人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-046-03 (11/09)

部分共有アーキテクチャを用いた深層学習ベースの音源同定の検討

Sound Source Identification based on Deep Learning with Partially-Shared Architecture

森戸隆之 *1，杉山治 *2，小島諒介 *1，中臺一博 *1,3
Takayuki MORITO*1, Osamu SUGIYAMA*2, Ryosuke KOJIMA*1, Kazuhiro NAKADAI*1,3
東京工業大学 *1，京都大学 *2，(株) ホンダ・リサーチ・インスティチュート・ジャパン *3
Tokyo Institute of Technology*1, Kyoto University*2, Honda Research Institute Japan Co., Ltd.*3
morito@cyb.mei.titech.ac.jp, sugiyama@kuhp.kyoto-u.ac.jp,
kojima@cyb.mei.titech.ac.jp, nakadai@jp.honda-ri.com

Abstract

災害地における要救助者の捜索を音源同定で実現するために，*Partially Shared Deep Neural Network* (PS-DNN) およびこの拡張版である *Partially Shared Convolutional Neural Network* (PS-CNN) を提案し，これで音源同定器を学習する手法を提案する．通常の深層学習には大量のデータにラベルを付与する作業が必要であるが，提案手法は音源同定器の学習にラベルが付与されていないデータを有効に利用することで，ラベルが付与されたデータのみで学習した場合と比べて高い同定精度が得られることを検証した．

1 序論

地震等の災害現場では，寸断された道路や散乱した瓦礫が要救助者の捜索活動の大きな妨げとなる．クアドロコプタを始めとする *Unmanned Aerial Vehicle* (UAV) で捜索すれば移動の問題は解消されるが，要救助者が瓦礫に埋もれている場合，カメラやレンジファインダ等の視覚的なセンサでの探知は困難である．このため，我々はセンサとしてマイクロホンアレイを用い，災害現場で発生する音の種類と発生位置を同定することで要救助者を探知する方法を研究している．

クアドロコプタにマイクロホンアレイを搭載する場合，風切り音やプロペラが発する雑音によって *Signal-to-Noise* (SN) 比が低下する．このような低 SN 比環境下で音源同定を行う手法として，我々はこれまでに多チャネル音響信号を元にした音源定位手法である *MUltiple SIgnal Classification based on incremental Generalized Singular Value Decomposition* (iGSVD-MUSIC) [Ohata 14] を用いてマイクロホンアレイの収録音から同定対象の音の定位と区間検出を行い，音源分離手法である *Geometric High-order Decorrelation-based Source Separation* (GHDSS) [Nakajima 10] を用いて SN 比の低い多チャネル音から信号成分のモノラル音を分離し，この分離音の種類を *Convolutional Neural Network* (CNN) [Lawrence 97] で識別する手法 [Uemura 15] を提案した．しかし，この手法では音源分離が識別器の最適化とは独立しているため，多チャネル音からの音源分離という大幅な低次元化の過程で識別に有用な情報までもが失われる可能性がある．明示的にノイズ抑圧等の処理を行わず，大規模な DNN を用いて原信号から直接識別する手法 [Hannun 14] も提案されているが，大規模な DNN の学習には大量の学習データが必要である．実環境音を収録して学習データセットを構築する場合，何の音がどの区間で鳴っているのかを示すラベルデータを付与する作業（アノテーション）を人力で行う必要があるり，データ量が多くなれば膨大な工数が発生する．

本稿では，全収録音の一部しかアノテーションされていないデータセットを用いて音源同定器を効率的に学習する手法を提案し，これを実際に DNN，および CNN に適用してその有効性を検証する．一般的な深層学習は学習に教師データ，つまり入力とそれに対応する望ましい出力の組み合わせが必要であるため，アノテーションされていないデータは学習に使用できない．提案手法はラベルデータに加え，信号処理的な音源分離手法で自動生成できる分離音を学習データとして用いることで，音源同定器の学習を効率的に行いつつ未アノテーションデータを有効に利用することができる．

2 部分共有型ニューラルネットワーク

本節では，*Multi-Task Learning* (MTL) [Caruana 97] の一種である *Partially Shared Deep Neural Network* (PS-DNN) およびこれを CNN に拡張した *Partially Shared Convolutional Neural Network* (PS-CNN) の構造につい

