ロボット聴覚の極限音響への展開

Development of Robot Audition under Severe Conditions

○奥乃　博（早大）　中臺　一博（東工大/ホンダ RI）　公文　誠（熊大）
糸山　克寿　　吉井　和佳　　坂東　宜昭（京大）　　佐々木洋子（産総研）

Hiroshi G. OKUNO, Waseda University, okuno@aoni.waseda.jp
Kazuhiro NAKADI, Tokyo Institute of Technology/Honda Research Institute Japan Co., Ltd.
Makoto KUMON, Kumamoto University
Katsutoshi ITOYAMA, Kazuyoshi YOSHII, Yoshiaki BANDO, Kyoto University
Yoko Sasaki, National Institute of Advanced Industrial Science and Technology

The ability of robots to listen to several things at once with their own “ears”, i.e., *robot audition*, is critical in improving the performance of search and rescue activities under severe conditions. This paper introduces “HARK” robot audition open-source software and its capabilities of suppressing ego-noise that is caused by robot’s own movements such as motor, propeller and/or flying noise. Then it describes three main applications of robot audition: 1) Unmanned Aerial Vehicle (UAV) with a microphone array to capture sounds can localize a sound source by suppressing ego-noise with either hovering, slow gliding or fast gliding. It can also recognize a sound source by CNN. 2) A serpentine robot with a microphone array can estimate its posture by sound. It can also enhance a voice by Online Robust PCA. 3) A robot with a LiDAR and 32-channel microphone can visualize a sound map by superimposing sound source directions on point clouds.

Key Words: Robot Audition, Unmanned Aerial Vehicle, Serpentine Robots, Sound Source Localization, Sound Source Separation, Sound Source Identification

1. なぜロボット聴覚が必要なのか

日常環境で私たちは複数の音源からの音が，時には残響の影響を受けて混ざった混合音を聞いている。システムやロボットを日常環境で使用するためには，混合音を聞き分ける機能が不可欠である。ロボット聴覚とは，ロボット自身に装着したマイクロフォンで収音し，音源定位，音源分離，分離音認識等を行う機能，および関連する研究領域を指す言葉である[1]。20 世紀のロボットにはマイクロフォンは装備されていても，その聴覚機能は貧弱で，音声対話ロボットの大部分は話者の口元にあるマイクロフォンを使用するなど，ロボット聴覚機能が実現されているとはいえなかった。実際，当時のロボット知覚研究は，画像処理が中心で，音響情報は活用されていなかったため，ロボット視覚研究と同義語であった。

21 世紀になり，我々は，複数のマイクロフォンからなるマイクロフォンアレイ（以下，マイクアレイ）をロボットに装着して，ロボット聴覚の要素技術である，混合音からの音源定位，音源分離，分離音認識という聞き分け技術を開発し，世界のロボット聴覚研究を先導してきた[2]。この結果，IEEE/RSJ 共催の IROS で robot audition が 2014 年から keyword として登録され，最近では音響系分野でもロボット聴覚研究が始まっている[3]。特に，我々がオープンソースとして公開しているロボット聴覚ソフトウエア HARK[1][4-5]は，世界中から 2015 年末時点で約 7 万件のダウンロードがされている。

ロボット聴覚の応用として，図 1 に示すように多面的な展開を行ってきた[2]。聖徳太子のように複数の訴えを同時に聞き分ける聖徳太子ロボット，人とロボットとのインタラクションとしてクイズ司会者ロボット HATTACK25, 音楽共演ロボットなどを開発し，その可能性を実証してきた。最近では，カエルの合唱の解明のための HARKFrog や野鳥の鳴き交わしを観測するための HARKBird などにも展開している。

本論文では，日常環境よりも条件が厳しい極限環境，例えば，空中やがれきの下で，ロボット聴覚を応用する「極限音響」での課題について指摘し，これまでに開発してきた要素技術と，具体的なシステム実装による評価について報告する。

Fig. 1 Development of Robot Audition Research

2. ロボット聴覚ソフトウエア HARK の概要

ロボット聴覚用ソフトウエアは，実時間処理とモジュール化による自由なカスタマイズ機能が不可欠である。HARK は，事前知識を最少にしたポータビリティの高いシステム，いわば「聴覚の OpenCV」を目指して，京都大学とホンダ・リサーチ・インスティチュート・ジャパンで共同研究・開発を行った成果である。HARK の特徴は以下のとおりである：

