}$$

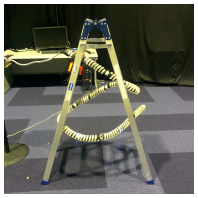
where $\mathbf{R}^{\psi}_k$ represents the covariance matrix of the measurement noise and tilt $\psi_m(\boldsymbol{z}_k)$ is calculated by accumulating the vertical link angles $\phi_{a,k}$ as follows:

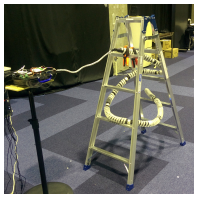
$$\psi_m(\boldsymbol{z}_k) = \frac{1}{2} \sum_{i=1}^{2m-2} \phi_{i,k} + \frac{1}{2} \sum_{i=1}^{2m-1} \phi_{i,k} \tag{6.14}$$

6.3 Evaluation

This section reports an experiment evaluating the proposed method of 3D posture estimation in rubble-containing environments.

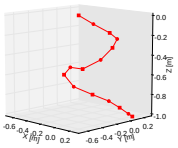

Ground-truth posture


Front

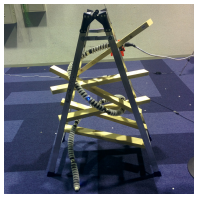

Left

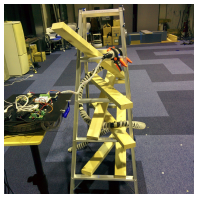

Right

(a) Condition 1: open space


Ground-truth posture

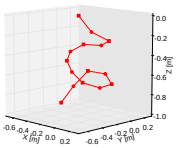

Front

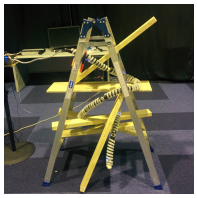

Back

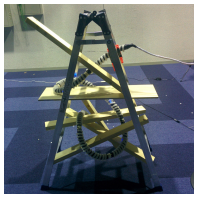

Left


Right

(b) Condition 2: sticks


Ground-truth posture


Front


Back


Left


Right

(c) Condition 3: sticks and plate

Figure 6.5: Three conditions for experimental evaluation. Ground-truth postures were measured using a motion capture system.

6.3.1 Experimental Settings

The proposed method was compared with a baseline method estimating the posture by using only microphone information. This experiment was conducted in an experimental room where the reverberation time RT_{60} was 800 ms. As shown in Figure 6.5, the robot postures were estimated in the following three conditions:

1. **Open space**: There was no rubble around the robot. The robot curved three-dimensionally on a stepladder 140 cm high.
2. **Sticks**: Six wooden sticks (91 cm $\times$ 9 cm $\times$ 4 cm) representing rubble were placed around the robot.
3. **Sticks and plate**: The six wooden sticks and a wooden plate (91 cm $\times$ 25 cm $\times$ 1.5 cm) were placed around the robot.

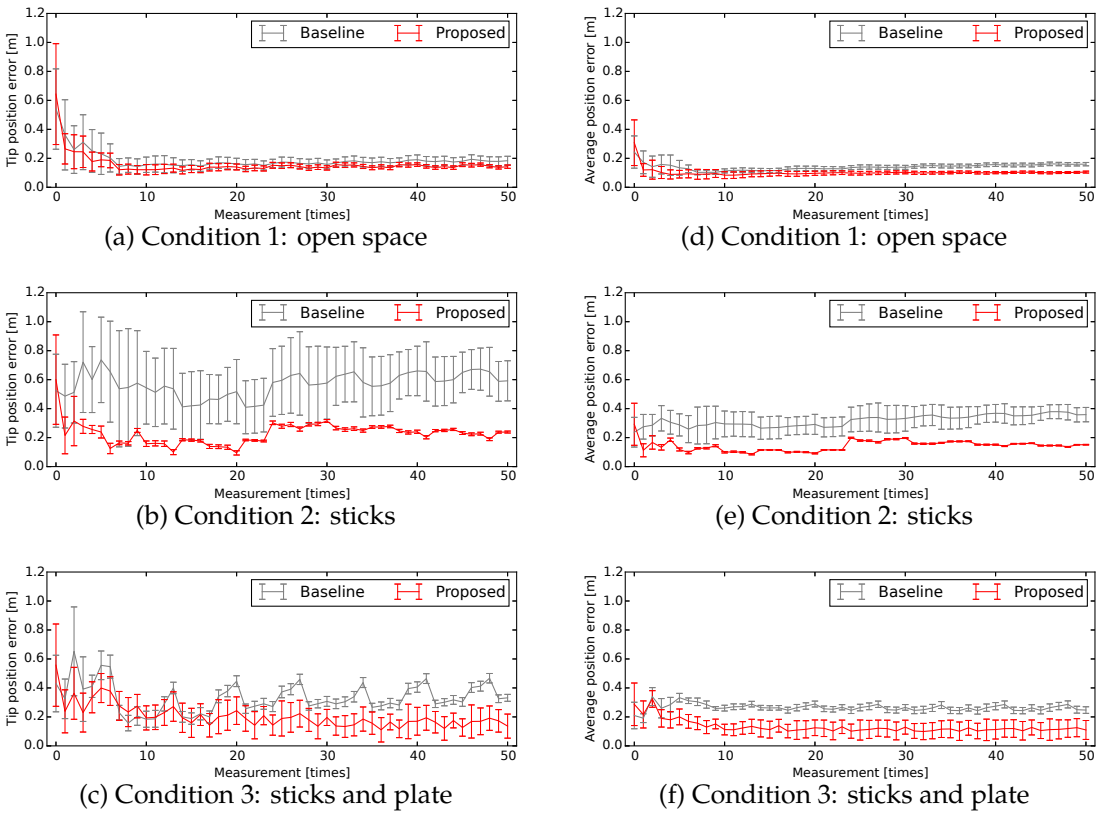
The TSP reference signals used in this experiment had a length of 8,192 samples (512 ms) at 16 kHz. The initial state $\boldsymbol{z}_0 = [\theta_{i,0}, \cdots, \phi_{i,0}, \cdots, l_{i,0}, \cdots]^{\mathrm{T}}$ of the UKF was determined in the following manner. The initial horizontal and vertical link angles $\theta_{i,0}$ and $\phi_{i,0}$ were sampled from a Gaussian distribution whose mean corresponded to the ground-truth posture and standard deviation was 6°. The link lengths $l_{i,0}$ were set to 0.2 m which was the distance between mic-acc and src-vib modules on the robot. The threshold of the TDOA estimation ϵ was set to $0.04/340$ sec. The other parameters were determined experimentally.