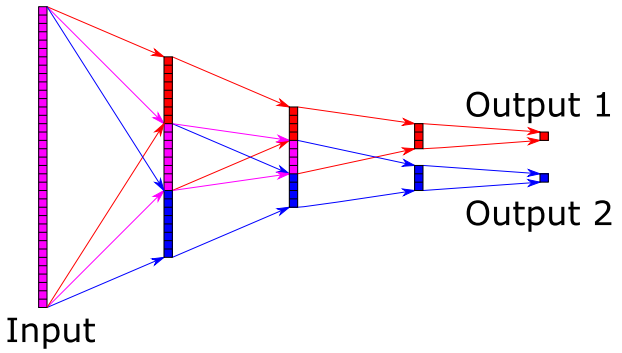

図 1: Partially Shared Deep Neural Network

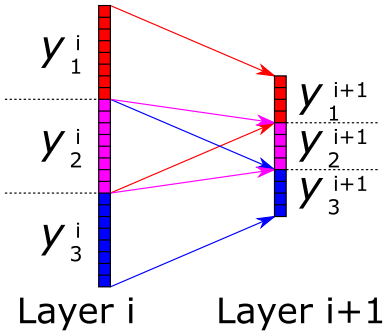

図 2: A hidden layer of PS-DNN

て述べ，多チャネル音響信号を元に音源同定を行う方法を説明する．

2.1 Partially Shared Deep Neural Network

PS-DNN の構造を図 1 に示す．PS-DNN は二つのサブネットワークから成るニューラルネットワークであり，本稿では片方はラベルデータを出力する音源同定器，もう片方は分離音を出力する音源分離器である．サブネットワーク間で入力層と隠れ層の一部が共有されており，共有された隠れ層は二種類の教師データを用いて学習される．

入力層と最初の隠れ層の間は全結合であるが，隠れ層間は全結合ではなく，$\boldsymbol{y}_1^i, \boldsymbol{y}_2^i, \boldsymbol{y}_3^i$ を図 2 における第 i 層の上側，中央の共有部分，下側の隠れ層の出力とすると，第 $i+1$ 層の出力 $\boldsymbol{y}_1^{i+1}, \boldsymbol{y}_2^{i+1}, \boldsymbol{y}_3^{i+1}$ は次の式 (1) で計算される．

$$\begin{pmatrix} \boldsymbol{y}_1^{i+1} \\ \boldsymbol{y}_2^{i+1} \\ \boldsymbol{y}_3^{i+1} \end{pmatrix} = \boldsymbol{\sigma}\left(\begin{pmatrix} \boldsymbol{W}_{11}^i & \boldsymbol{W}_{12}^i & 0 \\ 0 & \boldsymbol{W}_{22}^i & 0 \\ 0 & \boldsymbol{W}_{32}^i & \boldsymbol{W}_{33}^i \end{pmatrix}\begin{pmatrix} \boldsymbol{y}_1^i \\ \boldsymbol{y}_2^i \\ \boldsymbol{y}_3^i \end{pmatrix} + \begin{pmatrix} \boldsymbol{b}_1^i \\ \boldsymbol{b}_2^i \\ \boldsymbol{b}_3^i \end{pmatrix}\right) \quad (1)$$

ここで $\boldsymbol{W}_{jk}^i$ は $\boldsymbol{y}_k^i$ から $\boldsymbol{y}_j^{i+1}$ への重み行列，$\boldsymbol{b}_j^i$ はバイアスベクトル，$\boldsymbol{\sigma}(\cdot)$ は要素ごとの活性化関数である．

共有された隠れ層の出力は上層の全ネットワークに影響を与える．この構造は，音源同定と音源分離はある程度共通の処理で行えるという予想に基づいている．一方，非共有部分の出力はもう一方のサブネットワークの上層には影響を与えず，パラメータは一種類の教師データのみを用いて学習される．これは，音源同定と音源分離にはそれぞれに固有の処理も必要であるという予想に基づいている．この構造により，音源同定と音源分離に共通する処理を二種類の教師データを用いて効果的に学習しつつ，音源分離に固有の処理が音源同定の学習に悪影響を与えることを抑制することが期待される．

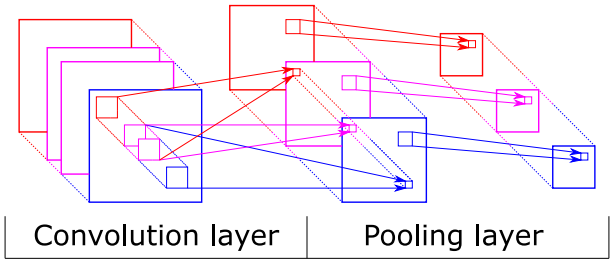

図 3: A convolution-pooling layer in PS-CNN

2.2 Partially Shared Convolutional Neural Network

PS-CNN は，CNN の畳み込み層に PS-DNN の構造を取り入れたものである．なお，本稿では CNN で音響信号を扱う際，1 チャネル分の音響特徴量ベクトルを時間方向に並べた 2 次元の配列を 1 枚の画像とみなす．つまり，一般的なカラー画像認識を行う CNN の入力は画素値を表す 2 次元配列を RGB の 3 チャネル分並べたものであるが，本稿で扱う CNN の入力は上述の 2 次元配列をマイク数分並べたものである．[1]

