人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-049

ＡＩチャレンジ研究会 (第４９回)

Proceedings of the 49th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　**2017** 年 **11** 月 **25** 日　　場　所　慶應義塾大学 矢上キャンパス **12** 棟　**108** 室
Keio University, Kanagawa, Nov. 25, 2017

社団法人 人工知能学会
Japanese Society for Artificial Intelligence

一般社団法人 人工知能学会
Japanese Society for Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-1 (11/25)

Non-Line-of-Sight Sound Localization in Unknown Indoor Environments

Tomonari FURUKAWA
Department of Mechanical Engineering
Virginia Tech
tomonari@vt.edu

Abstract

This paper presents a new approach that localizes Non-Line-of-Sight (NLOS) targets in unknown indoor environments. Sensors used in the approach are an auditory sensor, an visual sensor and sensors for Simultaneous Localization and Mapping (SLAM). The visual sensor localizes a target when the target is on its Line-of-Sight (LOS), but it is also used to recognize the environment and localize itself in conjunction with the SLAM sensors. The auditory sensor localizes the target even if the target is not on the LOS of the visual sensor. In order to localize a NLOS target in an unknown environment, the proposed approach extracts and analyzes the first-arrival diffraction signal and the first-arrival reflection signal. The estimation is performed within the Recursive Bayesian Estimation (RBE) framework where observations of the visual and auditory sensors are each converted into an observation likelihood. The ability of the proposed approach was experimentally validated in a controlled indoor environment.

1 Introduction

Indoor environments where humans stay and work are typically so complex with many structures or obstructions that the Line-of-Sight (LOS) region is significantly limited. Here is an example conversation that you could have when you are in such a Non-Line-of-Sight (NLOS) environment:

A: "Where are you?".
B: "I am here".
A: "I am getting close. Where are you?".
B: "I am here".
A: "I found you!".

If the target person is communicative, humans search for and find the target person who is not in the Field-of-View (FOV) by estimating the location of the sound. Audition is used not only as a means of communication but also as a sensor for target localization, and is as important as vision due to such multi-functional capabilities. Robotic audition, if the NLOS localization is made possible, becomes a useful tool for both the co-robots and people who are blind.

Past work on the localization of a NLOS target has been conducted in three different ways. The first approach, forming a Wireless Sensor Network (WSN), localizes the target by measuring the intensity of the transmitted signal at each wireless receiver and fusing the measurements of all the receivers under the LOS assumption [2, 4, 17, 27, 13, 7, 8]. Radio signals are commonly used since sound signals reflect excessively and do not create unique signal intensity. The approach is easy to install, but the accuracy depends on the validity of the LOS assumption [3, 18, 20, 10].

In the second approach, Time-of-Arrival (TOA) information of the received signal is used for target localization. The approach most commonly utilizes acoustic signals due to their slow speed compared to that of radio signals. The majority of sound localization challenges have been however focused on the direction of sound rather than its position due to complexity of sound wave propagation [25, 21]. For a NLOS target, Mak and Furukawa [14] located it by using the diffraction characteristics of low-frequency sound though the time of sound generation, which is often unknown, must be informed beforehand.

The last approach enhances the NLOS target localization numerically [15, 5, 6]. The approach localizes the target by updating and maintaining its probabilistic belief in the framework of Recursive Bayesian Estimation (RBE) and processing observations as likelihoods. In the use of an optical sensor, the event of "no detection" is converted into an observation likelihood describing no probability that the target exists. While the no-detection information is still useful, the belief, however, becomes highly unreliable unless the target is rediscovered within a short period after being lost. Takami et al. [24, 23] incorporated a stereo auditory sensor such that the NLOS target can be detected using positive in-

formation accordingly. Their approach however needs to collect acoustic cues of the environment in advance. It is not applicable if the environment is unknown.

This paper presents an extensive approach that localizes NLOS targets in unknown indoor environments. Sensors to be implemented in the proposed approach are an auditory sensor, an visual sensor and sensors for Simultaneous Localization and Mapping (SLAM). The visual sensor accurately localizes a target when the target is on its LOS, but it is also used to recognize the environment and localize itself in conjunction with the SLAM sensors. The auditory sensor localizes the target even if the target is not on the LOS of the visual sensor. In order to localize a NLOS target in an unknown environment, the first-arrival diffraction signal and the first-arrival reflection signal are extracted and analyzed. The estimation is performed within the RBE framework where observations of the visual and auditory sensors are each converted into an observation likelihood.

The paper is organized as follows. The following section presents the proposed approach that uses the visual and the auditory sensors and localizes NLOS targets in unknown indoor environments. Section 3 presents the proposed extraction of the first-arrival diffraction and reflection signals and the construction of the joint auditory observation likelihood. While its application as an assistive and training device for people who are blind or visually-impared is described in Section 4, Section 5 shows the results of the preliminary experimental study and demonstrates the efficacy of the proposed approach. Conclusions are summarized in the final section.

2 Hybrid Visual/Auditory RBE for Unknown Environments

2.1 Overview

Figure 1 shows a schematic diagram of the hybrid visual/auditory approach proposed in this paper. This is to localize a NLOS target in an unknown environment. Sensors used are an auditory sensor, a visual sensor and SLAM sensors. The auditory sensor is a microphone array whereas the visual sensor is a RGB-D sensor, which measures not only RGB information but also depth information and recognizes three-dimensional (3D) surroundings on the LOS. SLAM sensors are those to enable SLAM and include an Inertial Measurement Unit (IMU).

The proposed approach operates as follows. Similarly to the past work of the author, it deploys the framework of the grid-based RBE and estimates a sound target in terms of a non-Gaussian belief [5, 6, 11]. Since environments are assumed to be unknown, the proposed approach incorporates SLAM where the RGB-D sensor is utilized in addition to the SLAM sensors. The self-location is thus known, so only the target pose is estimated in the RBE. The RGB-D sensor detects and locates a sound target if it is on the LOS or in the FOV. Otherwise, a sound target is processed as "not detected", and the RGB-D image is used primarily to recognize 3D surroundings including geometry and material which influence NLOS sound propagation. The microphone array is the sensor that locates a NLOS target in addition to a LOS target. Based on the visual information on the surroundings, a NLOS target is identified by extracting and analyzing the first-arrival diffraction signal and the first-arrival reflection signal. The proposed approach allows localization in unknown environments because it is sound physics based and does not thus require spatial information. Fusion of the visual and auditory observation likelihoods results in a joint likelihood, which updates belief in the glsrbe.

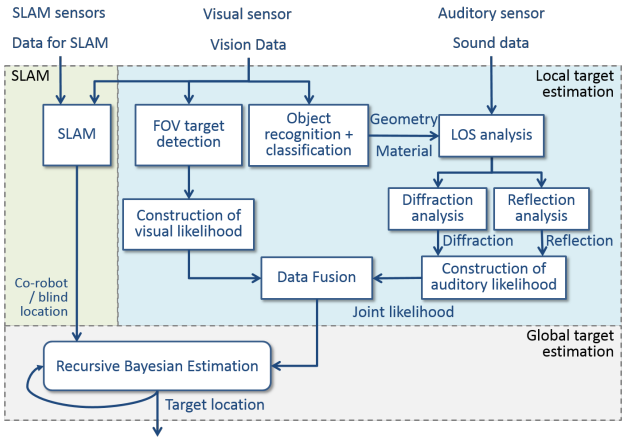


Figure 1: Schematic diagram of the proposed approach

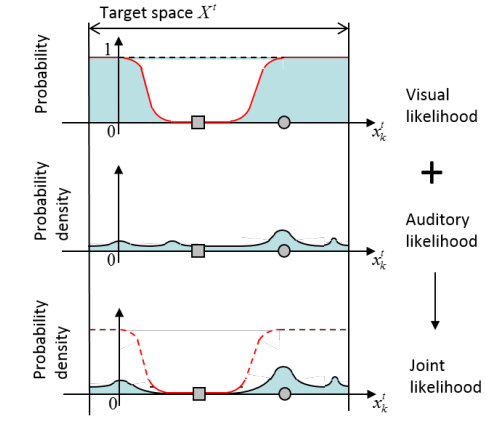
2.2 Mathematical Formulation

The mathematical framework of the hybrid visual/auditory RBE is as follows. Let the state of the robot s and the map updated by SLAM at time step $k-1$ be $\bar{\mathbf{x}}^s_{k-1} \in \mathcal{X}^s$ and $\bar{\mathbf{m}}_{k-1} \in \mathcal{M}$ respectively. Given a sequence of observations by the robot s from time step 1 to time step $k-1$, ${}^s\tilde{\mathbf{z}}^t_{1:k-1} \equiv \{{}^s\tilde{\mathbf{z}}^t_\kappa | \forall \kappa \in \{1, ..., k\}\}$, the RBE iteratively updates the belief on the state of a target t, $\mathbf{x}^t_k \in \mathcal{X}^t$, in time and observation. Let the belief given the sequence of observations and the robot state and the map at time step $k-1$ be $p\left(\mathbf{x}^t_{k-1} | {}^s\tilde{\mathbf{z}}^t_{1:k-1}, \bar{\mathbf{x}}^s_{k-1}, \bar{\mathbf{m}}_{k-1}\right)$. Chapman-Kolmogrov equation updates the prior belief in time, or predicts the belief at time step k, by the probabilistic motion model $p\left(\mathbf{x}^t_k | \mathbf{x}^t_{k-1}, \bar{\mathbf{x}}^s_{k-1}, \bar{\mathbf{m}}_{k-1}\right)$:

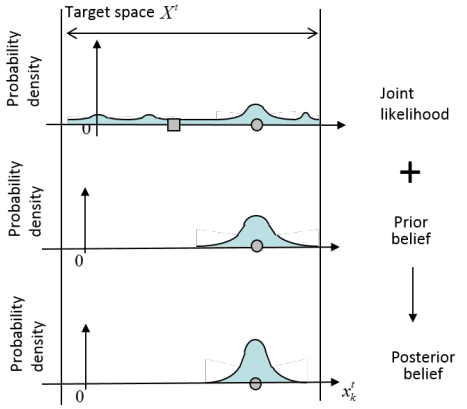
$$p\left(\mathbf{x}^t_k | {}^s\tilde{\mathbf{z}}^t_{1:k-1}, \bar{\mathbf{x}}^s_{k-1}, \bar{\mathbf{m}}_{k-1}\right) = \int_{\mathcal{X}^t} p\left(\mathbf{x}^t_k | \mathbf{x}^t_{k-1}, \bar{\mathbf{x}}^s_{k-1}, \bar{\mathbf{m}}_{k-1}\right) p\left(\mathbf{x}^t_{k-1} | {}^s\tilde{\mathbf{z}}^t_{1:k-1}, \bar{\mathbf{x}}^s_{k-1}, \bar{\mathbf{m}}_{k-1}\right) d\mathbf{x}^t_{k-1}. \quad (1)$$

The observation update, or the correction process, is performed using the Bayes theorem. The target belief is corrected using the new observation ${}^s\tilde{\mathbf{z}}^t_k$ as

$$p\left(\mathbf{x}^t_k | {}^s\tilde{\mathbf{z}}^t_{1:k}, \bar{\mathbf{x}}^s_k, \bar{\mathbf{m}}_k\right) = \frac{q\left(\mathbf{x}^t_k | {}^s\tilde{\mathbf{z}}^t_{1:k}, \bar{\mathbf{x}}^s_{k-1:k}, \bar{\mathbf{m}}_{k-1:k}\right)}{\int_{\mathcal{X}^t} q\left(\mathbf{x}^t_k | {}^s\tilde{\mathbf{z}}^t_{1:k}, \bar{\mathbf{x}}^s_{k-1:k}, \bar{\mathbf{m}}_{k-1:k}\right) d\mathbf{x}^t_k}, \quad (2)$$



(a) Joint visual/acoustic observation likelihood



(b) RBE incorporating an acoustic sensor

Figure 2: Hybrid visual/auditory target estimation

where $q(\cdot) = l\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) p\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_{1:k-1}^t, \bar{\mathbf{x}}_{k-1}^s, \bar{\mathbf{m}}_{k-1}\right)$, and $l\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right)$ represents the observation likelihood of $\mathbf{x}_k^t$ given ${}^s\tilde{\mathbf{z}}_k^t$, $\bar{\mathbf{x}}_k^s$ and $\bar{\mathbf{m}}_k$.

One of the core technologies proposed in this paper is the hybrid use of visual and auditory sensors. This is given by

$$l\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) = l^c\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) l^a\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) \tag{3}$$

where $l^c(\cdot)$ and $l^a(\cdot)$ are the likelihoods of the visual sensor (RGB-D camera) and the auditory sensor (microphone array). In order to maximize information, the camera observation is used not only to detect a target if it is in the FOV but also to construct the no-detection likelihood if the target is outside the FOV:

$$l^c\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) = \begin{cases} p\left({}^s\tilde{\mathbf{z}}_k^t | \mathbf{x}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) & \exists {}^s\tilde{\mathbf{z}}_k^t \in {}^s\mathcal{X}_d^t \\ 1 - P_d\left(\mathbf{x}_k^t | \bar{\mathbf{x}}_k^s\right) & \nexists {}^s\tilde{\mathbf{z}}_k^t \in {}^s\mathcal{X}_d^t, \end{cases} \tag{4}$$

where ${}^s\mathcal{X}_d^t$ is the camera FOV or, more precisely, the detectable region. The effectiveness of Equation (4) is thoroughly investigated by the author in the context of autonomous search and tracking.

While the derivation of $l^a(\cdot)$ is most challenging and thus will be dealt with separately in the next section, the advantage of Equation (3) in RBE is illustratively shown in Figure 2. The possible locations of the target are narrowed down even though the no-detection likelihood is used in visual sensing since the likelihood clears out the joint likelihood in the FOV and dropped some peak(s) as shown in Figure 2(a). Because sharpest and most Gaussian is the visual observation likelihood with detection, the prior belief is most determined by the last visual observation and remains a sharp Gaussian distribution as shown in Figure 2(b). The posterior belief with the joint observation likelihood inherits this characteristics since the joint likelihood most likely captures the target location with a peak and magnifies the confidence of the prior belief with the joint likelihood.

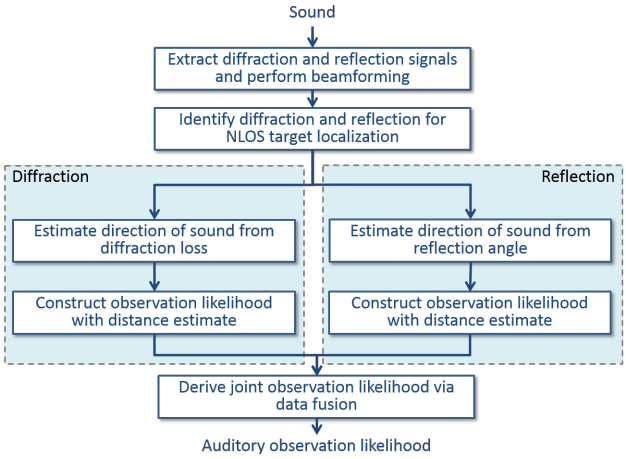
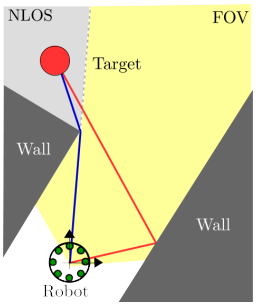


Figure 3: Construction of auditory NFOV target likelihood

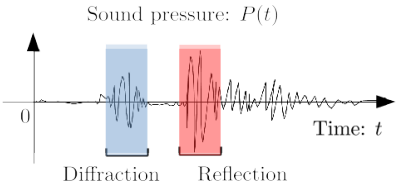
3 Physics based NLOS Auditory Observation Likelihood

3.1 Overview

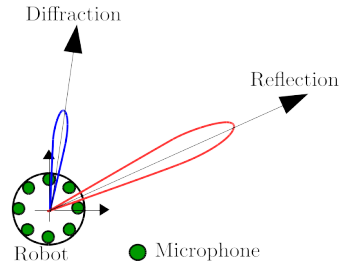
Figure 3 shows the overview of how to construct a NLOS auditory observation likelihood using the physics of sound wave propagation. Unlike radio signals, sound signals reflect significantly without penetrating into different media while they also diffract at low frequencies [1]. The proposed approach begins with obtaining a time-domain signal of a relatively impulsive sound at the microphone array. Notable signals are then extracted as candidate first-arrival diffraction and reflection sounds. The sound target is considered in a NLOS region if the diffraction and reflection signals behave as expected. An acoustic beamformer identifies the directions of the diffraction and reflection signals in the LOS region, and the diffraction and reflection points are identified to further localize the sound target in the NLOS region. The direction of the diffraction signal beyond the LOS is inferred by deriving the loss of sound energy through diffraction, or the diffraction loss. That of the reflection signal is identified by considering the orientation of the reflection wall. Diffraction and reflection observation likelihoods are eventually constructed by additionally incorporating knowledge on the distance from the sound magnitude and characteristics. An auditory observation likelihood is finally created by fusing the diffraction and reflection observation likelihoods.



(a) Acoustic signals from NFOV target

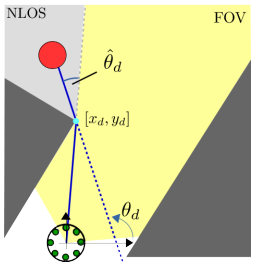


(b) Extraction of diffraction/reflection signals

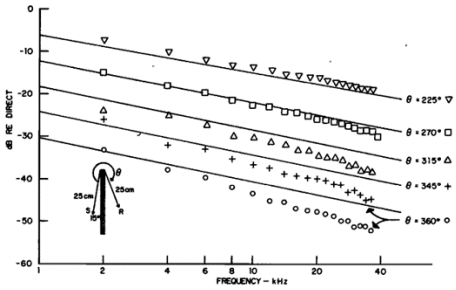


(c) Direction of signals identified though beamforming

Figure 4: Auditory NFOV target observation



(a) Diffraction sound estimation



(b) Magnitude with different orientation angles [16]

Figure 5: Diffraction sound estimation

3.2 Extraction of First-arrival Diffraction and Reflection Signals

Figure 4 shows the identification of a NLOS sound target and the extraction of the diffraction and reflection signals proposed in this paper illustratively in one of the simplest scenarios where a robot carrying a microphone array receives sound emitted by a target in the NLOS in a two-dimensional indoor environment with three walls (Figure 4(a)). Figure 4(b) shows the pressure of sound in the time domain, $P_i(t)$. As shown in the figures, sound waves emitted from a NLOS target reach the robot first through diffraction and second through reflection and, if the sound is relatively impulsive, the first-arrival diffraction and reflection signals can be extracted clearly. Because the sound energy loss from diffraction is larger than that from reflection, the sound target can be recognized NLOS if the absolute peak of the first signal is smaller than the second signal (The first-arrival signal is strongest with a LOS target as it is the LOS signal). Extraction becomes challenging for complex environments, but various existing techniques proposed to extract signals or select thresholds for extraction reportedly achieve successful extraction and identify candidate diffraction and reflection signals [12, 9, 26, 8]. The extraction results in the identification of directions of diffraction and reflection sounds through acoustic beamforming shown in Figure 4(c). Once it observes the sound directions, the RGB-D camera identifies the end of LOS and thus identifies the diffraction and reflection points as well as the orientation of the reflection wall and the reflection angle.

3.3 Estimation of Sound Direction from Diffraction Signals

Figure 5(a) shows the notations used for estimating sound direction from diffraction signals in the scenario introduced in the last subsection. The direction angle with respect to the robot frame to estimate is defined by θ_d whereas the NLOS angle is given by $\hat{\theta}_d$. The diffraction point is given by $[x_d, y_d]$. Of these, the diffraction point is known from the result of beamforming and the depth measurement, so the NLOS angle $\hat{\theta}_d$ must be further identified.

The proposed approach identifies the angle by analyzing the magnitudes of diffraction and reflection sounds, $M^d(\omega)$ and $M^r(\omega)$, which are extracted after representing the sound signals with respect to frequency ω. The loss of sound energy is assumed to be more if the NLOS angle is more. This assumption, in fact, has been found to be valid by the work of Medwin a quarter-century ago [16] shown in Figure 5(b). The magnitude of diffraction sound drops when the "level of NLOS" represented by the orientation angle is increased. Assuming that reflection is specular with negligible loss, this makes the proposed approach define the diffraction loss as

$$L_d = \int \left[M^r(\omega) - M^d(\omega) \right] \mathrm{d}\omega \geq 0 \tag{5}$$

and associates it with the level of NLOS. The work of Medwin also shows that the diffraction loss is approxi-

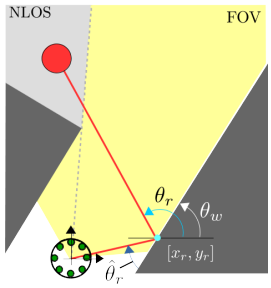


Figure 6: Reflection sound estimation

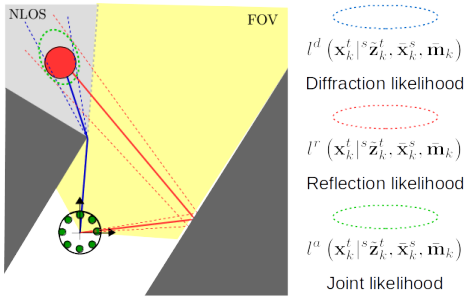


Figure 7: Joint acoustic observation likelihood

mately proportional to the level of NLOS. The nonlinear polynomial can be derived from the presented graph and enables more accurate identification of diffraction angle; $\hat{\theta}_d = f(L_d)$. Having the NLOS angle specified, the diffraction angle θ_d is derived as

$$\theta_d = \hat{\theta}_d + \tan^{-1}\frac{y_d}{x_d}. \tag{6}$$

3.4 Estimation of Sound Direction from Reflection Signals

Figure 6 shows the proposed approach for estimation of sound direction from reflection signals. Reflection makes the sound propagation and the subsequent target estimation complicated, but if the wall is smooth and yields specular reflection, the sound direction can be estimated easily [19]. Let the reflection angle with respect to the robot frame to derive be θ_r and the sound direction to the reflection wall and the reflection point be $\hat{\theta}_r$ and $[x_r, y_r]$ respectively. Since both $\hat{\theta}_r$ and $[x_r, y_r]$ are known from the preceding measurement, the orientation of the wall with respect to the robot frame is given by

$$\theta_w = \hat{\theta}_r + \tan^{-1}\frac{y_r}{x_r}. \tag{7}$$

The reflection angle is resultantly given by

$$\theta_r = \hat{\theta}_r + \theta_w = 2\hat{\theta}_r + \tan^{-1}\frac{y_r}{x_r}. \tag{8}$$

3.5 Construction of Joint Acoustic Observation Likelihood

While the sound can be better identified in direction rather than distance, it is also possible to make an estimate on how far the sound target is. The proposed approach makes the estimate by utilizing any available information including the magnitude, sound patterns stored in a database, or sound characteristics in a knowledge base and constructs an observation likelihood for each of the diffraction and reflection signals by modeling uncertainties. The diffraction and reflection likelihoods are then combined to create an auditory joint observation likelihood via the canonical data fusion formula:

$$\begin{aligned} l^a\left(\mathbf{x}_k^t|^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) \\ = l_j^d\left(\mathbf{x}_k^t|^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) l_j^r\left(\mathbf{x}_k^t|^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) \end{aligned} \tag{9}$$

where $l^d(\cdot)$ and $l^r(\cdot)$ are the diffraction and reflection observation likelihood.

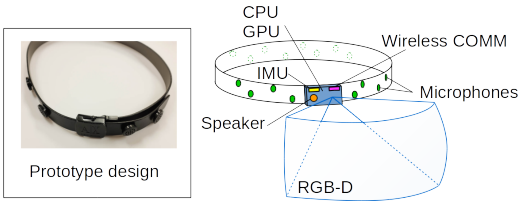
Figure 7 illustrates the diffraction and reflection observation likelihoods as well as the joint observation likelihood where the observation likelihood is represented by an ellipsoid indicating a probability distribution with a covariance. The diffraction and reflection likelihoods are shown to have high eccentricity due to more accuracy in direction than in distance. Since the difference of the diffraction and reflection likelihoods in orientation may not be significant, the resulting auditory joint likelihood could also be given by an ellipsoid with high eccentricity, but the proposed approach, utilizing the diffraction and reflection physics of sound, could estimate the location of the sound target.

4 Assistive/Training Devices for Blind and Visually Impaired People

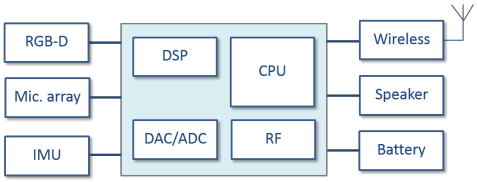
Figure 8 shows the schematic diagram of the wearable assistive/training device for blind and visually impaired people to be developed. The wearable device is chosen to be a belt partly because it is one of the most common wearings and partly because it can implement a ringed microphone array as well as other sensors/components with some weight including an RGB-D camera. The major components of the device are (1) a multi-story ringed microphone array, (2) an RGB-D camera, (3) an IMU, (4) a speaker, (5) a central unit and (6) a wireless module. Because of the multi-story design, the microphone array can form acoustic beams in not only the horizontal direction but also in the vertical direction. This allows the identification and removal of sound components coming from the floor and the ceiling and the extraction of the corresponding first-arrival diffraction and reflection signals. The RGB-D camera, which works based on the principle of time-of-flight, structured light or stereovision, is used not only to recognize LOS 3D environments for NLOS target estimation but also to perform SLAM together with an IMU. The speaker is used to provide feedback by sound for assistive and training use. The NLOS target estimation can be performed on the central unit since its computation can be made light in weight.

5 Preliminary Experimental Validation

This subsection presents a result of the preliminary proof-of-concept. Figure 9 shows the test environment developed and the actual test conducted in the anechoic

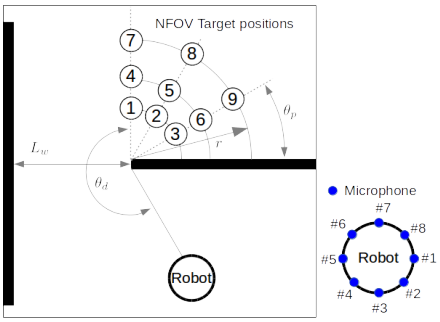


(a) Mechanical design



(b) Electronic design

Figure 8: Assistive/training devices for blind people



(a) Test environment

(b) NLOS test with Daniel Kish

Figure 9: Experimental proof-of-concept settings

chamber. The capability of the microphone array, as well as sighted and blind people including the pioneer of human echolocation, Daniel Kish, was tested in the environment. Only two walls were placed in the anechoic chamber to make the problem two-dimensional (2D) for the proof-of-concept purpose. A mechanical clicker created impulsive sound at positions labeled 1-9, and the robot with the microphone array and human testees were supposed to identify the correct sound direction and location [22].

Figure 10 shows the sound amplitude in the time domain collected at Microphones 6 and 8. The result shows that the diffraction signal and reflection signal are distinctly separable having enough time interval between the signals. More signal processing efforts will be necessary when more natural sounds are deployed, and the signal processing effort is ongoing while completing the proof-of-concept. Figure 11 shows the constructed joint

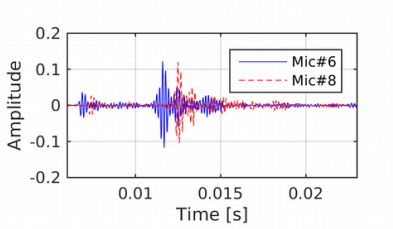
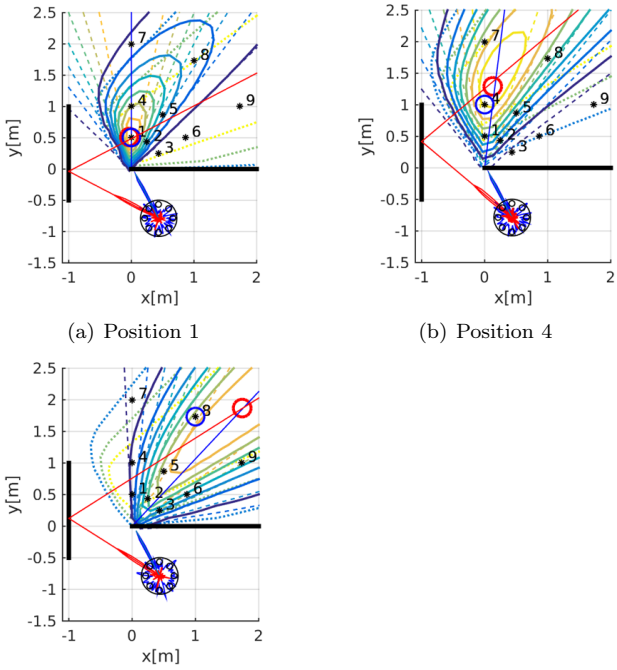


Figure 10: Time-domain signals



(a) Position 1

(b) Position 4

(c) Position 8

Figure 11: Result of proof-of-concept

observation likelihood together with the true target location colored blue when the sound source was located at Positions 1, 4 and 8. The maximum likelihood is shown by a red circle to compare to the true location. It can be first seen that the diffraction and reflection points are always identified well to indicate that the sound source is in the NLOS region. The accuracy drops particularly for targets in a severely NLOS region, but the proposed approach could still make a good estimate.

6 Conclusions

This paper has presented a new approach that localizes NLOS targets in unknown indoor environments. While the visual sensor localizes a target when the target is on its LOS, the auditory sensor stably localizes the target even if the target is not on the LOS of the visual sensor. In order for a NLOS target in an unknown environment, the proposed approach extracts and analyzes the first-arrival diffraction signal and the first-arrival reflection signal. The RBE localizes the NLOS most reliably and accurately. The ability of the proposed approach was experimentally validated in a controlled indoor environment.

The work presented in this paper is only the first step for the NLOS target localization in unknown indoor environments. Ongoing work includes the experimental validation in the uncontrolled indoor environment, the development of the assistive/training device for people who are blind or visually impaired and the use of human voice.

Acknowledgments

The work was primarily supported by National Science Foundation EAGER program (Award #1554961). The author also express his gratitude to Makoto Kumon, Kuya Takami and Hangxin Liu for their collaboration to the work.

References

[1] Giovanni Bellusci, Junlin Yan, Gerard JM Janssen, and Christian CJM Tiberius. An ultra-wideband positioning demonstrator using audio signals. In *Positioning, Navigation and Communication, 2007. WPNC'07. 4th Workshop on*, pages 71–76. IEEE, 2007.

[2] Marco Bertinato, Giulia Ortolan, Fabio Maran, Riccardo Marcon, Alessandro Marcassa, Filippo Zanella, Matrizio Zambotto, Luca Schenato, and Angelo Cenedese. Rf localization and tracking of mobile nodes in wireless sensors networks: Architectures, algorithms and experiments. 2008.

[3] Pi-Chun Chen. A non-line-of-sight error mitigation algorithm in location estimation. In *Wireless Communications and Networking Conference, 1999. WCNC. 1999 IEEE*, pages 316–320. IEEE, 1999.

[4] Huan Dai, Zhao-Min Zhu, and Xiao-Feng Gu. Multi-target indoor localization and tracking on video monitoring system in a wireless sensor network. *Journal of Network and Computer Applications*, 2012.

[5] Tomonari Furukawa, Frederic Bourgault, Benjamin Lavis, and Hugh F Durrant-Whyte. Recursive bayesian search-and-tracking using coordinated uavs for lost targets. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 2521–2526. IEEE, 2006.

[6] Tomonari Furukawa, Lin Chi Mak, Hugh Durrant-Whyte, and Rajmohan Madhavan. Autonomous bayesian search and tracking, and its experimental validation. *Advanced Robotics*, 26(5-6):461–485, 2012.

[7] Sinan Gezici. A survey on wireless position estimation. *Wireless Personal Communications*, 44(3):263–282, 2008.

[8] Ismail Guvenc and Chia-Chin Chong. A survey on toa based wireless localization and nlos mitigation techniques. *Communications Surveys & Tutorials, IEEE*, 11(3):107–124, 2009.

[9] Zoubir Irahhauten, Giovanni Bellusci, Gerard JM Janssen, Homayoun Nikookar, and CCJM Tiberius. Investigation of uwb ranging in dense indoor multipath environments. In *Communication systems, 2006. ICCS 2006. 10th IEEE Singapore International Conference on*, pages 1–5. IEEE, 2006.

[10] Hiam M Khoury and Vineet R Kamat. Evaluation of position tracking technologies for user localization in indoor construction environments. *Automation in Construction*, 18(4):444–457, 2009.

[11] Makoto Kumon, Daisuke Kimoto, Kuya Takami, and Tomonari Furukawa. Bayesian non-field-of-view target estimation incorporating an acoustic sensor,. In *Proceedings of 2013 IEEE/RSJ International Conference of Intelligent Robots and Systems*, November 1-3, 2013 2013.

[12] Joon-Yong Lee and Sungyul Yoo. Large error performance of uwb ranging in multipath and multiuser environments. *Microwave Theory and Techniques, IEEE Transactions on*, 54(4):1887–1895, 2006.

[13] Hui Liu, H. Darabi, P. Banerjee, and Jing Liu. Survey of wireless indoor positioning techniques and systems. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(6):1067–1080, 2007.

[14] Lin C Mak and Tomonari Furukawa. Non-line-of-sight localization of a controlled sound source. In *Advanced Intelligent Mechatronics, 2009. AIM 2009. IEEE/ASME International Conference on*, pages 475–480. IEEE, 2009.

[15] R. Mauler. *Recent Developments in Cooperative Control and Optimizatio*, chapter Objective Functions for Bayesian Control-Theoretic Sensor Management, II: MHC-Like Approximation, pages 273–316. Kluwer Academic Publishers, Norwell, MA, 2003.

[16] H Medwin. Shadowing by finite noise barriers. *The Journal of the Acoustical Society of America*, 69(4):1060–1064, 1981.

[17] Lionel M Ni, Yunhao Liu, Yiu Cho Lau, and Abhishek P Patil. Landmarc: indoor location sensing using active rfid. *Wireless networks*, 10(6):701–710, 2004.

[18] Eric A Prigge. *A positioning system with no line-of-sight restrictions for cluttered environments*. PhD thesis, Stanford University, 2004.

[19] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of Audio Engineering Society*, 45(6):456–466, June 1997.

[20] Chee Kiat Seow and Soon Yim Tan. Non-line-of-sight localization in multipath environments. *Mobile Computing, IEEE Transactions on*, 7(5):647–660, 2008.

[21] Piergiorgio Svaizer, Alessio Brutti, and Maurizio Omologo. Environment aware estimation of the orientation of acoustic sources using a line array. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1024–1028. IEEE, 2012.

[22] Kuya Takami, Tomonari Furukawa, Makoto Kumon, and Lin Chi Mak. Non-field-of-view indoor sound source localization based on reflection and diffraction. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2015 IEEE International Conference on*, pages 59–64. IEEE, 2015.

[23] T; Kumon M; Dissanayake G. Takami, K.; Furukawa. Non-field-of-view acoustic target estimation in complex indoor environment. In *Field and Service Robotics*, June 24-26 2015.

[24] Tomonari; Kumon Makoto; Dissanayake Gamini Takami, Kuya; Furukawa. Estimation of a nonvisible field-of-view mobile target incoorperating optical and acoustic sensors. *Autonomous Robots*, 7(26):1–17, 2015.

[25] JeanMarc Valin, François Michaud, Jean Rouat, and Dominic Létourneau. Robust sound source localization using a microphone array on a mobile robot. In *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 2, pages 1228–1233. IEEE, 2003.

[26] Chi Xu and Choi L Law. Delay-dependent threshold selection for uwb toa estimation. *IEEE communications letters*, 12(5):380–382, 2008.

[27] Dian Zhang, Yanyan Yang, Dachao Cheng, Siyuan Liu, and Lionel M Ni. Cocktail: an rf-based hybrid approach for indoor localization. In *Communications (ICC), 2010 IEEE International Conference on*, pages 1–5. IEEE, 2010.