The proposed method was implemented by using Python without multiprocessing. The estimation was conducted with a standard laptop computer with an Intel Core i7–3517U CPU (2-core, 1.9 GHz) and 4.0 GB of memory. The CPU time and elapsed time for the whole estimation algorithm with 50 measurements were 8.561 s and 9.129 s, respectively. These values were small enough compared with the whole signal length of the reference signals (25.6 s) that the method could work in real time.

As in Section 5, the tip position error and average estimation error were evaluated The tip position error was the distance between the ground-truth and estimated positions of the tip module (8-th mic-acc module). The average estimation error was the average distance between the ground-truth and estimated positions of all the modules. The ground-truth position of each module was measured using a motion capture system (OptiTrack, NaturalPoint Inc.). The estimation errors were evaluated with 32 different initial states. Since the conventional audio-only method, which does not consider the tilt information, has rotation ambiguity at the x-axis of the 1-st mic-acc module, the estimated posture was rotated to make the average estimation error as small as possible.

6.3.2 Experimental Results

As shown in Figure 6.6, in all conditions, the proposed method suppressed the tip position errors at the initial states to about 0.2 m and suppressed the average position errors there to less than 0.2 m. Moreover, when the robot was placed in rubble-containing environments (conditions 2 and 3), the baseline audio-based



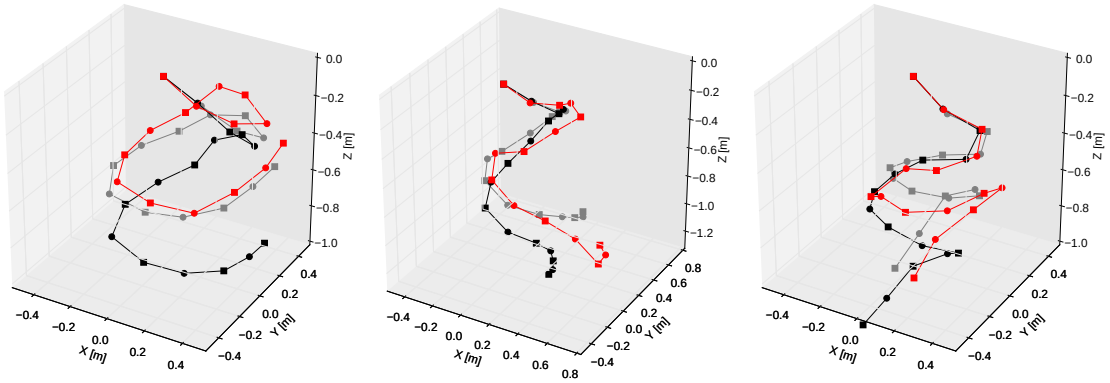
(a) Condition 1: open space
(d) Condition 1: open space
(b) Condition 2: sticks
(e) Condition 2: sticks
(c) Condition 3: sticks and plate
(f) Condition 3: sticks and plate

Figure 6.6: Tip and average position errors obtained by proposed and baseline methods in the three conditions. Polylines and error bars indicate the mean and standard deviation for 32 different initial states, respectively.

method failed to estimate the robot's posture. The proposed method, on the other hand, robustly suppressed the estimation errors

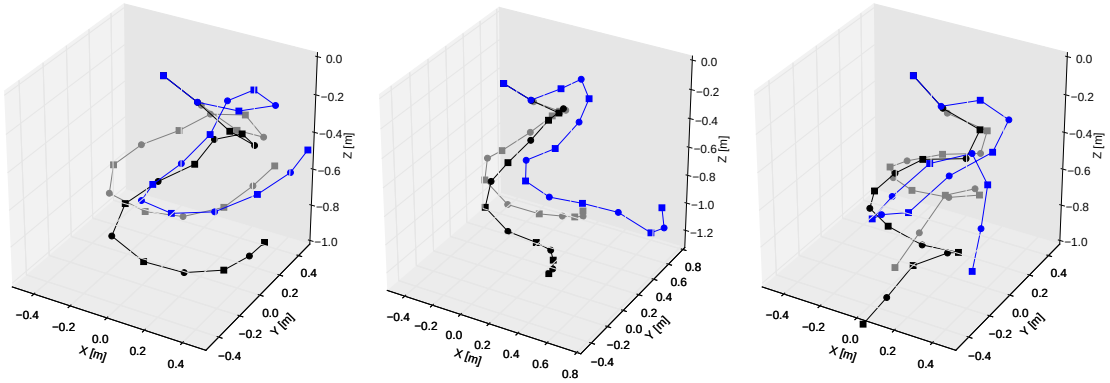
As shown in Figure 6.7, in the all conditions, the postures estimated by the proposed method were close to the ground-truth posture, whereas when the robot was placed in condition 2 or 3, the first joint angle estimated by the conventional method was significantly different from the ground-truth posture. Both of the rubble-containing environments had a wooden stick in front of the joint place (2nd src-vib module) to prevent estimation of the robot posture. This shows that in the proposed method the lack of information at the joint was compensated by the information obtained from the accelerometers.