- ロボットの形状，マイクの数・配置に非依存
- HarkDesigner による GUI プログラミング環境提供
- ミドルウエア BatchFlow[6] による低オーバーヘッドなモジュール統合
- 音源定位・音源分離・分離音認識等の実時間モジュール，ツール群の提供
- ネットワーク, python, ROS インタフェースによる他システムとの接続の容易性
- 300 ページ超の日英のマニュアル・クックブックの提供

マイクアレイは，市販品の Tamago，クラゲ君（システム・イン・フロンティア），Kinect（マイクロソフト），PS-EYE（ソニー），Microcone（DevAudio），に加えて，独自配置のアレイ

[1] http://www.hark.jp/ で公開。無料講習会毎年実施。

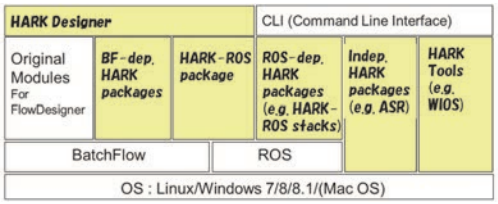

Fig. 2 Stack of HARK [3,4]

の使用も可能である。HARK は，図 2 に示すように，Windows・Ubuntu 上のミドルウエア上に構築されている。両耳聴（2 本のマイクを使用）用の HARK-Binaural+も提供している[7, 8]。

2.1 雑音への対応

HARK は，ロボットを対象としているため，雑音の扱いに多くの労力が費やされている。音源定位にはもともと MUSIC（Multiple Signal Classification）法ベースの雑音に頑健なアルゴリズムが実装されているが，中村らは，これをさらに発展させ，雑音に関する知識を雑音相関行列として導入して，雑音を白色化する GEVD-MUSIC（Generalized EigenValue Decomposition）法を開発し，音源定位性能が著しく向上できることを報告している[9]。また，一般固有値展開の代わりに一般特異値展開を用いることにより，計算コストを削減する GSVD-MUSIC（Generalized Singular Value Decomposition）もあわせて開発している[10]（図 3 参照）。

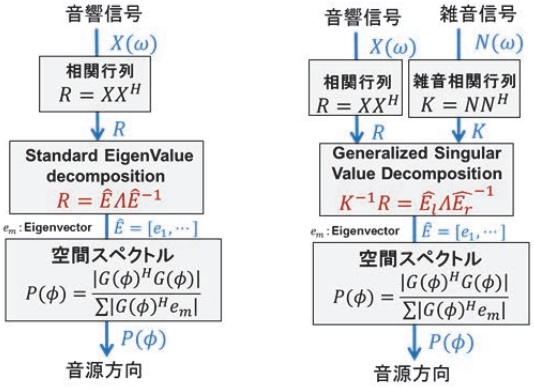

Fig.3 Algorithms for GEVD-MUSIC and GSVD-MUSIC

$\boldsymbol{X}$ and $\boldsymbol{N}$ show spectra of an input and noise signal. $\boldsymbol{R}$ and $\boldsymbol{K}$ show correlation matrices for the input and noise signal. $\boldsymbol{G}$ is a steering vector for sound directions ϕ. P is called MUSIC spectrum, and a peak with higher power than a threshold in P is regarded as a sound source.

音源分離には，適応ビームフォーマやブラインド分離など 11 種類のアルゴリズムが提供されている。音源分離に関しては，状況に応じてアルゴリズムを選択することが好ましいが，比較的 GHDSS-AS（Geometrically-constrained Higher-order Decorrelation-based Source Separation with Adaptive Step-size control）がロバストに動作するため，このアルゴリズムの使用を推奨している。

ロボットの自己雑音に対しては，Ince らによるテンプレートを用いてロボットの姿勢と自己雑音の関係をデータベース化し，これを用いて自己雑音を推定する手法が実装されている。また，この手法を HARK が提供する音源定位，分離，分離音認識それぞれに適用し，いずれも有効であることを報告している［11, 12］。

3. 極限環境下でのロボット聴覚の応用

本章では，極限環境として，空中，瓦礫下，大規模展示会場の 3 つを取り上げ，UAV による空中からの収音・音源定位・音源同定，索状ロボットの音による姿勢推定・音声強調，大規模展示会場での視聴覚情報統合による音マップ構築について報告する。