一般社団法人 人工知能学会
Japanese Society for Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-2 (11/25)

アクティブ周波数レンジフィルタを用いた雑音にロバストな音源定位手法の提案

Noise-robust sound source localization method using active frequency range filter

干場功太郎 *1, 中臺一博 *1,*2, 公文誠 *3, 奥乃博 *4

Kotaro HOSHIBA[*1], **Kazuhiro NAKADAI**[*1,2], **Makoto KUMON**[*3], **Hiroshi G. OKUNO**[*4]

東京工業大学 工学院 システム制御系 *1

(株) ホンダ・リサーチ・インスティチュート・ジャパン *2

熊本大学 大学院 先端科学研究部 *3

早稲田大学 実体情報学博士プログラム *4

Department of Systems and Control Engineering, School of Engineering

Tokyo Institute of Technology[*1]

Honda Research Institute Japan Co., Ltd.[*2]

Faculty of Advanced Science and Technology, Kumamoto University[*3]

Graduate Program for Embodiment Informatics, Waseda University[*4]

{hoshiba, nakadai}@ra.sc.e.titech.ac.jp, kumon@gpo.kumamoto-u.ac.jp, okuno@nue.org

Abstract

われわれは，被災地等での要救助者の探索を目的に，UAV (Unmanned Aerial Vehicle) 搭載マイクロホンアレイを用いた音源探査の研究を行っている．これまで，屋外実環境での使用を想定し，さまざまな音源定位手法を提案してきた．しかし，これらの手法は，雑音耐性とリアルタイム性がトレードオフの関係にあった．被災地等での音源探査の場合，雑音耐性とリアルタイム性の両者が重要であり，それらを同時に満たす手法の開発が必要である．本稿では，MUSIC (MUltiple Signal Classification) 法に基づく，アクティブ周波数レンジフィルタを用いた音源定位手法を提案する．本手法では，単純な四則演算のみを用いたフィルタを使用することで，雑音耐性とリアルタイム性を確保する．計算機シミュレーションによる評価を行った結果，以前の手法に比べ，定位性能・処理遅延ともに優位性を示すことができた．

1 はじめに

屋外環境での音響処理に関する研究は，計測分野などさまざまな応用が考えられるため，さかんに行われている．われわれは，JST ImPACT タフロボティクスチャレンジの極限音響チームにおける研究開発活動の一環として，これまでに培ってきたロボット聴覚技術 [1] を用いて，被災地等での要救助者の探索を目的に，UAV (Unmanned Aerial Vehicle) 搭載マイクロホンアレイを用いた音源探査の研究を行っている．こうした手法を確立できれば，上空から広範囲かつ迅速に被災者の音声を探査する有効な手段ができると考えられる．これまで，動的に変化する UAV の自己雑音を抑制するさまざまな音源定位手法の提案や [2, 3]，屋外実環境において実時間で音源探査を行うシステムの開発を行ってきた [4, 5]．しかし，これらの手法は，雑音耐性を高めると計算コストが大きくなるためリアルタイム性が失われ，リアルタイム性を高めると雑音耐性が低くなるといった，トレードオフの関係にある．また，機械学習の一種である LSTM (Long Short Term Memory) を用いて，周波数方向のマスクを作成するという手法も提案されているが [6]，計算コストが大きく，リアルタイム性を確保できない．本研究では，音源探査システムの組み込みシステム化の検討も行っているため [7]，組み込みボードでも処理が行えるよう，組み込みボードでも処理が行えるような計算コストの小さい，リアルタイム性と雑音耐性の両者を満たす音源定位手法の開発が必要である．そこで本稿では，MUSIC (MUltiple Signal Classification) 法 [8] に基づく，単純な四則演算のみを用いたアクティブ周波数レンジフィルタを用いた音源定位手法を提案し，従来手法と比較することで音源定位性能の評価と考察を行う．

2　音源定位手法

本章では，本稿で扱う音源定位手法について述べる．

2.1　SEVD-MUSIC 法

音源定位において多く用いられる，一般的な MUSIC 法である SEVD-MUSIC (MUSIC based on Standard Eigen Value Decomposition) 法のアルゴリズムを示す．M チャンネル入力音響信号の f フレーム目をフーリエ変換して得られる $Z(\omega, f)$ から，以下のように相関行列 $R(\omega, f)$ を定義する．

$$R(\omega, f) = \frac{1}{T_R} \sum_{\tau=f}^{f+T_R-1} Z(\omega, \tau) Z^*(\omega, \tau) \tag{1}$$

ここで，ω は周波数ビン番号，T_R は相関行列の計算に用いるフレーム数，Z^* は Z の共役転置である．次に，f 番目のフレームに対して，f_s 前のフレームから，T_N フレーム分の信号は雑音区間であると仮定して，雑音の相関行列 $K(\omega, f)$ を求める．SEVD-MUSIC 法では，こうして得られた $R(\omega, f)$ を固有値展開して固有ベクトルを計算する．

$$R(\omega, f) = E(\omega, f)\Lambda(\omega, f)E^*(\omega, f) \tag{2}$$

ただし，$\Lambda(\omega, f)$ は降順に並んだ固有値を対角成分に持つ行列，$E(\omega, f)$ は $\Lambda(\omega, f)$ に対応する固有ベクトルを並べた行列である．これと，UAV 座標系での音源方向 ψ に対応した伝達関数 $G(\omega, \psi)$ を用いて MUSIC 空間スペクトル $P(\omega, \psi, f)$ を計算する．

$$P(\omega, \psi, f) = \frac{|G^*(\omega, \psi)G(\omega, \psi)|}{\sum_{m=L+1}^{M} |G^*(\omega, \psi)e_m(\omega, \psi)|}. \tag{3}$$

ただし，L は目的音源数，e_m は，E に含まれる m 番目 $(1 \leq m \leq M)$ の固有値ベクトルを表す．また，ψ は UAV に対する方位角 θ，仰角 ϕ から $\psi = (\theta, \phi)$ と定義する．このようにして得られた $P(\omega, \psi, f)$ を，音源方向を推定するために以下のように ω 方向に平均する．

$$\bar{P}(\psi, f) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} P(\omega, \psi, f) \tag{4}$$

なお，ω_H，ω_L は使用する周波数ビンの上限と下限に対応したインデックスである．$\bar{P}(\psi, f)$ に対して閾値処理，ピーク検出を行い，得られたピークに対する ψ を音源方向として検出する．

SEVD-MUSIC 法では，雑音に対する耐性が低いが，計算コストが小さいため，処理遅延が少ないという特徴が挙げられる．また，ω_H，ω_L を目的音源に応じ狭帯域に設定することで雑音耐性を高めることもできるが，異なる周波数特性を持った目的音源が複数あった場合，対応することができない．

2.2　iGSVD-MUSIC 法

われわれは，これまで屋外音環境特有の問題を解決するため，MUSIC 法を拡張した音源定位手法を提案してきた．その一つとして，MUSIC 法に一般化特異値展開を導入した iGSVD-MUSIC ((MUSIC based on incremental Generalized Singular Value Decomposition) 法 [3] が挙げられる．iGSVD-MUSIC 法では，時間変化する雑音に対応するため，逐次的に雑音モデルを推定し，雑音モデルを用いた白色化により，音源定位性能の向上を図る．

以下にそのアルゴリズムを示す．iGSVD-MUSIC 法では，f 番目のフレームに対して，f_s 前のフレームから，T_N フレーム分の信号は雑音区間であると仮定して，雑音の相関行列 $K(\omega, f)$ を求める．

$$K(\omega, f) = \frac{1}{T_N} \sum_{\tau=f-f_s-T_N}^{f+f_s} Z(\omega, \tau) Z^*(\omega, \tau) \tag{5}$$

iGSVD-MUSIC 法では，フレームごとに雑音が推定できるため，動的な雑音変化に対応できることができる．R の左から K^{-1} を掛けることで，雑音成分を白色化する．こうして得られた $K^{-1}(\omega, f)R(\omega, f)$ を一般化特異値展開して特異値ベクトルを計算する．

$$K^{-1}(\omega, f)R(\omega, f) = Y_l(\omega, f)\Sigma(\omega, f)Y_r^*(\omega, f) \tag{6}$$

ただし，$\Sigma(\omega, f)$ は降順に並んだ特異値を対角成分に持つ行列，$Y_l(\omega, f)$，$Y_r(\omega, f)$ は $\Sigma(\omega, f)$ に対応する特異値ベクトルを並べた行列である．ここから，SEVD-MUSIC 法と同様に，MUSIC 空間スペクトル $P(\omega, \psi, f)$ を計算する．

$$P(\omega, \psi, f) = \frac{|G^*(\omega, \psi)G(\omega, \psi)|}{\sum_{m=L+1}^{M} |G^*(\omega, \psi)y_m(\omega, \psi)|} \tag{7}$$

ただし，y_m は，Y_l に含まれる m 番目の特異値ベクトルを表す．このようにして得られた $P(\omega, \psi, f)$ を，Eq. 4 と同様に，ω 方向に平均する．その後，閾値処理，ピーク検出を行う．

iGSVD-MUSIC 法では，UAV のローター音など時間変化する雑音を抑制することができる一方，計算コストが大きく，2 s 以上の遅延が発生することがわかった [5]

2.3　アクティブ周波数フィルタを用いた音源定位手法の提案

上記の 2 つの手法は，雑音耐性とリアルタイム性がトレードオフの関係にあり，実環境で音源探査を行う場合，両者を満たす手法の開発が必要である．そこで，MUSIC 法における，使用する周波数レンジをアクティブに変化させる手法を提案する．SEVD-MUSIC 法をベースに，Eq. 4 における ω_H，ω_L を状況に応じて変化させることで，雑音に対する耐性とリアルタイム性を確保する．

アルゴリズムを以下に示す．f 番目のフレームに対して，f_s 前のフレームから，T_N フレーム分の信号は雑音

区間であると仮定して，雑音の周波数スペクトル Z_n をフレーム間，チャンネル間の平均から求める．

$$Z_n(\omega, f) = \frac{1}{T_N \cdot M} \sum_{m} \sum_{\tau=f-f_s+1}^{f-f_s+T_N} Z(\omega, \tau) \tag{8}$$

得られた Z_n と現在の周波数スペクトル Z の差分から，評価関数 $J(\omega, f)$ を算出する．

$$J(\omega, f) = Z(\omega, f) - Z_n(\omega, f) \tag{9}$$

J は，雑音に対する各周波数のパワーの変化量を意味する．つまり，J の値が大きい周波数は目的音源によるものと考えられる．そこで，J が最も大きい周波数 $\omega_J(f)$ を求める．

$$\omega_J(f) = \underset{\omega}{\operatorname{argmax}}\, J(\omega, f) \tag{10}$$

得られた ω_J から，使用する周波数レンジを決定する．

$$\omega_H = \omega_J + \frac{f_w}{2} \tag{11}$$

$$\omega_L = \omega_J - \frac{f_w}{2} \tag{12}$$

ここで，f_w は周波数レンジの大きさである．f_w は目的音源の周波数特性を考慮し，可能な限り小さく設定する．そうすることで，雑音の影響を抑制する．このように算出された ω_H，ω_L を用いて，SEVD-MUSIC 法による処理を行う．以降，本手法を AFRF-MUSIC (MUSIC using Active Frequency Range Filter) 法と呼ぶ．

AFRF-MUSIC 法では，f_w を狭帯域に設定することで雑音への耐性が期待できる．また，アクティブに周波数レンジを変化させるため，異なる周波数特性を持った複数の音源が存在した場合にも対応することができる．さらに，本手法では，SEVD-MUSIC 法に加えて単純な四則演算のみを使用しているため，計算コストが小さく，処理遅延が少なくなり，リアルタイム性を確保できる．

3 検証実験

3.1 実験方法

計算機シミュレーションにより，音響信号を作成し，各音源定位手法を行うことで，定位性能評価を行った．探査対象の音源として，周波数特性の異なる，笛の音と人の声の 2 種類を用いた．これらの音源を用いて，方位角 θ を $-180°$〜$180°$，仰角 ϕ を $-90°$〜$0°$ の範囲で，5° 刻みで各方向から到達した場合の信号をシミュレーションにより作成する．また，雑音として実際のアレイで収録した UAV のローター音を加え，信号の SNR (Signal-to-Noise Ratio) はを 20〜-20 dB の間で変化させ，各音源定位手法により信号処理を行う．マイクロホンアレイには，Fig. 1(a) に示される，球形のマイクロホンアレイを用いる．本マイクロホンアレイはカスケード接続された 12 ch の MEMS マイクロホンで構成され，Fig. 1(b) のように UAV のアームの先に接続し，使用する．各マイクロホンのアレイは，Fig. 1(c) のように球形の筐体の下半球に配置されている．探査対象音源，アレイで収録したローター音のスペクトログラムを Fig. 2 に示す．笛の音は 2.5〜3 kHz にパワーが集中していることがわかる．また，声は 850 Hz 付近が最もパワーが大きく，約 4 kHz まで倍音が存在している．ローター音の周波数帯域は広域に渡っているが，特に 2 kHz 以下の成分が大きい．作成した音響信号を，前述の各音源定位法にて処理を行う．アルゴリズムの実装には，ロボット聴覚オープンソースソフトウェア HARK (Honda Research Institute Japan Audition for Robots

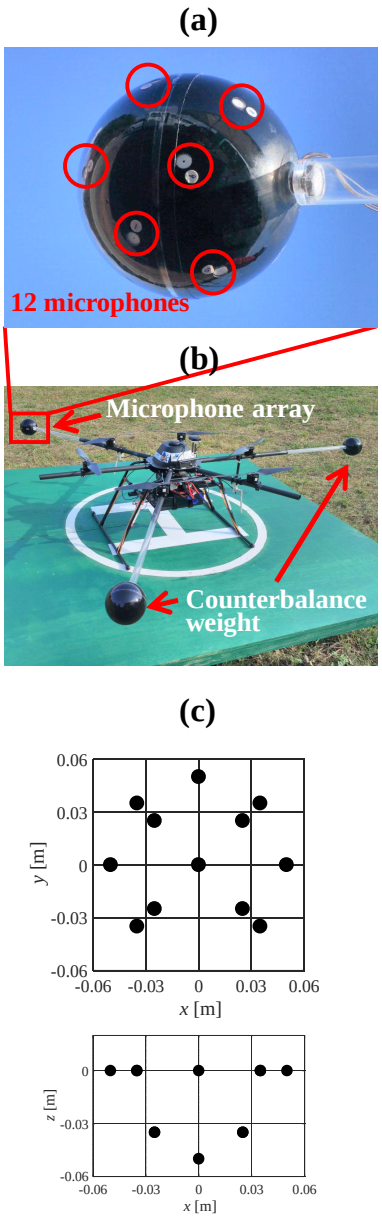


Fig. 1: (a) Microphone array, (b) UAV with microphone array, (c) coordinates of microphone positions in the microphone array.

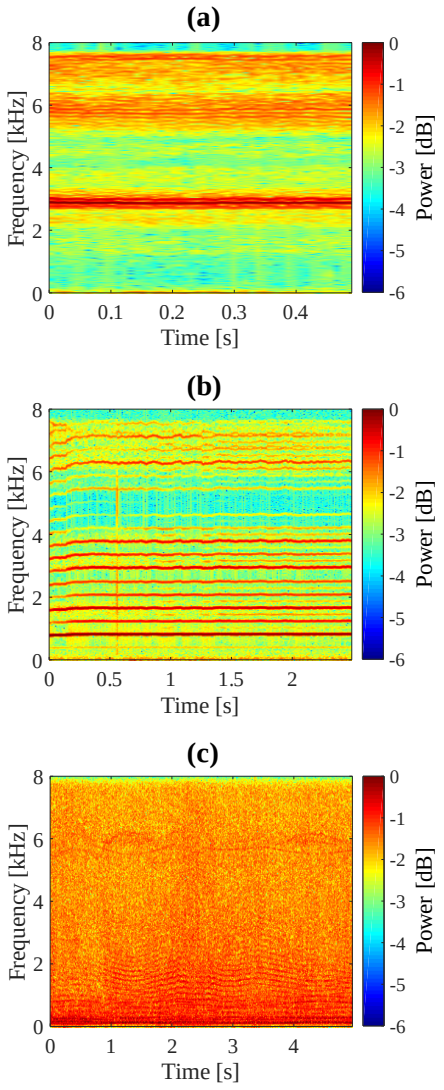


Fig. 2: Spectrograms. (a) Whistle, (b) voice, (c) noise of UAV recorded by microphone aray.

with Kyoto University)[1][9] を用いた．MUSIC 法で用いる伝達関数 (Eq. 3 における G) については，幾何計算で算出した．各方向・各音源につき，周波数レンジを 500–3000 Hz に設定した SEVD-MUSIC・iGSVD-MUSIC，レンジを 2500–3000 Hz に設定した SEVD-MUSIC，AFRF-MUSIC の 5 つの手法を用いて MUSIC スペクトル P を各 250 フレーム算出し，評価を行った．AFRF-MUSIC における f_w は 500 Hz に設定した．本稿では，Fig. 3 に示すように方位角 θ，仰角 ϕ を設定し，MUSIC スペクトルの評価を行う．

3.2 実験結果

実験結果について述べる．SNR が 0 dB の場合に算出された MUSIC スペクトルの一例を Fig. 4 に示す．(a) が SEVD-MUSIC (500–3000 Hz)，(b) が iGSVD-MUSIC (500–3000 Hz)，(c)，(d) が SEVD-MUSIC (2500–3000 Hz)，(e) が AFRF-MUSIC による結果である．また，音源として，(a)，(b)，(c)，(e) は笛の音，(d) は声を用いている．Fig. 3 に従ってプロットされており，カラーバーで各方向からの音のパワーを示している．また，音源の方向は $\theta = -30°$，$\phi = -45°$ に設定している．Fig. 4(a) に示されるように，周波数レンジを大きく設定した場合の SEVD-MUSIC では，音源方向にピークが確認できるものの，MUSIC スペクトルに UAV のローター雑音が大きく現れているため，定位精度に大きく影響が出ると考えられる．iGSVD-MUSIC を用いた場合，Fig. 4(b) のように雑音が大きく抑制され，音源方向に鋭いピークを確認することができる．また，SEVD-MUSIC における周波数レンジを，笛の周波数特性を考慮し，2500–3000 Hz と設定した場合，Fig. 4(c) のように，ピークの鋭さはないが，ローター雑音を抑制することができている．このことから，周波数レンジを狭帯域に設定することで雑音耐性を確保することが可能であるとわかる．しかし，このように固定の周波数レンジを用いると，周波数特性の異なる声が音源の場合，Fig. 4(d) のようにピークがほぼ確認できなくなることから，2 種類以上の異なる音源が存在した場合に定位性能が低下してしまう．AFRF-MUSIC を用いた場合，使用する狭帯域周波数レンジが適切に設定されるため，Fig. 4(e) に示されるように，レンジを 2500–3000 Hz に設定した SEVD-MUSIC と同様の結果になる．また，AFRF-MUSIC では音源の周波数特性に応じてレンジが変化するため，異なる音源が存在する場合であっても定位性能が低下しにくいと考えられる．

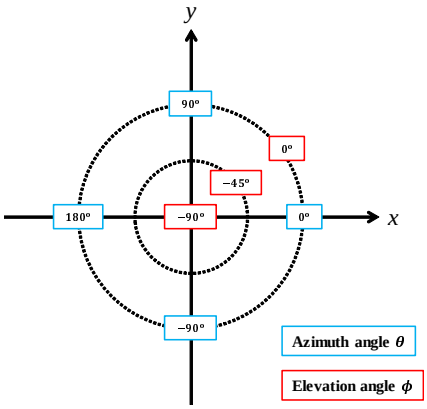


Fig. 3: Setting of azimuth angle θ and elevation angle ϕ.

3.3 考察

各手法を用いた場合の音源定位の正解率による定位性能の評価を行った．MUSIC スペクトルの最大ピークの方向が，音源の設定方向と一致した場合に正解とし，笛の音・声を用いたすべての方向からのシミュレーション音源にて処理・ピーク探査を行い，正解率を算出する．Fig. 5 が算出

[1] http://www.hark.jp/

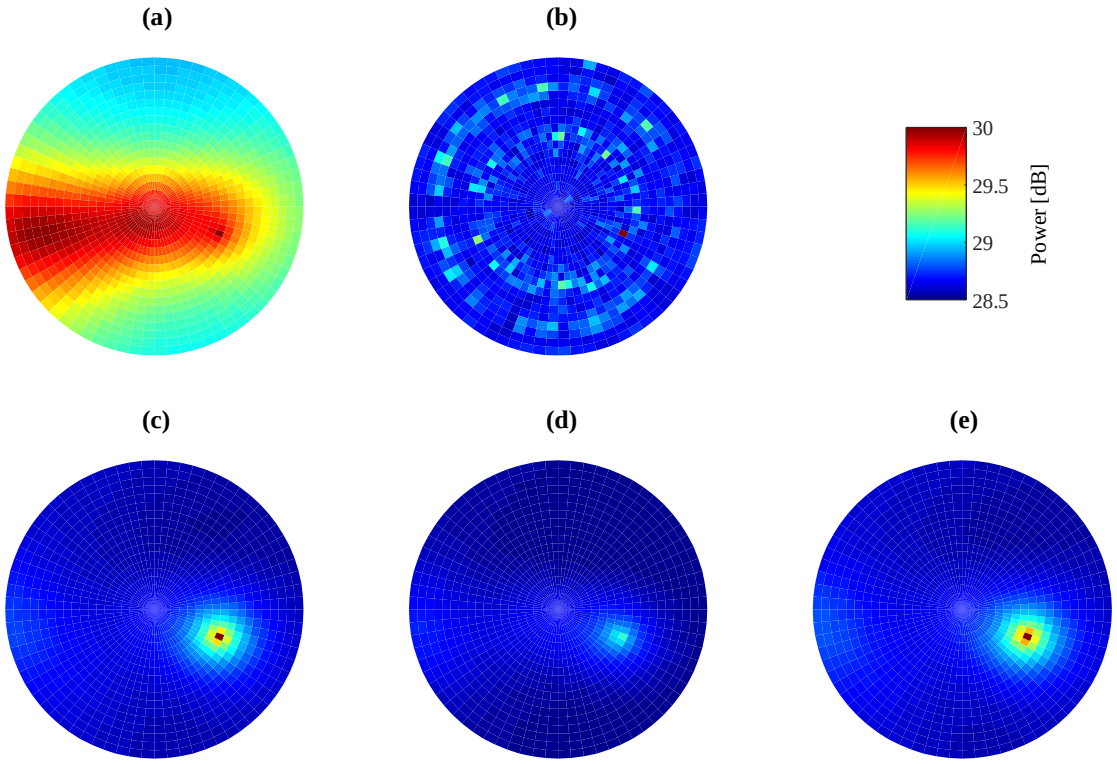


Fig. 4: MUSIC spectra. (a) whistle sound using SEVD-MUSIC (500–3000 Hz), (b) whistle sound using iGSVD-MUSIC (500–3000 Hz), (c) whistle sound using SEVD-MUSIC (2500–3000 Hz), (d) voice using SEVD-MUSIC (2500–3000 Hz), (e) whistle sound using AFRF-MUSIC.

した各手法の正解率である．横軸が SNR，縦軸が正解率である．周波数レンジを広帯域に設定した SEVD-MIUSIC では，SNR が低下するとともに，正答率も低下していることがわかる．iGSVD-MUSIC では，SNR が-10 dB までは正解率が 100 %に近く，-20 dB の場合でも 80 %を超えている．このことから，雑音耐性が非常に高いことがわかる．また，レンジを狭帯域に設定した SEVD-MUSIC では，SNR が 0 dB 以下の場合には広帯域の SEVD-MUSIC と比べ正解率が高いが，SNR が 0 dB 以上の場合には 70 %程度となっている．これは，周波数レンジを狭くすることで雑音を抑制するため，SNR が低い場合には有効的であるが，笛の音に応じてレンジを設定したため，音源が声の場合の正解率が低下し，SNR が高い場合でも 100 %にはならない．AFRF-MUSIC では，雑音を抑制し，さらに複数の種類の音源がある場合でも，それぞれに応じた周波数レンジが設定されるため，iGSVD-MUSIC よりは正解率は低いものの，近い値となっていることがわかる．

また，処理遅延を計測することで，リアルタイム性の評価を行った．各手法の処理遅延を Table 1 に示す．iGSVD が最も遅延が大きく，3 s 以上の遅延が発生していた．また，広帯域に設定した場合と狭帯域に設定した場合の SEVD-MUSIC について，遅延は使用する周波数レンジの大きさに比例して大きくなることがわかっている [5]．今回の実験の場合でも，広帯域に設定した場合と比べ，狭帯域に設定した場合は遅延が少なく，リアルタイム性が高くなっている．AFRF の場合，f_w を 500 Hz と設定しているため，計算コストは狭帯域に設定した SEVD-MUSIC とほぼ変わらず，遅延もほぼ同程度である．

これらの結果から，AFRF-MUSIC の定位性能は雑音耐性の高い iGSVD-MUSIC と同程度，遅延は計算量の少ない，狭帯域に設定した SEVD-MUSIC と同程度であることから，雑音耐性・リアルタイム性の両者を満たす音源定位手法として有用性が確認できた．本実験では単一のレンジを用いたが，複数のレンジを同時に使用するといった拡張もできるため，同時に発話されている複数音源への対応も可能であり，今後検討を行っていく予定である．

4 おわりに

本稿では，リアルタイム性と雑音耐性の両者を満たす音源定位手法の開発を目的に，MUSIC 法に基づく，アクティブ周波数レンジフィルタを用いた音源定位手法を提案した．最も計算コストの小さい SEVD-MUSIC をベースに，単純な四則演算のみで構成されるアクティブ周波数レンジフィルタを適用した AFRF-MUSIC を開発し，リアル

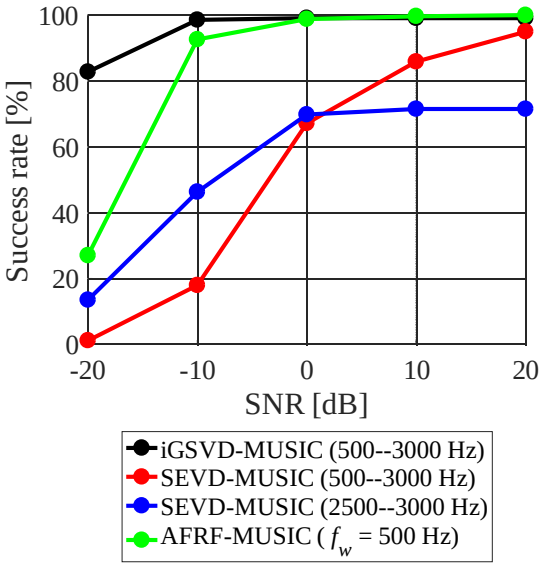


Fig. 5: Success rate of localization.

Table 1: Success rate of localization.

Algorithm	Delay [s]
iGSVD-MUSIC	3.219
SEVD-MUSIC (500–3000 Hz)	0.533
SEVD-MUSIC (2500–3000 Hz)	0.194
AFRF-MUSIC	0.204

タイム性と雑音耐性を確保した．評価実験により，その有用性が確認できた．しかし，本手法では f_s フレーム前の信号を雑音として扱うため，目的音源が f_s フレームを超えた場合，差分として目的音源が検出されない場合がある．また，UAV の飛行中のような環境が変わりやすい場面では，環境やノイズによるパラメータチューニングが難しく，定位性能が低下する可能性がある．今後は，パラメータチューニングの自動化について検討を行っていく予定である．

謝辞

本研究は，JSPS 科研費 16H02884，16K00294,，17K00365 および，JST ImPACT タフロボティクスチャレンジの助成をうけた．

参考文献

[1] K. Nakadai, T. Lourens, H. G. Okuno and H. Kitano, “Active audition for humanoid”, Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000), pp. 832-839, 2000.

[2] K. Okutani, T. Yoshida, K. Nakamura and K. Nakadai, “Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter”, Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS), pp. 3288-3293, 2012.

[3] T. Ohata, K. Nakamura, T. Mizumoto, T. Tezuka and K. Nakadai, “Improvement in outdoor sound source detection using a quadrotor-embedded microphone array”, Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS), pp. 1902-1907, 2014.

[4] K. Hoshiba, O. Sugiyama, A. Nagamine, R. Kojima, M. Kumon, K. Nakadai, “Design and assessment of sound source localization system with a UAV-embedded microphone array”, Journal of Robotics and Mechatronics, vol. 29, No. 1, pp. 154-167, 2017.

[5] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, H. G. Okuno, “Design of UAV-Embedded Microphone Array System for Sound Source Localization in Outdoor Environments”, Sensors, vol. 17, No. 11, pp. 1-16, 2017.

[6] C. Xu, X. Xiao, S. Sun, W. Rao, E. S. Chng, H. Li, “Weighted Spatial Covariance Matrix Estimation for MUSIC based TDOA Estimation of Speech Source”, Proceedings of the INTERSPEECH 2017, pp. 1894-1898, 2017.

[7] 干場功太郎, 若林瑞保, 鷲崎海, 石木隆洋, 公文誠, Daniel Gabriel, 中臺一博, 奥乃博, “UAV 搭載マイクロホンアレイを用いた組み込みシステムによる音源探査性能の評価”, 第 35 回日本ロボット学会学術講演会, pp. 1-4, 2017.

[8] R. O. Schmidt, “Multiple emitter location and signal parameter estimation”, IEEE Transactions on Antennas and Propagation, Vol. 34, No. 3, pp. 276-280, 1986.

[9] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa and H. Tsujino, “Design and Implementation of Robot Audition System ’HARK’ – Open Source Software for Listening to Three Simultaneous Speakers”, Advanced Robotics, Vol. 24, No. 5-6, pp. 739-761, 2010.

一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-3 (11/25)

マイクロホンアレイを有するマルチロータヘリコプタを用いた地上の複数音源の位置推定について

Position Estimation of Multiple Sound Sources on the Ground by Multirotor Helecopter with Microphone Array

若林瑞保 公文誠

Mizuho WAKABAYASHI, Makoto KUMON

熊本大学

Kumamoto University

wakabayashi@as.mech.kumamoto-u.ac.jp

Abstract

マルチロータヘリコプタは，環境認識のために様々なセンサを搭載し，実用的なタスクへの活用が期待されている．音情報はそのようなタスクを実行するうえで有用な情報の一つであり，本論文では，ヘリコプタで測定された信号を解析し音源の位置を推定することを検討する．ロータによって発生するノイズは目的信号を歪ませるので，推定される音源方向は不確かである．さらに，ヘリコプタは空中から幅広い範囲の複数の信号を受信するため，それらの情報を統合することは困難である．本論文では，誤検出が存在する下でデータアソシエーション手法を用いて複数の音源の検出を試みる．本論文では，開発したシステムが実際の飛行実験を通してリアルタイムで約 3m の精度で複数の音源を定位できることを示す．

1 はじめに

近年，無人航空機は計測や救難，物流など様々なタスクにおいて実用化への取り組みがなされている．より高度な無人航空機を実現するには，与えられたタスクを自律的に達成することが期待されており，無人航空機に搭載されたセンサを用いて地上の物体や事象を観測し環境を認識する必要がある．カメラ画像やレーザスキャナによる空間形状等の認識に関する研究はすでに多くなされているが，本論文では，要救助者の呼び声のような地上音源の検出について考えるものとする．

このようにロボットによる音環境理解は，Okuno, Nakadai[Nakadai 00] によって提案されたロボット聴覚として研究されている．この一例として，Sasaki[Sasaki 06] は，大型のマイクロホンアレイを用いて複数の音源を三角測量によって定位することを提案している．Kumon[Kumon 11] は，音源定位のための拡張カルマンフィルタの相関行列に基づいてモバイルカートの制御法を提案している．無人航空機については Basiri[Basiri 12] が小型の固定翼機に 3 つのマイクロホンを搭載し，地上にある音源や周囲を飛行する他の無人航空機をパーティクルフィルタを使って定位することを提案している．無人航空機の場合，対象音源について水平方向（以下方位角）だけではなく垂直方向（以下仰角）を含めて，三次元的に音源位置を推定する必要がある点は注意が必要である．

無人航空機の中でもマルチロータヘリコプタは特に注目を集めており，盛んに研究されている飛行プラットフォームである．一般的なマルチロータヘリコプタは固定ピッチのロータを組み合わせたシンプルな構造ながら，垂直離着陸，ホバリング，低速飛行が可能で，地上音源を観測に好適である．例えば，安定したホバリング飛行を行い，音情報を観測することで音源方向が得られれば，その際の機体位置と適当な仮定の下で音源位置を算出することができる．先行研究として，Okutani[Okutani 12] と Ohata[Ohata 14] は，雑音情報を逐次的に推定し音源方向を推定する MUSIC(MUltiple SIgnal Classification) 法 [Schmidt 86] によって地上音源の位置を推定することを提案している．また Washizaki[Washizaki 16] は飛行中に得られる音源方向情報を統合して音源位置を三次元的に推定する方法，著者ら [若林 16][山田 17] は推定された音源に接近してその位置をより正確に推定する方法などを提案している．

ところで，ロータの発生する騒音（以下，エゴノイズと呼ぶ）は大きなパワーを有する非定常な信号で，目標とする地上からの音信号を認識する上で障害となる．このため，音源方向の推定は不確かであり，誤検出や未検出などを含め誤った推定をする可能性がある．また，上空から収録する方法では，音源の探索範囲が広くなるため複数の地上音源の信号を観測することになる．従って，マルチロータによる地上音源の位置推定では，音源方向情報の

不確かさと，複数の対象が存在することへの配慮が必要となる点に特徴がある．

このような問題では，ロバストなデータアソシエーションを行う必要がある．不確かさのある信号でのデータアソシエーション手法には，例えば MHT（Multi Hypothesis Tracking）[Kim 15] や JPDA（Joint Probability Data Association）[Rezatofighi 15] などがある．本研究では，エゴノイズの影響で誤検出が多く発生し，このような環境での MHT では膨大な数の仮説を維持するため，計算量が増加しリアルタイムでの処理が困難となる．さらに，データアソシエーションが多く使われている画像やレーザースキャナを用いた物体の追跡に比べ，観測は不確かであり，不規則に得られる．そのため本論文では，計算量が少なく，複数の対象も追跡可能な GNN（Global Nearest Neighbor）法 [Konstantinova 03][Ozaki 12] を適用することを提案する．

また，手法を実現する上で，マルチロータヘリコプタの限られたペイロードや通信帯域，電力などの制約を考慮する必要がある．本研究では，著者らがこれまでに開発してきたマイクロホンアレイ搭載型マルチロータヘリコプタ [Kumon 14][Ishiki 15] を改良し，提案手法を実時間で計算可能なシステムを開発し，実際の飛行実験を通じて，提案手法が実現可能であることを示し，さらに得られた定位性能についても評価を行う．

2 マイクロホンアレイ搭載マルチロータヘリコプタ

2.1 マイクロホンアレイによる音源方向推定

マイクロホンアレイで収録したマルチチャンネルの音信号に対して，音源から各マイクまでの音の伝達特性に基づいて音到来方向を推定する方法の一つに MUSIC 法 [Schmidt 86] がある．MUSIC 法は信号空間と雑音空間の直交性を利用した部分空間法に基づいた推定法で，一定の構造の雑音に対してロバストな手法として知られており，本研究でも MUSIC 法によって音源方向の推定を行う．ここで，マルチロータヘリコプタでの収録信号においては，エゴノイズが非定常で大きな信号であること，目的音がロータに比べて遠方で低 SNR 条件であることから，十分な推定精度を得られない場合があることに注意が必要である．

空中から地上音源を探査する場合，地上で音源探査する場合とは異なった特性があることも留意しなければならない．空中にあるマイクロホンアレイと地上音源の間には通常遮蔽はなく，また屋外では残響も無視できるため，比較的シンプルな音伝達特性を想定できる．同時に，地上の広い範囲からの音信号の到来が考えられるので，目的音以外の音源からの音信号も到来する複数音源の場合を考える必要がある．上述の通り，エゴノイズの存在による低

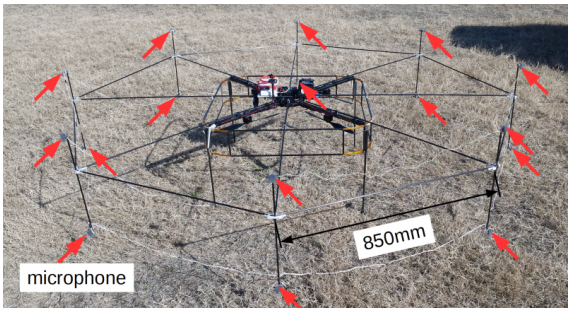


図 1: Quadrotor helicopter with microphone array

SNR 条件であることも加味すれば，音源方向が推定されたとしても，その情報は目的音か目的以外の音源なのか，あるいは誤検出であるかの全ての可能性を考えなければならない．