As shown in Figure 6.8, when the errors of the initial state of the Kalman filter were larger than those in the other conditions, the estimation error became



Condition 1: open space　Condition 2: Sticks　Condition 3: Sticks and plate

(a) Proposed

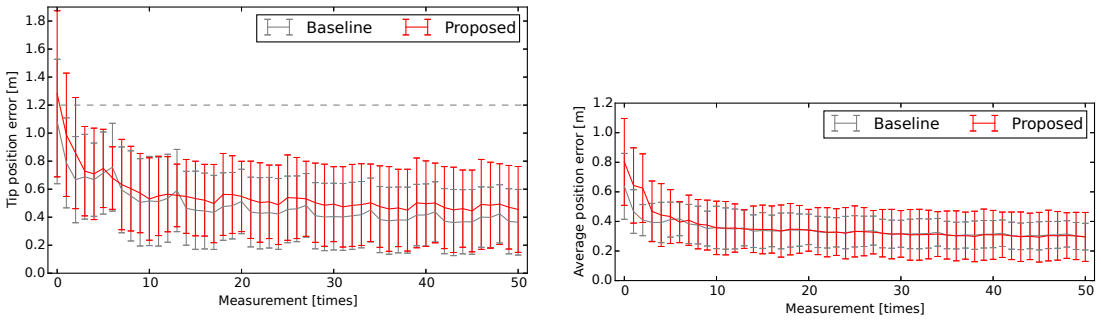


Condition 1: open space　Condition 2: Sticks　Condition 3: Sticks and plate

(b) Baseline

Figure 6.7: Examples of estimated postures at the 50-th measurement. Black and gray lines represent initial and ground-truth postures, respectively.

larger. In this condition the standard deviation of initial errors was set to 30 deg, whereas in the other conditions it was set to 6 deg. This shows that the proposed method is sensitive to the initial state. This is because mirror symmetrical ambiguity could not solved even if both TDOA and tilt information were used. A promising solution to this problem is to predict the time-varying posture of a moving robot in a dynamical manner. Since the posture at the initial insertion is given with a insertion-guide pipe [24], the current posture can be obtained by tracking the time-varying posture during the insertion. It was shown that a sound-based method can track the moving posture by considering the posture change rate in Chapter 5. Integration with sequential information



(a) Tip position error

(b) Average position error

Figure 6.8: Tip and average position errors with larger errors of initial states, in condition 3. The standard deviation of initial errors was set to 30 deg (it was set to 6 deg in the other conditions).

obtained by accelerometers and gyrometers would also be beneficial for further improvement of 3D time-varying posture estimation.

6.4 Summary

This chapter presented a 3D posture estimation method using microphones and accelerometers for a hose-shaped rescue robot. Since correct TDOAs are not always obtained at all microphones if a reference signal is blocked by some obstacles, the proposed method incorporates tilt information obtained by the accelerometers for estimating a robot posture robustly in rubble-containing environments. A nonlinear state-space model was formulated to integrate TDOA and tilt information, and the robot posture was estimated by using the unscented Kalman filter. Experiments using a 3 m hose-shaped robot with eight microphones and accelerometers and seven loudspeakers showed that the method successfully reduced the tip position errors of the initial states to about 0.2 m even when the robot was placed in rubble-containing environments. The future work includes the extension for estimating the 3D time-varying posture.

Chapter 7

Conclusion

This thesis addressed audio scene analysis for rescue robots that work in severely adverse environments. This chapter first reviews the contributions of this thesis, and then presents directions for future research.

7.1 Contributions

This work focused on audio scene analysis for a hose-shaped rescue robot, which is one of the ground rescue robots for penetrating into narrow gaps of collapsed buildings. Sets of microphones, inertial sensors, and loudspeakers on the robot were used for sensing the surrounding environments and the robot itself. This thesis addressed two fundamental functions for the audio scene analysis: speech enhancement and posture estimation that are robust against the dynamic configuration and partial occlusion of microphones.

7.1.1 Speech Enhancement

A speech enhancement method robust against both the dynamic configuration and partial occlusion of microphones was presented in Chapter 3. The proposed method called Bayesian RNTF works on the magnitude spectrogram instead of using phase information, which is sensitively affected by the array layout. By assuming speech and noise spectrograms to be sparse and low-rank, respectively, the Bayesian RNTF works without training data of both the noise and speech. It can also cope with the partial occlusion of microphones by estimating the speech

gain at each microphone. The Bayesian RNTF was extended to a mini-batch method so that the enhancement is conducted in a real-time. The experimental results showed that the noise signals were successfully suppressed regardless of SNR conditions and the layout of microphones and sources although the method often failed to extract speech in the severely low-SNR conditions. The results also showed that the method outperformed conventional multichannel methods even when half of the microphones are occluded.

To improve the enhancement performance, a prior distribution of speech signals based on a deep generative model was introduced in Chapter 4. Instead of using the simple sparse speech model, speech signals were modeled with a variational autoencoder (VAE) that is trained in advance with a sufficient amount of clean speech. The noise signal was modeled by assuming the low-rank structures as in the Bayesian RNTF. These two models were combined into a single probabilistic model called VAE-NMF, and a unified inference algorithm based on a Markov chain Monte Carlo (MCMC) algorithm was derived. The experimental results datasets of urban noise signals showed that VAE-NMF outperformed a method based on low-rank and sparse decomposition. In addition, VAE-NMF outperformed the conventional supervised DNN-based method in unseen noisy environments. It was also experimentally confirmed that the enhancement performance of VAE-NMF for a hose-shaped rescue robot was higher than that of the single-channel low-rank and sparse decomposition method when the SNR was relatively high.

7.1.2 Posture Estimation

An audio-based method that can deal with the dynamic configuration of microphones was presented in Chapter 5. A 2D time-varying posture was tracked by estimating the posture change rate and predicting the current posture. The conventional gyroscope-based method increased the estimation error monotonically over time [56]. The experiments using a 3 m moving hose-shaped robot showed that the audio-based method successfully suppressed the estimation error under 0.2 m. It was also revealed that the purely audio-based method of-

ten confuses mirror-symmetrical postures, depending on the initial value of the estimation. It is experimentally confirmed that the mirror-symmetrical problem of a moving robot is reduced by estimating the posture change rate.

For dealing with the partial occlusion of microphones, the audio-based method was extended to a 3D posture estimation method based on microphones and accelerometers in Chapter 6. The partial occlusion was tackled by integrating the TDOAs obtained from microphones and the tilt angles obtained from accelerometers. The proposed method excludes TDOAs distorted by obstacles and covers the lack of the TDOA measurements with the tilt information. It was experimentally confirmed that the method successfully reduced the tip position errors of the initial states to about 0.2 m even when the robot was placed in rubble-containing environments. When the initial errors of initial states are less than 20 %, the method can estimate the correct 3D posture in real-time.