3.1 UAV による空中からの収音・音源定位・音源同定

UAV からの収音では，①マイクアレイを装着，②音源定位，③音源同定，が課題となる。

マイクアレイの設置場所は，UAV の種類や使い方に依存する。現在，我々のグループでは，次に示す 3 種類の状況を検討している：

1. 通常の multirotor によるホバリング・飛行
2. 凧式 UAV による低速滑空
3. 無人グライダによる高速滑空

通常の UAV に対しては，ロータからできるだけマイクアレイを離すことによって，自己雑音の影響を軽減することを狙う必要がある。公文らは，Quadrotor（Zion PG560，enRoute 社）に 16-ch マイクアレイを装着するときの最適位置をシミュレーションで求めており，さらに実際にマイクアレイを搭載し，その性能評価を行っている[13]（図 4 参照）。

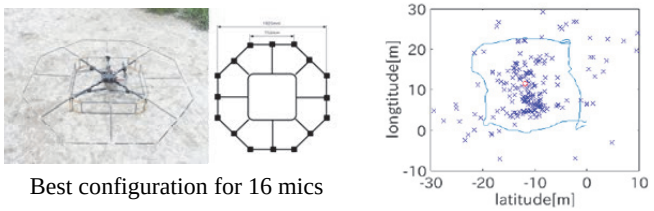

Fig.4 A quadrotor with 16 microphones and a result of sound source localization by flying around the sound source. Red and blue circles indicate the ground truth and estimated sound source localization, respectively [13].

凧式 UAV は，駆動用のモータで推力を生むが，ある程度の高度に達すると，駆動用モータを止め，プロペラによる自己雑音が発生するのを押さえて，滑空しながら音源探索を行う。公文らは，周期的なプロペラの回転・停止を行って探索する場合と，音源探索経路を与えた下での周期的なプロペラの回転・停止を行って音源探索する場合の性能を，シミュレーションおよび実機で実験しており，後者の方が音源探索性能がよいことを示している。また，探索経路を与えなくても，プロペラの停止により，音源探索性能が向上することを確認している［14］（図 5 参照。）

無人グライダによる高速滑空時には，外部からの推力はないので，プロペラによる自己雑音は発生せず，風切り音が最大の自己雑音の発生源となる。現在，マイクアレイの配置位置と収音の実験を行っている。

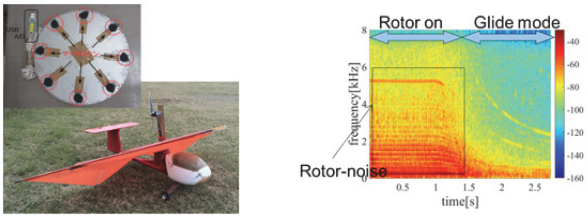

Fig.5 A kiteplane with a 8-channel microphone array. The spectroguram shows that ego-noise drastically disappeared when the rotor was off and the kiteplane flied in a glide mode [14].

音源定位手法については，ロータの出力や風雑音の動的変化に対応できるようにするため，GEVD-MUSIC に対して，一定時刻前の信号は雑音であることを仮定し，逐次的に雑音相関行列を推定できるように改良を行った incremental GEVD-MUSIC（iGEVD-MUSIC）法を提案している[15]。また，ロータの状態から雑音相関行列を回帰推定する手法も提案している[15]。さらに，iGEVD-MUSIC で問題であった計算量の削減と雑音相関行列の推定精度を向上させるため，GSVD-MUSIC をベースに拡張を行った incremental GSVD-MUSIC with Correlation Matrix Scaling（iGSVD-MUSIC-CMS）法を提案し，20m 程度離れた音源でも検出可能であることを報告している[16]。上述の 1-3 についても，これらの音源定位手法を用いて評価を行っている。

音源同定は，音源定位で検出される様々な音源の中から，人に由来する音源など目的の音源を効率的に発見するためには必須の機能である。音源定位はあくまでも音源の方向を検出するための手法であるため，検出された音自体にはロータや風雑音が含まれたままである。このため，上村らは，GHDSS-AS を適用し雑音抑圧を行うと共に，音響信号がサウンドスペクトログラム上では画像とみなせることに注目して，近年，画像識別で高い性能が得られることで知られている Convolutional Neural Network（CNN）を適用する手法を提案している[17]。この手法は，10 クラス問題で 90%程度の高い認識性能が得られるものの，予めアノテーションを行った学習データを大量に用意する必要があった。これに対して，森戸らは，音源分離という回帰問題と音源同定という識別問題の両方を一つのニューラルネットワークで扱うことにより，部分的にしかアノテーションされていないデータでも効率的に学習が可能な手法を提案している[18]。この手法を用いると，一般的な Deep Neural Network（DNN）を用いた end-to-end 学習よりも高い識別性能が得られることが示されている[18]。