2.2 対象とするヘリコプタシステム

本論文は地上の複数音源の位置推定法を扱うものであるが，エゴノイズによる不確かさの程度によってその難しさが影響されるため，まず想定しているヘリコプタシステムについて説明する．

本研究では，エゴノイズを抑制する上でマイクロホンアレイの適切な構造や配置について著者らの先行研究（Kumon[Kumon 14] と Ishiki[Ishiki 15]）を元に，音源の方位角と仰角を測定出来るようマイクロホンが立体的な配置にできるよう拡張したものを用いる．関連研究では，Nakadai[Nakadai 17] と Hoshiba[Hishiba 17] は球形に密集させたマイクロホンアレイを用いているが，こちらも音源の方位角と仰角を測定出来るので，前者のマイクロホンアレイと同様に扱うことができる．マイクロホンからの音響情報は，オンボードの音響処理ユニットである RASP-ZX（System Infrontier 社）[音響処理装置（RASP-ZX） n.d.] で収録され，地上局に 2.4GHz の無線 LAN で送信される．詳細は筆者らの先行研究 [Washizaki 16] を参照されたい．

マイクロホンアレイを搭載したヘリコプタ (enRoute 社 PG-560 [ZION-PG560-SPECS 16]) を図.1 に示す．ヘリコプタはフライトコントローラ (Pixhawk[Pixhawk 17]) によって GPS による自動飛行が可能で，機体の位置・姿勢情報をリアルタイムに取得することができる．本研究では MAVROS[ROS.org - mavros 17] の ROS ノードを機体情報のインターフェースとして基地局で扱い，音響データとともに後述する音源位置アルゴリズムによってリアルタイムに位置推定するよう ROS ノードと接続される．また，処理の結果は地上局で図.3 のように表示される．

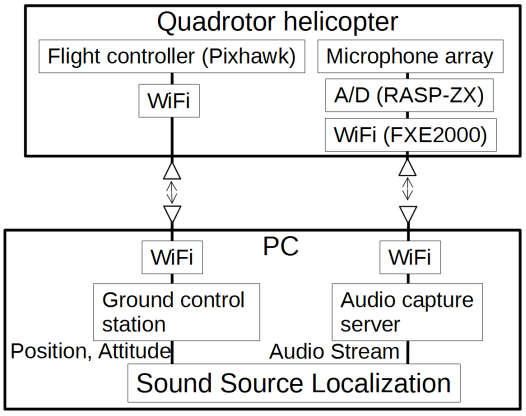


図 2: Structure of UAV based acoustic data capturing system

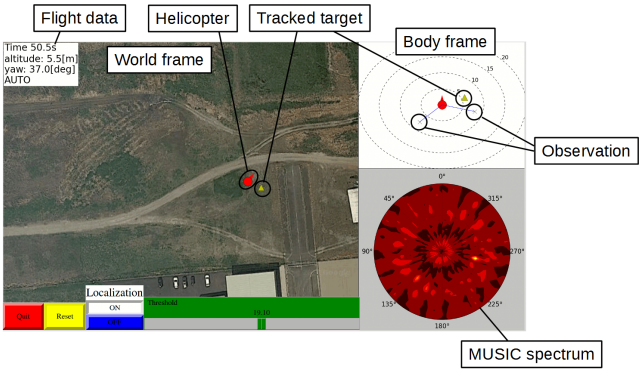


図 3: Screenshot of the developed sound localization system displayed at ground station

3 複数音源位置推定

3.1 カルマンフィルタによる音源位置推定

音源定位には，測定値の不確かさを考慮するためにカルマンフィルタ [Welch 01] に基づく位置推定を用いる．図.5 に示すように，機体の高度，ヨー角，音源の方位角，仰角をそれぞれ h，ψ，θ，ϕ とする．ここで，h，ψ，ϕ に不確かさがあると仮定し，$h = \tilde{h} + \delta h$ のように観測値 $\tilde{h}$ と不確かさ δh に分けて考える．さらに地表面が不整地である場合などによる不確かさ $\delta \boldsymbol{g}$ を考える．不確かさの信号が十分に小さいと仮定すると，時刻 k に m 個の観測が得られその j 番目 $(j = 1, 2 \cdots m)$ の観測から機体座標における音源位置 $\boldsymbol{y}_{k,j}$ は以下のように算出される．

$$\boldsymbol{y}_{k,j} = (\hat{h}_k + \delta h) \tan(\hat{\phi}_k + \delta\phi)\boldsymbol{b}_k + \delta \boldsymbol{g}_k, \tag{1}$$

ここで $\boldsymbol{b}_k = \left[\sin(\psi_k - (\hat{\theta}_k + \delta\theta)) \;\; \cos(\psi_k - (\hat{\theta}_k + \delta\theta))\right]^T$．$\delta h$，$\delta\phi$，$\delta\theta$，$\delta \boldsymbol{g}$ がそれぞれ $\mathcal{N}(0, \sigma_h)$，$\mathcal{N}(0, \sigma_\phi)$，$\mathcal{N}(0, \sigma_\theta)$，$\mathcal{N}(0, \Sigma_{\boldsymbol{g}})$ の正規分布に従うとすると，(1) 式は $\boldsymbol{y}_{k,j} = \boldsymbol{H}\hat{\boldsymbol{x}}_{k,j} + \boldsymbol{a}_k$ と書ける．ただし，$\hat{\boldsymbol{x}}_{k,j} = \boldsymbol{b}_k \hat{h}_k \tan\hat{\phi}_k$ は観測によって算出される音源位置で，$\boldsymbol{a}_k$ は $\sim \mathcal{N}(0, \boldsymbol{P})$ の正規分布に従い，$\boldsymbol{H}$ は観測行列であり，ここでは単位行列である．共分散行列 $\boldsymbol{P}$ は図.5 の幾何学的関係より算出される．

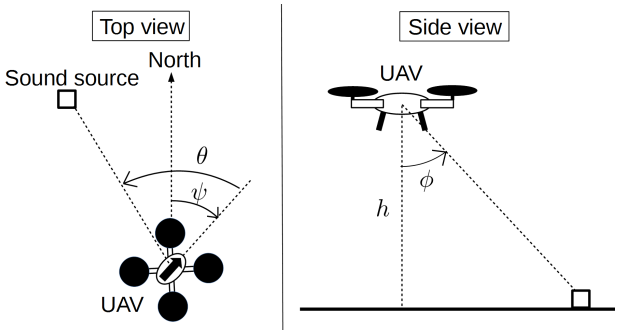


図 4: Geometric relationship between sound source position and its estimated direction

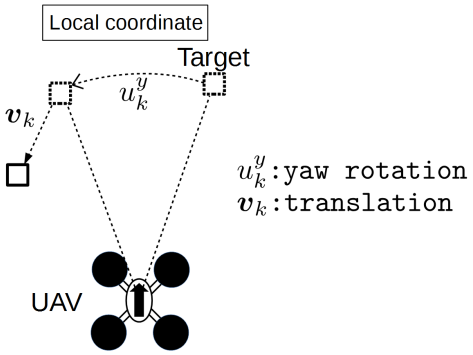


図 5: Update of the estimated target with UAV flight

機体座標で表現される音源位置は機体の位置・姿勢の変化によって変化する．ここで，絶対座標における機体の平行移動とヨー角の変化量をそれぞれ $\boldsymbol{u}_k^t$，u_k^y とする．時刻 $k-1$ から k において機体座標での音源位置は $\boldsymbol{x}_{k-1}$ から $\boldsymbol{x}_k$ に変化し，その移動ベクトルを $\boldsymbol{v}_k$ とする．$\boldsymbol{v}_k$ には不確かさがあるとすると，時刻 k における i 番目 $(i = 1, 2 \cdots n)$ の音源 $\boldsymbol{x}_{k,i}$ は以下のようにかける．

$$\boldsymbol{x}_{k,i} = \boldsymbol{R}(\hat{u}_k^y)\boldsymbol{x}_{k-1,i} + \hat{\boldsymbol{v}}_k + \delta\boldsymbol{v}, \tag{2}$$

ただし，$\boldsymbol{R}(\hat{u}_k^y)$ はヨー周りの回転角 $\hat{u}_k^y$ だけ回転させる回転行列を表す．$\delta\boldsymbol{v}$ が $\mathcal{N}(0, \boldsymbol{Q})$ の正規分布に従うと仮定すると，(2) 式は $\boldsymbol{x}_{k,i} = \boldsymbol{R}(\hat{u}_k^y)\boldsymbol{x}_{k-1,i} + \hat{\boldsymbol{v}}_k + \boldsymbol{c}_k$ と書ける．ただし，$\boldsymbol{c}_k \sim \mathcal{N}(0, \boldsymbol{Q})$．上記のカルマンフィルタで用いる分散は全て経験的に設定される．

上記のモデルを用いて，音源 i と観測 j のカルマンフィルタの予測ステップと更新ステップは以下のように書ける．

1. 予測ステップ

$$\bar{\boldsymbol{x}}_{k,i} = \boldsymbol{R}(\hat{u}_k^y)\boldsymbol{x}_{k-1,i} + \hat{\boldsymbol{v}}_k \tag{3}$$

$$\bar{\boldsymbol{\Sigma}}_{k,i} = \boldsymbol{Q} + \boldsymbol{R}(\hat{u}_k^y)\boldsymbol{\Sigma}_{k-1,i}\boldsymbol{R}(\hat{u}_k^y)^T \tag{4}$$

2. 更新ステップ

$$\tilde{\boldsymbol{y}}_{k,ij} = \boldsymbol{y}_{k,j} - \boldsymbol{H}\boldsymbol{x}_{k,i} \quad (5)$$

$$\boldsymbol{S}_k = \boldsymbol{H}\bar{\boldsymbol{\Sigma}}_{k,i}\boldsymbol{H}^T + \boldsymbol{P} \quad (6)$$

$$\boldsymbol{K}_k = \bar{\boldsymbol{\Sigma}}_{k,i}\boldsymbol{H}^T\boldsymbol{S}_k^{-1} \quad (7)$$

$$\boldsymbol{x}_{k,i} = \bar{\boldsymbol{x}}_{k,i} + \boldsymbol{K}_k\tilde{\boldsymbol{y}}_{k,ij} \quad (8)$$

$$\boldsymbol{\Sigma}_{k,i} = \bar{\boldsymbol{\Sigma}}_{k,i} - \boldsymbol{K}_k\boldsymbol{H}\bar{\boldsymbol{\Sigma}}_{k,i} \quad (9)$$

3.2 データアソシエーション

複数の音源が存在する環境下で音源定位を実現するには，追跡音源と観測を対応付けるデータアソシエーションが必要となる．本論文では，GNN（Global Nearest Neighbor）法 [Konstantinova 03][Ozaki 12] を用いている．

まず，追跡音源と観測の誤対応を減らすために，それぞれの追跡音源に以下で説明するような有効領域を導入する．観測 j が追跡音源 i の有効領域の中に存在する場合，追跡音源 i はこの観測 j を用いて更新される．追跡音源 i と観測 j の間のマハラノビス距離は (7) 式の観測残差と，(8) 式の共分散行列を用いて

$$d^2_{k,ij} = \tilde{\boldsymbol{y}}^T_{k,ij}\boldsymbol{S}^{-1}\tilde{\boldsymbol{y}}_{k,ij} \quad (10)$$

と計算でき，これを元に有効領域として閾値 G を定義する．

$$d^2_{k,ij} = \tilde{\boldsymbol{y}}^T_{k,ij}\boldsymbol{S}^{-1}\tilde{\boldsymbol{y}}_{k,ij} < G. \quad (11)$$

(11) 式が満足する場合，追跡音源 i に観測 j が割り当てられる．閾値 G はマハラノビス距離が χ 二乗分布に従うことから，2 自由度の χ 二乗分布表から決定される．本論文では，追跡音源の 95 %信頼楕円を有効領域とするために $G = 5.991$ と設定した．

一つの追跡音源の有効領域に一つの観測が存在する場合は直ちに割り当てることができる．しかし，一つの追跡音源の有効領域に複数の観測が存在する場合や，一つの観測が複数の追跡音源の有効領域の内部にある場合，直ちに割り当てることは困難である．時刻 $k-1$ に n 個の音源を追跡しているときに得られる観測は，現在追跡中の音源からの観測か，新たな音源からの観測のどちらかである．ここで，時刻 k に n 個の音源を追跡しており，m 個の観測が得られたと仮定する．(11) 式を満足しているかどうか，追跡音源と観測のすべての組み合わせで検証し，以下のコスト行列によって与えられる割り当て問題を解く．

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix} \quad (12)$$

$$c_{ij} = \begin{cases} E & d^2_{ij} > G \\ d^2_{ij} & d^2_{ij} < G \end{cases} \quad (13)$$

このコスト行列で，追跡音源と観測の組み合わせでマハラノビス距離の和を最小にする解を求める．本論文では，Munkres 法 [Munkres 57] を用いてこの問題を解いている．ただし，式 (13) の定数値 E はこの問題を距離の和が最小となる組み合わせを求める問題とするために，十分に大きな値を設定する．

3.3 追跡音源の管理方法

安定した音源の追跡のために，追跡音源の管理方法を導入する．

(a) 新規追跡音源の生成

追跡している音源が存在しない場合や，観測が現在追跡中の音源に割り当てられなかった場合，新たな追跡音源が生成され，カルマンフィルタによる追跡が開始される．このプロセスは，安定した観測が得られる場合に実行する必要があり，通常は多くの追跡音源が生成されてしまうので，観測が一定回数（N_1 とする）以上連続して得られた場合のみ追跡を継続する．さらに追跡音源に観測が十分な回数（ここでは N_1 よりも多い N_2 を設定する）得られた場合，その追跡音源は正しい音源とラベル付けされる．

(b) 追跡中の音源

音源は観測が得られている間は追跡し続けるが，機体が音源から離れた時や，音が鳴らなくなった場合など長時間観測が得られない場合，機体の移動による不確かさが積み重なり，当該の音源の有効領域が拡大し続けてしまう問題がある．これを避けるために，観測が正当とラベル付けされた音源がさらに十分な回数（N_3 回）観測されたにも関わらず，T_1(sec) の間に観測が得られなかった場合，カルマンフィルタによる追跡を停止し，その際の追跡音源の情報を保存して，更新を停止する．この場合，有効領域は半径が一定値 r(m) の円と仮定する．再び観測が得られた場合は，保存状態からカルマンフィルタによる追跡を再開する．

(c) 追跡の終了

観測が得られた回数が，N_1 以上 N_3 以下で T_2(sec) の間に観測が得られない場合，追跡を終了する．

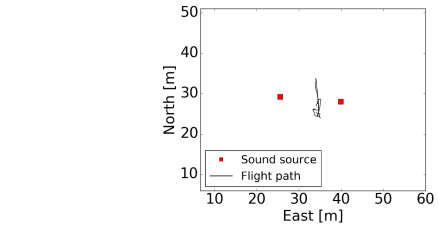
4 検証

提案システムを検証するために，飛行実験を行った．

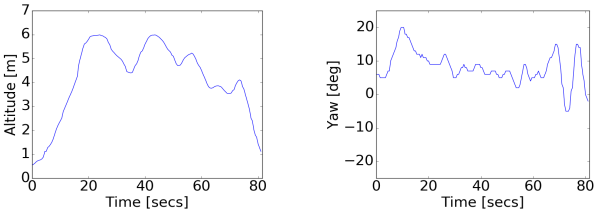
4.1 実験設定

マイクロホンアレイを搭載したクアッドロータヘリコプタをマニュアル操縦とオートパイロットにより飛行させる．この実験では，オンボードの制御装置による姿勢安定化によってアシストされるマニュアル操縦によるホバリングと，オートパイロットによるウェイポイントの追従を行った．それぞれのフライトは，機首方向を北に向け，高度を地上から約 5m に維持した．

図 6: Photo of the experiment



(a) Flight path and position of the target



(b) Altitude (c) Yaw

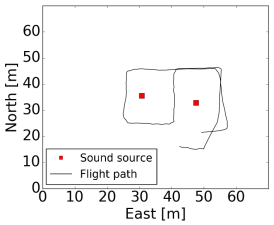
図 7: UAV state (1st flight)

音源には，人の叫び声を発するスピーカと，笛の音を使用した．スピーカは地上に設置したが，笛は人が立って吹いているので，音源は地上から約 1.6m の高さに位置している．マニュアル操縦によるホバリング飛行時は，スピーカを西側に，笛を東側に，それぞれ機体のホバリング位置から約 10m 離れた位置に設置した．オートパイロットによる飛行時は，スピーカを西側に，笛を東側に，音源間の距離が約 20m となるように設置し，飛行経路は機体がそれらの音源の周囲を飛行するよう設定した．図.6 に実験の状況を示している．

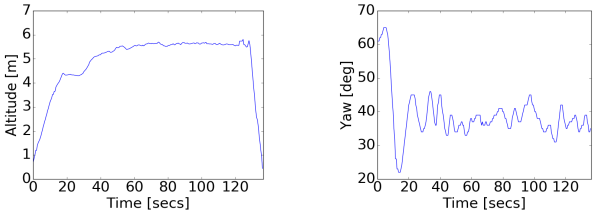
4.2 結果

4.2.1 フライト結果

飛行経路を図.7(a) と図.8(a) に示す．図中の赤い四角が音源の位置を示している．次に，飛行高度を図.7(b) と図.8(b) に示す．ホバリングでは約 80 秒間，オートパイロットでは約 130 秒間飛行している．ホバリング飛行はマニュアル操縦であったため高度にばらつきがある．オートパイロットでは，機首方向が北向き（0 度）になるように設定したが，実際は北向きから 40 度ずれた結果になっている．この偏差は，(2) 式の観測モデルで機体のヨー方向を考慮しているのでこの実験では問題とならない．



(a) Flight path and position of the target



(b) Altitude (c) Yaw

図 8: UAV state (2nd flight)

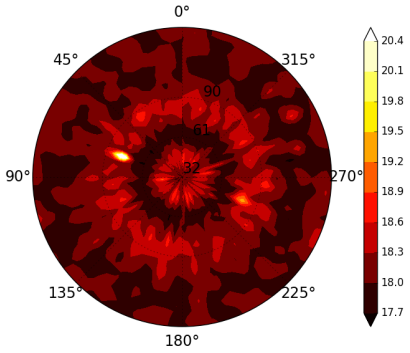
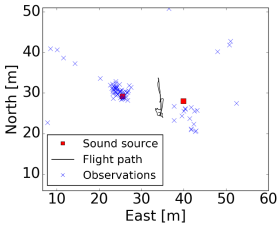


図 9: Example of MUSIC spectrum when a source was found (yellow peak)

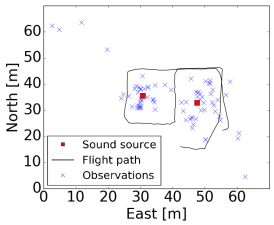
4.2.2 複数音源の定位

音源定位は，MUSIC 法を 0.5 秒周期でリアルタイムに計算して音源方向を推定し，この情報と機体の位置・姿勢を統合して実施する．また，離陸中や着陸中はロータの雑音が大きくなり，誤検出が多く発生してしまうので，GNN 法による定位は高度 3.5m 以上の時のみ実行されるよう設定した．3.3 章で説明されているパラメータは，$N_1 = 2$, $N_2 = 4$, $N_3 = 6$, $T_1 = 10$sec, $T_2 = 10$sec, $r = 1.5$m とした．

図.9 に機体を真上から見て円周方向を方位角，半径方向を仰角とした MUSIC スペクトルを示す．この図では，80 度方向と 260 度方向に MUSIC スペクトルのピークが立っているのがわかる．本論文の目的は，複数の音源を定位することであるため，MUSIC スペクトルの複数のピークを観測として用いる．観測として用いられるピークの数は，MUSIC の処理で仮定する音源数と等しく設定しており，本実験では 3 としている．さらに誤検出を減らすために閾値を設定し，パワーの小さいピークは観測として用いない．この閾値は，経験的に値を設定しており，この

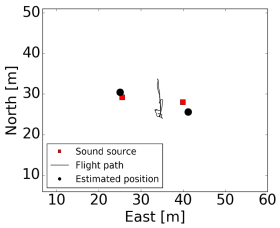


(a)First flight

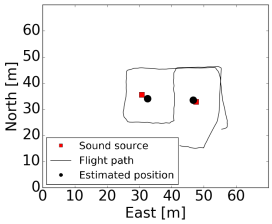


(b)Second flight

図 10: Observations projected on the ground



(a)First flight



(b)Second flight

図 11: Final estimated sound source position

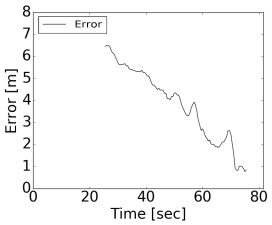
実験では 19.1 とした．

図.10 にそれぞれの飛行での観測を示す．図.10(a) ではスピーカ付近で密集した観測が得られているが，笛付近の観測はばらつきがあり実際の位置とはわずかにずれた位置に得られている．さらに，この図では左下と右下に誤検出があることがわかる．図.10(b) では，機体が移動しながらの観測であったため観測にばらつきがあり，左上と右下に誤検出も見られるが，実際の音源位置の周囲に多く観測が得られていることがわかる．

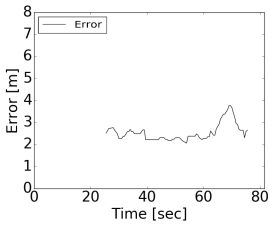
図.11 には各飛行終了時のカルマンフィルタによる推定音源位置を黒い丸で示しており，両方の飛行で正解音源位置に近い結果が得られているのがわかる．

図.12 と図.13 には追跡音源の推定位置の誤差を示している．図.12(a) では最初の観測では誤差が大きかったが，音源に近づくにつれて誤差が減少していく結果となった．図.13 の上部に現れている誤差は誤検出によるものであり，3 章で提案された追跡音源の管理方法 (c) により誤検出と判断し途中で削除されている．図.13 で誤差が一定の区間があるのは追跡音源の管理方法 (b) によるもので，カルマンフィルタによる予測を停止し，絶対座標で保存された位置を用いてデータアソシエーションによるマッチングを行っているためである．図.14 には飛行中に得られた誤検出を示している．これらの誤検出は提案システムにより誤りだと判断され削除されている．

表 1 に各飛行後の推定音源位置の誤差を示す．結果として誤差は 3m 以内に収まった．提案した音源定位法と音響データ伝送システムにより，複数音源の定位が実用的であると結論付けることができる．

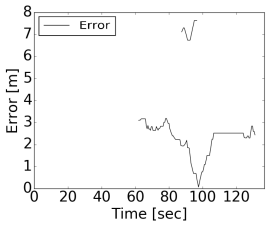


(a)Speaker

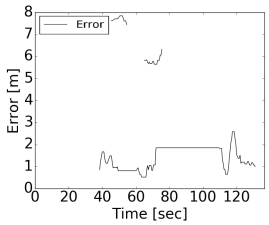


(b)Whistle

図 12: Position estimation error (1st flight)



(a)Speaker



(b)Whistle

図 13: Position estimation error (2nd flight)

表 1: Estimation error of the targets

Flight	Sound source	Error[m]
First	Speaker	0.84
First	Whistle	2.65
Second	Speaker	2.44
Second	Whistle	1.00

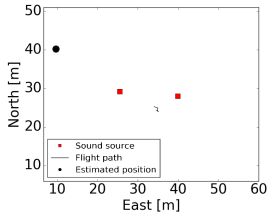
5 おわりに

本論文では，マイクロホンアレイを搭載したクアッドロータヘリコプタで得られた音響情報を用いて地上の複数の音源の位置推定を行った．各追跡音源の有効領域は追跡音源と観測の誤対応を防ぐことに有効である．

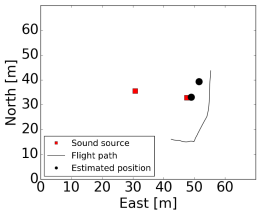
ホバリング飛行と自律飛行による実験を通して性能評価を行った．機体が音源から遠く離れていた場合，誤差は大きかったが，目標に近づくと誤差は 3m 以内にまで減少した．本論文での機体の飛行経路は，音源の周囲を飛行するように指定されていたが，実際に音源を探査する環境では音源の位置は不明であり，そのような飛行経路を指定することは不可能である．今後の課題として，音源が機体から離れた場所にある場合，音源に近づくことで正確な音源位置を得るために，推定された音源位置から音源に近づくような経路を機体に送信するシステムを開発することが挙げられる．

謝辞

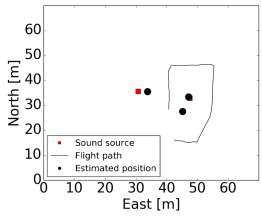
本研究の一部は，総合科学技術・イノベーション会議により制度設計された革新的研究開発推進プログラム（ImPACT）タフ・ロボティクス・チャレンジならびに科学研究費補助金 17K00365 の助成を受けたものです．



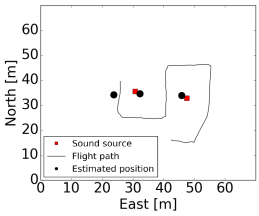
(a) Case1 in First flight



(b) Case2 in Second flight



(c) Case3 in Second flight



(d) Case4 in Second flight

図 14: False positives detected during the experiments

参考文献

[Nakadai 00] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, “Active audition for humanoid.” in AAAI/IAAI, H. A. Kautz and B. W. Porter, Eds. AAAI Press / The MIT Press, 2000, pp. 832-839.

[Sasaki 06] Y. Sasaki, S. Kagami, and H. Mizoguchi, “Multiple sound source mapping for a mobile robot by self-motion triangulation” in Proceedings of 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006, pp. 380-385.

[Kumon 11] M. Kumon and S. Uozumi, “Binaural localization for a mobile sound source” Journal of Biomechanical Science and Engineering, vol. 6, no. 1, 2011, pp. 26-39.

[Basiri 12] M. Basiri, F. S. Schill, P. Lima U., and D. Floreano, “Robust acoustic source localization of emergency signals from micro air vehicles, ” in Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 4737-4742.

[Okutani 12] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, “Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter.” in Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 3288-3293.

[Ohata 14] T. Ohata, et al. “Improvement in outdoor sound source detection using a quadrotorembedded microphone array” in Proceedings of 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2014, pp. 1902-1907.

[Washizaki 16] K. Washizaki, M. Wakabayashi and M. Kumon, “Position Estimation of Sound Source on Ground by Multirotor Helicopter with Microphone Array,” Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2016, pp. 1980-1985.

[若林 16] 若林瑞保, 鷲崎海, 公文誠, “クアッドロータヘリコプタを用いた音源探査”, 第 34 回日本ロボット学会学術講演会論文集, 2016, RSJ20161C3-04.

[山田 17] 山田健志郎, 公文誠, “Grid based Recursive Bayes Filter に基づくマルチロータヘリコプタによる音源探査における地図管理,” 第 35 回日本ロボット学会学術講演会論文集, 2017, RSJ2017AC3AC2-06.

[Rezatofighi 15] S. H. Rezatofighi, et al. “Joint Probabilistic Data Association Revisited” in ICCV. IEEE, 2015, pp. 3047-3055.

[Kim 15] C. Kim, et al. “Multiple Hypothesis Tracking Revisited” in ICCV. IEEE, 2015, pp. 4696-4704.

[Kumon 14] M. Kumon and T. Ishiki, “A microphone array configuration for an auditory quadrotor helicopter system,” in Proceesings of the 12 th IEEE International Symposium on Safety, Security and Rescue Robotics, 2014, p. 34.

[Ishiki 15] T. Ishiki and M. Kumon, “Design model of microphone arrays for multirotor helicopters.” in IROS. IEEE, 2015, pp. 6143-6148.

[Nakadai 17] K. Nakadai, et al. “Development of Microphone-Array-Embedded UAV for Search and Rescue Task” in Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017, pp. 5985-5990

[Hishiba 17] K. Hoshiba, et al. “Design of UAV-embedded Microphone Array System for Sound Source Localization in Outdoor Environments” in Sensors, 17(11), 2017, doi:10.3390/s17112535.

[ZION-PG560-SPECS 16] enRoute Inc., “ZION-PG560-SPECS.” Available: https://enroute1.com/portfolio-posts/zion-pg-700/zion-pg560-specs/

[Pixhawk 17] 3D Robotics Inc., “Pixhawk.” Available: https://store.3dr.com/t/pixhawk

[ROS.org - mavros 17] “ROS.org - mavros” Available:http://wiki.ros.org/mavros

[音響処理装置（RASP-ZX） n.d.] System Infrontier Inc., “音響処理装置（RASP-ZX）.” Available: http://www.sifi.co.jp/system/modules/pico/index-.php?content_id=36

[Washizaki 16] K. Washizaki, M. Wakabayashi and M. Kumon, “Position estimation of sound source on ground by multirotor helicopter with microphone array.” in Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2016, pp. 1980-1985.

[Schmidt 86] R.Schmidt, “Multiple emitter location and signal parameter estimation” IEEE Transactions on Antennas and Propagation, vol.34, no.3, 1986, pp276-280.

[Welch 01] G. Welch, G. Bishop, “An Introduction to the Kalman Filter” in SIGGRAPH Course Notes, ACM, 2001, Course 8.

[Konstantinova 03] P. Konstantinova, A. Udvarev, T. Semerdjiev, “A Study of a Target Tracking Algorithm Using Global Nearest Neighbor Approach” in CompSysTech, 2003, pp. 290-295.

[Ozaki 12] M. Ozaki, K. Kakinuma, M. Hashimoto. and K. Takahashi, “Laser-based pedestrian tracking in outdoor environments by multiple mobile robots” in Sensors, Vol. 12 2012, pp. 14489-14507.

[Munkres 57] J. Munkres, “Algorithms for the Assignment and Transportation Probrems” in Journal of the Society for Industrial and Applied Mathematics, Vol. 5, No. 1 1957, pp. 32-38.

一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-4 (11/25)

外来種ソウシチョウが在来種の歌行動へ与える影響を探る：マイクロフォンアレイを用いた森林性鳥類の観測実例

Exploring the effect of invasive Red-billed Leiotrix(*Leiothrix lutea*) on songs of native birds: An example of observing forest birds using microphone arrays

松林志保[1]・斉藤史之[2]・鈴木麗璽[3]・千葉尚彬[4]・中臺一博[5]・奥乃博[6]

[1]大阪大学大学院工学研究科付属オープンイノベーション教育研究センター
[2]いであ株式会社大阪支社生態保全部
[3]名古屋大学大学院情報学研究科
[4]名古屋大学大学院情報科学研究科
[5]東京工業大学工学院システム制御系、Honda Research Institute Japan Co., Ltd.
[6]早稲田大学理工学術院創造理工学研究科

要旨

本稿は、マイクロフォンアレイとロボット聴覚ソフトウエア HARK を用いた森林性鳥類の観測実例として「日本の侵略的外来種 100」に選定されるソウシチョウと近傍他種の歌行動に関する予備的調査結果を報告する。音の到来方向情報を伴った歌の始まりと終わりのタイミング情報から、ウグイス同士は同時に鳴くことを避ける重複回避行動を示すことが明らかになった。異種間ではヒヨドリとソウシチョウ間では歌行動に相互関係は見受けられなかったのに対し、ウグイスとソウシチョウ間では有意な重複回避行が確認された。種間による相互関係の違いは、周波数が近く、かつマイクロハビタットが似た種間では歌空間をめぐる競争が発生する可能性を示唆する。音源を 6 つのクラスに分けて試行した簡易音源分類ツールの総合的な分類精度は 76.1%、k 係数は 0.66（Substantial agreement）であり、分類エラーは鳥の歌行動の違いを反映した。

1 はじめに

鳥類の種類や位置などの正確なデータの蓄積は、生態学者や環境実務者、鳥類保護に携わる政策者にとって喫緊の課題である。鳥類生態学者にとって、ターゲットとする種の位置や数を正確に把握することは重要だが、野外調査は今もなお、目と耳による種類の同定や位置の推定が主流である。この基本的な観測技法の内、特に歌を歌う種に関しては、IC レコーダーの導入は、長時間にわたる観測を実現し、録音を後から再生できるためデータの再現性を高めた。しかし単一のマイクロフォンによる録音では歌の位置の推定には限界がある。この問題を解決するため、近年、複数のマイクロフォンで構成されるマイクロフォンアレイを用いて音の到来方向（Direction of arrival: DOA）を推定する技術を野外環境に応用する研究例が徐々に増えつつある[1]。マイクロフォンアレイ技術の鳥類観測への応用実例は、独自機材を用いた熱帯雨林に生息するムナジロカンムリアリドリの 2 次元位置推定[2]や、6 つの 2 チャンネル長期録音型録音機材[1]から成る 12 チャネルのアレイを用いたハシナガハチドリの個体間距離の算出[3]などがあげられるが、機材の入手および収録作業が一般の生態学者にとって容易ではないため、まだその利用は限定的である。

我々は野鳥の歌行動の情報収集を効率よく行うことを目的として、ロボット聴覚オープンソースソフトウェアである HARK（Honda Research Institute Japan Audition for Robots with Kyoto University）[4]、市販のマイクロフォンアレイとノート PC を用いた簡易な録音システム HARKBird[2]を構築し、比較的廉価で汎用性の高いデータ収集を試みている。3 つのマイクロフォンアレイと HARKBird を用いて揖斐川河川敷で実施したオオヨシキリの観測実験では、高頻度で定位された音源の位置と、実際に人間の観測者が記録した鳥の 2 次元空間内での位置分布パターンは高い類似性が示された[5]。また、定位された歌情報から得た各個体の総歌時間を、縄張り争いにおける個体間の優劣の指標と仮定した上で、観測した個体間の歌空間をめぐる優劣が示唆された[6]。この観測実験は、音の伝達環境のよい河川敷において定点で比較的長時間大声でさえずるオオヨシキリのような鳥類を対象とした場合の当該技術の高い実用性を実証した。

本研究は、観測地を遮蔽物が少ない河川敷から、多様な植物相が遮蔽物として音の減衰率を高める森林に移し、より実践的な環境で HARKBird の実用可能性を試した。観測対象としては日本の侵略的外来種 100」[7]にも選定されるソウシチョウ*(Leiothrix lutea)* と同所的に生息する在来周辺鳥類（鳴禽類）を採択した。

[1] SongMeter, SM2, Wildlife Acoustics, Inc, U.S.A.
[2] http://www.alife.cs.is.nagoya-u.ac.jp/~reiji/HARKBird/

ソウシチョウ（Figure 1[3]）は元来中国南部や東南アジアに生息し、その美しい鳴き声から日本では江戸時代には飼い鳥として輸入が始まり古くから親しまれていた。その分散由来には諸説あるが、1980 年代になると日本各地の山林でソウシチョウの生息及び繁殖が報告されるようになり、現在では特にニホンジカによって下層植生が衰退した森林内を中心に、主に九州から南関東にかけての広範囲で爆発的に数を増やしている[8][9]。九州地方の落葉広葉樹林では、採餌方法の違いから餌資源に関しては在来種との直接的な競争は少ない [10]との報告がある一方で、大きな歌声によりカケスなどの捕食者を誘引する可能性がある[11]など、在来鳥類への間接的な悪影響が懸念されている。しかしながら、その影響の実態は未解明な点が多い。

Figure 1　ソウシチョウ

本研究では、ソウシチョウが他鳥種へ与える影響の一例として歌行動に着目する。ウグイスなどソウシチョウと似通った生息環境に住み、かつ似通った周波数帯のさえずりを持つ種間では、より歌を伝達しやすい周波数帯をめぐる競争が行われ劣位の種の歌行動に変化があるのではないかという可能性を探る。複数のマイクロフォンアレイと HARKBird で 2 次元音源定位を、そして深層学習に基づく簡易な教師無し分類を試み、この生態学的な仮説検証のために音源定位技術がいかに貢献しうるかを検討した。

2 手法

2.1 録音方法

2017 年 7 月のほぼ無風の気象条件の下、複数のマイクロフォンアレイを用いて、林内で鳥の歌を夜明けから 3〜6 時間録音した。調査地は、標高約 300 m の落葉広葉樹林（室池園地・大阪府四條畷市）に位置する。元来は薪炭林であり、現在は混生するコナラ、リョウブ、ケヤキの林床に高さ約 0.5〜1 m の笹類を主とする下層植生が繁茂する（Figure 2）。調査地ではソウシチョウは留鳥で、その密度は一年を通して非常に高く年に複数回の繁殖が推定される。