7.2 Remaining Issues and Future Directions

The author concludes by presenting some directions for future research.

7.2.1 Remaining Issues

The Bayesian RNTF presented in Chapter 3 will be further improved by putting the VAE-based deep prior distribution on speech signals. The speech model with the sparse assumption in the Bayesian RNTF can extract only the speech component clearly appeared in a noisy amplitude spectrogram. The VAE-based speech model can restore the missing speech part overlapped by ego-noise because it represents the joint distribution of a speech signal over frequency bins. Since the source signal models and their mixing system were separately formulated in VAE-NMF, the mixing system can be replaced for the multichannel scenario. The VAE-NMF currently has two drawbacks. One is its low suppression performance of noise signals. The VAE-NMF presented in Chapter 4 outputs parts of noise signals as speech in low-SNR conditions. The VAE-based speech model would be improved by introducing a sparse constraint. The other

is its high computational time due to the Metropolis algorithm. Its inference will be accelerated by using the encoder of a VAE and the Hamiltonian Monte-Carlo (HMC) algorithm [62].

For tracking a 3D time-varying posture, the future work includes the integration of the audio-based and gyroscope-based methods. Although the gyroscope is robust against external environments and has high time resolution, it accumulates its errors as time passes [56]. As presented in Chapters 5 and 6, the integration of microphones and accelerometers can avoid the accumulative error problem even when the robot is in rubble-existing environments or moves. The audio-based method, however, has the mirror-symmetrical problem although it can be reduced by estimating the posture change rate. By integrating these sensors and estimating the drift errors of the gyroscopes, these problems can be complementarily solved.

Evaluation in real or simulated disaster environments is important future work. There are several stations where rescue robots can be evaluated in simulated disaster environments. International Rescue System Institute in Kobe, for example, provides a collapsed house simulation facility [24]. Disaster City in Texas, USA provides many kinds of disaster sites such as piles of rubble and collapsed buildings [27]. It is also important to ask a real rescue team to use the robot in such environments, and evaluate the usability of speech enhancement and posture estimation in rescue missions.

7.2.2 Use of Posterior Estimates

All the methods presented in this thesis provide posterior distributions of the latent variables (e.g., speech signals and posture). The operator of a hose-shaped rescue robot can use this information as reliability of each output. A user interface that effectively visualizes the reliability would be useful for the operator. The posterior estimates can also be used for planning actions automatically [160]. The posture estimation, for example, can reduce the frequency for submitting the reference sound. The loud sounds for posture estimation prevent the operator from searching for victims. By monitoring the posterior distribution of the

posture, the reference sound can be submitted only when the estimated posture becomes unreliable.

7.2.3 Higher-Level Audio Scene Analysis

Speech enhancement and posture estimation addressed in this thesis enable the higher-level audio scene analysis as discussed in Chapter 1. The Bayesian RNTF can be extended to localize a victim by using the estimated posture. Since the Bayesian RNTF calculates a simple distribution of estimated speech gain at each microphone, it would be able to roughly estimate the location of a victim by using the gain differences across microphones. The operator of the robot currently has to manually detect the speech included in the enhancement results. Voice activity detection should be tackled to improve the usability of this system. The robot movements can be predicted and controlled from the current posture information [161]. These techniques will enable the robot to find and approach victims automatically by detecting and localizing the victim's speech sound. It is also important to develop an auditory display that effectively visualizes the audio scene in the complex rubble-containing environments.

7.2.4 Applications for Other Rescue Robots

Application for other rescue robots is another interesting direction for future research. Speech enhancement and posture estimation are important problems not only for a hose-shaped rescue robot. Drone robots, for example, have large and continuous flight noise [20]. Most of ground robots generate ego-noise from their wheels or actuators [34]. The proposed enhancement methods have the portability to adapt various robots because they not only allow the dynamic configuration and partial-occlusion of microphones but also work without the array layout and training data of the noise signals. On the other hand, there are various robots having flexible and soft bodies that have no joints [162–164]. Since the proposed posture model is formulated as a simple link model, it will be easily extended for such robots. The speech enhancement and posture estimation for these robots will also enable the higher-level audio scene analysis for the robots.

Bibliography

[1] R. R. Murphy, *Disaster Robotics*. MIT Press, 2014.

[2] R. R. Murphy, "Navigational and Mission Usability in Rescue Robots," *Journal of the Robotics Society of Japan*, vol. 28, no. 2, pp. 142–146, 2010.

[3] K. Nagatani, S. Kiribayashi, Y. Okada, S. Tadokoro, T. Nishimura, T. Yoshida, E. Koyanagi, and Y. Hada, "Redesign of rescue mobile robot Quince," in *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 13–18, 2011.

[4] K. Ohno, S. Kawatsuma, T. Okada, E. Takeuchi, K. Higashi, and S. Tadokoro, "Robotic control vehicle for measuring radiation in Fukushima Daiichi Nuclear Power Plant," in *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 38–43, 2011.

[5] K. Nagatani, S. Kiribayashi, Y. Okada, K. Otake, K. Yoshida, S. Tadokoro, T. Nishimura, T. Yoshida, E. Koyanagi, M. Fukushima, *et al.*, "Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots," *Journal of Field Robotics*, vol. 30, no. 1, pp. 44–63, 2013.

[6] R. Voyles and G. Jiang, "Hexrotor UAV platform enabling dextrous interaction with structures – preliminary work," in *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR) 2012*, pp. 1–7.

[7] V. Baiocchi, D. Dominici, M. V. Milone, and M. Mormile, "Development of a Software to Plan UAVs Stereoscopic Flight: An Application on Post Earthquake Scenario in L'Aquila Ciy," in *International Conference on Computational Science and Its Applications (ICCSA)*, pp. 150–165, Springer, 2013.

[8] M. Onosato, F. Takemura, K. Nonami, K. Kawabata, K. Miura, and

H. Nakanishi, "Aerial robots for quick information gathering in USAR," in *SICE-ICASE International Joint Conference*, pp. 3435–3438, 2006.