PS-CNN の畳み込み・プーリング層の構造を Fig. 3 に示す．PS-DNN では各層の出力ベクトルの要素を共有部分と非共有部分に分けたのに対し，PS-CNN ではチャネルを共有チャネルと非共有チャネルに分ける．つまり，一般的な CNN では一つのフィルタは前の層の全てのチャネルを入力とするのに対し，PS-CNN では各サブネットワークに固有のチャネルおよび共有チャネルのみを入力とする．この構造により，PS-DNN と同様に二種類の教師データを有効に利用しつつ，CNN の構造を取り入れることができる．プーリング層は一般的な CNN と同様にチャネルごとにプーリングを行う．出力層の前の全結合層の構造は PS-DNN と同様である．

第 i 層の出力の，一つ目のサブネットワークに固有のチャネルの数を $K_{i,1}$，共有チャネルの数を $K_{i,2}$，二つ目のサブネットワークに固有のチャネル数を $K_{i,3}$ とする．第 i 層の出力を $[\boldsymbol{X}_1^{(i,1)}, \cdots, \boldsymbol{X}_3^{(i,K_{i,3})}]$，$\boldsymbol{X}_1^{(i,1)} = [x_{1,1,1}^{(i,1)}, \cdots, x_{1,V,H}^{(i,1)}]$，第 $i+1$ 層の第 j チャネルの出力のサイズを $V \times H$，出力

[1] CNN で音響信号を扱う別の方法として，音響特徴量ベクトルの次元数をチャネル数とし，音響特徴量ベクトルの要素をマイク数，フレーム数分並べた 2 次元配列を 1 枚の画像とみなす方法も考えられる．この方法の検討は今後の課題とする．

を $\boldsymbol{C}^{(i+1,j)} = [c_{1,1}^{(i+1,j)}, \cdots, c_{v,h}^{(i+1,j)}, \cdots, c_{V,H}^{(i+1,j)}]$ とすると，$c_{v,h}^{(i+1,j)}$ は第 j チャネルが一つ目のサブネットワークに固有のチャネルである場合は式 2 で，共有チャネルである場合は式 3 で，二つ目のサブネットワークに固有のチャネルである場合は式 4 で求められる．

$$c_{v,h}^{(i+1,j)} = \sigma \quad (\sum_{k=1}^{K_{i,1}} \sum_{s=1}^{m} \sum_{t=1}^{n} w_{k,s,t} x_{1,v+s,h+t}^{(i,k)} + \sum_{k=1}^{K_{i,2}} \sum_{s=1}^{m} \sum_{t=1}^{n} w_{k,s,t} x_{2,v+s,h+t}^{(i,k)} + b^{(i,j)}) \tag{2}$$

$$c_{v,h}^{(i+1,j)} = \sigma \quad (\sum_{k=1}^{K_{i,2}} \sum_{s=1}^{m} \sum_{t=1}^{n} w_{k,s,t} x_{2,v+s,h+t}^{(i,k)} + b^{(i,j)}) \tag{3}$$

$$c_{v,h}^{(i+1,j)} = \sigma \quad (\sum_{k=1}^{K_{i,2}} \sum_{s=1}^{m} \sum_{t=1}^{n} w_{k,s,t} x_{2,v+s,h+t}^{(i,k)} + \sum_{k=1}^{K_{i,3}} \sum_{s=1}^{m} \sum_{t=1}^{n} w_{k,s,t} x_{3,v+s,h+t}^{(i,k)} + b^{(i,j)}) \tag{4}$$

ここで m, n はフィルタサイズ，$w_{k,s,t}$ は重み，$b^{(i,j)}$ はバイアス，$\sigma(\cdot)$ は活性化関数である．

3　評価実験

提案手法の有効性を示すため，各手法で音源同定器を構成し，同定精度を比較した．同定精度はフレームごとの正解率とした．各ネットワークは Python のライブラリである TensorFlow version 0.8.0 [Abadi 15] で実装した．