鳥の歌の録音には、8 チャネル（TAMAGO-02 システムインフォロンティア社製）を 2 つ用いた。このマイクロフォンアレイの卵形のケースの胴体部分には各マイクロフォンが 45° 間隔で水平に配置され、24 ビット、16K Hz の集音が可能である。各マイクロフォンアレイを約 10m 離れた高さ約 1.5 m の三脚の上に設置し、USB ケーブルで野外調査用 PC(TOUGHBOOK, CF-C2,Panasonic)に接続して録音を行った（Figure 2）。また、各マイクロフォンアレイの位置を GPS で測定した。

Figure 2 観測風景

2.2 定位推定方法

各マイクロフォンアレイで収録した音声信号から方向・分離音を抽出するために、HARKBird[4] を用いて音源定位・分離を行った。まずそれぞれのマイクロフォンアレイにおいて、8 チャネルの音声信号を読み込み、短時間フーリエ変換によって得た各チャネルのスペクトログラムから MUSIC 法[12] を用いて音源定位を行った。次にその定位結果をもとに Geometric High order Decorrelation based Source Separation(GHDSS)法[13]を用いて各音源方向に対応した音源を分離する音源分離を行った。なお後述のように、今回の録音では周波数の高いセミの鳴き声が含まれていたためこれは定位せず、主な観測対象であるウグイス、ソウシチョウ、ヒヨドリの 3 種の鳥の歌の周波数帯を含む 1800~5000 Hz を定位対象とした。

2.3 2 次元統合方法

鳥の歌を録音後、人間による直接観測結果と比較するため、各マイクロフォンアレイで計測した音の到来方向を統合することで音源の 2 次元空間内での位置を推定した。具体的には、2 つのマイクロフォンアレイで同時間帯に定位され、かつ後述する簡易音源分類ツールで同一の音源の種類と判定された音源のペアに関し、各マイクロフォンアレイから各定位音源の方向に伸びる半直線の交点を 2 次元定位置と推定した。定位時刻に関しては、定位位置に近いマイクロフォンアレイによる音源を定位時刻として採用した。

2.4 人間による鳥類観測方法と歌情報の抽出

録音開始と同時に、録音機材の付近に立った人間の観測者が、周辺でさえずる鳥の種類や数、大まかな位置、歌い初めと終わりの時間を 5 分ごとに分単位で記録した。鳥種の識別や位置の推定は目視と耳による観測に基づく(Figure 3)。次に録音の中で複数種が鳴きあう時間帯の録音 1 時間分に着目し、音声分

[3] 四條畷市室池園地にて衣川直美氏撮影

析ソフトウエア Praat[5]を用いて、鳥種および個々の歌の始まりと終わりのタイミングを手動によるアノテート作業で抽出した。種の判別や歌のタイミングに関する判断はスペクトログラム（声紋）と音声再生に基づく。

2.5 重複回避

手動で抽出した鳥の歌の始まりと終わり時間のタイミングを観測値とし、観測値に基づく歌の長さをランダムに入れ替えた時系列を 1000 回作成し計測した期待値を比較することで、各鳥の歌のタイミングが他個体の歌のタイミングに与える影響を調べた。分析には、統計分析フリーソフト R パッケージ Song Overlap - Null-model Generator for animal communication (SONG)[14]を用いて各鳥が近傍個体と同時に鳴くことを積極的に回避する場合（重複回避：p≧0.975）、反対に積極的に重複（p≦0.025）する場合、もしくは各鳥の歌うタイミングは近傍個体の歌のタイミングとは無関係（0.025<p<0.975）であるという関係性の統計的有意性を調べた。期待値算出の際には、各鳥種の歌の順番は固定した。これは観測されたウグイスが複数の歌のレパートリーを持ち、各歌は求愛や威嚇など異なった役割を果たすこと[15]、そしてその歌の順番には何かしらの行動学的意味（文脈）を持つ可能性があるためである。さらに、同種の複数個体が同時に鳴いていた場合、各歌がどの個体によって発せられたかは、定位された音の到来方向と人間による直接観測に基づく位置や歌のタイミング情報から総合的に判断した。

2.6 簡易音源分類

録音した音を人間が聞き返し、スペクトログラムでチェックするアノテート作業は、鳥の種類や歌のタイミングに関する分析に不可欠なステップではあるが、莫大な時間的コストがかかる。また、機械学習を実施する際の教師データを作成する際にも同様のコストがかかる。一方、録音データには分析対象である種や個体特有の特徴を持つ固有の歌が繰り返し出現するため、周辺の環境音に起因する多様なノイズや、数回偶発的に鳴くのみの種や個体に比べて定位がしやすい可能性がある。これまで、分離音源のスペクトログラムを Deep auto-encoder による教師なし学習を用いて特徴量を抽出し、2 次元平面上に落とし込むことで分析対象の種の歌を大まかにグルーピングする HARKBird の簡易音源分類機能はウグイスの歌の抽出に試行されている[16]。本研究ではこの機能を用いて、耳で分離音を聞き、さらに目で分離音のスペクトログラムを確認しながら音源を 6 つのクラス（ウグイス、ソウシチョウ、ヒヨドリ、セミ、セミと他鳥種の混合音、環境音などのその他の音源やセミで断片化された音源）に識別した。

2.7 簡易音源分類性能の検証

録音に用いたマイクロフォンアレイのうち 1 つが定位・分離・分類した 10 分間のデータに着目し、6 つのクラスへの分類結果と、分類音を再生し耳で確認した結果に基づいて作成したエラーマトリックスから、全体の精度(overall accuracy)と各クラスの分類エラーを算出した。分類エラーは過誤認(Commission error)および認識漏れ(Omission error)の双方を考慮した。また全体の精度はエラーマトリックス上で偶然による一致率(chance agreement)を勘案していないため、偶然ではない一致率として k 係数(Kappa coefficient)を以下の数式から算出した。

$$\widehat{k} = \frac{N\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k}(x_{i+} \times x_{+i})}{N^{2} - \sum_{i=1}^{k}(x_{i+} \times x_{+i})}$$

数式中 k は分類クラス数、x_{ii} は行iと列i のクラスの音源がそれぞれ観測された回数、x_{i+} と x_{+i} は行 $_i$ と列 $_i$ の各周辺合計、そして Nは総観測回数であり[17]、k 係数は 1 に近いほど一致率が偶然に起因しないことを示す[18]。

3 結果と考察

3.1 簡易音源分類・定位例

Figure 3 は約 3 時間分の分離音源の特微量を 2 次元平面上に表示したものであり、図中各色点は 6 つの分類クラスに対応する。ウグイス（JBWA）、ソウシチョウ（RBLE）、およびヒヨドリ（BEBU）の鳥 3 種は垂直座標の高いエリアから低いエリアへ順に位置すると同時に、セミおよびセミと各鳥種の混合音と

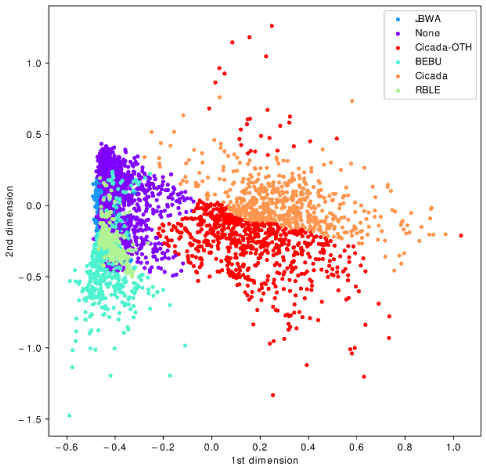


Figure 3　簡易音源分類一例

[5] Boersma, Paul & Weenink, David (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.34, retrieved 10 October 2017 from http://www.praat.org/

比べて水平座標において比較的低いエリアに分類された。

この簡易音源分類結果を反映させたマイクロフォンアレイのうちの 1 つが定位した音源のスペクトログラム（上パネル）、それぞれの音源の到来方向に対応した MUSIC スペクトラムおよびその分類結果（下パネル）を Figure 4 に示す。分類結果及び分離音の耳による確認により、図中はじめの約 20 秒間はセミ（Cicada）が鳴き、これと同時にソウシチョウ 1 個体がマイクロフォンアレイからおよそ -50 度（北東）方向で鳴いていることがわかる。このソウシチョウはこの位置で 1 分間に計 8 回鳴いたがそのうち 5 回は該当種に分類された。一方 1 回目の歌の一部は環境音として、2 回目はセミとの混声として、4 回目はヒヨドリと誤検出された。また、ウグイス 1 個体がおよそ 20 度（北北東）方向で 2 回鳴いたことも確認された。

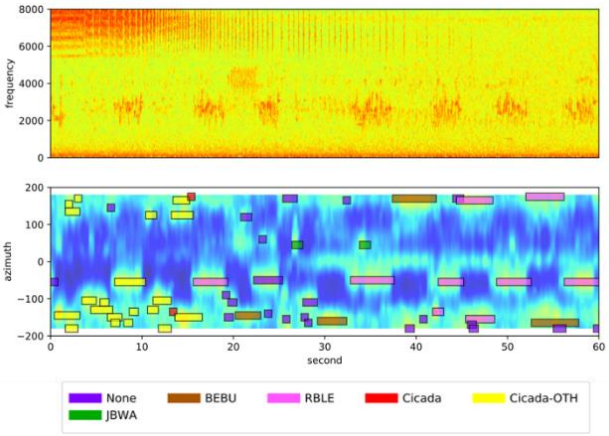


Figure 4 定位例（１０時から１分間の録音データの解析）

3.2 2 次元定位結果

録音に用いた 2 つのマイクロフォンアレイから得た 20 分間の情報を統合し、その定位音源の位置を 2 次元平面上に示したものを Figure 5 に、同時刻帯に

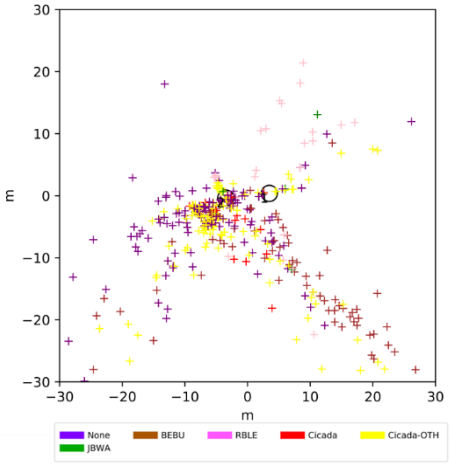


Figure 5 定位音源の 2 次元分布．図中黒丸はマイクロフォンアレイを示す．

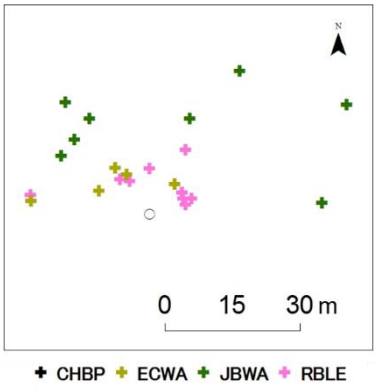


Figure 6 人間が観測した鳥の位置．図中黒丸は観測者に最寄りのマイクロフォンアレイを示す．

人間が観測した鳥の大まかな位置を Figure 6 に示す。定位結果は 20 分間にわたって歌毎に分析し続けた結果なので、全方向に多数の音源が観測された。ただしこれらの音源にはセミやウシガエル、周辺の環境音など鳥以外の音源が多数含まれる。一方人間が直接観測した結果は、当初分析の対象になり得ると期待した種や方向に注意して、歌っている鳥の位置を 5 分ごとに主に音に基づいて大まかにプロットしたものであるため、定位結果のようにすべての歌を詳細にカバーしたものではない。そのためこれらの図を直接比較することはできないが、鳥種ごとにいくつかの傾向が明らかになった。

まず観測対象のソウシチョウ（RBLE）に関しては、北方向およびマイクロフォンアレイ周辺で定位された様子は直接観測結果と類似した分布パターンを示した。これはソウシチョウの歌が比較的長く今回定位対象とした周波数帯とも合致すること、そして移動はしつつも定点でさえずり続け定位がしやすかったことが原因と考えられる。また南東方向では、フィールド調査中は観測対象とはしなかった一般種ヒヨドリ（BEBU）が定位されており、各マイクの定位結果からも存在が確認された。一方ウグイス（JBWA）に関しては、人間は観測したが定位確率は著しく低かった。これは、この観測時間帯ではウグイスが頻繁に移動しながら比較的遠方で控えめに鳴いていたため、その歌の「ケキョ」部分のみが短く聞こえがちで、2 つのマイクで同時に定位される確率が低かったためと考えられる。同様に、センダイムシクイ（ECWA）も短時間に移動を繰り返しながら単発の歌を歌っていたため人間は観測ものの 2 次元上では定位されなかった。また、分析に用いたパラメータ設定および周辺音環境下で観測された鳥の歌声の定位限界距離は 20 ~ 30 m 弱であった。この距離限界を超えると人間の耳では容易に識別できるコジュケイ（CHBP）の甲高く数分間繰り返される歌なども定位されなかった。

3.3 簡易音源分類精度の検証

10 分間の録音で定位された音源を 6 つのクラスに分類した結果と耳で確認した後に分類した音源のエラーマトリックスを Table 1 に、それぞれのクラスの

認識漏れおよび過誤認識エラーを Table 2 にまとめる。エラーマトリックスから得られる全体的な分類精度は 76.1%、偶然による一致を排除した k 係数は 0.66（Substantial Agreement）[18]であった。

Table 1　定位音源を 6 つのクラスに分類したエラーマトリックス

Class types determined from classified source	Class types determined from reference source					
	None	BEBU	RBLE	Cicada	Cicada-OTH	JBWA
None	134	23	6	0	0	13
BEBU	5	109	11	0	0	0
RBLE	10	7	24	0	1	5
Cicada	1	0	0	1	0	0
Cicada-OTH	3	0	1	0	15	1
JBWA	10	2	1	0	0	35

Table 2 各クラスの分類エラー

	Omission error	Comission error
None	17.8	23.9
BEBU	22.7	12.8
RBLE	44.2	48.9
Cicada	0.0	50.0
Cicada-OTH	6.3	25.0
JBWA	35.2	27.1

鳥以外の自然環境音や、セミで断片化された鳥の歌で構成されるクラス（None）の認識漏れおよび過誤認識エラーはともに他の鳥種に比べて低いことから、鳥の歌とその他の音源の識別は容易であったことがわかる。鳥 3 種のうち、ソウシチョウの両分類エラーが比較的高かった要因は、個々の歌の長さと抑揚の幅の広さ、そして複雑な歌構造が、セミや他鳥種と混声もしくは断片化を招いたと推定される。同様にウグイスも単発の歌を頻繁に移動しながら発した場合が多かったこと高い認識漏れにつながった可能性がある。一方ヒヨドリは多彩なバリエーションの歌を発したものの、一つ一つの歌は単純な構造を持つこと、さらに大声量で歌っていたため他 2 種に比べ過誤認識が低かったと考えられる。

3.4 鳥種間・個体間の歌を介した相互関係

1 時間の録音時間中、調査地では、ウグイスやソウシチョウ、センダイムシクイに加えハシブトガラスやヒヨドリ、そして外来種コジュケイなど住宅地から低山まで幅広い生息地に住み環境適応力の高い種が観測された。各種の鳥の数および総歌時間に占める割合を Figure 7 に示す。このうち歌数が 20 以上、かつ各個体が総歌時間に占める割合が 5 % 以上の個体に着目すると、ウグイス、ヒヨドリ、ソウシチョウが総歌空間の 92 %を占有した。

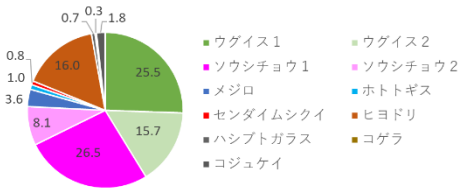


Figure 7 各鳥の歌が総歌時間に占める割合

各個体の歌のタイミングから、各個体の歌によるやりとりを介した相互関係を Figure 8 にまとめる。同種間の相互関係に着目すると、ウグイス同士は重複を有意に回避する傾向が示唆された。この結果はこれまで鳥類の他種ハシナガハチドリ[3] やオオヨシキリ[5] [6]で確認された重複回避行動と一致する。ソウシチョウに関しては、1 個体がもう 1 個体を回避したが、その逆の回避行動は確認されなかった。ソウシチョウの種内での重複回避行動に関しては、当該種は主に雄のみが歌う他鳥種と異なり、雌雄が交互にデュエットを行う歌行動の影響もあるため追跡調査が必要である。

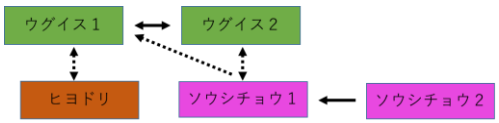


Figure 8 各鳥の歌を介した相互関係。矢印実線は同種間、破線は異種間で、矢印元の種が矢印先の種の歌のタイミングに有意な影響を与えた関係を示す。

異種間では各個体・鳥種は場合に応じた相互関係を示した。観測ターゲットであるソウシチョウとウグイス 1 ペアは双方で回避し合った。観測されたウグイスとソウシチョウ各 2 個体の相互関係を総合的に見ると、ソウシチョウがウグイスを回避する確率に比べ回避される確率のほうが低い。つまりソウシチョウがウグイスを効率よく回避しているのに対し、ウグイスはソウシチョウを回避しきれていないのであるならば歌空間をめぐる 2 者の競争でソウシチョウが有意に立っている可能性が示唆された。歌空間をめぐる競争で有意に立つ種は共存する生息環境をめぐる競争でも有意に立つ可能性が高いため、ソウシチョウはウグイスに負の影響を与えている可能性がある。一方ウグイスとヒヨドリの在来種間では、歌のタイミングに関する相互関係は観測されなかった。

今後のフィールド調査の課題として、ソウシチョウの歌の性差の追跡調査に加え、密度や近傍個体との距離が重複回避行動に与える影響の解明があげられる。また、ソウシチョウが周辺他種の歌行動に与える影響を検証する際、巣や卵、抱卵班の確認などにより各種の繁殖タイミングが一致しているかの確認も必須である。

4 おわりに

本稿は、マイクロフォンアレイとロボット聴覚を使った森林性鳥類の観測実例として、侵略的外来種ソウシチョウがその周辺在来種種の歌行動に与える影響の解明を試みた。鳥類の歌は、求愛や縄張り宣言、警戒、威嚇など一日の時間帯やシチュエーションにより多様な機能を持つが、特に繁殖期においては歌でいかに効率よくメッセージを送るかが繁殖成功率に大きな影響を及ぼす。重複回避の解析からソウシチョウが歌の周波数帯の近いウグイスに負の影響を及ぼす可能性が示唆された。

位置情報付きの音声データを解析することは鳥類の歌行動を理解するにあたり大きな意義を持つ。当該技術の活用により、音の到来方向に裏付けされた個体間識別が実現し、鳥がいつどこで鳴いたかの情報が明確になった。また、当初観測対象としていなかったヒヨドリとソウシチョウの歌を介した相互関係の解析が可能になるなど、当該技術を用いる利点のひとつであるデータの再現性・検証性の高さが実証された。

2 次元における定位の試みに関しては、予測していた音の伝達を妨げる植物などの遮蔽物の課題に加え、セミなど観測対象以外の音声を含む多様な周辺音の問題が浮き彫りになった。一方で、多様な障壁にもかかわらず、歌の特徴によっては森林内でも大まかな位置の推定が可能であることも示唆された。歌数の多い種は検出されやすい中でも特に、定位対象の周波数帯の歌を定点で長時間繰り返すソウシチョウやヒヨドリは 2 次元上でも定位されやすく、頻繁に移動し歌の頻度や音量が低い種はそもそも複数のマイクで同時に定位される確率が低いため、その結果として 2 次元上で定位されにくい傾向を示した。

また、簡易音源分類機能の全体的な分類精度は 76.1%、偶然による一致を排除した k 係数は 0.66（Substantial Agreement）であった。この分類機能は、特に鳥の歌とその他の音源を識別する際に有効であった。鳥の歌の分類精度は、2 次元定位と同様に種毎の歌行動や歌の特徴を反映した。定点に留まり大音量で単純な構造の歌を繰り返すヒヨドリに比べ、複雑な歌構造を持つソウシチョウや、頻繁に移動を繰り返したウグイスはセミと混声そして断片化されやすく、その結果高い分類エラーを示した。

マイクロフォンアレイと HARKBird を用いた野鳥観測とその生態理解への今後の課題として、自動分類機能のさらなる洗練があげられる。自動分類精度の向上により歌の分析コストが大幅に軽減し、長期間にわたる観測が実現するため、鳥類の位置的そして時間的棲み分けに関する生態の解明に向け応用の可能性が高まると考えられる。

謝辞

炭谷晋司氏、森松健充氏（名古屋大学）のデータ分析、衣川直美氏（NPO 法人里山サロン）のフィールド調査における協力に深く感謝申し上げる。本研究の一部は JSPS 科研費 JPA17H068410、JP15K00335、JP16K00294、JP24220006 の助成を受けたものである。

参考文献

[1] Mennill,D.J., Battiston, M., Wilson, D.R., Foote, J.R., & Doucet, S.T.(2012).Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour. *Methods in Ecology and Evolution, 3*,704-712.

[2] Collier,T.C., Kirschel, A.N.G., & Taylor, C.E. (2010). Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network. *Journal of Acoustical Society of America, 128*(1), 182-189.

[3] Araya-Sales, M. Wojczulanis-Jakubas, K., Phillips, E.M., Mennill, D.J., & Wright, T.F. (2017). To overlap or not to overlap: context-dependent coordinated singing in lekking long-billed hermit. *Animal Behaviour, 124*, 57-64.

[4] Nakadai, K., Takahashi, T., Okuno, H.G., Nakajima, H., Hasegawa, Y., & Tsujino, H. (2010). Design and implementation of robot audition system “HARK”- open source software for listening to three simultaneous speakers. *Advanced Robotics, 24*,739-76.1.

[5]Suzuki,R.Sumitani,S. Naren, Matsubayashi, S., Arita,T., Nakadai,K., & Okuno,H.G. Field observations and virtual experiences of ecoacoustic dynamics of bird songs using an open-source software for robot audition HARK. (accepted).

[6] Matsubayashi, S., Suzuki, R., Saito, F., Murate, T., Masuda, T., Yamamoto, K., & Okuno, H. G. (2017). Acoustic monitoring of the great reed warbler using multiple microphone arrays and robot audition. *Journal of Robotics and Mechatronics, 29*(1), 224-235.

[7] 鷲谷いずみ(監修)日本生態学会 (編) .(2002).「外来種ハンドブック」.地人書館, 東京.

[8] 江口和洋,増田智久(1994).「九州におけるソウシチョウ Leiothrix lutea の生息環境」日本鳥学誌 43, 91-100.

[9] 東條一史.(1994).「筑波山塊におけるソウシチョウ Leiothrix lutea の増加」日本鳥学誌 43,39-42.

[10] 江口和洋,天野一葉.(2008).「ソウシチョウの間接効果によるウグイスの繁殖成功の低下」:日本鳥学会誌 57（1),3-10.

[11] Amano, H.E. & Eguchi K.(2002).The foraging niches of introduced red-billed leiothrix and native species in Japan. *Ornithological Science,1*, 123-131.

[12] Schmidt, R.O.(1986). Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions, 34*(3), 276-280.

[13] Nakajima,H., Nakadai,K.,Hasegawa,Y.& Tsujino,H. (2008). Adaptive step-size parameter control for real world blind source separation, *In Proc. ICASSP*, 149-152.

[14] Masco, C., Allesina, S. Mennill, D.J., & Pruett-Jones, S.(2016).The song overlap null model generator(SONG) - a new tool for distinguishing between random and non-random song overlap. *Bioacoustics, 25* (1), 29-40.

[15] Hamao, S. (2013). Acoustic structure of songs in island population of the Japanese-bush warbler, Cettia diphone, in relation to sexual selection. *Journal of Ethology,31*, 9-15.

[16] 千葉尚彬,炭谷普司,松林志保,鈴木麗璽,有田隆也、中臺一博,奥乃博.（2017）.「ロボット聴覚を活用した野鳥の歌行動分析のためのツール HARKBird の機能強化」. 第 35 回ロボット学会学術講演会講演概要集, RSJ2017ACA3-03.

[17] Jensen,J.R.(2005). Introductory Digital image processing. A remote sensing perspective.NJ:Pearson Prentice Hall.

[18] Landis, J.R., & Koch, G.G.(1977). The measurements of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-5 (11/25)

可聴音を用いた周波数選択に基づく距離推定法の実環境利用に向けた評価

Assessment of distance estimation using audible sound based on spectral selection.

高尾麻衣子 [1] ，干場功太郎 [1] ，中臺一博 [1,2]

Maiko TAKAO[1], Kotaro HOSHIBA[1], Kazuhiro NAKADAI[1,2]

東京工業大学 工学院 システム制御系 [1]

(株) ホンダ・リサーチ・インスティチュート・ジャパン [2]

Tokyo Institute of Technology[1]

Honda Research Institute Japan Co., Ltd.[2]

{mtakao,hoshiba,nakadai}@ra.sc.e.titech.ac.jp

Abstract

環境認識は長年にわたって研究されており，自動運転やロボットを利用した災害救助など様々な目的に利用可能である．本稿では，人間に不快感を与えない形で周辺環境認識を行うため，音楽など日常環境に存在する音響信号を用いた距離計測手法を提案する．既存の音響信号を用いた距離計測手法は分解能が不十分，狭帯域信号に対するノイズ耐性が低いといった問題がある．提案手法では周波数ごとに重みを設定することにより，これらの問題の解決を図る．実際に，室内で距離計測実験を行った結果，既存手法と比べ雑音環境下での検出性能で有効性を確認した．

1 はじめに

環境認識は長年にわたって研究されており，自動運転やロボットを利用した災害救助など様々な目的に利用可能である．周辺環境認識技術のうち，音響信号を用いた距離計測は，環境光の影響を受けにくいこと，比較的安価で構成できることなどのメリットがあるため，レーザーやカメラでは計測が難しい暗所やガラス面などでも利用可能である．

音響信号は人間の耳に聴こえない超音波信号と，人間の耳に聴こえる可聴音信号の 2 つにわけることができる．超音波信号による計測は，音響信号による計測の中では波長が短いため分解能がよく，低周波域に多く存在する環境ノイズ影響を受けにくいため，広く用いられている．しかし，出力の際に耳障りな立ち上がり音が伴ってしまうことや，実際に人間の耳では超音波を感知することができないため，超音波曝露の危険性 [1] が考えられる．そのため，人間のいる環境において超音波を使用することは不向きといえる．可聴音を信号として使用する場合，送信信号を工夫すれば耳障りでない音響信号を利用した距離計測が可能になる．また誤って大きな音を出してしまってもすぐに気づくことができ，曝露の危険性も少なくなる．一方で，超音波に比べて周波数が低いことによる分解能の劣化や，環境ノイズに影響を受けてしまうというデメリットがある．例えば，移動ロボットに搭載して足元の状態を確認する際に使用する場合を想定すると，段差を検知できる程度の分解能が必要である．また，実環境で使用することを考えるとノイズ耐性も必要となる．そこで本稿では可聴音を用いた周波数選択に基づく距離推定法を提案する．耳障りでない音として，送信信号に音楽信号を用いることを考える．実環境での実験を通し，既存手法と比較することで提案手法の有用性を評価する．

2 音響信号を用いた距離計測

音響信号を用いた距離計測の手法として，パルス波を用いて伝搬時間を計測する手法 [2]，相互相関により伝搬時間を計測する手法 [3]，定在波を利用する手法 [4]，受信信号の振幅により正規化した相互相関関数を用いた白色化相互相関（Cross-power Spectrum Phase Analysis : CSP）法 [5] などが挙げられる．パルス波を用いたパルスエコー法は最も単純な手法であるが，時間方向のパワーが小さいため雑音の影響を受けやすく，計測距離の分解能はその波長に依存するため，可聴領域では超音波領域と比べ分解能が低下してしまう．また，相互相関を用いた場合，時間方向に出力されたエネルギーを圧縮するため，雑音耐性が高くなる．しかし，分解能はサンプリング周期に依存するため，音声の解析などでよく用いられる 16 kHz 程度のサンプリングレートでは十分な分解能が得られるとは言えない．定在波を用いる場合，正確なキャリブレーションが必要となってくるため，ロボットのような移動体への実装が困難であると考えられる．CSP 法は，高い雑音耐性を有しているが，広帯域の雑音がある場合は，性能が劣化してしまう．また基本的には相互相関処理に基づく手法であるため前述のように分解能はサンプリング周波

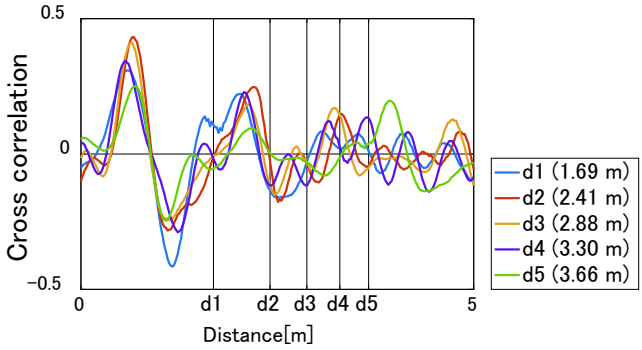


図 1: Distance estimation using cross correlation.

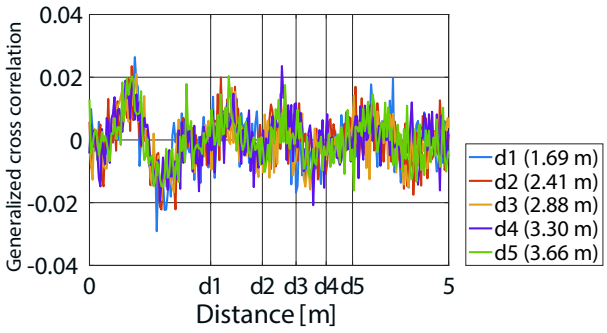


図 2: Distance estimation using CSP.

数に依存する．

実際に，上記の既存手法を利用した距離計測の問題点を確かめるため，2 つの実験を試みた．まず，相互相関処理を用いた際の問題点を確認する．SNR が 30 dB 程度の雑音が少ない環境で，スピーカから部屋の壁に向けて，音楽信号を出力し，その反射音を受信した．スピーカと壁の距離については 5 パターンの計測を行った．Fig. 1 は，送信信号と受信信号の相互相関を計算した結果である．横軸は伝搬時間を距離に換算したもので，レーザー距離計による距離計測値を d1 から d5 で示す．0.5 m 程度の場所に存在するピークは直接波による相互相関値を示し，最大ピークとして観測できる．一方で反射波のピークは雑音に埋もれてしまって，検出が難しいことがわかる．次に，CSP 法の問題点を確認する．CSP 法は前述の通り，広帯域の雑音がある場合や計測に用いる信号の周波数帯域が狭い場合，性能が劣化してしまう．以下のような実験でこれを確認する．先ほどの実験と同じ環境で，スピーカーから部屋の壁に向けて，単一周波数の正弦波を送信し，その反射音を受信した．音源の周波数は 3 種類，音源と壁の距離は 3 パターンの計測を行った．距離ごとに計測した 3 種類の周波数の受信波を PC 上で足し合わせ，疑似的な受信波を作成し，これについて CSP 法で距離推定を行う．Fig. 2 に結果を示す．横軸は Fig. 1 と同様に伝搬時間を距離に換算した値である．正しく計測ができている場合，d1〜d5 それぞれの位置にピークが現れるはずであるが，Fig. 2 ではその周辺にピークを観測することができず，正しい距離推定ができていない．CSP 法は全周波数を参照して相関処理を行う手法であるため，本実験のように計測で使用できる周波数が限られている場合には正しく計測できないことがわかる．

3 提案手法

前節で述べたように，既存手法では可聴領域で狭帯域の送信信号に対し雑音耐性を備えた距離計測を行うことは難しい．そこで，これらの問題を解決するため，周波数選択に基づいた距離推定法を提案する．

ターゲットからの反射波を含む受信信号を $s(t)$，参照信号を $s_{ref}(t)$ とし，これらの周波数領域での表現を $S(\omega)$，$S_{ref}(\omega)$ とする．$S(\omega)$ は周波数 ω におけるスペクトルの振幅と位相情報を持った複素信号となるため，以下のように参照信号との位相差 $\phi(\omega)$ を算出する．

$$\phi(\omega) = \mathrm{angle}(S(\omega) \cdot \mathrm{conj}(S_{ref}(\omega))) \tag{1}$$

ここで $\phi(\omega)$ は $0 \leq \phi(\omega) < 2\pi$ であり，各周波数成分の伝搬距離による位相のずれを表す．得られた位相差を，以下のように伝搬距離 $d_n(\omega)$ に変換する．

$$d_n(\omega) = \frac{2n\pi + \phi(\omega)}{2\pi}\lambda, \qquad (n \in \mathbb{Z}_+) \tag{2}$$

ただし，λ は周波数 ω の波長，$\mathbb{Z}_+$ は 0 以上の整数である．このように，周期が λ の周期関数として $d_n(\omega)$ を定義することができ，これを Fig. 3(a) に示す．横軸は距離で，青，赤，黄色の点はそれぞれ異なる周波数 $\omega 1, \omega 2, \omega 3$ の成分から求められた伝搬距離 $d_n(\omega 1), d_n(\omega 2), d_n(\omega 3)$ を示す．これは，上記に挙げた，位相差を用いる場合には伝搬距離が波長以内でないと距離を一意に定めることができないという問題に対応するため，位相の周回を考慮した伝搬距離候補を挙げるという意味を持つ．理想環境では，送信信号に含まれるすべての周波数に共通する $d_n(\omega)$ が存在し，その値が伝搬距離となる．Fig. 3(a) では 3 つの周波数の $d_n(\omega)$ が真値の位置で一致している．しかし実環境では雑音の影響で誤差が発生する．Fig. 4(a) に，計測に誤差が発生してしまった場合の計測における，Fig. 3(a) と同様の図を示す．ここでは計測誤差によって推定距離が真値と一致していないことがわかる．そこで，得られた $d_n(\omega)$ に対し，ガウス分布を畳み込んだ周期関数であると仮定して，伝搬距離尤度を以下のように定義する．

$$p_\omega(x) = \sum_\omega \frac{1}{\sqrt{2\pi\sigma^2(\omega)}} \exp\left(-\frac{(x - d_n(\omega))^2}{2\sigma^2(\omega)}\right) \tag{3}$$

ここで，$\sigma^2(\omega)$ はガウス分布の分散，x は距離である．このようにすることで，雑音による計測誤差を考慮した推定ができる．これにより，尤度を持った伝搬距離の候補を定義できる．Fig. 3(b) では理想環境での計測に対して伝

搬距離尤度を求めた結果を，Fig. 4(b) では誤差のある計測に対して伝搬距離尤度を求めた結果を示す．横軸が伝搬距離，縦軸が尤度であり，3 つの周波数それぞれで求めた結果を異なる色で示す．Fig. 3(b) ではノイズによる計測誤差がないため，全ての $p_\omega(x)$ が真値の位置にピークを持っている．Fig. 4(b) では誤差が発生しているため，ピークが真値と少しずれている．しかし，伝搬距離尤度としてガウス分布を畳み込むことで，どの周波数でも真値の位置に十分大きな値を持っている．さらに，各周波数で得られる $p_\omega(x)$ に対し周波数重みを定義し，全周波数を統合したブロードバンド尤度関数を以下のように定義する．

$$p(x) = \sum_{\omega} p_\omega(x) \cdot W(\omega) \tag{4}$$

ここで，$W(\omega)$ は各周波数に対する重み関数である．重みを定義することにより，参照信号の持たない周波数を除外したり，計測しやすい周波数の影響を強く受けるように設定できるため，より雑音の影響を受けにくい距離推定が可能となる．このようにして求められたブロードバンド尤度関数に対し，伝搬距離 D を以下のように算出する．

$$D = \underset{x}{\operatorname{argmax}}\, p(x) \tag{5}$$

つまり，ブロードバンド尤度関数の最大ピーク位置が求めたい伝搬距離 D である．Fig. 3(c) では理想環境での計測に対して求めたブロードバンド尤度関数，Fig. 4(c) では誤差のある計測に対して求めたブロードバンド尤度関数を示す．横軸が伝搬距離，縦軸が尤度を示す．Fig. 3(c) では理想環境での計測なので，真値の位置に問題なく最大ピークが観測できる．また，Fig. 4(c) でも同様に真値付近で最大ピークを観測することができる．伝搬距離尤度を求める際にガウス分布を想定したことにより，この時最大ピークのほかにも大きなピークが見られるが，サンプルに使用する周波数を増やせばピーク以外の部分がランダム分布となり平滑化されるので，最大ピークがより強調される．重みを調整することでも，計測に関係のないピークを抑制することができる．

本手法の特徴は，$\sigma(\omega)$ や $W(\omega)$ を状況に応じて適切に定めることができる点であり，広帯域にわたる特徴的なノイズの中であっても雑音耐性を持った距離推定が可能であることが期待できる．例えば，広帯域のノイズに対し一部の周波数域しか持たない狭帯域の信号を送信する場合，送信信号に含まれる周波数に対する重みを $W(\omega) = 1$ とし，それ以外の周波数に対する重みを 0 とすれば，計測に必要な周波数のみを使用できるので，雑音の影響を受けにくい距離推定を行うことができる．

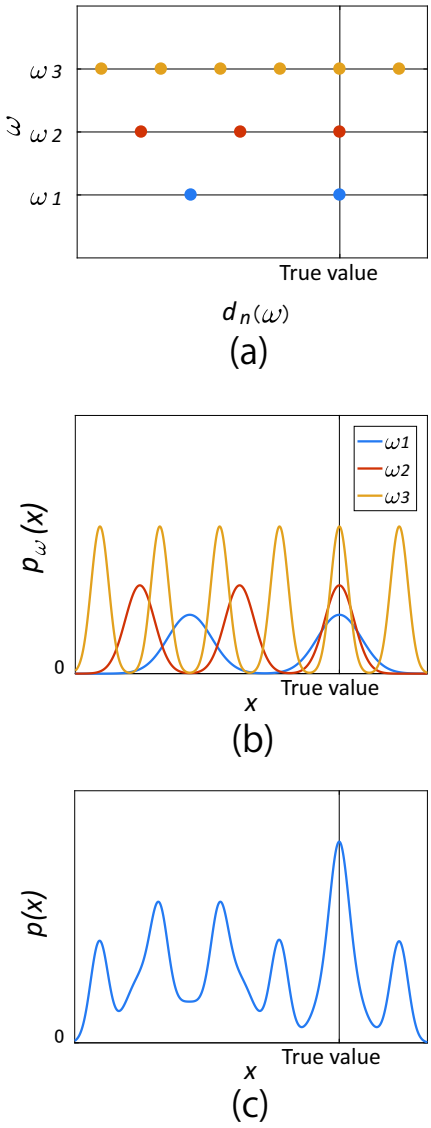


図 3: Algorithm of proposed method in noiseless environment. (a) $d_n(\omega)$, (b) $p_\omega(x)$, (c) $p(x)$.