3.2 索状ロボットの音による姿勢推定・音声強調

東北大学田所・昆陽研究室で進めている Active Scope Camera ロボットの原型である索状ロボットは，瓦礫の下や複雑に入り組んだ場所での探索活動に期待がかかっている。索状ロボットの主たる 2 つの課題は，①ロボットのナビゲーションの効率化，②オペレータの探索の効率化，である。我々は，索状ロボットの胴体にマイクを分散的に配置するマイクアレイを装着することで 2 つの課題に取り組んでいる。索状ロボットの駆動用の振動モータのそばにスピーカを装着し 2 つの振動モータ間に MEMS マイクと IMU を装着し，スピーカとマイクとを交互に配している[19, 20]（図 6 参照）。

索状ロボットの姿勢推定では，スピーカから鳴らした TSP（Time Streched Pulse）信号を再生し，それを各マイクで受音し，各マイクでのスピーカからの音の到達時間差を求める。TSP 信号は逆 TSP 信号を重畳するとパルス信号が得られるので，到達時間差を精度高く求めることができる。マルチチャネル装置は，すべてのマイク入力と IMU のデータを，スピーカからの再生と同期して，取得することができる。どれか 1 つのスピーカから TSP 信号を再生し，マイク間の到達時間差を求める。この時に，スピーカからの反射音がある場合複数の到達時間差が得られるので，最短のものを選択する。さらに，障害物で直接音が取れず，反射音しか取れない場合には，実際のマイク間距離よりも到達時間差が大きくなるので，そのような外れ値は除去する。次々に異なるスピーカから TSP 信号を再生して，有意な到達時間差と IMU 情報を，単一の状態空間で表現し，switching Kalman filter を用いて，姿勢推定を行う。全長 3m の索状ロボットの先頭部分の位置推定の精度は，2D だと 20cm 程度である[21]。3D の場合には，鏡対称の曖昧性が残るので，IMU から得られる重力方向の情報を利用して，曖昧性を解消して，位置推定を行っている。先頭位置の誤差は，20cm 以下である[19, 20]。

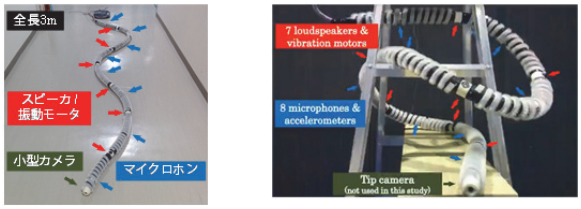

Fig.6 A Serpentine Robot with microphones, IMU and loudspeakers. A pair of a microphone and an IMU and a loudspeaker are alternately allocated along the body of the robot [19,20,21].

オペレータの探索を効率化するために，被災者の声だけを強調する音声強調にも取り組んでいる。理想的な場合には，上記の姿勢推定からマイク位置が分かるので，マイクアレイ処理，例えば，ビームフォーマで音源定位と音源分離が可能となる。しかし，瓦礫下では，信頼できるマイクの推定が不可欠である。坂東らは，個々のマイクから得られる入力音に対して，FFT をかけてスペクトログラムを求め，Online Robust PCA（ORPCA）を適用して調波構造を強調し，索状ロボットの走行音を抑制する。次にすべての ORPCA の結果を median で統合し，得られたスペクトルを逆 FFT で音声を再合成する[22]。本手法で，従来の自己雑音抑制法よりも SDR が 7.4 dB，SIR が 17.2 dB 向上する結果を得ている[22]。

高田らは，事前に収録した走行音の自己雑音を学習データとして，マルチチャネル NMF による統計的な音声強調手法を開発しており，坂東らと同程度の性能を得ている[23]。ただし， ORPCA はオンライン処理であるのに対して，本手法はオフライン処理である。