[9] R. R. Murphy, K. S. Pratt, and J. L. Burke, "Crew roles and operational protocols for rotary-wing micro-UAVs in close urban environments," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 73–80, 2008.

[10] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4737–4742, IEEE, 2012.

[11] E. T. Steimle, R. R. Murphy, M. Lindemuth, and M. L. Hall, "Unmanned marine vehicle use at hurricanes wilma and ike," in *OCEANS 2009, MTS/IEEE Biloxi-Marine Technology for Our Future: Global and Local Challenges*, pp. 1–6, IEEE, 2009.

[12] S. Arnold and K. Yamazaki, "Real-time scene parsing by means of a convolutional neural network for mobile robots in disaster scenarios," in *IEEE International Conference on Information and Automation*, pp. 201–207, 2017.

[13] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based slam," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 4, pp. 31–43, 2010.

[14] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[15] D. F. Rosenthal and H. G. Okuno, *Computational auditory scene analysis.* Lawrence Erlbaum Associates Publishers, 1998.

[16] 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, 2011.

[17] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound sepa-

ration on the Texai telepresence system using HARK," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2130–2137, 2011.

[18] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 161–164, 2009.

[19] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1902–1907, 2014.

[20] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3943–3948, 2013.

[21] J. Melo and A. Matos, "Guidance and control of an asv in auv tracking operations," in *OCEANS 2008*, pp. 1–7, IEEE, 2008.

[22] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based Online Calibration for Asynchronous Microphone Array," *Advanced Robotics*, vol. 26, no. 17, pp. 1941–1965, 2012.

[23] A. Kitagawa, H. Tsukagoshi, and M. Igarashi, "Development of Small Diameter Active Hose-II for Search and Life-prolongation of Victims under Debris," *Journal of Robotics and Mechatronics*, vol. 15, no. 5, pp. 474–481, 2003.

[24] K. Hatazaki, M. Konyo, K. Isaki, S. Tadokoro, and F. Takemura, "Active scope camera for urban search and rescue," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2596–2602, 2007.

[25] H. Namari, K. Wakana, M. Ishikura, M. Konyo, and S. Tadokoro, "Tube-type active scope camera with high mobility and practical functionality," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3679–3686, 2012.

[26] J. Fukuda, M. Konyo, E. Takeuchi, and S. Tadokoro, "Remote vertical explo-

ration by active scope camera into collapsed buildings," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1882–1888, IEEE, 2014.

[27] 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," in *IEEE International Workshop on Advanced Robotics and its Social Impacts (ARSO)* , pp. 47–50, 2009.

[28] S. Uemura, O. Sugiyama, R. Kojima, and K. Nakadai, "Outdoor acoustic event identification using sound source separation and deep learning with a quadrotor-embedded microphone array," in *The Abstracts of the international conference on advanced mechatronics: toward evolutionary fusion of IT and mechatronics: ICAM 2015.6*, pp. 329–330, The Japan Society of Mechanical Engineers, 2015.

[29] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[30] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[31] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, no. EPFL-CONF-192584, 2011.

[32] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *EUROSPEECH*, pp. 1691–1694, 2001.

[33] N. Cho and E.-K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, 2011.

[34] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *IEEE International*

Conference on Robotics and Automation (ICRA), pp. 3517–3522, 2011.

[35] B. Cauchi, S. Goetze, and S. Doclo, "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization," in *Workshop on Speech and Multimodal Interaction in Assistive Environments. Association for Computational Linguistics*, pp. 28–33, 2012.

[36] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*. PTR Prentice Hall Englewood Cliffs, 1993.

[37] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 189–192, 2011.

[38] T. Tezuka, T. Yoshida, and K. Nakadai, "Ego-motion noise suppression for robots based on semi-blind infite non-negative matrix factorization," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6293–6298, 2014.

[39] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors Journal*, vol. 17, no. 8, pp. 2447–2455, 2017.

[40] S. Araki, M. Okada, T. Higuchi, A. Ogawa, and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 385–389, 2016.

[41] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 2, pp. 1228–1233, 2003.

[42] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 380–385, 2006.

[43] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, 1997.

[44] K. Audenaert, H. Peremans, Y. Kawahara, and J. Van Campenhout, "Accurate ranging of multiple objects using ultrasonic sensors," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1733–1738, 1992.

[45] L. Kleeman and R. Kuc, "Mobile robot sonar for target localization and classification," *The International Journal of Robotics Research*, vol. 14, no. 4, pp. 295–318, 1995.

[46] C. Evers, A. H. Moore, and P. A. Naylor, "Acoustic simultaneous localization and mapping (a-slam) of a moving microphone array and its surrounding speakers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6–10, 2016.

[47] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano, "Active audition system and humanoid exterior design," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 2, pp. 1453–1461, 2000.

[48] T. Iyama, O. Sugiyama, T. Otsuka, K. Itoyama, and H. G. Okuno, "Visualization of auditory awareness based on sound source positions estimated by depth sensor and microphone array," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1908–1913, 2014.

[49] E. Vincent, A. Sini, and F. Charpillet, "Audio source localization by optimal control of a mobile robot," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5630–5634, IEEE, 2015.

[50] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5610–5614, 2015.

[51] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech*, pp. 436–440, 2013.

[52] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 196–200, 2016.

[53] T. Kim, "Real-time independent vector analysis for convolutive blind

source separation," *IEEE Transactions on Circuits and Systems I*, vol. 57, no. 7, pp. 1431–1438, 2010.

[54] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, no. 3, pp. 550–563, 2010.

[55] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 276–280, 2015.

[56] M. Ishikura, E. Takeuchi, M. Konyo, and S. Tadokoro, "Shape estimation of flexible cable," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2539–2546, 2012.

[57] J. Lee, G. Ukawa, S. Doho, Z. Lin, H. Ishii, M. Zecca, and A. Takanishi, "Non visual sensor based shape perception method for gait control of flexible colonoscopy robot," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 577–582, 2011.

[58] D. Simon and D. L. Simon, "Constrained kalman filtering via density function truncation for turbofan engine health estimation," *International Journal of Systems Science*, vol. 41, no. 2, pp. 159–171, 2010.