4 評価実験

4.1 実験状況

提案手法の有用性を確認するため，評価実験を行った．実験状況を Fig. 5 に，計測機器の正面写真を Fig. 6 に示す．送波器 1 から信号を送波し，計測対象である壁からの反射波を受波器 1 にて受信する．また，参照信号 s_{ref} として，送波器 1 の近傍に設置した受波器 2 にて送波器 1 からの直接波を受信する．これは，送波器，受波器の周波数特性が計測に影響を与えないようにするためである．送波器 1 から受波器 1 への直接波を低減させるため，送波器 1 と受波器 2 の間に吸音材を挟み，送波器 1 をダンボールの覆いで囲った．計測は SNR が 30 dB 程度の雑音の少ない環境において，対象までの距離を 5 通りに変化させて実験を行う．送波器 1 には GENELEC 8010APM，受波器には MEMS マイクロホンを使用した．また，送信信号

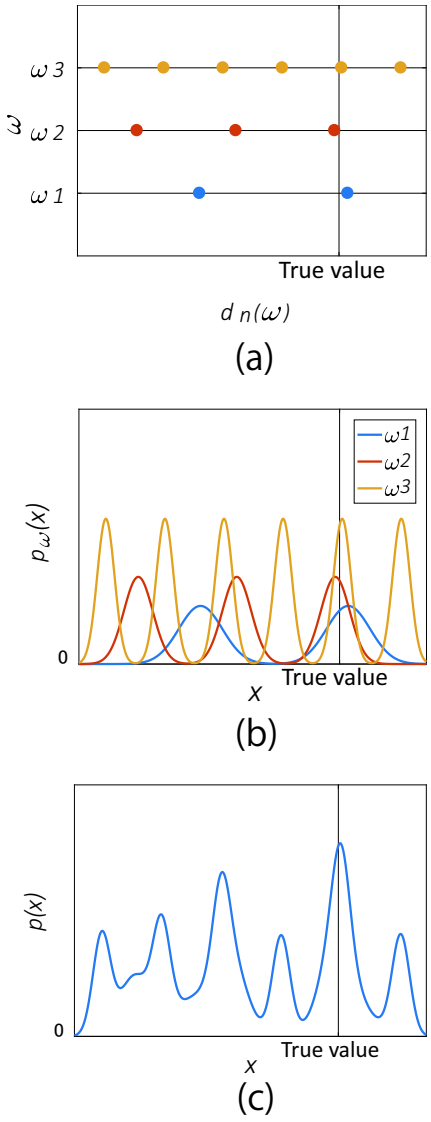


図 4: Algorithm of proposed method in noisy environment. (a) $d_n(\omega)$, (b) $p_\omega(x)$,(c) $p(x)$.

として，RWC 研究用音楽データベース（RWC-MDB-P-2001-M01NO.11）の男声楽曲を利用した．送波器 1，受波器 1，2 は，システムインフロンティア社製の多チャンネル音響信号収録装置 RASP-24[1] に接続され，同期収録を行う．音響信号は，サンプリング周波数 16 kHz，量子化ビット数 16 bit で収録される．収録した音響信号は，PC にて処理を行う．受信信号の 0〜0.5 kHz, 4〜8 kHz の周波数帯域に対して，中心値が $\pi/4$，標準偏差が $\pi/6$ のガウスノイズに置き換えた．これは，0.5〜4 kHz 以外の周波数を参照すると，方向性のある雑音として大きなピークが生じるように，疑似的に設定したものである．特に CSP 法を利用した場合，すべての周波数帯を参照するので，このノイズによる影響を強く受け，計測とは関係のないピークによって正しい値が計測できないような状況になると考えられる．

[1] http://www.sifi.co.jp/system/modules/pico2/index.php?content_id=4

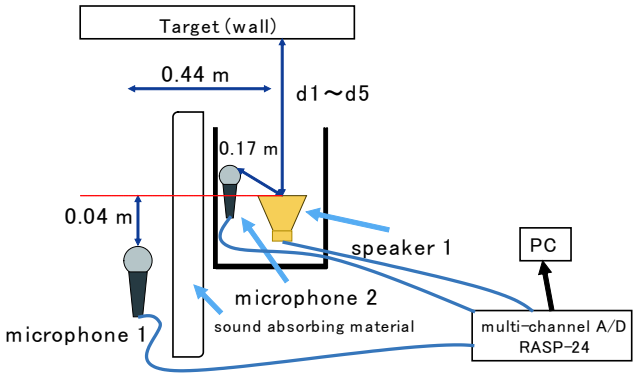


図 5: Experimental configuration.

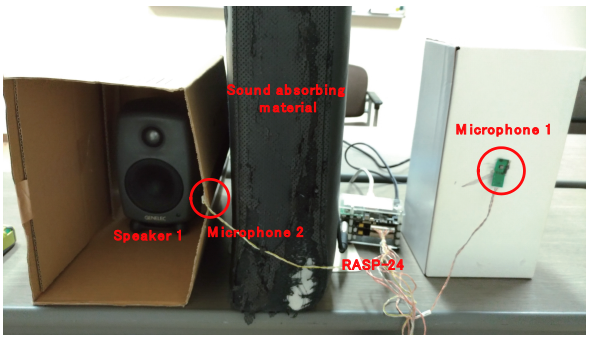


図 6: Measurement system for proposed method.

4.2 実験結果

実験を行った結果について述べる．受波器 2 にて受信した信号の周波数スペクトルの例を Fig. 7 に示す．周波数スペクトルから，送信信号の周波数領域がほぼ 3 kHz 以下におさまっていることが確認できる．また，500 Hz 以下の低周波において，回り込みの影響で直接波が受波器 1 に到達しやすくなることがわかっている．これらを加味し，以下のように σ, $W(\omega)$ を設定した．

$$\sigma = \frac{30}{\omega}, W(\omega) = \begin{cases} \sqrt{A} \cdot \frac{\sqrt{2\pi\sigma^2}}{\omega} & (0.5 < \omega < 3) \\ 0 & (otherwise) \end{cases} \quad (6)$$

ここで，A は ω における送信信号の周波数スペクトルの大きさを表す．周波数が高くなるほど波長は短くなるので，σ は周波数に反比例する値とした．ガウス分布の特性上，周波数ごとに伝搬尤度 $p_\omega(x)$ のピークの値が変わってしまうので，ピークの値を統一するために $W(\omega)$ として $\sqrt{2\pi\sigma^2}$ をかける．また，周波数が低いと波長が長くなるため $d_n(\omega)$ の間隔が大きくなる．おおまかな距離推定のためには $d_n(\omega)$ の間隔が大きいもの，詳細な距離推定のためには $d_n(\omega)$ の間隔が小さいものが有効である．そのため $1/\omega$ をかけることで，低い周波数になるほど評価関数に強く影響するように周波数に反比例する値とした．さらに $\sqrt{A}$ をかけることにより送信信号がより大きなパ

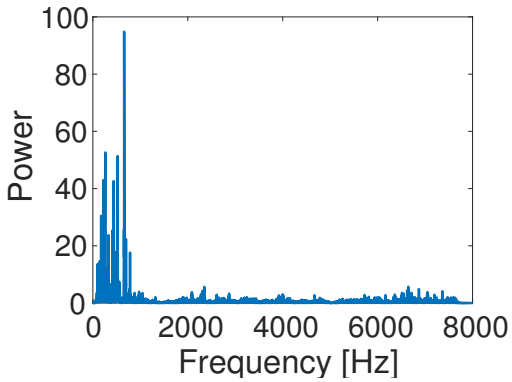


図 7: Spectrum of reference signal.

ワーを持つ周波数を重点的に見ることができる．

以上のパラメータを用い，CSP 法によって距離を算出した結果を Fig. 8(a) に示す．同様の実験について，実際に評価関数を算出した結果を Fig. 8(b) に示す．横軸は伝搬距離を示す．レーザー距離計によって計測した伝搬距離は 1.69 m, 2.41 m, 2.88 m, 3.30 m, 3.66 m であり，それぞれ d1, d2, d3, d4, d5 と表す．伝搬距離が異なる 5 通りの計測の結果を異なる色で示す．次に，同様の実験を同じ環境下で 20 回行った際の，レーザー距離計での計測結果との誤差の分布を算出した結果を Fig. 9 に示す．CSP 法では最大ピークが $x = 0$ 付近であり CSP 法そのものとの精度比較ができないため，$x = 0.5$ 以降の結果からピーク探査し，伝搬推定距離 D を求める．それぞれの実験環境における誤差の大きさを縦軸に示す．赤が提案手法，青が CSP 法での結果である．

4.3 考察

Fig. 8(a) に示された，位相の偏ったノイズをかけたときの CSP 変換の結果を見ると，推定される伝搬距離 D となる最大ピークは $x = 0$ 付近に存在する．これはノイズ部分に疑似的に指定した位相差によって生じるものであり，正しい計測の弊害となることがわかる．計測距離付近にはピークが見られるものの，相対的には小さくなってしまっているので，ノイズによるピークの位置によっては計測がより困難になるといえる．したがって，参照信号が狭帯域の場合，その他の周波数帯域のノイズの影響により性能が劣化することが読み取れる．一方，Fig. 8(b) に示された提案手法での結果によると，それぞれレーザー距離計での計測距離と同じ位置に最大ピーク D が確認できる．CSP 法で発生したようなノイズ部分の影響はここでは見られないため，この条件の下で提案手法はより有効な計測となっているといえる．

また，Fig. 9 より，どの距離においてもレーザー距離計での計測値により近い結果を示しているのは提案手法である．CSP 法の分解能は 0.04 m 程度であり，提案手法では推定結果の分散が大きいが，分解能は 0.01 m 程度と十

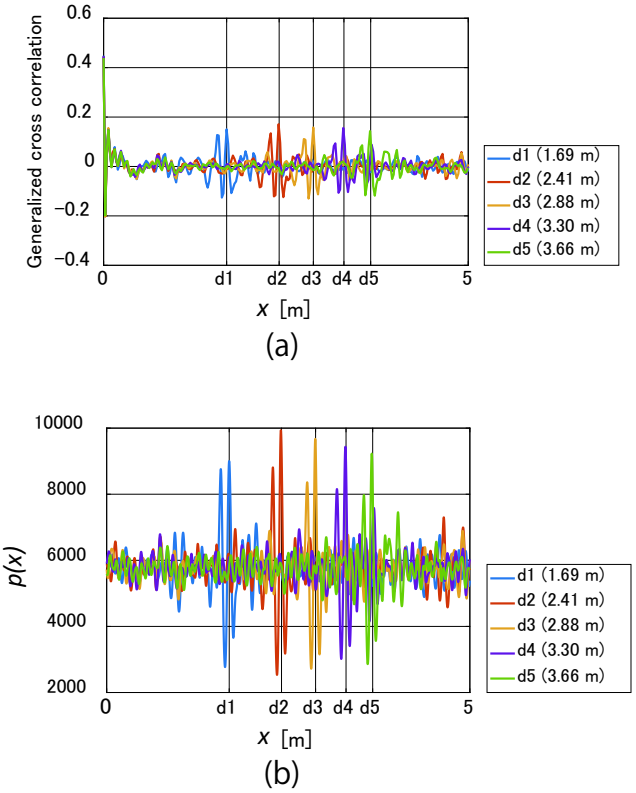


図 8: Distance estimation result using (a) CSP, (b) proposed method.

分小さくなっている．実環境において，例えば移動ロボットが足元の状態を確認するために使用する場合に必要となる分解能を考える．移動ロボットは壁や段差を検知し，進行方向を変えるなどの選択が必要である．特に段差の検知を見落としてしまうと，ロボットの転倒や故障の恐れがある．分解能が 0.04 m の場合，階段のような大きな段差は検知できるが，路肩のような段差を検知することはできない．一方，分解能が 0.01 m の場合，そのような段差でも検知することができる．したがって，この場合 0.01 m の分解能を持つ提案手法がより優れているといえる．

一方で，提案手法では各周波数についてそれぞれ処理を行うため，計算に時間がかかる．また，使用信号に低周波が含まれていない場合に計測が難しいなどの欠点もあり，実用化には改良の必要がある．

5 おわりに

本稿では，将来的なロボット聴覚への応用を目標に，可聴音領域の音楽音響信号をもとに距離を計測する技術について述べた．従来から用いられる手法や超音波計測で用いられる手法をそのまま適用しただけでは，可聴音領域信号での距離計測には問題があることを示し，これを解決する手法として，周波数重みを用いる手法を提案した．参照信号や環境ノイズ，周波数特性に応じた適切な周波数重みを用いることで，参照信号が狭帯域であってもロ

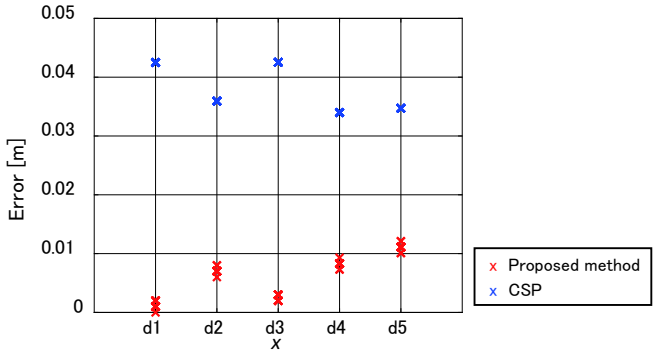


図 9: Error distribution.

バストに，かつ高分解能で距離推定が可能であることを示した．

今後は距離の計測だけではなく，二次元的，あるいは三次元的な地図の作成，システム（送受波器）が移動する場合や周囲の環境が動的に変化する場合への対応などの課題への取り組みを通じて，実用的な技術の構築を目指したい．

謝辞

本研究は，JSPS 科研費 16H02884，16K00294，17K00365 および，JST ImPACT タフロボティクスチャレンジの助成をうけた．

参考文献

[1] Wesley L. Nyborg. Biological effects of ultrasound: development of safety guidelines: Part i: personal histories. *Ultrasound in Medicine & Biology*, Vol. 26, No. 6, pp. 911–964, 2000.

[2] 松野保久, 山中有一. 2 球体の超音波 (50khz) 反射パルス波形情報利用のための実験に基づく検討. 日本水産学会誌, Vol. 56, No. 8, pp. 1219–1224, 1990.

[3] 松尾成光, 平野憲雄, 片尾浩, 安藤雅孝. 超音波を利用した精密音響測距装置の開発. 東京大学地震研究所技術研究報告, No. 5, pp. 65–73. 東京大学地震研究所, 1999.

[4] Hanabusa Shimpei, Uebo Tetsuji, Tsuchida Yuuta, Shinohara Toshihiro, and Nakasako Noboru. Distance estimation based on interference of audible linear chirp signal. *IEEJ Transactions on Electronics, Information and Systems*, Vol. 129, No. 11, pp. 2027–2033, 2009.

[5] M Omologo and P Svaizer. Use of the crosspower-spectrum phase in acoustic event location. *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 3, pp. 288–292, 1997.

一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-6 (11/25)

ブラインド音源分離
～時空間スモールデータの非ガウス・低ランクモデリングとその最適化の数理～

Blind Source Separation – Non-Gaussian and Low-Rank Modeling for Time-Spatial Small Data and Its Optimization –

猿渡　洋
Hiroshi SARUWATARI
東京大学・大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

Abstract

ブラインド音源分離（BSS）技術は，音の空間伝搬及び時間周波数スペクトログラムをモデリングする音響分野において独自の発展を遂げてきた．本稿では，その歴史と潮流について振り返る．特に，優決定 BSS 問題における代表的なアルゴリズムとして周波数領域独立成分分析・独立ベクトル分析及び独立低ランク行列分析を取り上げ，それらが何をモデリングしどのような数理アルゴリズムによって信号推定するのか考察する．また，空間モデルの推定に関し，多チャネル非負値行列因子分解に見られるような「生成モデル型」と BSS で用いられる「分離モデル型」の違い・得失に関しても解説を行い，最適化パラメータのドメイン変更によって性能が大きく改善されていった歴史を紹介する．

1　はじめに

ブラインド音源分離 (blind source separation: BSS) とは，音源位置や混合系が未知の条件で観測された信号のみから混合前の元信号を推定する信号処理技術である．優決定条件 (音源数 ≤ 観測チャネル数) における BSS では，独立成分分析 (independent component analysis: ICA) [1] に基づく手法が主流であり，盛んに研究されてきた [2]–[7]．一方，モノラル信号等を対象とした劣決定条件 (音源数 > 観測チャネル数) 下では，非負値行列因子分解 (nonnegative matrix factorization: NMF) [8] を応用した手法が注目を集めており，多チャネル信号用に拡張した多チャネル NMF (multichannel NMF: MNMF) [9]–[11] も提案されている．

優決定条件における周波数領域 ICA (frequency-domain ICA: FDICA) や ICA の多変量モデルである独立ベクトル分析 (independent vector analysis: IVA) [12]–[14] では，時間周波数領域での線形時不変混合を仮定する．この仮定は，多チャネル観測信号の空間相関行列のランクが 1 になることから,「ランク 1 空間近似」と呼ばれ，複素スペクトログラムの各時間フレーム内で複数の音源が瞬時混合されているという混合系を想定したものである．このような仮定は，各音源から各マイクロフォンまでのインパルス応答が，短時間フーリエ変換の窓関数と比べて十分に短い場合に成立する．Kitamura らは近年，IVA を拡張し，任意ランクの非負値行列積で音源スペクトログラムをモデリングする独立低ランク行列分析（independent low-rank matrix analysis: ILRMA） [15] を提案しており，従来の手法を凌駕する分離性能と高速な最適化アルゴリズムを実現している．本稿ではこれら優決定 BSS の歴史を踏まえ，ランク 1 空間近似を用いた 3 つの代表的な BSS アルゴリズム (FDICA, IVA, ILRMA) を取り上げ，それぞれの手法が仮定する音源モデルと空間モデル及びその最適化手法について概観する．

2　BSS 概観：何をモデリングするのか？

2.1　ランク 1 空間近似

音源数と観測チャネル数をそれぞれ N, M とし，各時間周波数における多チャネル音源信号，多チャネル観測信号，分離信号をそれぞれ

$$\boldsymbol{s}_{ij} = (s_{ij,1} \cdots s_{ij,N})^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij,1} \cdots x_{ij,M})^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij,1} \cdots y_{ij,N})^{\mathrm{T}} \tag{3}$$

と表す (要素はすべて複素数)．ここで，$i = 1, \cdots, I$ は周波数インデックス，$j = 1, \cdots, J$ は時間インデックス，$n = 1, \cdots, N$ は音源インデックス，$m = 1, \cdots, M$ はチャネルインデックスを示し，$^{\mathrm{T}}$ は転置を表す．

混合系が線形時不変であり，時間周波数領域での複素瞬時混合で表現できると仮定すると，各時間フレームにおいて周波数毎の複素混合行列 $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \cdots \boldsymbol{a}_{i,N})$ ($\boldsymbol{a}_{i,n}$は各

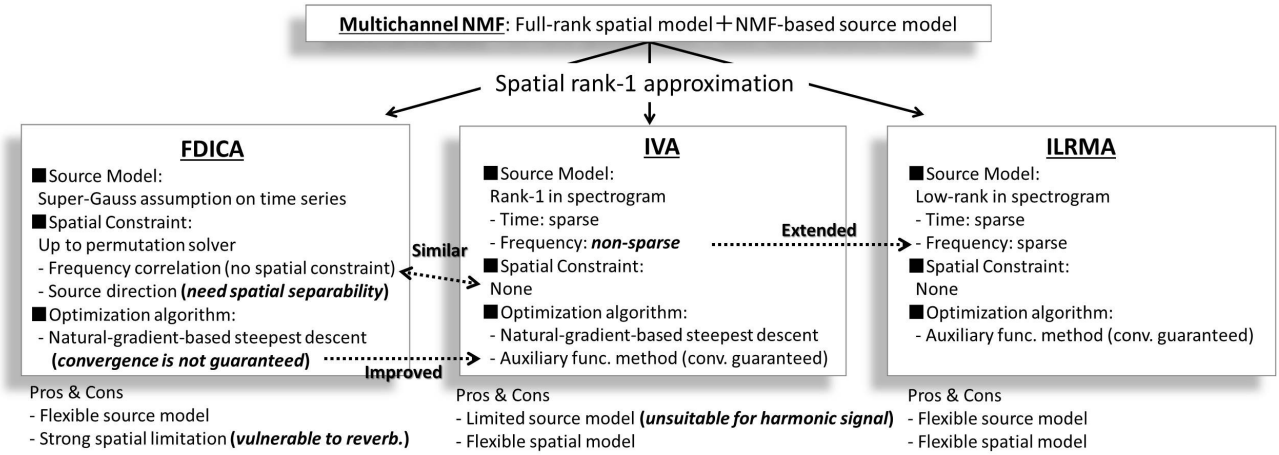


図 1: Overview and relationship between typical acoustic BSS algorithms.

音源のステアリングベクトル) が定義でき，多チャネル観測信号を次式で表現できる．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{4}$$

このとき，観測信号 $\boldsymbol{x}_{ij}$ に含まれる各音源の空間相関行列のランクは必ず 1 となる [16]．すなわち，「混合系が線形時不変かつ複素瞬時混合」という仮定は，ランク 1 空間近似と等価であり，各音源の伝達系が周波数毎の時不変なステアリングベクトル $\boldsymbol{a}_{i,n}$1 本で表現できるという近似を与えている．

式 (4) の混合系において $\boldsymbol{A}_i$ をフルランクとすれば，分離ベクトル $\boldsymbol{w}_{i,n}$ で表現される分離行列 $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1} \cdots \boldsymbol{w}_{i,N})^{\mathrm{H}}$ が存在し，分離信号は次式となる．

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{5}$$

但し，$^{\mathrm{H}}$ はエルミート転置を示す．

ランク 1 空間近似を用いた BSS では，式 (5) 中の分離行列 $\boldsymbol{W}_i$ を推定することが最終的な目標となる．様々なアルゴリズムが提案されているが，大きく分けて「複素時系列の非ガウス性に着目した FDICA」,「FDICA を多変量モデルへ拡張した IVA」及び「音源スペクトログラムを時間周波数低ランク非負値行列としてモデリングする ILRMA」の 3 種類が代表的である（Fig. 1 に相関図を示す）．以降の節では，これらのアルゴリズムについて解説する．

2.2 FDICA の仮定する音源及び空間モデル

FDICA における最適化問題は，以下のコストを最小化する分離行列 $\boldsymbol{W}_i$ を見つける問題に帰着する．これは，音源時系列（周波数ビン i 毎の時系列）の確率密度関数 $p(\cdot)$ をモデルとする対数尤度関数の負値である．

$$Q_{\mathrm{FDICA}} = -2 \log |\det \boldsymbol{W}_i| - \frac{1}{J} \sum_j \sum_n \log p(y_{ij,n}) \tag{6}$$

本式と「非ガウス性」の関連について簡単に解説してみる．本式右辺第一項は音源間の関連度合を表す結合エントロピーを制御し，主に分離の精度に依存する．一方，右辺第二項は個々の音源に関する周辺エントロピー和を制御し，これを最小化するということは「より非ガウスな分離信号へ帰着させる」ことを意味する．つまり，FDICA は，音源の非ガウス信号モデリングと言うことが出来る．

時間周波数領域で各周波数成分に独立な ICA を施す FDICA では，パーミュテーション問題の解決が極めて重要であり，これまでに多くの手法が提案されてきた．代表的なパーミュテーション問題の解決法の一つとして，周波数成分間の相関を用いる手法 [4] がある．これは，後述の IVA と本質的に等価であり，IVA が分離行列の推定と同時にパーミュテーションを解くのに対して，本手法はポスト処理としてパーミュテーションを解いている．もう一つの代表的な解決法は，音源の到来方向 (direction of arrival: DOA) の違いを活用する手法 [3] である．本手法では，推定した周波数毎の分離行列から各音源のステアリングベクトルを逆算し，位相差及び振幅比から DOA を算出して音源毎にクラスタリングすることでパーミュテーションを解いている．この手法の音源モデルは，IVA や周波数間の相関を用いたパーミュテーション解決法とは異なり，時間方向の非ガウス性制約のみである．一方で，FDICA で推定した DOA をパーミュテーション解決に用いる為，空間モデルに関する制約を与えている．複数の音源位置が空間的に接近した場合や残響による拡散の影響が強い場合等，音源 DOA のクラスタリングが困難な状況では分離性能が劣化する．

2.3 IVA の仮定する音源及び空間モデル

IVA は複数の周波数成分を同時に取り扱う為に，ICA を多変量モデルへと拡張した手法である．周波数成分間の高次相関を考慮することで，FDICA におけるパーミュテーション問題 [3, 4] を解決しながら同時に分離行列 $\boldsymbol{W}_i$ を推定する．ICA が非ガウス性の分布を仮定するように，IVA も非ガウスな多変量分布を仮定する．このとき，変数間の高次相関を考慮する為に，球対称の多変量分布を仮定

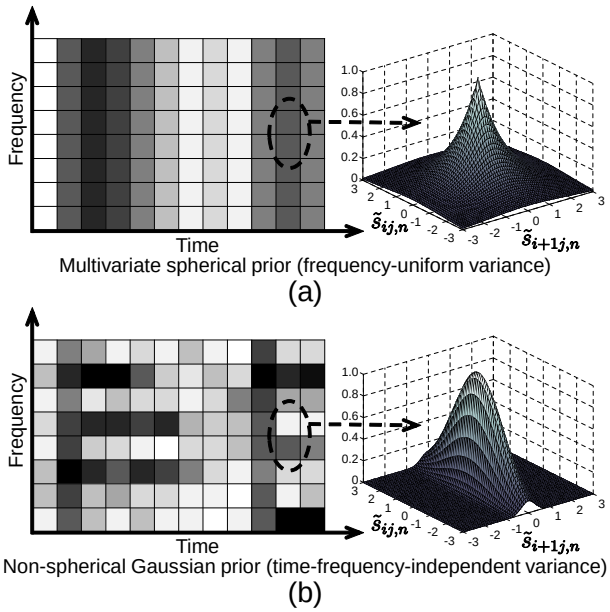


図 2: Illustration of source models (model spectrograms) for one source in (a) IVA and (b) ILRMA, where gray scale of each time-frequency slot indicates value of variance and $\tilde{s}$ denotes only real or imaginary part of complex-valued component s.

することが重要である [13]．最もよく用いられる分布は，Fig. 2 (a) の右側に示す球状ラプラス分布である．この図では，二つの周波数成分の同時分布を示しており，原点を中心に球対称となっている．この性質から，二つの変数間に高次の相関が保証される．

IVA が仮定している音源モデルは，球状多変量分布そのものと解釈できる．この音源モデルを Fig. 2 (a) の左側に示す．各音源は周波数方向に一定の分散値を持っており，それらが時間的に変化するようなパワースペクトログラムを仮定している．従って，複数の周波数で同時に生起する成分を同一音源としてまとめる傾向がある．さらに，音源モデルのパワースペクトログラムを行列とみたとき，1 本の基底ベクトルで表現できる．これは 1 つの音源に対して 1 本のスペクトル基底を与えた NMF と解釈することもできる．

一方，IVA は空間の性質に関して具体的なモデルを与えていない．音源やマイクの位置条件に関係なく，音源モデルの統計的独立性及び多チャネルの観測信号のみから分離行列の推定を行う．

2.4 ILRMA の仮定する音源及び空間モデル

ILRMA のコスト関数は以下で定義される [15].

$$Q_{\mathrm{ILRMA}} = \sum_{i,j}\left[\sum_n \frac{|y_{ij,n}|^2}{\sum_l t_{il,n}v_{lj,n}} - 2\log|\det \boldsymbol{W}_i| + \sum_n \log\sum_l t_{il,n}v_{lj,n}\right] \tag{7}$$

ここで，$t_{il,n}, v_{lj,n}$ は n 番目の音源モデルに対応する非負値スペクトル基底とアクティベーションであり，$l=1,\cdots,L$ は基底のインデックスを示す．すなわち，$\sum_l t_{il,n}v_{lj,n}$ は n 番目の音源のモデルパワースペクトログラムとなる．また，観測チャネル数と音源数の関係は $M=N$ としている．このとき，ILRMA のコスト関数は IVA のコスト関数 (式 (7) の第一項及び第二項) と単一チャネル NMF のコスト関数 (式 (7) の第一項及び第三項) を重ね合わせた形をしている．これらの事実から，IVA は ILRMA においてスペクトル基底数が 1 の特殊ケースに相当しており，その意味で ILRMA は IVA の自然な拡張となっていると解釈できる．

ILRMA の仮定する音源モデルを Fig. 2 (b) に示す．IVA と比較して，1 つの音源に対して L 本のスペクトル基底を用いることができる為，より複雑なパワースペクトログラムを表現可能となっている．また，各時間周波数スロットで独立な複素ガウス分布を音源モデルとして仮定しており，コスト関数 (7) は板倉斎藤擬距離の行列版である log-determinant divergence となっている．従って，時間と周波数いずれの方向にも分散が変動する分布を定義でき，より複雑な時間周波数構造を，限られた基底数で低ランク分解される音源モデルとして表現できる．

一方，空間モデルに関して，ILRMA は，IVA と同様に具体的なモデルを与えていない．音源やマイクの位置に依存せず，観測信号と前述のモデルスペクトログラムの独立性から分離行列を推定する．

3 アルゴリズム概観：どう最適化するのか？

3.1 生成モデル or 分離モデル？

前章を眺めてお気づきの読者もいるとは思うが，先に示した BSS アルゴリズムは全て「音源のパラメタライズ」と「空間分離フィルタ」を推定する問題となっている．つまり，「分離モデル型」であると言える．しかし，一般に統計的モデル推定の観点から眺めると，生成モデル（音源モデル＋空間混合モデル $\boldsymbol{A}_i$）の同時推定問題を解く手法の方がポピュラーかもしれない．実際，MNMF をはじめとする他の音源分離手法は，「生成モデル型」であることが多い．では，BSS における「分離モデル型」の利点は何であろうか？実は，最適化アルゴリズムの発展と深く関係がある．以下の節にて具体的な例を挙げて説明を試みる．

3.2 FDICA における最適化

FDICA におけるコスト関数 Q_{FDICA} の最小化問題を直接解くことは出来ないため，反復法を用いて分離行列及び音源モデルパラメータを求める．様々なものが提案されたが，音声音響信号処理にて最も普及していたアルゴリズムは，分布 $p(\cdot)$ をラプラス分布等で固定化した最急降下法に基づくものであり，Amari らによって提案された自

然勾配 (natural gradient) [17] が有名である．

$$-\frac{\partial Q_{\mathrm{FDICA}}}{\partial \boldsymbol{W}_i}\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_i = \left(\boldsymbol{W}^{-\mathrm{T}} - \frac{1}{J}\sum_j \boldsymbol{\Phi}(\boldsymbol{y}_{ij})\boldsymbol{x}_{ij}^{\mathrm{T}}\right)\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_i = \left(\boldsymbol{I} - \frac{1}{J}\sum_j \boldsymbol{\Phi}(\boldsymbol{y}_{ij})\boldsymbol{y}_{ij}^{\mathrm{T}}\right)\boldsymbol{W}_i \tag{8}$$

ここで $\boldsymbol{\Phi}(\cdot)$ は適当なベクトル関数であり，シグモイド関数等が用いられる．

この自然勾配法の意義について，音響信号処理の視点で考察する．式 (8) 左辺を見ると，単純な勾配（$\partial Q_{\mathrm{FDICA}}/\partial \boldsymbol{W}_i$）にリーマン計量 $\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_i$ が乗じられていることが分かる．これにより，単純な勾配に現れる「分離行列の逆行列 $\boldsymbol{W}^{-\mathrm{T}}$（つまり生成系 $\boldsymbol{A}_i^{\mathrm{T}}$）」を打ち消し，一切の逆行列演算を行うことなく分離音源を求めることが出来る．この恩恵は，空間混合が畳み込みで表される音響信号処理においては非常に大きなものであった．また，分離行列そのものを（逆行列演算せずに）反復更新するので，それをビームフォーミング等の空間フィルタと解釈すれば，過去の音響信号処理研究で得られた事前情報を盛り込むことが容易となり，様々な融合手法が産み出されるに至った [5, 7].