3.3 大規模展示会場での音マップ構築

2013 年に開催された ET-2013（図 7-a)）の会場において，産

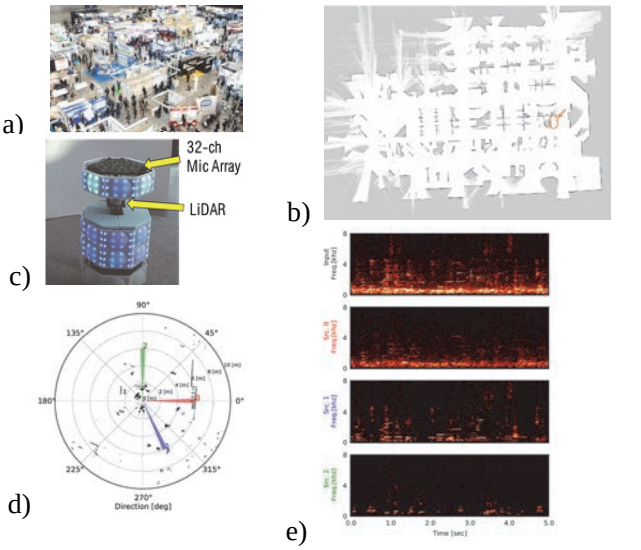

Fig.7 a) ET-2013 Exhibition Hall, b) Map generated by SLAM with LiDAR, and c) Peacock robot placed at the red circle in c). d) Point clouds (black points) obtained from LiDAR and the directions of separated sounds at ET-2013. E) Spectrograms of captured sound, separated sounds 0, 1, and 2 from up to down, which correspond to background noise, broadcasted female voice, and talking male voice near the robot, respectively [24].

総研の佐々木らは LiDAR と 32-ch マイクアレイを装着したロボット Peacock（図 7-c)）を使用し，LiDAR 計測とマルチチャネル録音を行った。LiDAR から得られたポイントクラウドマップを図 7-b) に示す[24]。図中の右側中央の赤い円に Peacock がいたときに収音された音響データを，Bayesian Non-Parametrics for Microphone Array Processing（BNP-MAP）[25]で音源定位・音源分離を行った。音源定位をポイントクラウド上に表示した動画のスナップショットを図 7-d)に示す。図 6-d) の時点では，３つの音源が存在しており，それぞれのビームは，音源方向と音量を示す。このうち，分離音 0, 1, 2 は各々0 度，300 度，90 度から来ており，そのスペクトログラムを図 7-c) に示す。分離音を聞くと，各々，背景雑音，館内スピーカから放送される女性の声，ロボットのそばを話している男性の声であることが分かる。

HARK を用いて音源定位・分離した場合にも，雑音は白色化され，男女の声が分離できている。HARK と BNP-MAP の性能はどちらが良いということは一概に言えないものの，オンライン処理では HARK を使った可視化が，オフライン処理では，HARK と BNP-MAP の両方から得られるデータを使用した可視化が，音マップの構築や音環境理解に有効であろうと期待している。

4. ロボット聴覚の普及に向けて

本稿では，ロボット聴覚の現在の技術レベルについて報告し，極限環境でロボット聴覚が展開できるための基礎技術の概要を報告するとともに，UAV による空中からの音源定位・音源同定実験，索状ロボットによる音による 3D 姿勢推定，マイクアレイによる音声強調，さらに，大規模展示会場での LiDAR とマイクアレイによる音環境可視化について報告した。

これまで，ロボット知覚は，ロボット視覚が中心であった。近年 IEEE/RSJ IROS 等の学会では，Alternative Sensing for Robot Perception や Multimodal Sensor-based Robot Control といったワークショップが提案され，視覚センサ以外のセンサにも徐々に注目が集まってきている．そうした流れの中で，ロボット聴覚も少しずつ注目されるようになってきているものの，いまだ視覚センサや LiDAR 等の Depth を提供するセンサがメジャーである状況に変わりはない。上記で紹介したようにロボット聴覚も使える技術に向けて進化を続けている。本稿がロボット研究者にロボット聴覚の普及の契機となることを期待する。

謝辞 HARK の更改・保守を担当されるホンダ RI の水本武志氏，中村圭佑氏を中心とする HARK 開発メンバ，極限音響応用を担当される ImPACT 極限音響チームの皆さんに感謝します。ロボット聴覚の要素技術開発は科研費基盤研究(S) No.24220006，ロボット聴覚フィールド実験は ImPACT-TRC の支援を受けた。