音源同定器は 4 つの手法で構成し，それぞれ DNN, PS-DNN, CNN, PS-CNN と表記する．DNN, CNN はそれぞれ典型的なフルコネクテッド，畳み込みニューラルネットワークで構成した音源同定器で，PS-DNN と PS-CNN は第 2 節で述べた学習手法で構成した音源同定器である．DNN, CNN では学習用データセットの内アノテーション済みのものしか学習に使用しないが，PS-DNN, PS-CNN では未アノテーションデータも音源分離器の学習に使用する．

学習・評価用の音源として，DCASE2016 [Mesaros 16] の Acoustic scene classification に収録されている 15 種類合計 35100 秒分の音データを用いた．5 分割交差検証を行うために合計 1170 個の Wave ファイルを 5 つのグループに分け，3.1-3.2 に示す手順で合計約 465 万個のデータベクトルを生成した．

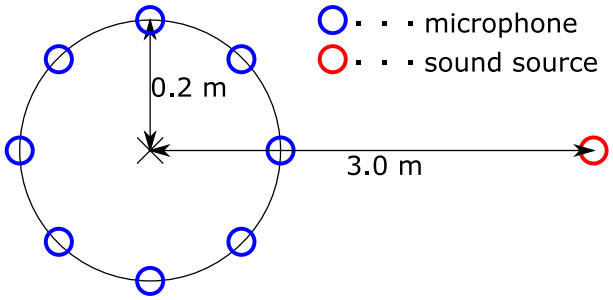

図 4: The layout of the microphones and the sound source

図 5: UAV (Parrot Bebop Drone)

3.1　音響信号の合成

実験に使用した多チャネル音響信号は数値シミュレーションで合成した．まず，コーパスの収録音を元にマイクロホンアレイと音源の位置関係が図に示す通りであるときの 8 チャネルの音を合成した．その後，図 5 に示すクアドロコプタで実際に収録したノイズを適当な重みを付けて足し合わせることで，一定の SN 比の多チャネル音響信号を生成した．なお，本実験で使用したデータセットは場面認識のベンチマーク用のものであるため，収録されている音は既に様々なノイズを含んでいるが，SN 比を計算する際は純信号として扱った．

本稿で使用した SN 比の計算式を式 5 に示す．SN 比は 0 dB に統一した．

$$SNR = 20 \log(S_p / N_p) \tag{5}$$

ここで S_p, N_p はそれぞれ信号成分の最大振幅，雑音成分の最大振幅である．SN 比は信号のエネルギの比率で計算されることもあるが，このような計算方法では有音区間の定め方によって求められる SN 比が大きく変化する場合があるため，本稿では最大振幅で定義した．

表 1: Dimensions for the DNN

Hidden layer	Units
1	2000
2	1000
3	400

表 2: Dimensions for the PS-DNN

Hidden layer	Units		
	Identify	Shared	Separate
1	1500	1500	1500
2	800	400	800
3	400	0	800

3.2 音響特徴量の算出

各学習器への入力としてメルフィルタバンク特徴量を使用した．各音のサンプリングレートは 16 kHz に統一し，フレーム幅 512 sample (32 ms)，フレームシフト 120 sample (7.5 ms) でフレーム化し，窓関数として複素窓を掛けて短時間フーリエ変換で複素スペクトルを求めた．これの絶対値から，下限周波数 63 Hz, 上限周波数 8 kHz, 次元数 20 のメルフィルタバンク特徴量を算出した．以上の処理は，ロボット聴覚ソフトウェア *Honda Research Institute Japan Audition for Robots with Kyoro University* (HARK) [Nakadai 10] で実装した．

各学習器への入力は，20 次元の音響特徴量を 8 チャネル各 20 フレーム分並べた，合計 3200 次元のベクトルである．また，PS-DNN, PS-CNN の音源分離側の出力は，多チャネル音の合成に使用したモノラル音から同様に算出した 400 次元のベクトルである．

3.3 学習器の条件

各学習器の層構成を表 1-4 に示す．全ての場合で入力は 3.2 で述べた 3200 次元のベクトルである．音源同定器の出力層は 15 次元のソフトマックス層であり，PS-DNN, PS-CNN の音源分離側の出力層は 400 次元の全結合層である．各パラメータは 0 に近い正の値で初期化し，pre-training を行わずに Adam で学習した．隠れ層に対しては Dropout を使用し，drop rate は畳み込み層で 0.2，プーリング層で 0，その他の層で 0.4 とした．畳み込み層のフィルタは

表 3: Dimensions for the CNN

Hidden layer	Type	Channels	Size
1	Conv	40	20×20
2	Pool	40	10×10
3	Conv	80	10×10
4	Pool	80	5×5
5	Full	400	1×1