3.3 IVA 及び ILRMA における最適化

FDICA のパーミュテーション問題を解決するために提案された IVA においても，当初は自然勾配の形で分離行列を反復更新するアルゴリズムが使われていた．前述の通り，演算量の少なさは大きな魅力であったが，基本は最急降下法であるため，その収束性（コスト関数の単調減少性）は保証されないという問題があった．この収束性の保証問題に関し，2011 年 Ono らは補助関数法と Iterative Projection (IP) に基づく IVA 更新式を提案した [14]．これは分離行列自体を補助関数と IP で更新するものであり，分離モデル型ならではの特徴を活かしつつ，かつコスト関数の単調減少性を保証する画期的な発明であった．

この補助関数法及び IP に基づく IVA の発明は，続く ILRMA の発明にも大きく影響している．2.4 節にて述べた通り，ILRMA は IVA の自然な低ランク行列拡張であるが，そのコスト関数式 (7) は板倉斎藤擬距離基準 NMF と IVA との結合となっている．つまり，このコスト関数全体を補助関数法で最適化することができ，全パラメータ（分離行列 $\boldsymbol{W}_i$ 及び音源パラメータ $t_{il,n}, v_{lj,n}$）に関して収束性の保証が与えられた反復更新式を得ることが出来る．その場合，音源低ランクモデル $t_{il,n}, v_{lj,n}$ は乗算更新の形となり，非負値性も保証される．

以下，具体的な最適化アルゴリズム [15] について説明する．まず，音源パラメータ $t_{il,n}, v_{lj,n}$ が固定された元で，分離行列 $\boldsymbol{W}_i$ の更新を行う．ここでは $r_{ij,n} = \sum_k t_{ik,n}v_{kj,n}$ とおき，以下のステップに従って IP を実行する．

$$\boldsymbol{U}_{i,n} \leftarrow \frac{1}{J}\sum_j \frac{1}{r_{ij,n}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} \tag{9}$$

$$\boldsymbol{w}_{i,n} \leftarrow (\boldsymbol{W}_i\boldsymbol{U}_{i,n})^{-1}\boldsymbol{e}_n \tag{10}$$

$$\boldsymbol{w}_{i,n} \leftarrow \frac{\boldsymbol{w}_{i,n}}{\sqrt{\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{U}_{i,n}\boldsymbol{w}_{i,n}}} \tag{11}$$

$$y_{ij,n} \leftarrow \boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij} \tag{12}$$

ここで，$\boldsymbol{e}_n$ は n 番目の要素が 1 である単位ベクトルである．このアルゴリズムを繰り返すことにより，コスト関数 (7) が単調に減少するよう分離行列 $\boldsymbol{W}_n$ が更新される．

次に，音源の低ランクモデルパラメータ $t_{il,n}$ 及び $v_{lj,n}$ の更新を行う．これは，その形式より，通常の板倉斎藤擬距離基準 NMF と同様な最適化アルゴリズムが適用できる．具体的には，補助関数法を用いて以下のように更新される．

$$t_{ik,n} \leftarrow t_{ik,n}\left[\frac{\sum_j \frac{|y_{ij,n}|^2}{\left(\sum_k t_{ik,n}v_{kj,n}\right)^2}v_{kj,n}}{\sum_j \frac{1}{\sum_k t_{ik,n}v_{kj,n}}v_{kj,n}}\right]^{\frac{1}{2}} \tag{13}$$

$$v_{kj,n} \leftarrow v_{kj,n}\left[\frac{\sum_i \frac{|y_{ij,n}|^2}{\left(\sum_k t_{ik,n}v_{kj,n}\right)^2}t_{ik,n}}{\sum_i \frac{1}{\sum_k t_{ik,n}v_{kj,n}}t_{ik,n}}\right]^{\frac{1}{2}} \tag{14}$$

$$r_{ij,n} \leftarrow \sum_k t_{ik,n}v_{kj,n} \tag{15}$$

上記は大規模な逆行列演算を多数回行う必要もなく，複雑な代数方程式を解く必要も無い．以上より，非常に少ない演算量で ILRMA の全パラメータを更新可能であることが分かる．

3.4 MNMF と ILRMA のミッシングリンク

本節では，BSS の拡張として提案された ILRMA と低ランクモデリングとして先に発展した MNMF とを比較し，各最適化アルゴリズムの関連性・相違について述べ，分離モデル型の優位性について概説する．まず，MNMF [11] においては，多チャネル観測信号の相関行列 $\mathsf{X}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}}$ を定義し，それを個々の音源に関する空間相関 $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ と NMF 音源パラメータによる近似

$$\mathsf{X}_{ij} \approx \hat{\mathsf{X}}_{ij} = \sum_k \left(\sum_n \boldsymbol{R}_{i,n}^{(\boldsymbol{s})}d_{nk}\right)t_{ik}v_{kj} \tag{16}$$

により時空間のモデリングを行う．ここで d_{nk} は各音源へ基底を分配する変数である．本モデルに基づく MNMF のコスト関数は以下で与えられる．

$$Q_{\mathrm{MNMF}} = \sum_{i,j}\left[\mathrm{tr}\left(\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}\right) + \log\det\hat{\mathsf{X}}_{ij}\right] \tag{17}$$

前述の通り，MNMF は信号の生成系を定義して観測信号を記述しており,「生成モデル型」と言うことが出来る．本

モデリングは柔軟であり，空間がランク 1 でない場合も記述することが出来るが，一方でその最適化は容易ではなく莫大な演算量と解の不安定性が大きな問題とされていた．

ここで大変興味深いことは，この MNMF のコスト関数式 (17) において空間ランク 1 仮説，つまり $\boldsymbol{R}_{i,n}^{(s)} = \boldsymbol{a}_{i,n}\boldsymbol{a}_{i,n}^{\mathrm{H}}$ とおき，更に $\mathsf{D}_{ij} = \mathrm{diag}[\sum_k d_{1k}t_{ik}v_{kj}, \cdots, \sum_k d_{Nk}t_{ik}v_{kj}]$ とおけば，$\boldsymbol{A}_i = \boldsymbol{W}_i^{-1}$ より

$$
\begin{aligned}
Q &= \sum_{i,j}\Big[\mathrm{tr}\Big(\boldsymbol{W}_i^{-1}\boldsymbol{y}_{ij}\boldsymbol{y}_{ij}^{\mathrm{H}}\big(\boldsymbol{W}_i^{\mathrm{H}}\big)^{-1}\boldsymbol{W}_i^{\mathrm{H}}\mathsf{D}_{ij}^{-1}\boldsymbol{W}_i\Big) \\
&\qquad + \log\left(\det \boldsymbol{A}_i\right)\left(\det \mathsf{D}_{ij}\right)\left(\det \boldsymbol{A}_i^{\mathrm{H}}\right)\Big] \\
&= \sum_{i,j}\Big[\mathrm{tr}\Big(\boldsymbol{W}_i\boldsymbol{W}_i^{-1}\boldsymbol{y}_{ij}\boldsymbol{y}_{ij}^{\mathrm{H}}\big(\boldsymbol{W}_i^{\mathrm{H}}\big)^{-1}\boldsymbol{W}_i^{\mathrm{H}}\mathsf{D}_{ij}^{-1}\Big) \\
&\qquad + 2\log|\det \boldsymbol{A}_i| + \log\det \mathsf{D}_{ij}\Big] \\
&= \sum_{i,j}\Bigg[\sum_n \frac{|y_{ij,n}|^2}{\sum_k d_{nk}t_{ik}v_{kj}} - 2\log|\det \boldsymbol{W}_i| \\
&\qquad + \sum_m \log\sum_k d_{nk}t_{ik}v_{kj}\Bigg] \qquad (18)
\end{aligned}
$$

となり，これは ILRMA のコスト関数式 (7) に基底分配変数 d_{nk} を含めたものと一致する．つまり，ILRMA のコスト関数は，MNMF に空間ランク 1 仮説を置き，更に空間モデルのパラメータを生成モデル型（$\boldsymbol{A}_i$）から分離モデル型（$\boldsymbol{W}_i$）へ変更したものだと言うことが出来る．たかが逆行列の関係と思えるかもしれないが，最適化問題においてどちらのドメインで計算するかは非常に大きな差異を生み出す．この空間パラメータ最適化ドメインの変更こそが ILRMA の本質的な新規性であり，それによって全変数の補助関数法（特に分離行列は IP による更新）による高速最適化及び収束性の保証が達成されたと言える．

3.5 実験的比較例

前節で考察した各手法の音源及び空間モデルの違いを例示する為に，人工的に作成した音源を用いた実験結果を Fig. 3 に示す．ここでは，音源スペクトログラムのランク R を変え，FDICA（パーミュテーション解決は DOA 利用），IVA，及び ILRMA による分離を行った（実験条件の詳細は [15] 参照）．本図より，FDICA は音源のランクに影響されにくいがパーミュテーションエラーにより音源分離性能が低く，IVA は逆に音源のランクに強く影響されることが分かる．一方で，ILRMA は適切な基底数を与えてやれば，高い分離性能を維持できることも示されている．

次に，実際の音響データ（SiSEC [19]）を用いた分離実験結果例を Table 1 に示す（実験条件の詳細は [15] 参照）．従来の手法と比べ，ILRMA の分離精度が高く，演算時間の面でも効率的であることが示されている．

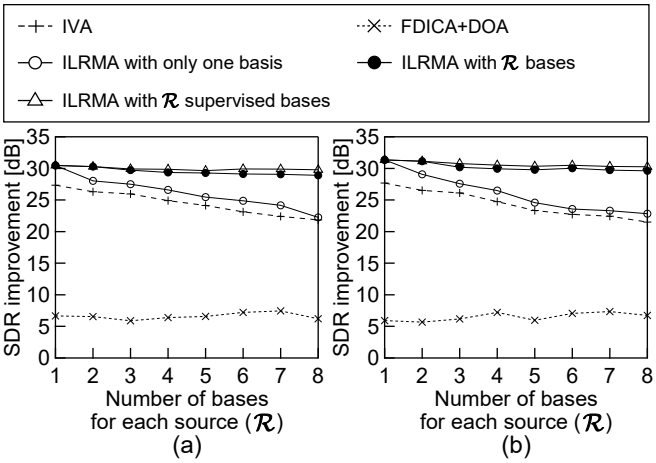


図 3: SDR [18] results of (a) source 1 and (b) source 2 for various numbers of bases.

表 1: Averaged SDR improvement in dB under SiSEC conditions and computational time normalized by IVA's one

Algorithm	SDR improv.	Comp. time
Soft masking [20]	-0.1	-
IVA [14]	2.6	1.0
Ozerov's MNMF [9]	1.2	-
Sawada's MNMF [11]	5.0	49.1
ILRMA	8.7	1.3

4 スパースな生成モデルに基づく BSS

4.1 複素 Student's t 分布に基づく ILRMA

前述の通り，従来の ILRMA はその生成確率モデルとして時変複素ガウス分布を仮定していた．一方で，音声や音楽信号等に関し，ガウス分布よりもさらに尖度の高い「スパース」な分布を仮定することも出来る．ここでは，その一例として，複素 Student's t 分布に基づくものを紹介する [21]．以降ではオリジナルな ILRMA[15] と区別するため，本アルゴリズムを t-ILRMA と呼ぶ．ここでは以下の生成モデルを考える．

$$
\prod_{i,j} p(y_{ij,n}) = \prod_{i,j}\frac{1}{\pi\sigma_{ij,n}^2}\left(1+\frac{2}{\nu}\frac{|y_{ijn}|^2}{\sigma_{ij,n}^2}\right)^{-\frac{2+\nu}{2}} \qquad (19)
$$

$$
\sigma_{ij,n}^{p} = \sum_l t_{il,n}v_{lj,n} \qquad (20)
$$

ここで，分布 $p(y_{ij,n})$ は球状（原点対称）複素 Student's t 分布であり，$\sigma_{ij,n}$ は時間周波数において振幅スペクトル $|y_{ij,n}|$ に対応する時変な非負値スケールである．また，ν は分布の形状を制御する自由度パラメータ，p はスペクトログラムの指数乗ドメインを定めるドメインパラメータであり，$1 \le p \le 2$ を満たす．ここで $\nu \to \infty$ かつ $p=2$ とするならば，式 (19) は時変複素ガウス分布に基づく生成モデルに一致し，$\nu=1$ かつ $p=1$ とするならば，式 (19) は時

変複素コーシー分布に基づく生成モデルに一致する．信号間の独立性の仮定とともに，式 (19) の負対数尤度は以下で与えられる．

$$\mathcal{L}_t = \text{const.} - 2J \sum_i \log|\det \boldsymbol{W}_i| + \sum_{i,j,n} \left[\left(1 + \frac{\nu}{2}\right) \log\left(1 + \frac{2}{\nu} \frac{|y_{ij,n}|^2}{\sigma_{ij,n}^2}\right) + 2\log\sigma_{ij,n} \right] \quad (21)$$

ここで，$\nu \to \infty$ かつ $p=2$ ならば，式 (21) は ILRMA のコスト関数式 (7) に一致する．

4.2 *t*-ILRMA における分離行列の更新式

ここでは分離行列 $\boldsymbol{W}_i$ の更新について考える．一般に，前述の複素ガウス分布由来の ILRMA において用いられていた IP は，logdet 項と $|y_{ij,n}|^2 = |\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}|^2$ の項の和に対してのみ適用できたが，*t*-ILRMA におけるコスト関数 (21) においては $|y_{ij,n}|^2$ の項が対数関数内にあるため，IP を直接適用することが不可能である．そこで，補助関数法を適用するため，以下の接線不等式を考える．

$$\log\left(\sum_q z_q\right) \le \frac{1}{\lambda}\left(\sum_q z_q - \lambda\right) + \log\lambda \quad (22)$$

ここで，z_q は元の変数，$\lambda > 0$ は補助変数である．式 (22) の等号は $\lambda = \sum_q z_q$ の時に限り成立する．式 (22) を式 (21) の第三項及び第四項に適用することにより，以下の補助関数 $\mathcal{L}_t^+$ を得ることが出来る．

$$\begin{aligned} \mathcal{L}_t \le \text{const.} &- 2J \sum_i \log|\det \boldsymbol{W}_i| \\ &+ \sum_{i,j,n} \left[\left(1 + \frac{\nu}{2}\right) \frac{1}{\alpha_{ij,n}} \left(1 + \frac{2}{\nu} \frac{|y_{ij,n}|^2}{\sigma_{ij,n}^2} - \alpha_{ij,n}\right) \right. \\ &+ \left(1 + \frac{\nu}{2}\right) \log\alpha_{ij,n} + \frac{2}{p\beta_{ij,n}} \left(\sum_l t_{il,n} v_{lj,n} - \beta_{ij,n}\right) \\ &\left. + \frac{2}{p} \log\beta_{ij,n} \right] \\ &\equiv \mathcal{L}_t^+ \end{aligned} \quad (23)$$

ここでは $\sigma_{ij,n} = (\sum_l t_{il,n} v_{lj,n})^{1/p}$ を代入した．$\alpha_{ij,n}, \beta_{ij,n} > 0$ は補助変数であり，$\mathcal{L}_t$ と $\mathcal{L}_t^+$ は以下の条件の時に限り等しくなる．

$$\alpha_{ij,n} = 1 + \frac{2}{\nu} \frac{|y_{ij,n}|^2}{\sigma_{ij,n}^2} \quad (24)$$

$$\beta_{ij,n} = \sum_l t_{il,n} v_{lj,n} \quad (25)$$

式 (23) においては $|y_{ij,n}|^2 = |\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}|^2$ が対数関数の外にあるので，IP を直接適用することが出来る．さらに補助関数 (23) を整理し，以下の式を得る．

$$\begin{aligned} \mathcal{L}_t^+ = \text{const.} &- 2J \sum_i \log|\det \boldsymbol{W}_i| + J \sum_{i,n} \boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{U}_{i,n} \boldsymbol{w}_{i,n} \\ &+ \sum_{i,j,n} \left[\left(1 + \frac{\nu}{2}\right) \left(\alpha_{ij,n}^{-1} - 1 + \log\alpha_{ij,n}\right) \right. \\ &\left. + \frac{2}{p\beta_{ij,n}} \left(\sum_l t_{il,n} v_{lj,n} - \beta_{ij,n}\right) + \frac{2}{p} \log\beta_{ij,n} \right] \end{aligned} \quad (26)$$

$$\boldsymbol{U}_{i,n} = \frac{1}{J}\left(\frac{2}{\nu} + 1\right) \sum_j \frac{1}{\alpha_{ij,n}\sigma_{ij,n}^2} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{H}} \quad (27)$$

式 (26) の停留点を $\boldsymbol{w}_{i,n}$ に関して求めることは，以下の連立方程式を解くことと等価である．

$$\boldsymbol{w}_{i,k}^{\mathrm{H}} \boldsymbol{U}_{i,n} \boldsymbol{w}_{i,n} = \delta_{kn} \quad (28)$$

ここで，$\delta_{kn} = 1$ $(k=n)$ および $\delta_{kn} = 0$ $(k \ne n)$ である．式 (28) に IP を適用することにより，分離行列に関して以下の反復更新式を得ることが出来る．

$$\boldsymbol{w}_{i,n} \leftarrow (\boldsymbol{W}_i \boldsymbol{U}_{i,n})^{-1} \boldsymbol{e}_n \quad (29)$$

$$\boldsymbol{w}_{i,n} \leftarrow \frac{\boldsymbol{w}_{i,n}}{\sqrt{\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{U}_{i,n} \boldsymbol{w}_{i,n}}} \quad (30)$$

$\boldsymbol{W}_i$ の更新の後，分離信号 $\boldsymbol{y}_{ij}$ は $y_{ij,n} \leftarrow \boldsymbol{w}_{ij,n}^{\mathrm{H}} \boldsymbol{x}_{ij}$ のように更新される．

4.3 *t*-ILRMA における NMF パラメータの更新式

t-ILRMA の低ランク音源モデル（NMF）に関するパラメータ $t_{il,n}$ 及び $v_{lj,n}$ も，補助関数法を用いて求めることが可能である．式 (23) の NMF パラメータに関する補助関数を得るため，以下の Jensen の不等式を $\sigma_{ij,n}^{-2} = (\sum_l t_{il,n} v_{lj,n})^{-2/p}$ に適用する．

$$\left(\sum_q z_q\right)^{-2/p} = \left(\sum_q \mu_q \frac{z_q}{\mu_q}\right)^{-2/p} \le \sum_q \mu_q \left(\frac{z_q}{\mu_q}\right)^{-2/p} = \sum_q \mu_q^{\frac{2}{p}+1} z_q^{-\frac{2}{p}} \quad (31)$$

ここで $\mu_q > 0$ は $\sum_q \mu_q = 1$ を満たす補助変数である．なお，$1 \le p \le 2$ を想定しているので，式 (31) は変数 z_q に関して凸関数であることに留意する．式 (31) の等号は $\mu_q = z_q / \sum_{q'} z_{q'}$ の時に限り成立する．式 (31) を式 (21) の $\sigma_{ij,n}^{-2} = (\sum_l t_{il,n} v_{lj,n})^{-2/p}$ に適用することにより，以下の補助関数 $\mathcal{L}_t^{++}$ を得る．

$$\begin{aligned} \mathcal{L}_t^+ \le \text{const.} &- 2J \sum_i \log|\det \boldsymbol{W}_i| \\ &+ \sum_{i,j,n} \left[\left(1 + \frac{\nu}{2}\right) \frac{1}{\alpha_{ij,n}} \left(1 + \frac{2}{\nu} |y_{ij,n}|^2 \sum_l \gamma_{ij,nl}^{\frac{2}{p}+1} t_{il,n}^{-\frac{2}{p}} v_{lj,n}^{-\frac{2}{p}} - \alpha_{ij,n}\right) \right. \\ &+ \left(1 + \frac{\nu}{2}\right) \log\alpha_{ij,n} + \frac{2}{p\beta_{ij,n}} \left(\sum_l t_{il,n} v_{lj,n} - \beta_{ij,n}\right) \\ &\left. + \frac{2}{p} \log\beta_{ij,n} \right] \\ &\equiv \mathcal{L}_t^{++} \end{aligned} \quad (32)$$

ここで，$\gamma_{ij,nl} > 0$ は補助変数であり，$\mathcal{L}_t^+$ と $\mathcal{L}_t^{++}$ は以下の条件の時に限り等しくなる．

$$\gamma_{ij,nl} = \frac{t_{il,n} v_{lj,n}}{\sum_{l'} t_{il',n} v_{l'j,n}} \tag{33}$$

式 (32) の最小点を求めるため $\partial \mathcal{L}_t^{++} / \partial t_{il,n} = 0$ を計算し，以下の式を得る．

$$t_{il,n} = \left[\frac{\left(\frac{2}{\nu} + 1\right) \sum_j \frac{1}{\alpha_{ij,n}} |y_{ij,n}|^2 \gamma_{ij,nl}^{\frac{2}{p}+1} v_{lj,n}^{-\frac{2}{p}}}{\sum_j \frac{1}{\beta_{ij,n}} v_{lj,n}} \right]^{\frac{p}{p+2}} \tag{34}$$

式 (25) と (33) を式 (34) に代入することにより，基底 $t_{il,n}$ に関する以下の更新式を得る．

$$t_{il,n} \leftarrow t_{il,n} \left[\frac{\sum_j |y_{ij,n}|^2 \left(\frac{\nu}{\nu+2}\sigma_{ij,n}^2 + \frac{2}{\nu+2}|y_{ij,n}|^2\right)^{-1} \sigma_{ij,n}^{-p} v_{lj,n}}{\sum_j \sigma_{ij,n}^{-p} v_{lj,n}} \right]^{\frac{p}{p+2}} \tag{35}$$

式 (35) と同様にして，アクティベーション $v_{lj,n}$ に関する以下の更新式を得る．

$$v_{lj,n} \leftarrow v_{lj,n} \left[\frac{\sum_i |y_{ij,n}|^2 \left(\frac{\nu}{\nu+2}\sigma_{ij,n}^2 + \frac{2}{\nu+2}|y_{ij,n}|^2\right)^{-1} \sigma_{ij,n}^{-p} t_{il,n}}{\sum_i \sigma_{ij,n}^{-p} t_{il,n}} \right]^{\frac{p}{p+2}} \tag{36}$$

パラメータ $t_{il,n}$ と $v_{lj,n}$ を更新した後, 低ランクモデル $\sigma_{ij,n}^p$ が式 (20) に従って更新される．

4.4 *t*-ILRMA におけるスパース性と低ランク性の関係

複素 Student's *t* 分布を生成モデルに持つことによりスパースな信号を仮定することが出来たが，そのことと低ランクモデリングとの関係はどうなっているのであろうか．この疑問に答えるため，前節にて導出された低ランク音源モデル（NMF）パラメータ更新式を従来の板倉斎藤擬距離基準 NMF で解釈してみる．例えば，基底 $t_{il,n}$ に関する更新式 (35) は，以下のように書き換えることが出来る．

$$t_{ik,n} \leftarrow t_{ik,n} \left[\frac{\sum_j \frac{z_{ij,n}^p}{\left(\sum_{k'} t_{ik',n} v_{k'j,n}\right)^2} v_{kj,n}}{\sum_j \frac{1}{\sum_{k'} t_{ik',n} v_{k'j,n}} v_{kj,n}} \right]^{\frac{p}{p+2}} \tag{37}$$

$$\begin{aligned} z_{ij,n} &= \left(\sum_{k'} t_{ik',n} v_{k'j,n} \right)^{1-\frac{2}{p}} \left[\frac{\nu}{\nu+2} |y_{ij,n}|^{-2} \right. \\ &\quad \left. + \frac{2}{\nu+2} \left(\sum_{k'} t_{ik',n} v_{k'j,n} \right)^{-\frac{2}{p}} \right]^{-1} \\ &= \sigma_{ij,n}^{p-2} \left(\frac{\nu}{\nu+2} |y_{ij,n}|^{-2} + \frac{2}{\nu+2} \sigma_{ij,n}^{-2} \right)^{-1} \end{aligned} \tag{38}$$

式 (37) は,「板倉斎藤擬距離基準 NMF において指数乗に関して一般化された更新式 [22]」の形になるよう整式されている．ここでは，各パラメータの挙動に関して以下のように解釈することが出来る．

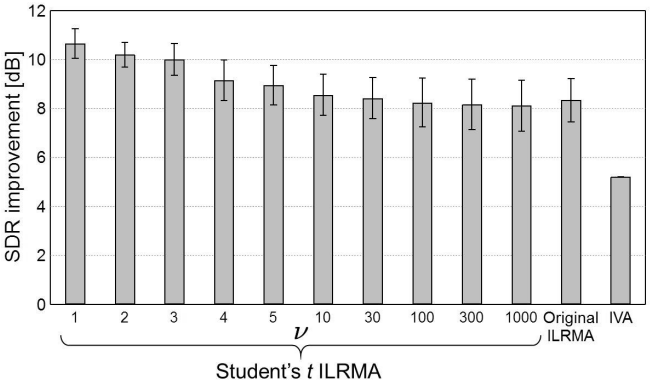


図 4: SDR results of *t*-ILRMA for various parameter ν.

- 式 (37) は，仮想的な観測信号 $z_{ij,n}$ に関する板倉斎藤擬距離基準 NMF だと見なせる．また，その $z_{ij,n}$ は，真の観測信号 $y_{ij,n}$ と低ランクモデル $\sigma_{ij,n}$ の調和平均で表される（式 (38) 参照）．上記の調和平均の比率は ν 対 2 であることより，自由度パラメータ ν を小さくする（よりスパースな信号を仮定する）につれて低ランク性が強調されることになる．

- ドメインパラメータ p は主に NMF における乗算更新量の指数 $p/(p+2)$ を制御する．p を小さくするほどこの指数は小さくなり，NMF の更新速度が遅くなる．一般に ILRMA においては，分離行列の更新速度とのバランスによって局所解への停留確率が変化することより，特に反復初期においてこの指数は小さい方が好ましいとされている [22].

- また，式 (38) には $\sigma_{ij,n}^{p-2}$ が含まれる．これは，$p < 2$ といった場合，低ランクモデルでの除算を意味し，低ランク性をディスカウントする効果が生じる．これにより，過度な低ランク性の強調が抑制される．

以上より，自由度パラメータ ν とドメインパラメータ p の設定についてまとめると，比較的小さな p 及び大きな ν で反復を開始し，徐々に ν を下げて低ランク性を強調していく tempering（焼き戻し）が有効であると言える．

4.5 実験的比較例

t-ILRMA におけるスパース音源モデルによる性能改善を示す為に，実際の音響データを用いた分離実験結果例を Fig. 4 に示す（実験条件の詳細は [21] 参照）．ここでは，前節で述べた tempering を行っており，パラメータの反復更新前半では $\nu = \infty$ とし，後半では図の横軸に示す値にしている．本図より，複素ガウス分布由来の手法と比べて ν の小さな *t*-ILRMA の分離精度が高く，スパースな音源モデルの導入が効果的であることが示されている．

5 事前分布・正則化の導入

2.4 節にて解説した ILRMA において，分離行列や低ランク音源モデルパラメータに事前分布を入れることで分離

性能を安定化させることも可能である．これは，各パラメータに関する正則化項を主コスト関数 Q_{ILRMA} へ加えることで実現される．以下にその実装例を示す．

• **低ランク音源モデル：** 従来の板倉斎藤擬距離基準 NMF と同様にして，スパース性を誘導する正則化項を付与することで音源モデルパラメータの事前分布を与える．例えば，文献 [23] では，アクティベーション $v_{lj,n}$ の事前分布としてラプラス分布を仮定し，L1 ノルムを正則化として加えるものが提案されている．また，Mitsui らによって，低ランク表現された音源スペクトログラムの事前分布として周波数毎に独立なカイ分布を仮定するものが提案されており，ILRMA における性能向上が報告されている [24].

• **分離行列：** FDICA 等の最急降下法においてはその勾配に適切な正則化項の勾配を加算するだけで実装できるため，分離行列の正則化に関して多くの手法が提案されてきた [5]．一方で，IVA や ILRMA 等の補助関数法に基づく手法では，正則化項の加算によって補助関数の最小点が変化するため，解析解の導出が困難になるという問題があった．近年，三井らによって，分離行列の事前分布としてガウス分布を仮定する場合（分離行列に関する L2 正則化）に限り，それを分離行列の行ベクトル毎のブロック座標降下法で解くアルゴリズムが提案されており，ILRMA における有効性が確認されている [25].

6 おわりに

本稿では，ランク 1 空間近似を用いた 3 つの BSS について，それらの音源及び空間モデルに関して考察した．特に，空間モデルパラメータの最適化に関し，MNMF 等との比較を通じて，分離モデル型に利があることを解説した．更に，実験的な比較例を通じ，ILRMA が分離精度及び演算量の両面において優れていることを示した．また，ILRMA の拡張として，スパース音源分布モデルを仮定した手法に関しても解説を行った．

謝辞 本稿を執筆するにあたり，貴重なご助言及び資料をご提供頂いた東京大・北村大地氏に深く感謝いたします．本研究の一部は，セコム科学技術振興財団及び総合科学技術・イノベーション会議による革新的研究開発推進プログラム（ImPACT）の支援を受けた．

References

[1] P. Comon, “Independent component analysis, a new concept?,” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[3] S. Kurita, H. Saruwatari, S. Kajita K. Takeda and F. Itakura, “Evaluation of blind signal separation method using directivity pattern under reverberant conditions,” *Proc. ICASSP*, vol. 5, pp. 3140–3143, 2000.

[4] N. Murata, S. Ikeda and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[5] L. Parra and C. V. Alvino, “Geometric source separation: Merging convolutive source separation with geometric beamforming,” *IEEE Trans. SAP*, vol. 10, no. 6, pp. 352–362, 2002.

[6] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, “The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech,” *IEEE Trans. SAP*, vol. 11, no. 2, pp. 109–116, 2003.

[7] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee and K. Shikano, “Blind source separation based on a fast-convergence algorithm combining ICA and beamforming,” *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[8] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Proc. Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[9] A. Ozerov and C. Fevotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[10] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux and K. Kashino, “Statistical model of speech signals based on composite autoregressive system with application to blind source separation,” *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, pp. 245–253, 2010.

[11] H. Sawada, H. Kameoka, S. Araki and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[12] A. Hiroe, “Solution of permutation problem in frequency domain ICA using multivariate probability density functions,” *Proc. ICA*, pp. 601–608, 2006.

[13] T. Kim, H. T. Attias, S.-Y. Lee and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[14] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. WASPAA*, pp. 189–192, 2011.

[15] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[16] N. Q. K. Duong, E. Vincent, and R. Gribonval, ”Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[17] S. Amari, S. Douglas, A. Cichocki, H. Yang, “Multichannel blind deconvolution and equalization using the natural gradient,” *Proc. IEEE International Workshop on Wireless Commun.*, pp. 101–104, 1997.

[18] E. Vincent, R. Gribonval and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe and A. Benichoux, “The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation,” *Proc. LVA/ICA*, pp. 414–422, 2012.

[20] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[21] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, “Independent low-rank matrix analysis based on complex Student’s t-distribution for blind audio source separation,” *Proc. MLSP*, 2017.

[22] Y. Mitsui, D. Kitamura, N. Takamune, H. Saruwatari, Y. Takahashi, K. Kondo, “Independent low-rank matrix analysis based on parametric majorization-equalization algorithm,” *Proc. CAMSAP*, 2017.

[23] W. Liu, N. Zheng, X. Lu, “Non-negative matrix factorization for visual coding,” *Proc. ICASSP*, vol. 3, pp. 293–296, 2003.

[24] Y. Mitsui, D. Kitamura, S. Takamichi, N. Ono, H. Saruwatari, “Blind source separation based on independent low-rank matrix analysis with sparse regularization for timeseries activity,” *Proc. ICASSP*, pp. 21–25, 2017.

[25] 三井祥幹, 高宗典玄, 北村大地, 猿渡洋, 高橋祐, 近藤多伸, “空間事前情報を用いた独立低ランク行列分析,” 第 32 回 SIP シンポジウム, no. B8-2, pp. 360–365, 2017.

一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-7 (11/25)

DNN Based Pitch Estimation Using Microphone Array*

Jani Even, Carlos Toshinori Ishi, Hiroshi Ishiguro
Hiroshi Ishiguro Laboratories, Advanced Telecommunications Research Institute International, Japan.
even@atr.jp *

Abstract

This paper presents some preliminary experiment for pitch classification of distant speech recorded with a microphone array. The pitch classification is performed by a deep neural network. Using the microphone array to perform beamforming is beneficial to the pitch classification. However it requires a larger amount of data for training the network. The network seems to be robust to data miss-matched as pretraining with close speech data improved the results for distant speech.

1 INTRODUCTION

This paper presents some preliminary results on the use of deep neural network for pitch classification. In particular, the goal is to investigate the possible improvement obtained by applying beamforming when considering distant speech. Pitch classification using neural network was applied by the authors of [1] using hand engineered features. Recent advance in neural networks make it possible to train deep architecture [2] that learns the features. In [3], the authors proposed different deep neural networks to estimate pitch. However, distant speech and the use of microphone array was not investigated.

2 OVERVIEW

Figure 1 shows the overview of the training phase and the testing phase. In the two phases, the voice of the subject was recorded using a linear microphone array (8 microphones with a spacing of 0.02 m) and a tie microphone. All the microphones are similar omni-directional microphones.

In order to access the improvement obtained by using the microphone array, the features are either extracted from one of the microphone of the microphone array or from a delay and sum beamformer (see [4] for microphone array processing). In the remainder of the paper, the features extracted from the single microphone are denoted as "far" features. The features extracted from the output of the delay and sum beamformer are denoted as "DS" features.

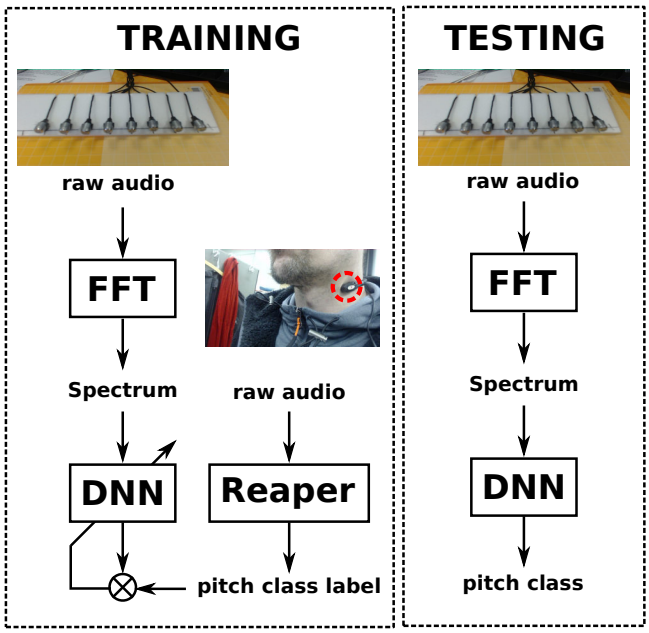


Figure 1: Overview of the training (left) and the testing (right).

The labels for pitch are extracted from a tie microphone by the "reaper" software [5]. This software simultaneously estimates the location of glottal closure instants, voicing state and pitch. Only the pitch estimation was used to label the data. Some experiments were conducted using the PRAAT software [6] to estimate the pitch label. The reaper software was preferred because it was easier to automate the labeling task with it.

3 DATA COLLECTION

The data corpus consists of approximately 42 minutes of speech data recorded from a single male speaker.

*Research supported by the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

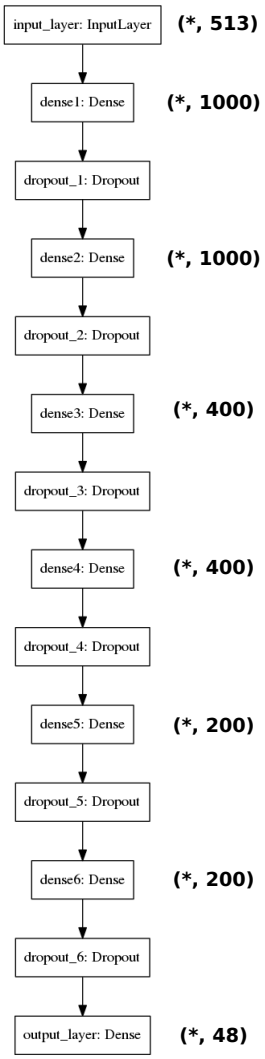


Figure 2: Architecture of the network.

The sampling frequency is 16 kHz. The audio data is transformed in the frequency domain with using a sliding hanning window of 320 ms with half overlap. The FFT size is 1024. The feature vector for each frame is the modulus of the 513 positive frequency components.

During extraction with the reaper software, the pitch values were limited to the range $[50, 300]$ Hz. The range $[50, 280]$ Hz was linearly divided in 46 bins (5 Hz per bin). A bin for non voiced frames (bin 1) and another bin (bin 48) for voiced frames over 280 Hz were also created. Thus the total number of pitch classes is 48.

The data was separated in testing set and development set. The development data is further split in training and validation set.