[59] D. Su, T. Vidal-Calleja, and J. V. Miro, "Simultaneous asynchronous microphone array calibration and sound source localisation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5561–5567, 2015.

[60] M. Hennecke, T. Plotz, G. A. Fink, J. Schmalenstroer, and R. Hab-Umbach, "A hierarchical approach to unsupervised shape calibration of microphone array networks," in *IEEE/SP 15th Workshop on Statistical Signal Processing*, pp. 257–260, 2009.

[61] S. Thrun, "Affine structure from sound," in *Neural Information Processing Systems (NIPS)*, pp. 1353–1360, 2006.

[62] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[63] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems (NIPS)*, pp. 2672–2680, 2014.

[64] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[65] O. Fabius and J. R. van Amersfoort, "Variational recurrent auto-encoders," *arXiv preprint arXiv:1412.6581*, 2014.

[66] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Interspeech*, pp. 1273–1277, 2017.

[67] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders," in *Interspeech*, pp. 1770–1774, 2016.

[68] M. Ishikura, E. Takeuchi, M. Konyo, and S. Tadokoro, "Vision-based localization using active scope camera—accuracy evaluation for structure from motion in disaster environment," in *IEEE/SICE International Symposium on System Integration (SII)*, pp. 25–30, 2010.

[69] S. Weiss *et al.*, "Monocular-SLAM–based navigation for autonomous micro helicopters in GPS-denied environments," *Journal of Field Robotics*, vol. 28, no. 6, pp. 854–874, 2011.

[70] K. Schmid *et al.*, "Stereo vision based indoor/outdoor navigation for flying robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3955–3962, 2013.

[71] M. Basiri, F. Schill, D. Floreano, and P. Lima, "Audio-based relative positioning system for multiple micro air vehicle systems," in *Robotics: Science and Systems RSS2013*, no. EPFL-CONF-191181, 2013.

[72] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-board relative bearing estimation for teams of drones using sound," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 820–827, 2016.

[73] D. A. Gray, B. D. Anderson, and R. R. Bitmead, "Towed array shape estimation using kalman filters-theoretical models," *IEEE journal of oceanic engineering*, vol. 18, no. 4, pp. 543–556, 1993.

[74] J. L. Odom and J. L. Krolik, "Passive towed array shape estimation using

heading and acoustic data," *IEEE Journal of Oceanic Engineering*, vol. 40, no. 2, pp. 465–474, 2015.

[75] K. Chuengsatiansup, K. Sajjapongse, P. Kruapraditsiri, C. Chanma, N. Termthanasombat, Y. Suttasupa, S. Sattaratnamai, E. Pongkaew, P. Udsatid, B. Hattha, *et al.*, "Plasma-rx: Autonomous rescue robots," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1986–1990, 2009.

[76] H. Sun, P. Yang, L. Zu, and Q. Xu, "A far field sound source localization system for rescue robot," in *International Conference on Control, Automation and Systems Engineering (CASE)*, pp. 1–4, 2011.

[77] M. W. Kadous, R. K.-M. Sheh, and C. Sammut, "Effective user interface design for rescue robotics," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 250–257, ACM, 2006.

[78] J. Suthakorn, S. S. H. Shah, S. Jantarajit, W. Onprasert, W. Saensupo, S. Saeung, S. Nakdhamabhorn, V. Sa-Ing, and S. Reaungamornrat, "On the design and development of a rough terrain robot for rescue missions," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1830–1835, 2009.

[79] J. Gao, X. Gao, J. Zhu, W. Zhu, B. Wei, and S. Wang, "Coal mine detect and rescue robot technique research," in *IEEE International Conference on Information and Automation*, pp. 1068–1073, 2009.

[80] M. W. Kadous, R. K.-M. Sheh, and C. Sammut, "Caster: A robot for urban search and rescue," in *Proceedings of the 2005 Australasian Conference on Robotics and Automation*, pp. 1–10, 2005.

[81] S. H. Young and M. V. Scanlon, "Detection and localization with an acoustic array on a small robotic platform in urban environments," Tech. Rep. ARL-TR-2575, ARMY RESEARCH LAB ADELPHI MD, 2003.

[82] H. Sun, P. Yang, Z. Liu, L. Zu, and Q. Xu, "Microphone array based auditory localization for rescue robot," in *IEEE Chinese Control and Decision Conference (CCDC)*, pp. 606–609, 2011.

[83] S. T. Roweis, "One microphone source separation," in *Neural Information*

Processing Systems (NIPS), pp. 793–799, 2001.

[84] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[85] S. Mohammed and I. Tashev, "A statistical approach to semi-supervised speech enhancement with low-order non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 546–550, 2017.

[86] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.

[87] C. Sun, Q. Zhang, J. Wang, and J. Xie, "Noise reduction based on robust principal component analysis," *Journal of Computational Information Systems*, vol. 10, no. 10, pp. 4403–4410, 2014.

[88] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, p. 11, 2011.

[89] Z. Chen and D. P. Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4, 2013.

[90] N. Dobigeon and C. Févotte, "Robust nonnegative matrix factorization for nonlinear unmixing of hyperspectral images," in *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–4, 2013.

[91] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 7, pp. 1233–1242, 2015.

[92] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3419–3430, 2011.

[93] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 3964–3977, 2012.

[94] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[95] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[96] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 504–511, 2015.

[97] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 22, no. 3, pp. 727–739, 2014.

[98] N. Mae, M. Ishimura, S. Makino, D. Kitamura, N. Ono, T. Yamada, and H. Saruwatari, "Ego noise reduction for hose-shaped rescue robot combining independent low-rank matrix analysis and multichannel noise cancellation," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 141–151, 2017.

[99] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2015.

[100] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound source separation using shifted non-negative tensor factorisation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. V, pp. 653–656, 2006.

[101] N. Murata, H. Kameoka, K. Kinoshita, S. Araki, T. Nakatani, S. Koyama,

and H. Saruwatari, "Reverberation-robust underdetermined source separation with non-negative tensor double deconvolution," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, pp. 1648–1652, 2016.

[102] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues," in *International Symposium on Computer Music Modeling and Retrieval*, pp. 102–115, 2010.