参考文献

[1] Nakadai, K., et al., “Active Audition for Humanoid,” *AAAI-2000*, pp. 832-839.
[2] 奥乃 博, “聞き分ける技術の水平展開”, *人工知能学会誌*, Vol.30, No.3, pp366-376, 2015.
[3] Okuno, H.G. and Nakadai, K., “Robot Audition: Its Rise and Perspectives,” *IEEE ICASSP-2015*, pp.5610-5614.
[4] Nakadai, K., et al., “Design and Implementation of Robot Audition System “HARK”,” *Advanced Robotics*, vol.24, pp.739-761, 2010.
[5] 中臺 一博 他, “HARK 2.2 の新機能とその組み込み，SaaS への展開”, *SI2015*, pp.1835-1838.
[6] Côté, C., et al., “Reusability tools for programming mobile robots,” *IEEE/RSJ IROS-2004*, pp.1820-1825, 2004.
[7] Kim, U-H, et al., “Improved Sound Source Localization in Horizontal Plane for Binaural Robot Audition,” *Applied Intelligence*, Vol.42, Issue 1, pp.63-74, 2015.
[8] 坂東 宜昭 他, “両耳聴ロボット聴覚ソフトウエア HARK-Binaural の紹介と Raspberry Pi 2 を用いたヒューマノイドロボットへの適用”, *音学シンポジウム*, 情報処理学会, 2015.
[9] Nakamura, K., et al., “Intelligent sound source localization for dynamic environments,” *IEEE/RSJ IROS-2009*, pp.664-669.
[10] Nakamura, K., et al., “A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition,” *Advanced Robotic*s, Vol.27, Issue 12, pp.933-945, 2013.
[11] Ince, G. et al., “Assessment of General Applicability of Ego Noise Estimation – Applications to Automatic Speech Recognition and Sound Source Localization –,” *IEEE-RAS ICRA-2011*, pp.3517-3522.
[12] Furukawa, K. et al., “Noise Correlation Matrix Estimation for Improving Sound Source Localization by Multirotor UAV,” *IEEE/RSJ IROS-2013, pp. 3943-3948.*
[13] Ishiki, T. and Kumon, M., “Design Method of Microphone Arrays for Multirotor Helicopters,” *IEEE/RSJ IROS-2015*, pp.6143-6148.
[14] 公文 誠 他, 凧型無人航空機を用いた音源探査, 第 43 回 AI チャレンジ研究会, pp.48-53, 2015.
[15] 奥谷 啓太 他, “クワドロコプター搭載のマイクロホンアレイを用いた 屋外音環境理解の逐次雑音推定による向上,” *日本ロボット学会誌*, Vol.31, No.7, pp.676-683, 2013.
[16] Ohata, T., et al., “Improvement in Outdoor Sound Source Detection Using a Quadrotor-Embedded Microphone Array,” *IEEE/RSJ IROS-2014*, pp.1902-1907.
[17] Uemura, S. et al., “Outdoor Acoustic Event Identification using Sound Source Separation and Deep Learning with a Quadrotor-Embedded Microphone Array,” *ICAM 2015*, pp.329-330, JSME.
[18] 森戸 隆之他, “部分共有型 Deep Neural Network を用いた音源同定”, *ロボッティクス・メカトロニクス講演会 Robomech 2016*.
[19] Bando, Y., et al., “Microphone-accelerometer based 3D posture estimation for a hose-shaped rescue robot,” *IEEE/RSJ IROS-2015*, pp. 5580-5586.
[20] 坂東 宜昭 他, “柔軟索状レスキューロボットのためのマイクロホン・加速度センサアレイを用いた 3 次元姿勢推定”, *ロボッティクス・メカトロニクス講演会 Robomch 2016*, 日本機械学会.
[21] Bando, Y., et al., “Posture estimation of hose-shaped robot using microphone array localization,” *Advanced Robotics, Vol.29, Issue 1*, pp. 35-49, 2015.
[22] Bando, Y., et al., “Human-Voice Enhancement based on Online RPCA for a Hose-shaped Rescue Robot with a Microphone Array,” *IEEE SSRR-2015*, 6p., 2015.
[23] 高田 一真 他, “教師あり多チャネル NMF と統計的音声強調を用いた柔軟索状ロボットにおける音源分離”, 音響学会, Mar, 2016..
[24] Bando, Y., et al., “Challenges in deploying a microphone array to localize and separate sound sources in real auditory scenes,” *IEEE ICASSP-2015*, pp. 723-727.
[25] Otsuka, T., et al., “Bayesian Nonparametrics for Microphone Array Processing,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, Vol.22, Issue 2 pp.493-504, 2014.