4 NEURAL NETWORK TRAINING

The pitch classifier is composed of 6 fully connected layers. The input layer has the feature vector size $F = 513$. The output layer is of size $C = 48$ corresponding to the number of pitch classes. Figure 2 shows the network and the number of units in the different layers. All the activations in the layers are "softplus" except for the output

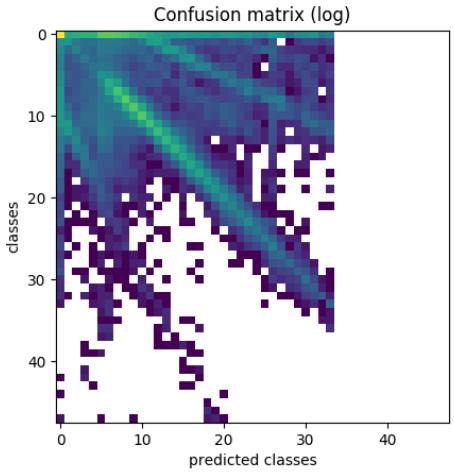


Figure 3: Confusion matrix in log scale for the "far" features without pre-training.

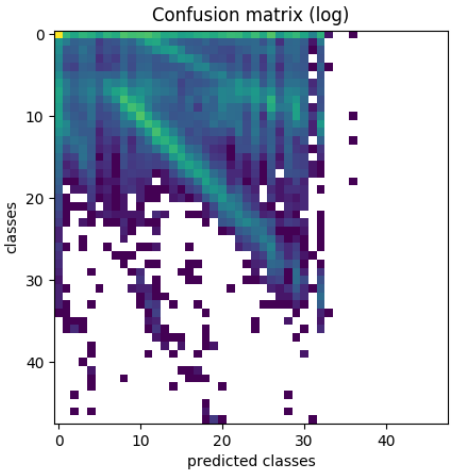


Figure 4: Confusion matrix in log scale for the "DS" features without pre-training.

layer that is using a "softmax". This neural network part is implemented using Keras [7] with Tensorflow backend [8]. The network has a total of $2,205,848$ parameters.

The network was either initialized randomly or using some pre-trained weights. The pre-trained weights were obtained by training the network with audio data from the Librispeech database [9]. LibriSpeech is a corpus of 16kHz read English speech. We used 100 hours of clean speech to train the model. During the pre-training, the pitch labels and the features are extracted from the same signal. Namely, both the features and the pitch labels are from close speech.

Since we are considering a classification problem, the network was trained to optimize the categorical cross entropy. When training from random weights, the adaptive subgradient method ("Adagrad" in Keras) was used to update the weights [10]. When using pre-trained weights, in order to move slowly away from the initial weights, a stochastic gradient descent with a small step size was used. For all the operations, batch of 100 samples were used.

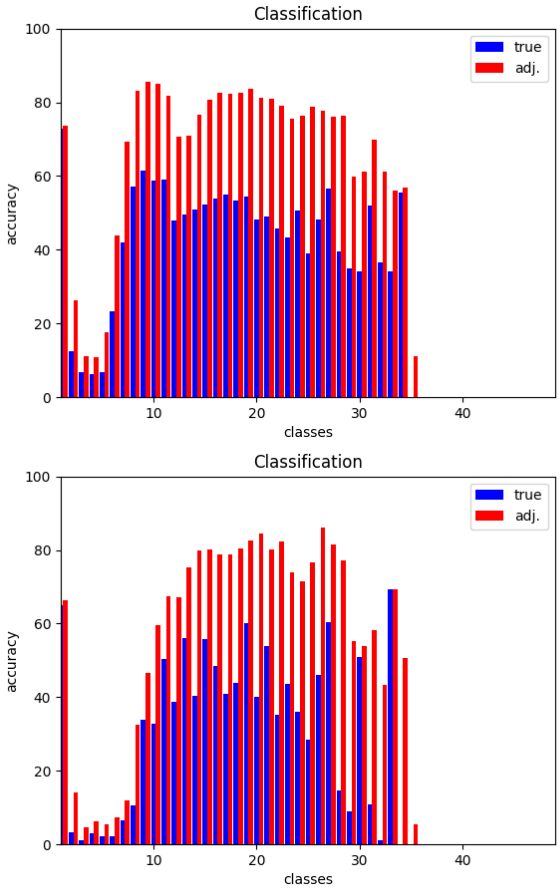


Figure 5: Classification results for the "far" (top) and "DS" (bottom) features without pre-training.

After the training, in each of the cases, the network having the best score on the validation set was selected.

5 CLASSIFICATION PERFORMANCE

5.1 Without pre-traing

Figure 3 shows the confusion matrix (log scale) for the classification of the "far" features and figure 4 shows the corresponding confusion matrix for the "DS" features.

The confusion matrices clearly show that the classification error is usually because of assignment to the neighboring bins. Thus, we give the results in term of true classification accuracy (classification in the true bin) and adjacent classification accuracy (classification in the true bin or its two immediate neighboring bins).

Figure 5 shows the classification results for both sets of features. The classification results when using the output of the delay and sum are worse. Moreover for both cases, the higher bins are not well classified.

5.2 With pre-training

The pre-training phase was added in order to improve the poor results obtained by direct training of the network.

The confusions matrices in figures 6 and 7 are still showing some errors between neighboring bins.

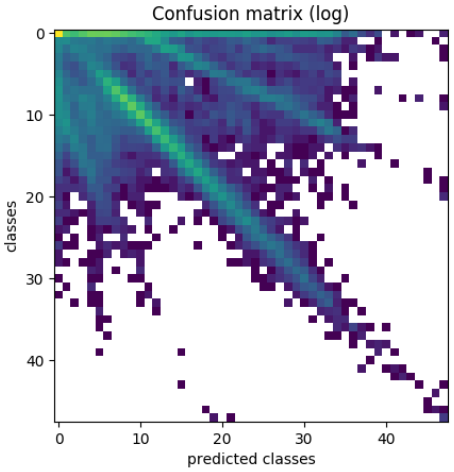


Figure 6: Confusion matrix in log scale for the "far" features with pre-training.

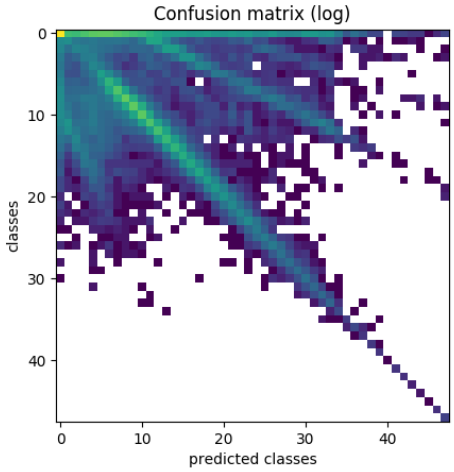


Figure 7: Confusion matrix in log scale for the "DS" features with pre-training.

However, the classification results in 8 are greatly improved by the addition of the pre-training phase. In particular, the results for the microphone array ("DS" features) are better than those for the "far" features.

The higher classes that were not well classified are now perfectly classified. The reason is that these classes were under represented in the dataset but appear in the data used for pre-training. This can be observed in the confusion matrices.

It is important to notice that the pre-training was performed using completely miss-matched data as the LibriSpeech corpus contains close talking speech recorded with a single microphone. This is an interesting results as it suggests that the DNN based pitch classifier is quite robust to data miss-match.

6 CONCLUSIONS

The experiment presented in this paper shows that using a microphone array to perform delay and sum beamforming improves the DNN based pitch classification of

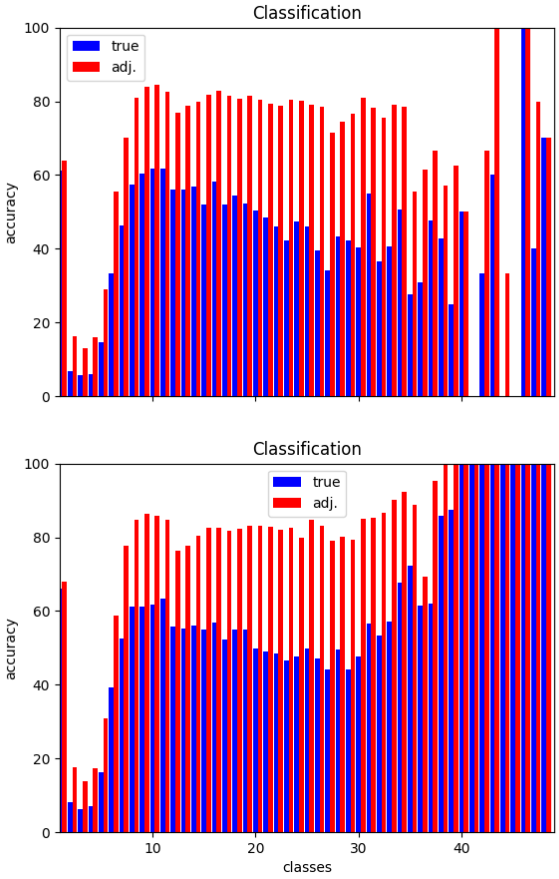


Figure 8: Classification results for the "far" (top) and "DS" (bottom) features with pre-training.

distant speech. However, it seems that a larger amount of training data is necessary as the microphone array results were only better than those of the single distant microphone when a large amount of data was used for pre-training. That improvement was very clear even if the pre-training was done with miss-matched conditions. The future research is testing this in controlled noisy environments in order to clearly assess the performance. Another point of interest is to access the robustness of the DNN based pitch classifier to data miss-match.

References

[1] E. Barnard, R. A. Cole, M. P. Vea, and F. A. Alleva, "Pitch detection with a neural-net classifier," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 298–307, 1991.

[2] G. E. Hinton, S. Osindero, and Y.W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[3] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2158–2168, 2014.

[4] J. DiBiase, H. Silverman, and M. Brandstein, *Microphone arrays : Signal Processing Techniques and Applications*, Springer-Verlag, 2007.

[5] David Talkin, "Reaper," `https://github.com/google/REAPER`, 2015.

[6] P. Boersma and D. Weenink, "Praat version 5.3.02," `http://www.praat.org/`, 2011.

[7] François Chollet et al., "Keras," `https://github.com/fchollet/keras`, 2015.

[8] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[10] John Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.

一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-8 (11/25)

アンドロイドの動作生成に向けた自然対話中のジェスチャの認識および分類に関する検討

町屋敷　大地 , 石井カルロス寿憲, 劉超然, 石黒浩
Daichi Machiyashiki, Carlos Toshinori Ishi, Chaoran Liu, Hiroshi Ishiguro
ATR(石黒特別研究所)/大阪大学, ATR(石黒特別研究所), ATR(石黒特別研究所), ATR(石黒特別研究所)
ATR HIL/Osaka University, ATR HIL, ATR HIL, ATR HIL
machiyashiki.daichi@irl.sys.es.osaka-u.ac.jp

Abstract

ロボットの動作生成を目指して, 人の対話と同時に起こるジェスチャーの分類とそれらのジェスチャー中の手の位置や, 発話との関係を調査した. ジェスチャーの分類はアノテーターによってラベルづけされ,k-means 法によってジェスチャーの手の動きのクラスタを生成した. また Wordnet からジェスチャーとともに現れる発話の上位概念を取得した. これらの取得したデータの関わりを今後も調べていき, ロボットの動作生成へとつなげていく.

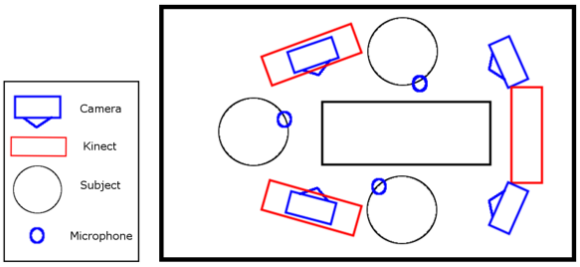


図 1: Environment setup for 3 person dialogue experiment.

1　はじめに

近い未来, アンドロイドなどのロボットがより大衆的になり, 社会の中での人に役割の一部を代替する場面が増加すると思われる. その役割の一つとして, 人と向き合って対話を行う人間の役割の代替が考えられる. その場合テキストのみによる対話システムと異なり, 対話の中に言語情報だけでなく非言語情報も含まれる. 人間とロボットが対面して自然な対話を行うためには, それらの情報からロボットが相手の伝えたいことを正しく理解したり, 自分の表現したいことが相手に伝わるように表現できることが望ましい. 当研究室では, 発話に伴う口唇・頭部・表情・腰部の動作生成に関する研究をこれまで数多く報告してきた [2][3][4][5][10][11]. 当研究では, 対話の中で出現する非言語情報のうち発話とともに現れるジェスチャーに焦点を当て, その動作生成を目的として, 人間の対面対話中に現れるジェスチャーの分析を行った.

対話の中におけるジェスチャーにはすでにたくさんの研究が行われているが, ジェスチャーをその機能ごとに分類する方法として有名なものは McNeill によるものである [13]. McNeill はジェスチャーのうち同時に話された言葉に関連する動きを映像的なジェスチャー, 隠喩的なジェスチャー, 拍子のジェスチャー, 指示を表すジェスチャーの 4 種類に分類した. 映像的なジェスチャーとは具体的な物体を表すものであり, 例えばリンゴの形を表すためにリンゴの形を手で表現するジェスチャーが当てはまる. 隠喩的なジェスチャーは抽象的な概念を表出するジェスチャーであり,「行く」などの動作や炎や心などの不定形なものを表すのに用いられる. 拍子のジェスチャーは対話の強調など目的とした手や指の振動で, 指示を表すジェスチャーは指さしジェスチャーが該当する. また複数の種類の重複が許されている. この分類のうち映像的なものを用いて CG エージェントの動作を生成する取り組みを LÜcking らが, 隠喩的なものに関しては Wordnet を用いて門野らが行っている [12] [6].

人間の代替を目的とする場合, 自然なふるまいを行うには, 特定の種類のジェスチャーだけでなく全ての種類のジェスチャーの生成を行う必要があり, さらに対話中の人間の動作には言語と関係のないアダプターと呼ばれる癖の動作が含まれる. この研究では, それらのジェスチャーの種類と癖がそれぞれどのようなときに現れ, 実際の手の動きとどのような関係があるのか, 個人差の影響はどの程度あるのかを分析し, ハンドジェスチャーが可能なアンドロイドのジェスチャーを生成することを最終目的とする.

2　対話データの取得

人間同士による自然な対話データの収集を目的に対話実験を行った. 様々な対話状態に対応したシステムを作るため,3 人による自由対話を記録し, そのジェスチャーを分析した. これは,2 人での対話では発話相手が決まっていて 1 人が発話するともう 1 人が必ず対応しなければならないなど状況が限定されるからである. 今回分析したのは 3 組による対話で, 被験者は互いに面識がある. 年齢層は 20 代から 40 代の男女からなる. 実験を複数回行った. 本稿ではそのうち拍子に関するもの以外のアノテーションが完了した 5 人分のデータを使用する.　実験環境は図 1 に示されるように四角いテーブルの 3 面に一人ずつ座った状態で, 約 45 分間自由対話をしてもらった. 途中, 座っている席の位置によるデータの偏りを防ぐため席順を 2 度変更した.4 台のカメラによって各被験者を撮影し,3 台のキネクトにより骨格情報および映像を, 各被験者が各々に装着したマイクロフォンと机に設置したマイクロフォンアレイから音声情報を記録した. また, 実験に使うテーブルは被験者の手が見えるように十分低くなっている.　実験後, 記録した動画に Openpose[14] を使用し, 全身の骨格 18ヶ所および, 手 21ヶ所, 顔 70ヶ所の 2 次元位置データを取得, 記録した. 座標系は水平軸を x 軸, 垂直軸を y 軸とし, データの範囲は 0 から対象となる動画の画素の各軸に対する数までとなる.

3　データ処理

3.1　対話中に現れる手の動きの分類

本研究では得られた実験データ中に現れる手に関連する動きを次の要素にしたがって定義し, 分類した.

1. ジェスチャーと癖
   対話と関連して現れる動きをジェスチャー, それ以外を癖とする. 例えば対話と関係なく自分の体を触る動きや, 姿勢の変更に関連して発生する手の動きは癖に分類される. また, 動きがない状態をホーム, その時の手の位置をホームポジションと呼ぶ.

2. ジェスチャーのフェーズ
   ジェスチャー動作には一連の流れが存在する. 本研究ではそれらを構成する要素をジェスチャーのフェーズと呼び,Kendon[8] の分類法に従って以下の 4 つのフェーズに分類する.

   - 準備
     ジェスチャーを行うために, ホームポジションから手を動かすフェーズ. 存在しない場合もある.
   - ストローク
     ジェスチャーのメインとなる意味がある動きを行うフェーズ.
   - 終わり
     準備フェーズとは反対に, 手の位置をホームポジションに戻す動き.
   - ホールド
     ストロークが起きる前後に手がホームポジションではない場所にとどまっているフェーズ.

3. ストローク内でのジェスチャーの機能
   McNeill[13] の分類のうち映像的, 隠喩的, 拍子, 指示とエンブレムの 5 つのカテゴリに分類する. エンブルムとは OK サインや V サイン, バイバイの動作など形と意味が社会的に定まっている動きである. さらに, 映像的, 隠喩的, 指示のジェスチャーに関して, 先行研究 [6] や書籍 [9] を参考に一部変更を加え, より詳細なサブカテゴリを定義した. 詳細を Figure 2 に示す. McNeill の定義では, 映像的, 隠喩的, 拍子, 指示の各カゴテリは重複可能であるが, 本研究では, 映像的, 隠喩的,, 指示それぞれと拍子との重複のみを考える. 拍子に関するラベルは以下のとおりである.

   - 動き
     - 単発
       上下運動一回のみ
     - 連続
       上下運動が連続して起こる
   - 意味
     - 強調
       強調部分で行われるビート
     - リズム
       単にテンポよく行われるビート

3.2　対話中に現れる手の動きのアノテーション

実験で得られたデータに対して 1 人のアノテーターが映像と音声によるジェスチャーの区間と機能のラベル付けを 3.1 章で定義したカテゴリに従って行った. 以降このラベルをジェスチャーの機能ラベルと呼ぶ. また, ジェスチャーと発話が同時に発生した場合, その発話を含む一文を書き起こした.

3.3　ジェスチャーの動作の特徴抽出

実験で得られたデータから各ジェスチャーのストロークフェイズにおける手の振る舞いについて調査する. 手の動きは複雑かつ多岐にわたるため基準となるクラスタの生成を試みた. Openpose から得られるデータは各関節に対応した画素の絶対位置であるため, 話者ごとにカメラからの距離や体格差によって同じ動きが違ったデータになること

映像的	図像	物の形や大きさを形取る
	図像2	実在する人などを表すが、形をなぞっていない（丸く描く等）もの
	動作（映像的）	イメージがつきやすい、具体的な動作（行く、食べる、歩く）
暗喩的（隠喩的）	名詞（名_隠喩的）	形の定まらない名詞、単語の概念（記憶、余裕、経験）、みんな等集団を表すもの
	動作（動_隠喩的）	イメージがつきにくい、抽象的な動作（...になる、夜が明ける、経験する）
	修飾（修飾）	「優しい」、「きれい」などの修飾語
	時間（期間）	「1年間」などの一定期間
	時間（時点）	「今」、「昨日」などの時間の一時点
	考え（否定）	「否定」を表すジェスチャー
	考え（考_その他）	その他の考え
	関係（関係）	物と物の関係性（年齢が上、先輩で...）
	様子（程度）	「ちょっと」、「いっぱい」などの物事の程度
	様子（擬音）	「ぐちゃぐちゃ」、「ぶりぶり」、「スペスペ」などの擬音、擬態
	様子（様_その他）	その他の様子
指示	指示（自分）	指さし動作。指さしのストロークは区切りにくい、指示する単語の発話のタイミングを考えながら区切る方法をとる。 自分を差すときは指示（自分）とし、相手を指差す場合に相手の被験者ナンバーで差した相手を判断する。
	指示（C01～C03）	
	指示（指_その他）	
慣習的（エンブレム）	慣習的（エンブレム）	形と意味が社会的慣習として定まっている（OKサイン、バイバイ、指で数を表す）

図 2:　Gesture categoriesThe list of categories and subcategories of gesture meanings in stroke phase

を防ぐため, 平滑化によるノイズ除去を行った後, すべてのデータを首元からの相対位置に変換した. さらに, 文献 [1] を参考にして x 軸に対して右肩から左肩の距離を 1,y 軸に対してのど元から腰の長さが 1 となるようにスケーリングを行った. 本稿では, 各腕の動きとして各手首の軌跡を用いる.

アノテートされたラベルから各ジェスチャーのストロークフェイズでの軌跡を取得する. 取得した軌跡は各腕に対して x 方向, y 方向, 時間の 3 軸からなる 3 次元データとなる.　ジェスチャーの動作をその機能や同時に発言された言葉と対応づけるためには, 何らかの手法で, ジェスチャーの動きを分類しなければならない. 本研究では k-means によるクラスタリングを行う. k-means とは d 次元の n 個のベクトルの観測 $(x_1, x_2, ..., x_n) \in$ x と k 個のクラスタ $(c_1, c_2, ..., c_k) \in$ C が存在する場合

$$argmin_C \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{C_i}} cost(\mathbf{x}, \boldsymbol{\mu}_\mathbf{i}) \qquad (1)$$

となるクラスタを計算アルゴリズムである. ただし,$cost(\mathbf{a}, \mathbf{b})$ はベクトル $\mathbf{a}$ と $\mathbf{b}$ の距離を測る何らかのコスト関数で,$\boldsymbol{\mu}_i$ はクラスタ c_i に含まれるベクトルすべての平均値である. k-means に用いるコスト関数として, 一般的にはユーグリッド距離が使用される. ジェスチャーの軌跡を扱うにあたって, 時系列データである軌跡に一般的なユーグリッド距離を用いた方法はジェスチャーには時間的は伸縮が考えられるため有効ではない. 次に時間的な伸縮にロバストな Dynamic Time Warping(DTW) を使用することが考えられるが, ジェスチャー開始時点での手の位置や動きの大きさによる影響によりうまくいかなかった. そこで, 本研究では, ジェスチャーの軌跡を x-y 平面に射影し, その写像の類似度をコスト関数とする k-means を用いることでクラスタの生成を試る. 写像する平面の範囲は,x 軸に関して肩幅の 2 倍,y 軸に関して肩から腰の下までとし, 計算量の観点から写像した平面の画像を 64*48Pixel に縮小した, また, 縮小する際軌跡が途切れないよう縮小前の軌跡の幅を図 3 で示したものの 10 倍とした. 写像の類似度の計算には比較する点の位置のずれにある程度ロバストな SSIM を用いた. この操作により生成したクラスタをジェスチャー動作クラスタと呼ぶ.

3.4　WordNet による概念抽出

ジェスチャーの機能ラベルや動作の特徴と対話に出現する単語の関係を調べるため, 先行研究に従って WordNet[7] による単語の概念抽出を行った.WordNet は単語を synset と呼ばれる類義関係のセットでグループ化していて, 一つの synset が一つの概念に対応している.Wordnet にはこれらの一つ一つの概念に上位の概念や下位の概念といったほかの概念との関係性が記録されており, これを利用することで上位語 (Hypernym) や下位語 (Hyponym) などを取得することができる. 本研究では先行研究と同様に単語どうしの関連度を測るために国立研究開発法人情報通信研究機構が提供する WordNet(http://compling.hss.ntu.edu.sg/wnja/) を用いる. 本稿ではジェスチャーとともに現れた発話一文すべてをジェスチャーと同時に起こったとして解析を行った. アノテートされた文に形態素解析を行ったのち, 各単語を原形に変換し,WordNet を用いて概念を抽出する. 一単語には複数の概念を持つものがあるが, 統計的な問題からそのうち一つを選ばなければならない. また上位の概念も複数存在するのでどの上位概念を取得するかを考える必要がある. 今回は, その方法として単語から 2 段階上の概念を取得し, もし複数の概念が存在していたならば, 過去にすでに出現したことのあるものを優先的に選んだ.

3.5　関係性の計算

3.3 章と 3.4 章で生成したデータとジェスチャー機能ラベルの関係を計算した. 次章はその結果を示す.

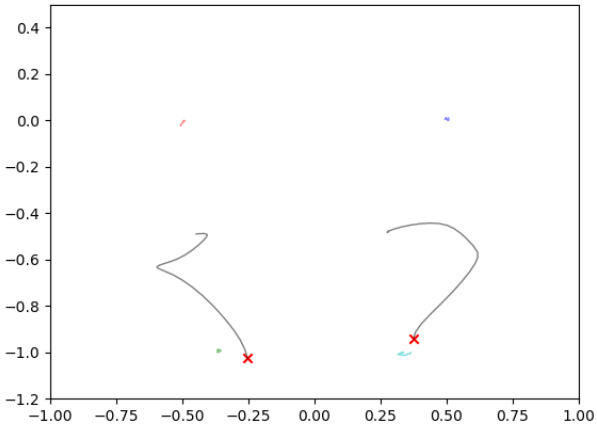


図 3: An example of projected gesture trajectories of both hands.
Cross points indicate start points of the gesture.
X-axis and y-axis are normalized to the shoulder and hip joints drawn by light dots.

4 結果

この章ではジェスチャー機能ラベル, ジェスチャーの動き, ジェスチャーと同時に出現する単語の関係に関する 3 調査結果を示す. データは実験を行ったグループのうちアノテーションが完了している 1 グループのデータを用いて計算した.

Fig 4 に示す通り, ジェスチャー機能のラベルと癖の動きの分布は, 映像的ジェスチャーが 57 回, 暗喩的ジェスチャーが 221 回, 拍子ジェスチャーが 504 回, 指示的ジェスチャーが 110 回, 癖が 146 回だった. なお, 拍子ジェスチャーはほ

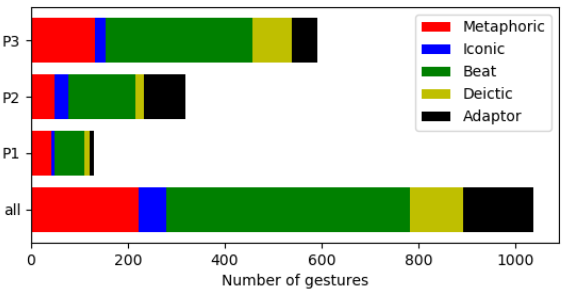


図 4: Number of occurrences of gesture meaning categories.

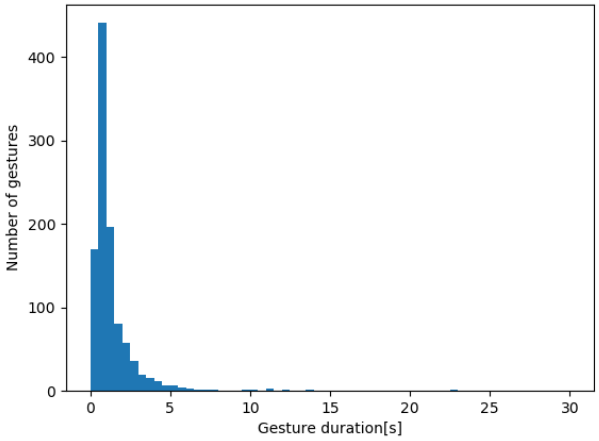


図 5: Histogram of gesture durations in stroke phases.

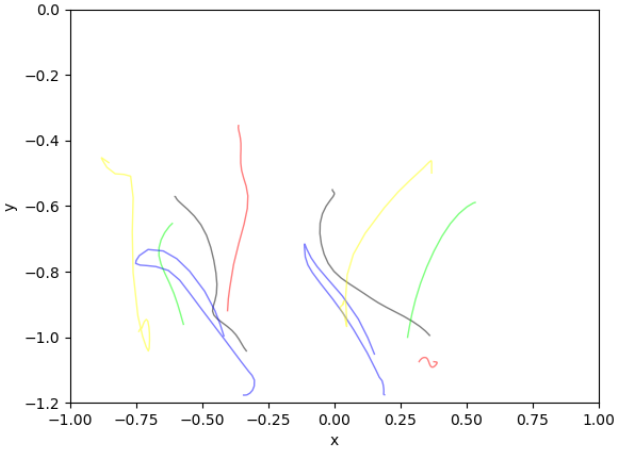


図 6: Example of a gesture cluster potentially expressing up-down movements.

かのジェスチャーと重複を許すが, 今回は, 重複していないものだけを拍子ジェスチャーとして計算した. また, 各ジェスチャーのストローク部分の継続時間は Fig 5 となり,0.5 秒から 1 秒の間が一番多かった.

4.1 ジェスチャー動作のクラスタリング

3.3 章で説明した方法を用いて,3 人分のデータを 100 クラスに分類した時の結果を示す.Fig 6,Fig 7 に示したものは 3 人分のデータを 100 クラスに分類した時のあるクラスタ内に含まれる一部のジェスチャーの軌跡の射影図である. 同じ色で示される 2 曲線が対となる左右の手の動きとなる.Fig 6 に示されるクラスタは, 腕の上下運動を表すと思われる. このクラスタは, 上または下への両手の動き 3 例と上または下への片手の動き 1 例, 両手を上にあげてから下に戻す動きを含んでいる.

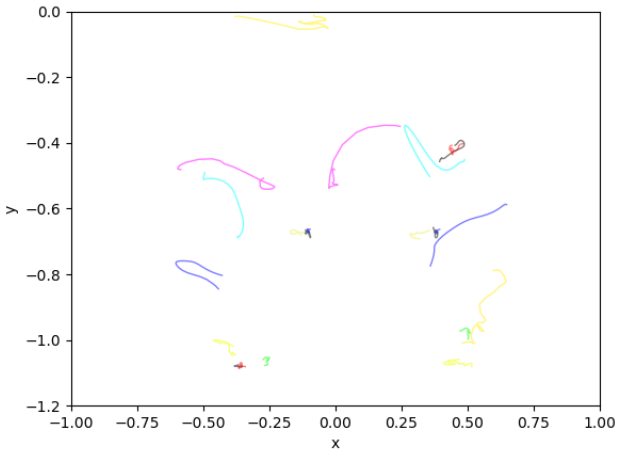


図 7: Example of an inefficient gesture cluster containing different types of trajectories.

表 1: Word concepts mostly appeared along with gestures.

出現回数	
概念	回数
交流	131
個人	124
鑑定	113
動静	102
原因物	77
考える	63
性状	57
属性	57
事	53
オーガナイゼーション	46

4.2 ジェスチャー機能ラベルと出現する単語の関係

Table 1 はジェスチャーとともに出現する概念数を,Table 3 に示したものは各ジェスチャー機能ラベルとそれと同時に出現する単語の概念の関係を表したものである. 左から出現した概念名, 出現回数, 出現した回数のうち, 各ジェスチャー機能ラベルとともに現れる確率を示す. 例えば, 指示カテゴリの中の聴覚コミュニケーションの概念は, 全会話中に 72 回出現し, その中の 36 回は指示的なジェスチャーとともに現れたということを表している. また,Table 2 で示されるのは, ジェスチャー機能のカテゴリごとの名詞の出現度を示す.

5 考察と今後の展望

Fig 4 より, ジェスチャーの機能の割合やジェスチャーが起きる頻度には, 同じ対話に参加していても個人差があることが分かる. ジェスチャー動作のクラスタリングについて,Fig 6 のようにある程度クラスタリングができているクラスタもあったが, できていないものも散見された. その原因として以下のことが考えられる. 一つ目はある点にとどまっているジェスチャーと動いている点では動いている点の影響が大きいことがあげられる. これは射影平面上でジェスチャーの軌跡が通る距離が変化するためでありその結果 Fig 6 内でみられるような, 片一方の手のジェスチャーは類似しているがもう片一方の手の動きは類似していないものが同一のクラスタになる現象が発生すると思われる. 次に, 往復動作と往のみの動作の区別がほとんどできないことも原因として考えられる. さらに, 今回の実験データでは指先の情報を使用しておらず, ジェスチャーが発生しているはずの区間でも軌跡では動いてないように見えるものが多数存在した. これらの改善のために, クラスタリングの手法の見直しや, 指の情報を追加するなど対策を今後行っていく.

表 2: Nouns mostly appeared in different gesture meaning categories.

映像隠喩		指示		エンブレム	
単語	回数	単語	回数	単語	回数
人	34	自分	13	人	9
何	23	私	9	一	6
感じ	9	人	8	二	5
みたい	8	相手	5	年間	3
一	7	一	5	三	3
異性	7	ここ	4	片手	2
友達	6	タイプ	4	感じ	1
それ	6	何	4	対	1
二	5	こと	3	五	1
の	5	みたい	3	日	1

次に, ジェスチャーの機能と単語の対応について,Table 2 では, 名詞に限定して単語の調査をし,Table 1 や Table 3 では,Wordnet を用いて概念を検索したものである. 前者では, 映像隠喩に感じやみたいといった比喩表現, 指示に人物, エンブレムに数字が多い系傾向にあることが分かるのに対し後者のほうは傾向がつかみにくい. これは Wordnet で上位概念に上りすぎたため一つの概念がカバーする範囲が広くなりすぎている可能性が考えられる. 以降は Wordnet をどこまでたどるか, どの品詞をたどるべきかなどを検討していく. また,Wordnet には一つの単語に複数の上位概念が候補として出てくることがあり, 前後の文脈などで正しいものを選択す機構も必要と思われる. これらの課題を解決しながら, 自然な動作が行えるよう動作生成へのモデルの検討も進めていく.

6 謝辞

この研究は JST, ERATO（グラント番号：JPMJER1401）の一環として行われたものです. ジェスチャーの分類やラベリングに協力いただいた三方瑠祐氏, 村瀬妙子氏, 奥野

表 3: Word concepts mostly appeared in different gesture meaning categories.

暗喩的			映像的			振動			指示		
概念	回数	確率	概念	回数	確率	概念	回数	確率	概念	回数	確率
行ない	21	0.428571	個人	124	0.169355	オーガナイゼーション	46	1	ピリオド	19	0.333333
遷移	40	0.425	動く	27	0.115385	抽象的実体	42	0.804878	聴覚コミュニケーション	45	0.272727
考え	12	0.416667	ピリオド	19	0.111111	時	25	0.708333	原因物	77	0.263158
通信	19	0.368421	原因物	77	0.105263	動静	102	0.67	考える	63	0.193548
属性	57	0.333333	交流	131	0.1	機器	12	0.666667	属性	57	0.157895
性質	24	0.333333	容態	32	0.096774	交流	131	0.653846	性状	57	0.142857
機器	12	0.333333	性状	57	0.089286	容態	32	0.645161	個人	124	0.137097
時	25	0.291667	事	53	0.078431	もの	36	0.638889	交流	131	0.130769
結付き	21	0.285714	抽象的実体	42	0.073171	事	53	0.627451	鑑定	113	0.125
内容	29	0.275862	属性	57	0.070175	動く	27	0.615385	事	53	0.117647

美紀氏に感謝する.

参考文献

[1] Hwang, S. J., et al. "Ada-Boost based Gesture Recognition using Time Interval Window." 2015

[2] C. Ishi, C. Liu, H. Ishiguro, N. Hagita. "Evaluation of formant-based lip motion generation in tele-operated humanoid robots," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012), pp. 2377-2382, October, 2012.

[3] C.T. Ishi, C. Liu, H. Ishiguro, and N. Hagita."Head motion during dialogue speech and nod timing control in humanoid robots," Proc. of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010), pp. 293-300, 2010.

[4] C. Ishi, T. Funayama, T. Minato, and H. Ishiguro (2016). "Motion generation in android robots during laughing speech," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016), pp. 3327-3332, Oct., 2016.

[5] C.T. Ishi, T. Minato, H. Ishiguro. (2017). "Motion analysis in vocalized surprise expressions and motion generation in android robots," IEEE Robotics and Automation Letters, Vol.2, No.3, 1748 - 1754, July 2017.

[6] 門野友城, 高瀬裕, and 中野有紀子. "隠喩的ジェスチャの分析とジェスチャ自動付与に向けた検討." 人工知能学会全国大会論文集 29 , 2015: 1-3.

[7] Kyoko Kanzaki, Francis Bond, Noriko Tomuro and Hitoshi Isahara, "Extraction of Attribute Concepts from Japanese Adjectives",.LREC-2008, Marrakech, 2008

[8] Kendon Adam, "Gesticulation and speech: two aspects of the process of utterance" ,The Relationship of Verbal and Nonverbal Communication,pp. 207-227, 1980

[9] 喜多　壮太郎, 身体とシステム　ジェスチャー　考えるからだ, 金子書房, 2002 .

[10] S. Kurima, C. Ishi, T. Minato, and H. Ishiguro. Online Speech-Driven Head Motion Generating System and Evaluation on a Tele-Operated Robot, IEEE International Symposium on Robot and Human Interactive Communication (ROMAN 2015), pp. 529-534, 2015.