[103] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 203–207, 2014.

[104] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 9, pp. 1652–1664, 2016.

[105] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7092–7096, 2013.

[106] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *Interspeech*, pp. 3642–3646, 2017.

[107] Z.-Q. Wang and D. Wang, "Recurrent deep stacking networks for supervised speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 71–75, 2017.

[108] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.

[109] T. T. Vu, B. Bigot, and E. S. Chng, "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 499–503, 2016.

[110] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian wavenet," *Interspeech*, pp. 2013–2017, 2017.

[111] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3734–3738, 2014.

[112] M. Sun, X. Zhang, and T. F. Zheng, "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 1, pp. 93–104, 2016.

[113] Y. Saito, S. Takamichi, H. Saruwatari, Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. PP, no. 99, pp. 1–1, 2017.

[114] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 11, pp. 2032–2045, 2016.

[115] S. Tully, G. Kantor, and H. Choset, "Inequality constrained kalman filtering for the localization and registration of a surgical robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5147–5152, 2011.

[116] J. J. Smith, Y. H. Leung, and A. Cantoni, "The cramér-rao lower bound for towed array shape estimation with a single source," *IEEE Transactions on signal processing*, vol. 44, no. 4, pp. 1033–1036, 1996.

[117] Y. Rockah and P. Schultheiss, "Array shape calibration using sources in unknown locations–part i: Far-field sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 3, pp. 286–299, 1987.

[118] J. C. Gower, "Euclidean distance geometry," *Math. Sci*, vol. 7, no. 1, pp. 1–14, 1982.

[119] M. Chen, Z. Liu, L.-W. He, P. Chou, and Z. Zhang, "Energy-based position

estimation of microphones and speakers for ad hoc microphone arrays," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 22–25, IEEE, 2007.

[120] Y. Kuang and K. Astrom, "Stratified sensor network self-calibration from tdoa measurements," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2013.

[121] T.-K. Le and N. Ono, "Closed-form and near closed-form solutions for tdoa-based joint source and sensor localization," *IEEE Transactions on Signal Processing*, vol. 65, no. 5, pp. 1207–1221, 2017.

[122] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing (TASSP)*, vol. 24, no. 4, pp. 320–327, 1976.

[123] Y. Bando, T. Mizumoto, K. Itoyama, K. Nakadai, and H. G. Okuno, "Posture estimation of hose-shaped robot using microphone array localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3446–3451, 2013.

[124] S. A. Raczyński, Ł. Grzymkowski, and K. Główczewski, "Distributed mobile microphone arrays for robot navigation and acoustic source localization," in *International Conference on Control Automation Robotics & Vision (ICARCV)*, pp. 1039–1044, 2014.

[125] M. Parviainen and P. Pertilä, "Self-localization of dynamic user-worn microphones from observed speech," *Applied Acoustics*, vol. 117, pp. 76–85, 2017.

[126] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[127] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 10, pp. 2140–2151, 2013.

[128] Y. Li *et al.*, "Speech enhancement based on robust NMF solved by alternating direction method of multipliers," in *IEEE MMSP*, pp. 1–5, 2015.

[129] G. Min, X. Zhang, X. Zou, and M. Sun, "Mask estimate through Itakura-Saito nonnegative RPCA for speech enhancement," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2016.

[130] C. Févotte and N. Dobigeon, "Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4810–4819, 2015.

[131] A. Deleforge and W. Kellermann, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 355–359, 2015.

[132] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP*, vol. 61, no. 6, 2009.

[133] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Neural Information Processing Systems (NIPS)*, pp. 2080–2088, 2009.

[134] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational intelligence and neuroscience*, vol. 2009, no. 785152, pp. 1–17, 2009.

[135] M. D. Hoffman, "Poisson-uniform nonnegative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5361–5364, 2012.

[136] D. Gamerman, T. R. Santos, and G. C. Franco, "A non-gaussian family of state-space models with exact marginal likelihood," *Journal of Time Series Analysis*, vol. 34, no. 6, pp. 625–645, 2013.

[137] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese

large vocabulary continuous speech recognition corpus," in *International Conference on Spoken Language Processing*, 1998.

[138] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 14, no. 4, pp. 1462–1469, 2006.

[139] E. Contal, V. Perchet, and N. Vayatis, "Gaussian process optimization with mutual information," in *International Conference on Machine Learning*, pp. 253–261, 2014.

[140] 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," in *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 1–6, 2015.

[141] H. Nakajima, G. Ince, K. Nakadai, and Y. Hasegawa, "An easily-configurable robot audition system using histogram-based recursive level estimation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 958–963, 2010.

[142] D. Liang and M. D. Hoffman, "Beta process non-negative matrix factorization with stochastic structured mean-field variational inference," *arXiv preprint arXiv:1411.1804*, 2014.

[143] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan, "Streaming variational Bayes," in *Neural Information Processing Systems (NIPS)*, pp. 1727–1735, 2013.

[144] A. T. Cemgil and O. Dikmen, "Conjugate gamma markov random fields for modelling nonstationary sources," in *ICA*, pp. 697–705, 2007.

[145] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[146] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 57–60, 2012.

[147] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[148] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[149] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Neural Information Processing Systems (NIPS)*, pp. 2980–2988, 2015.

[150] Y. Kim, Y. Kim, C. Lee, and S. Kwon, "Thin polysilicon gauge for strain measurement of structural elements," *IEEE Sensors Journal*, vol. 10, no. 8, pp. 1320–1327, 2010.

[151] E. A. Wan *et al.*, "The unscented kalman filter for nonlinear estimation," in *The IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pp. 153–158, 2000.

[152] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[153] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *The Journal of the Acoustical Society of America*, vol. 97, p. 1119, 1995.

[154] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2565–2568, 2008.

[155] R. van der Merwe *et al.*, "The unscented particle filter," in *Neural Information Processing Systems (NIPS)*, pp. 584–590, 2000.

[156] S. Lynen *et al.*, "A robust and modular multi-sensor fusion approach applied to MAV navigation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3923–3929, 2013.