[11] C. Liu, C. Ishi, H. Ishiguro, and N. Hagita. Generation of nodding, head tilting and gazing for human-robot speech interaction. International Journal of Humanoid Robotics (IJHR), vol. 10, no. 1, January 2013.

[12] Lücking, Andy, et al. "The Bielefeld speech and gesture alignment corpus (SaGA)." LREC 2010 workshop: Multimodal corpora?advances in capturing, coding and analyzing multimodality. 2010.

[13] McNeill, David. Hand and mind: What gestures reveal about thought. University of Chicago press, 1992.

[14] Shih-En Wei and Varun Ramakrishna and Takeo Kanade and Yaser Sheikh ,"Convolutional pose machines", CVPR, 2016

一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-9 (11/25)

Quad-directional LSTMを用いた音楽音響信号修復法の提案
Musical Audio Signal Restoration using Quad-directional LSTM

谷口亮輔 *1, 干場功太郎 *1, 中臺一博 *1,2
Ryosuke TANIGUCHI*1, Kotaro HOSHIBA*1, Kazuhiro NAKADAI*1,2
東京工業大学 工学院 システム制御系 *1
（株）ホンダ・リサーチ・インスティチュート・ジャパン *2
Tokyo Institute of Technology*1, Honda Reserch Institute Japan*2
{taniguchi, hoshiba, nakadai}@ra.sc.e.titech.ac.jp

Abstract

本稿ではLSTM（Long Short-Term Memory）を用いた音楽音響信号の修復法を提案する．実際にLSTMを適用した場合，情報が比較的スパースである高域の学習が十分でなくなり，修復性能が劣化してしまう．この問題に対し，我々は，入力信号に対して高域を強調するような周波数フィルタを用いて，その解決を試みた．また，この手法の拡張として，順方向の時系列情報だけではなく，逆方向の時系列情報も考慮したBLSTM（Bi-directional LSTM）を用いる方法を提案した．今回，そのさらなる拡張として，時間方向のみではなく，周波数方向の系列情報も考慮することが可能なQDLSTM（Quad-directional LSTM）を用いることを提案し，評価を行う．その結果，時間方向BLSTMのみと比較してより高音域での修復性能が向上することを確認した．

1 はじめに

近年，深層学習が多くの分野に活用され，その有用性が示されている．本稿では，その一手法であるLSTM（Long Short-term memory）を音楽音響信号修復に適用することを検討する．一般的に，深層学習を用いて性能の高いモデルを学習するためには，大量のデータが必要である．実際の音楽音響信号修復にLSTMを適用した場合，学習データ量が少ないと情報が比較的スパースである高域の学習が十分でなくなり，修復性能が劣化してしまう．この問題に対し，これまで我々は，入力信号に対して高域を強調するような周波数フィルタを用いて，その解決を試みた[1]．また，その拡張として，順方向の時系列のみではなく，逆方向の時系列情報も考慮したBLSTM（Bi-directional LSTM）を用いることを提案した[2]．本稿では時間方向のみではなく，周波数方向のBLSTMを構成し，その両方を用いるQDLSTM（Quad-directional LSTM）を用いることを提案する．提案手法では，時間方向のみでは難しい周波数方向の系列を考慮することが可能であるため，より高音域までの詳細な修復が可能になると考えられる．

2 音楽音響信号修復に用いる深層学習

はじめに，音楽音響信号修復に用いる深層学習手法について説明する．

2.1 Recurrent Newral Network（RNN）

RNNは，音楽音響信号のような系列データを扱うのに適した深層学習手法である[3]．特徴としては，Recurrentの名前の通り，系列方向に対して再帰構造を持つことにある．系列番号tにおける，RNNの入力を$\boldsymbol{x}_t$，出力y_tの関係式は以下のように表せる．

$$\boldsymbol{y}_t = f(\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{x}_t + \boldsymbol{W}_{rec}\boldsymbol{y}_{t-1} + \boldsymbol{b}) \quad (1)$$

ここで，fは活性化関数と呼ばれる非線形関数であり，$\boldsymbol{W}_{\boldsymbol{x}}$は入力に対する重み，$\boldsymbol{W}_{rec}$は以前の出力に対する重み，$\boldsymbol{b}$はバイアスを表す．この構造により，RNNは現在と過去の情報を合わせて考慮に入れた出力をすることが可能となっている．しかし，単純な再帰構造のみであるため，データが長くなると，学習時に用いる誤差の伝播の際，消失や発散を起こしてしまう．これを解決するため，RNNの改良型として，LSTMが提案されている[4]．

2.2 Long Short-Term Memory (LSTM)

RNNでは難しい長期依存関係のデータを扱うため，LSTMは，内部に情報を保持する機能，および各種入出力にゲートを持った構造となっている．実際のLSTMの実装には，複数のバージョンが存在するが，今回は1999年に提案されたGersらによるForget Gate (忘却ゲート) をもつ構造のLSTM[4]を用いた．LSTMの構造は図1のようになっており，式(2)~(4)として記述することができる．

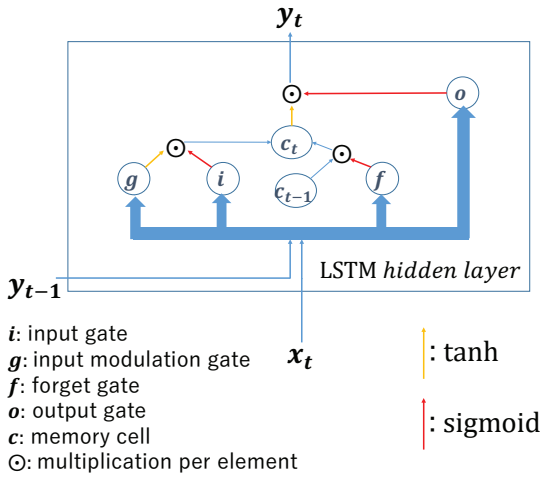


図 1: LSTM の構造

$$\begin{pmatrix} \boldsymbol{i} \\ \boldsymbol{f} \\ \boldsymbol{o} \\ \boldsymbol{g} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \boldsymbol{T}_{2n,4n} \begin{pmatrix} x_t \\ \boldsymbol{y}_{t-1} \end{pmatrix} \tag{2}$$

$$\boldsymbol{c}_t = \boldsymbol{f} \odot \boldsymbol{c}_{t-1} + \boldsymbol{i} \odot \boldsymbol{g} \tag{3}$$

$$\boldsymbol{y}_t = \boldsymbol{o} \odot \tanh(\boldsymbol{c}_t) \tag{4}$$

$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{i}, \boldsymbol{f}, \boldsymbol{o}, \boldsymbol{g}, \boldsymbol{c} \in \Re^n$
σ はシグモイド関数，$\odot$ は要素ごとの積をあらわす．
$\boldsymbol{T}_{2n,4n}$ は $2n$ 次元ベクトルから $4n$ 次元ベクトルを作る写像．

$\boldsymbol{i}$ のインプットゲートは入力を表し，$\boldsymbol{g}$ の入力判断ゲートによって調節されることで直近のデータをどれだけ用いるかが判断できる．$\boldsymbol{c}$ は記憶領域のメモリセルであり，忘却判断ゲート $\boldsymbol{f}$ を通して次の時間での依存関係を表すために用いられる．出力判断ゲート $\boldsymbol{o}$ は次の層へ渡す信号の量を調節する．それぞれの判断ゲートは，直前，現在のメモリ状態，隣接する層からの入力値を用いて，その時々に情報を用いるかという判断のみを行う．そのため，判断ゲートの学習では誤差の消失，異常な増加に対する影響を受けにくいため，長期の依存関係を扱うことができる．

2.3 Bi-directional LSTM (BLSTM)

単一の LSTM のみでは，順方向のみの系列しか考慮できない．そこに，逆順の系列を入力とすることで，逆方向の系列を考慮することができる LSTM を合わせて用いることにより，正負両方向（Bi-directional）での系列を考慮できるようにしたものが BLSTM[5] である．BLSTM は単純に二つの LSTM を組み合わせたものであるため，順方向の処理，逆方向の処理二つの LSTM を用いる．つまり，入力が N 個の系列 x の場合，順方向 LSTM に対しての入力は $\{x_1, x_2, x_3, \cdots, x_N\}$ の順になる．一方，逆方向 LSTM には $\{x_N, x_{N-1}, x_{N-2}, \cdots, x_1\}$ の順で入力をする．その後，順方向，逆方向それぞれの出力 $\boldsymbol{y}_F$ と $\boldsymbol{y}_B$ とを統合する．これらを式で表すと，

$$\boldsymbol{y}(t) = \boldsymbol{y}_F(t) \oplus \boldsymbol{y}_B(N - t + 1) \tag{5}$$

となる．順方向と逆方向それぞれの出力系列の位置を合わせるために，二つの LSTM の出力を $\oplus$ によって統合する．

3 システム構成

3.1 LSTM を用いた音楽音響信号修復モデル

LSTM を用いた場合の音楽音響信号修復モデルを述べる．構成としては，図 2 a) に示すように，入力層，線型結合（全結合）層，LSTM，線型結合層，出力層からなる．入出力としては音楽音響信号に STFT（Short-TimeFourier transform）をかけて得られる各フレームの振幅スペクトルを用いる．音楽音響信号修復は，過去数フレームの情報を入力として，次の 1 フレームを予測するという回帰問題として扱う．このモデル関数を f_{LSTM} と表記すると，時刻 t での出力 y_t は，時刻 1 から t までの入力 x_1^{t-1} より，以下のように定義できる．

$$\boldsymbol{y}_t = f_{LSTM}(\boldsymbol{x}_1^{t-1}) \tag{6}$$

3.2 フィルタ内包型 LSTM

LSTM を用いた音楽音響信号修復モデルを図 2 b) に示す．構成としては，図 2 a) の音楽音響信号修復モデルに対し，入力の直後にフィルタ層を，出力の直前に逆フィルタ層を挿入した形である．この内，逆フィルタ層の係数は，必ず，フィルタ層での逆数を取るようフィルタ層と逆フィルタ層の係数を一体で更新するように設計した．周波数フィルタがネットワークに内包された構造になっているため，最初に初期値の設定は必要なものの，それ以降は，LSTM の学習と同時に学習できるため，データから最適なフィルタを学習することが可能である．フィルタを W_{filt} とし，逆フィルタを W^*_{filt} と表すと，

$$\boldsymbol{y}_t = \boldsymbol{W}^*_{filt} f_{LSTM}(\boldsymbol{W}_{filt} \boldsymbol{x}_1^{t-1}) \tag{7}$$
$$(\text{ただし}, \boldsymbol{W}_{filt} \boldsymbol{W}^*_{filt} = (1, 1, 1, 1, 1, \cdots)^T)$$

となる．

3.3 フィルタ内包型 BLSTM

フィルタ内包型 LSTM を Bi-directional LSTM に適用した場合について述べる．先に述べたフィルタ内包型 LSTM の LSTM を Bi-directional LSTM に拡張した音楽音響信号修復モデルを図 2 c) に示す．線形結合層，LSTM 層，線形結合層の三層をひとまとめとして，順方向を F-LSTM，逆方向を B-LSTM と呼ぶことにする．F-LSTM に対しては，入力 $\boldsymbol{x}$ は $\{x_1, x_2, x_3, \cdots, x_N\}$ の順になる．一方，B-LSTM には $\{x_N, x_{N-1}, x_{N-2}, \cdots, x_1\}$ の順で入力をす

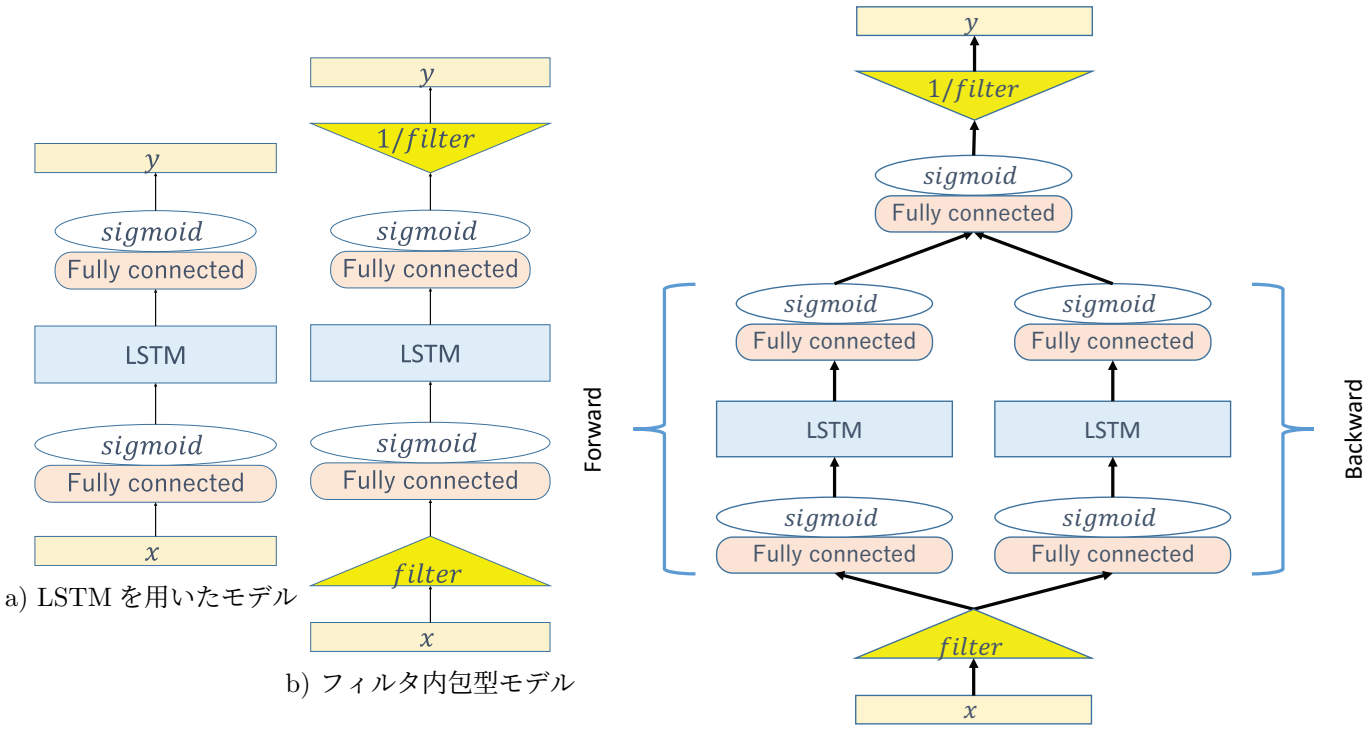


図 2: 音楽音響信号修復用のニューラルネットワーク

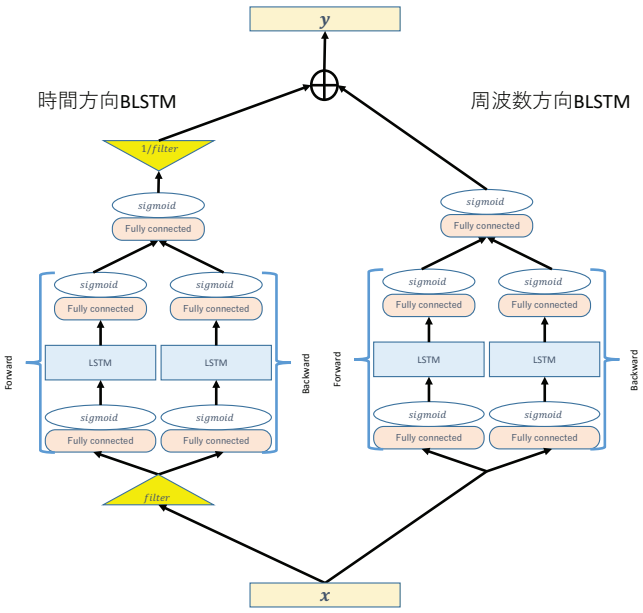


図 3: 提案モデル

る．その後，全結合層によりそれぞれからの出力をどのように合わせるかを判断する．入力 $\boldsymbol{x}$ が N 個の系列であるとし，全結合層を W_l とすると，

$$\boldsymbol{y}_t = \boldsymbol{W}^*_{filt}\boldsymbol{W}_l \begin{pmatrix} f_{F-LSTM}(\boldsymbol{W}_{filt}\boldsymbol{x}_1^{t-1}) \\ f_{B-LSTM}(\boldsymbol{W}_{filt}\boldsymbol{x}_N^{t+1}) \end{pmatrix} \quad (8)$$

となる．

3.4 Quad-directional LSTM

時間方向のみの BLSTM では，倍音など周波数方向での系列情報が保持されず，高音域での修復性能が依然として十分ではない．その問題を解決するため，本稿で拡張した点である Quad-directional LSTM について述べる．音楽音響信号は，時間方向に関して関係性をもつ系列信号であるが，倍音などから周波数方向にも関係性を持つ．そのため，STFT をかけたスペクトログラムにおいては，周波数方向にも系列データであるとみなすことができる．この周波数方向について LSTM を合わせて用いることで，時間方向，周波数方向の二次元での系列を考慮できるようにしたものが QDLSTM である [6]．本稿で用いたモデルは図 3 となる．時間方向に関しては，先のフィルタ内包型 BLSTM で用いたものを利用するが，周波数方向ではフィルタを用いない BLSTM を構成した．これら二つを独立に学習させ，その出力を足し合わせるという構成である．時間方向の系列 N 個，周波数方向の系列 F 個の信号に対して，時間方向 BLSTM の出力 $\boldsymbol{y}^{time}$，周波数方向 BLSTM の出力 $\boldsymbol{y}^{freq}$ とし，t, f をそれぞれ時間，周波数のインデックスとすると，式は以下のようになる．

$$\boldsymbol{y}_t^{time} = \boldsymbol{W}^*_{filt}\boldsymbol{W}_l^{time} \begin{pmatrix} f^{time}_{F-LSTM}(\boldsymbol{W}_{filt}\boldsymbol{x}_1^{t-1}) \\ f^{time}_{B-LSTM}(\boldsymbol{W}_{filt}\boldsymbol{x}_N^{t+1}) \end{pmatrix} \quad (9)$$

$$\boldsymbol{y}_f^{freq} = \boldsymbol{W}_l^{freq} \begin{pmatrix} f^{freq}_{F-LSTM}(\boldsymbol{x}_1^{f-1}) \\ f^{freq}_{B-LSTM}(\boldsymbol{x}_F^{f+1}) \end{pmatrix} \quad (10)$$

Algorithm 1 時間方向 BLSTM の学習

for $t = 1, 2, \ldots, l - 35$ **do**
  $x = W_{filt}\, data\{t, t+1, \ldots, t+35\}$
  $true = data\{t+3, t+4, \ldots, t+32\}$
  for $n = 1, 2, \ldots, 30$ **do**
    $y^*_{for}[n] = f^{time}_{F-LSTM}\{x^{n+2}_n\}$
    $y^*_{back}[31-n] = f^{time}_{B-LSTM}\{x^{35-n}_{37-n}\}$
  end for
  for $m = 1, 2, \ldots, 30$ **do**
    $y[m] = W^*_{filt} W^{time}_l \begin{pmatrix} y^*_{for}[m] \\ y^*_{back}[m] \end{pmatrix}$
    compute $loss$ of $y[m]$ and $true[m]$
    $loss_{sum} += loss$
  end for
  propagate $loss_{sum}$ backwards
end for

Algorithm 2 周波数方向 BLSTM の学習

for $t = 1, 2, \ldots, l - 11$ **do**
  $x = data\{t, t+1, t+2\} \& data\{t+9, t+10, t+11\}$
  $true = data\{t+3, t+4, t+5, t+6, t+7, t+8\}$
  for $f = 1, 2, \ldots, 257$ **do**
    $y^*_{for}[t, f] = f^{freq}_{F-LSTM}\{x_{t,f}\}$
    $y^*_{back}[t, 258-f] = f^{freq}_{B-LSTM}\{x_{t,258-f}\}$
  end for
  for $m = 1, 2, \ldots, 257$ **do**
    $y[t, m] = W^{freq}_l \begin{pmatrix} y^*_{for}[t, m] \\ y^*_{back}[t, m] \end{pmatrix}$
    compute $loss$ of $y[t]$ and $true[t]$
    $loss_{sum} += loss$
  end for
  propagate $loss_{sum}$ backwards
end for

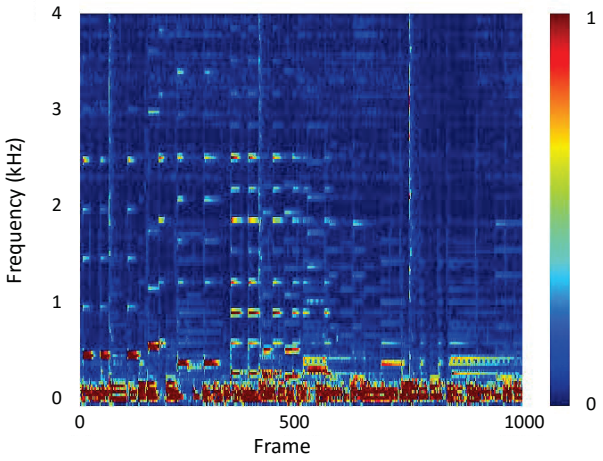
$$\boldsymbol{y}_{t,f} = \boldsymbol{y}^{freq}_{t,f} \oplus \boldsymbol{y}^{time}_{t,f} \tag{11}$$

4 評価実験

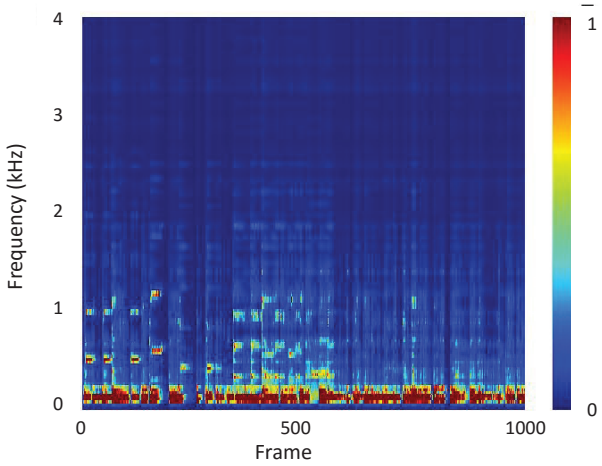
楽曲データ (サンプリングレート 16 kHz) に対し，フレーム長 512 サンプル，シフト長 128 サンプル，ハミング窓を窓関数として用いた STFT を行うことで振幅スペクトルを得る．各楽曲での最大振幅値を用いて正規化を行う．ニューラルネットワーク学習の際には，連続する 3 フレーム分 ($t \sim t+2$ フレーム) の振幅スペクトルを入力し，次のフレーム ($t+3$ フレーム) の振幅スペクトルを予測するよう学習を行う．評価の際には，欠損のない振幅スペクトル列を学習したニューラルネットワークに入力し，出力を時間方向に並べた振幅スペクトル列と，入力に用いた元の振幅スペクトル列との比較を行う．つまり，欠損のないデータが 3 フレーム連続して続き，その後の 4 フレーム目が欠損している信号に対する修復タスクを評価していることに相当する．学習には，ジャズ楽曲 6 曲を用い，図 2 a), b), c)，図 3 それぞれについて学習した 3 種類のニューラルネットワークを構築した．最小化する損失関数には MSE（Mean Square Error）を用いた．また，図 2 b), c)，図 3 におけるフィルタ層の初期値には，人間の聴覚を元にした周波数重み付けである A 特性 [7] を用いた．学習率は Adam (Adaptive Moment Estimation) [8] を用いて最適化を行った．また，評価では学習とは別のジャズ楽曲 1 曲の一部（1006 フレーム，約 8 秒分）に対して予測を行い，予測結果の内 1000 フレームを比較に用いた．順方向のみの深層学習モデルの場合，学習をするためには単純に t を 1 ずつ増加させ，そのたびごとの出力と正解との誤差を算出し，逆伝播させることが可能である．しかし，BLSTM の場合，順方向からの出力と逆方向からの出力との時間を合わせるために，ある程度の長さの範囲を指定する必要がある．そのため，本稿では，BLSTM の学習は，Algorithm1，Algorithm2 にしたがうものとした．ここで，l は学習に用いる音楽データのフレーム長である．また，QDLSTM における時間方向，周波数方向それぞれの BLSTM からの出力の統合には，二つの出力の要素の平均をとるという方法で行った．

5 比較結果

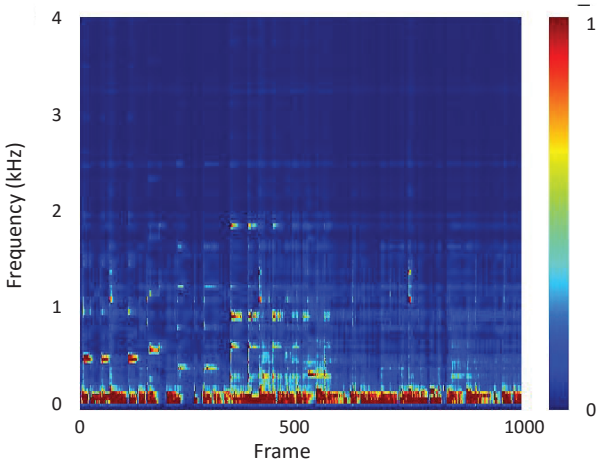
各モデルからの出力結果として得られた振幅スペクトログラム（0∼4 kHz）を図 4 b), c), d),e),f) に示す．また，図 4 a) に正解データの振幅スペクトログラムを示す．フィルタを内包した場合，通常の LSTM モデルではできない 2 kHz 付近の修復が可能となっている．また，通常の LSTM モデルでは 1 kHz 部分に大きく出ている誤差がフィルタ内包型では改善されている．しかし，音の輪郭（オンセット・オフセット）の関係が不明瞭なままとなっている．一方時間方向 Bi-directional LSTM の拡張では，2.5 kHz 付近の修復が可能となっている部分があり，また，音の輪郭部分が明瞭となっている．しかし，3 kHz 以上の高音域に関しては修復性能が十分でなく，倍音成分をうまく出力できていない．周波数方向 Bi-directional LSTM モデルでは，時間方向の BLSTM では修復ができていない 3 kHz 付近での修復が可能となり，より高音域までの周波数方向の関係を処理することが可能となっている．しかし，時間方向での関係性は保たれておらず，音信号のつながりが不明瞭となっている．これは，スペクトログラムに逆 STFT をかけ，音楽データとして再構成した際に顕著に感じらる．ピアノなどの倍音は修復されているため，音色としては豊かなものとなっている．しかし，各発音や残響音に違和感がのこり，結果として聞き心地の悪いものとなってい



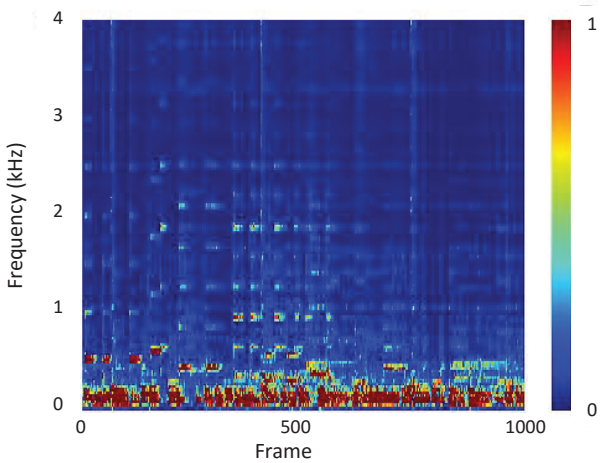
a) 正解データ



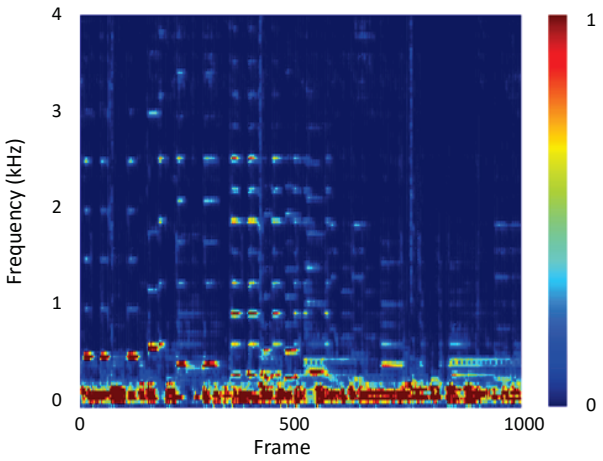
b) LSTM を用いたモデル



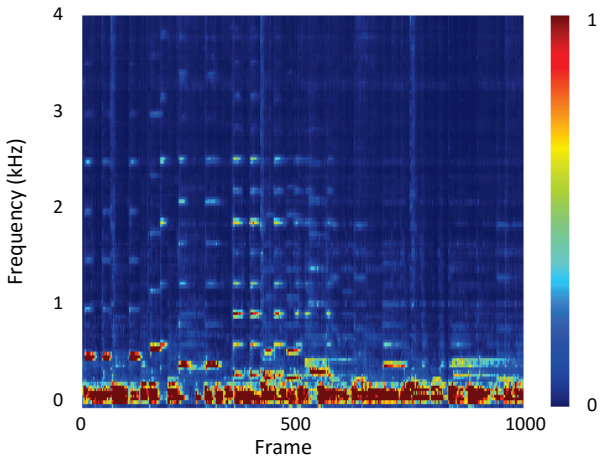
c) フィルタ内包型モデル



d) フィルタ内包型時間方向 BLSTM モデル



e) 周波数方向 BLSTM モデル



f) Quad-directional LSTM

図 4: 各モデルによる予測結果と正解データ

る．これらに対して，本稿での拡張である QDLSTM として再構成したものが図 4 f) である．時間方向 BLSTM の音の輪郭の明瞭さ，残響音関係の修復に加え，周波数方向 BLSTM の 2.5 kHz 以上での修復性能が向上している．音

表 1: 各モデル出力の SDR

モデル	SDR
LSTM	5.00
フィルタ内包型 LSTM	5.26
フィルタ内包型時間方向 BLSTM	8.77
周波数方向 BLSTM	6.51
QDLSTM	8.67

楽データとして再構成した場合，時間方向 BLSTM の際に感じられたピアノの倍音部分の表現の弱さが改善されており，より生楽器のピアノ音色として自然なものとなっている．それに加え，周波数方向 BLSTM の結果で感じられた発音間の不自然さが軽減され，自然な演奏として感じられる．しかし，依然として 3 kHz 付近から 4 kHz 付近にあるシンバルの倍音成分や，クラッシュシンバル，ライドシンバルのレガートなどの細かな残響による時間関係の修復はできておらず，ノイズのように感じられるものとなっている．

表 1 は，式 (12) によって算出した SDR（signal-to-distortion ratio）である．SDR は，修復音全体における正解との差と，正解との比をとったものであり，修復結果の歪の指標となる．s は正解，y は出力であり，数値が大きいほど歪が少ないことを示す．

$$SDR = 10\log_{10}\frac{\sum s^2}{\sum (s-y)^2} \tag{12}$$

周波数方向 BLSTM は，スペクトログラム上では最も高周波まで修復が可能であるが，聴覚上でも感じられた通り歪の大きい結果となっている．一方時間方向 BLSTM では歪が最も少ない結果となっている．QDLSTM は，SDR の値では時間方向 BLSTM に対してやや劣ったものとなっているが，周波数方向 BLSTM の高音域修復性能を持ったまま，歪を抑えられるということがわかる．

6 考察

周波数方向 BLSTM において，修復フレーム 6 フレームに対して前 3 フレーム，後ろ 3 フレームからの修復のため，時間関係の修復が十分にされていない．そのため，周波数方向での強い関係をもつ倍音成分に対しては系列情報として認識されるため，ピアノなどの倍音が強く出ているものには対応ができているのだと考えられる．しかし，残響に重なって新しく発音された場合などでは，残響音の信号は強度が弱いためうまく認識がされず，新しく発音された部分の倍音のみが修復されている．結果として，時間方向での系列がうまく処理されず，音楽として不自然な修復結果となっていると考えられる．また，QDLSTM として時間方向，周波数方向の二つの BLSTM を統合した場合，今回の統合方法では単純に各要素の平均を取るという方法をとっている．そのため，周波数方向 BLSTM には出ていない残響部分は時間方向 BLSTM の出力によって補完され，その逆に時間方向 BLSTM では出ていない高音域の倍音が周波数方向 BLSTM によって補完されている．これにより，フロント楽器であるため信号が強く出ているピアノや，低音域であるベースなどは音色が豊かになり，聞き心地のよい修復ができていると考えられる．しかし，シンバルレガートの部分など，残響音によって表現をされている奏法などでは，その弱い信号をどちらの BLSTM においても修復がされていないため，ノイズのような信号のままとして残されている．これらにより，今回用いた統合方法での QDLSTM では，聴覚的に自然なピアノ，ベースなど音階楽器と，ノイズのように感じられるシンバルなど打楽器とで乖離したような修復結果となっていると考えられる．また，単純な平均を取っている統合方法では，片方の BLSTM では出ているがもう一方では出ていない信号部分などが薄まってしまい，結果として正解から離れてしまうという部分も確認できる．これらの問題に対して，音階楽器と打楽器とでの特徴を深層学習に対して与える方法を検討する必要がある．また，二つの BLSTM の統合を，周波数ビンごと，時間ごとで比率を変えることができれえば，より詳細な修復が可能になると考えられる．

7 終わりに

本稿では，深層学習の一手法である LSTM を用いた音楽音響信号修復法として，周波数フィルタを内包するモデルを提案し，その拡張として時間方向，周波数方向の二つの Bi-directional LSTM を用いた Quad-directional LSTM に適用した．提案手法は，過去の手法に対してより高い修復性能を示すことを，修復タスクを想定した評価実験によって示した．

謝辞 本研究は、JSPS 科研費 16H02884, 16K00294, 17K00365 および，JST ImPACT タフロボティクスチャレンジの助成をうけた．

参考文献

[1] 谷口他, “LSTM による音楽音響信号の修復法の提案–周波数フィルタ導入による学習データ量削減の検討”, 第 79 回情報処理学会全国大会, 2017

[2] 谷口他, “Bi-directional LSTM を用いた音楽音響信号修復法の提案”, 第 35 回 日本ロボット学会学術講演会 , 2017

[3] Jerey L Elman. “Finding structure in time.” Cognitive science, 14(2):179211, 1990.

[4] F.A.Gers *et al.*,“Learning to forget: Continual prediction with LSTM.” ICANN ’99, 1999 p. 850 855

[5] A. Graves, J. Schmidhuber. “Framewise Phoneme Classification with Bidirectional LSTM Networks.” IJCNN 2005, Montreal, Canada, pp. 2047-2052.

[6] A. Graves, S. Fernndez, “J. Schmidhuber. Multi-Dimensional Recurrent Neural Networks.” ICANN 2007, Porto, Portugal, pp. 549-558.

[7] 西山他, “音響振動工学”, コロナ社, 1979.

[8] Kingma, D. P. *et al.*, “ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION”, International Conference on Learning Representations, 2015, pp.1-13.

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

ＡＩチャレンジ研究会
主 査 / 担 当 幹 事
公文 誠
熊本大学 大学院 先端科学研究部

Executive Committee Chair
Makoto Kumon
Faculty of Advanced Science and Technology,
Kumamoto University
kumon @ gpo.kumamoto-u.ac.jp

主 幹 事
光永 法明
大阪教育大学 教員養成課程 技術教育講座

Secretary
Noriaki Mitsunaga
Department of Technology Education,
Osaka Kyoiku University

担 当 幹 事
鈴木 麗璽
名古屋大学 大学院情報学研究科 複雑系科学専攻

Reiji Suzuki
Department of Complex Systems Science,
Graduate School of Informatics,
Nagoya University

中臺 一博
(株) ホンダ・リサーチ・インスティチュート
・ジャパン / 東京工業大学 工学院
システム制御系

Kazuhiro Nakadai
Honda Research Institute Japan Co., Ltd./
Department of Systems and Control
Engineering, School of Engineering,
Tokyo Institute of Technology

幹 事
植村 渉
龍谷大学 理工学部 電子情報学科

Wataru Uemura
Department of Electronics and Informatics,
Faculty of Science and Technology,
Ryukoku University

SIG-AI-Challenges home page (WWW): http://www.osaka-kyoiku.ac.jp/~challeng/