[157] M. Tailanian *et al.*, "Design and implementation of sensor data fusion for an autonomous quadrotor," in *IEEE International Instrumentation and Measurement Technology Conference*, pp. 1431–1436, 2014.

[158] J. M. Santos *et al.*, "A sensor fusion layer to cope with reduced visibility in SLAM," *Journal of Intelligent & Robotic Systems*, pp. 1–22, 2015.

[159] Y. Bando *et al.*, "Posture estimation of hose-shaped robot by using active

microphone array," *Advanced Robotics*, vol. 29, no. 1, pp. 35–49, 2015.

[160] S. Thrun, *Probabilistic robotics*. MIT Press, 2005.

[161] K. Sawata, M. Konyo, S. Saga, S. Tadokoro, and K. Osuka, "Sliding motion control of active flexible cable using simple shape information," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pp. 3736–3742, IEEE, 2009.

[162] C. Laschi, M. Cianchetti, B. Mazzolai, L. Margheri, M. Follador, and P. Dario, "Soft robot arm inspired by the octopus," *Advanced Robotics*, vol. 26, no. 7, pp. 709–727, 2012.

[163] M. T. Tolley, R. F. Shepherd, B. Mosadegh, K. C. Galloway, M. Wehner, M. Karpelson, R. J. Wood, and G. M. Whitesides, "A resilient, untethered soft robot," *Soft Robotics*, vol. 1, no. 3, pp. 213–223, 2014.

[164] E. W. Hawkes, L. H. Blumenschein, J. D. Greer, and A. M. Okamura, "A soft robot that navigates its environment through growth," *Science Robotics*, vol. 2, no. 8, p. eaan3028, 2017.

List of Publications

Refereed international journal paper

1) **Yoshiaki Bando**, Katsutoshi Itoyama, Masashi Konyo, Satoshi Tadokoro, Kazuhiro Nakadai, Kazuyoshi Yoshii, Tatsuya Kawahara, and Hiroshi G. Okuno: Speech Enhancement Based on Bayesian Low-Rank and Sparse Decomposition of Multichannel Magnitude Spectrogram, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp.215-230 (2018) → **Chapter 3**

2) **Yoshiaki Bando**, Hiroshi Saruwatari, Nobutaka Ono, Shoji Makino, Katustoshi Itoyama, Daichi Kitamura, Masaru Ishimura,Moe Takakusaki, Narumi Mae, Kouei Yamaoka, Yutaro Matsui, Yuichi Ambe, Masashi Konyo, Satoshi Tadokoro, Kazuyoshi Yoshii, and Hiroshi G. Okuno: Low-Latency and High-Quality Two-Stage Human-Voice-Enhancement System for a Hose-Shaped Rescue Robot, *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp.198–212 (2017)

3) **Yoshiaki Bando**, Takuma Otsuka, Takeshi Mizumoto, Katsutoshi Itoyama, Masashi Konyo, Satoshi Tadokoro, Kazuhiro Nakadai, and Hiroshi G. Okuno: Posture estimation of hose-shaped robot by using active microphoe array, *Advanced Robotics*, vol. 29, no. 1, pp. 35–49 (2015)

Refereed international conference papers

4) **Yoshiaki Bando**, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, Tatsuya Kawahara: Statistical Speech Enhancement Based on Probabilistic

Integration of Variational Autoencoder and Non-Negative Matrix Factorization, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, to appear→ **Chapter 4**.

5) **Yoshiaki Bando**, Hiroki Suhara, Motoyasu Tanaka, Tetsushi Kamegawa, Katsutoshi Itoyama, Kazuyoshi Yoshii, Fumitoshi Matsuno, Hiroshi G. Okuno: Sound-based Online Localization for an In-pipe Snake Robot, *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR*), pp.207–213, 2016.

6) **Yoshiaki Bando**, Katsutoshi Itoyama, Masashi Konyo, Satoshi Tadokoro, Kazuhiro Nakadai, Kazuyoshi Yoshii, Hiroshi G. Okuno: Variational Bayesian Multi-channel Robust NMF for Human-voice Enhancement with a Deformable and Partially-occluded Microphone Array, *European Signal Processing Conference (EUSIPCO)*, pp. 1018–1022, 2016 → **Chapter 3**.

7) **Yoshiaki Bando**, Katsutoshi Itoyama, Masashi Konyo, Satoshi Tadokoro, Kazuhiro Nakadai, Kazuyoshi Yoshii, Hiroshi G. Okuno: Human-Voice Enhancement based on Online RPCA for a Hose-shaped Rescue Robot with a Microphone Array, *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp.1–6, 2015.

8) **Yoshiaki Bando**, Katsutoshi Itoyama, Masashi Konyo, Satoshi Tadokoro, Kazuhiro Nakadai, Kazuyoshi Yoshii, Hiroshi G. Okuno: Microphone-accelerometer based 3D posture estimation for a hose-shaped rescue robot, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.5580–5586, 2015 → **Chapter 6**.

9) **Yoshiaki Bando**, Takuma Otsuka, Katsutoshi Itoyama, Kazuyoshi Yoshii, Yoko Sasaki, Satoshi Kagami, Hiroshi G. Okuno: Challenges in deploying a microphone array to localize and separate sound sources in real auditory scenes, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.723–727, 2015.

10) **Yoshiaki Bando**, Takuma Otsuka, Ikkyu Aihara, Hiromitsu Awano, Katsu-

toshi Itoyama, Kazuyoshi Yoshii, Hiroshi G. Okuno: Recognition of in-field frog chorusing using Bayesian nonparametric microphone array processing, *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-WS)*, pp.2–6, 2015 .

11) **Yoshiaki Bando**, Katsutoshi Itoyama, Masashi Konyo, Satoshi Tadokoro, Kazuhiro Nakadai, Kazuyoshi Yoshii, Hiroshi G. Okuno: A sound-based online method for estimating the time-varying posture of a hose-shaped robot, *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp.1-6, 2014 → **Chapter 5**.

12) **Yoshiaki Bando**, Takeshi Mizumoto, Katsutoshi Itoyama, Kazuhiro Nakadai, Hiroshi G. Okuno: Posture estimation of hose-shaped robot using microphone array localization, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.3446-3451, 2013.