表 4: Dimensions for the PS-CNN

Hidden layer	Type	Channels			Size
		Identify	Shared	Separate	
1	Conv	40	40	40	20×20
2	Pool	40	40	40	10×10
3	Conv	80	40	80	10×10
4	Pool	80	40	80	5×5
5	Full	400	0	800	1×1

表 5: Accuracy of Sound Source Identification

		DNN	PS-DNN	CNN	PS-CNN
100%	Avg.	55.88	**56.27**	55.79	**56.75**
	S.E.	0.6756	0.5742	0.5348	0.5275
75%	Avg.	54.07	**54.57**	54.36	**55.09**
	S.E.	0.6599	0.6584	0.5055	0.3268
50%	Avg.	51.63	**51.91**	51.95	**52.71**
	S.E.	0.5786	0.6525	0.5332	0.5357
25%	Avg.	47.84	**48.04**	48.35	**48.74**
	S.E.	0.6477	0.6526	0.5566	0.5898

全ての場合で 5 × 5 とし，zero padding を使用した．プーリング層では 2 × 2 の範囲で最大値プーリングを行った．全ての場合でバッチサイズは 100 とし，学習は 10 epoch 行った．

3.4 実験結果

実験結果を表 5 および図 6-9 に示す．識別精度はフレームごとの識別正解率とし，5 分割交差検証の平均値 (Avg.) と標本標準誤差 (S.E.) を求めた．5 分割されたデータセットの内 4 つを学習に用い，その 4 つの内の所定の数についてはラベルデータを使用しないことで，アノテーション率が 100%, 75%, 50%, 25%の場合の実験を行った．

表 5 より，識別精度はアノテーション率に依らず DNN < PS-DNN，また CNN < PS-CNN となった．いくつかの場合で両側 t 検定の検定の p 値が $p > 0.05$ で有意な差があった．

識別精度に大きな差が出なかったのは，実験で用いた音

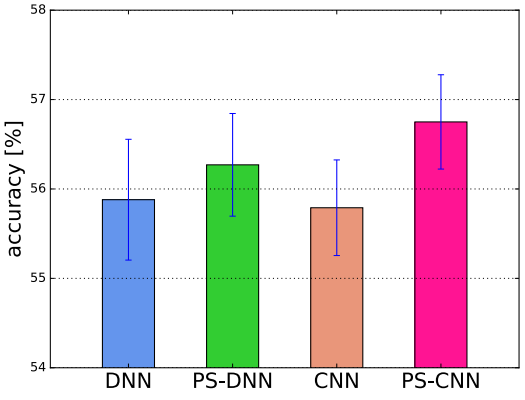

図 6: Trained with 100% annotated data

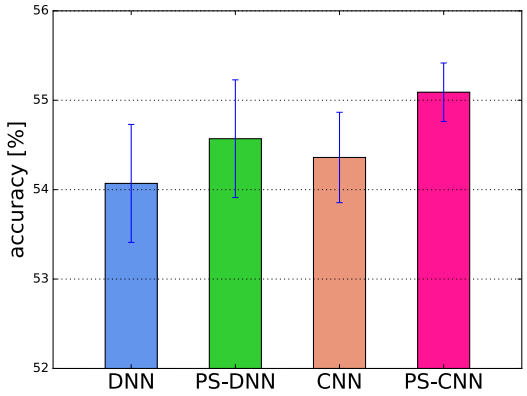

図 7: Trained with 75% annotated data

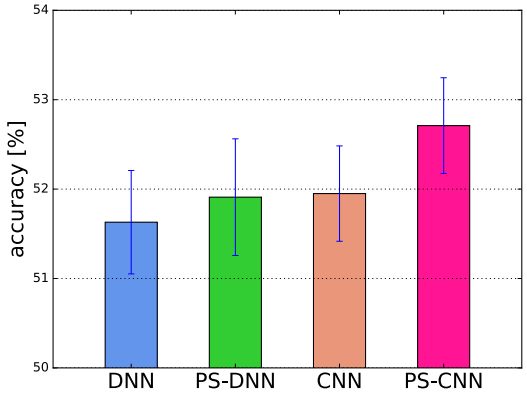

図 8: Trained with 50% annotated data

源分離の教師データに雑音が始めから含まれてしまっていたことが一因であろう．本稿で提案した手法は，低 SN 比環境下で音源同定を行うニューラルネットワークは雑音抑圧の処理を学習しているという推測の下，音源同定器に雑音抑圧の処理を効率的に学習させることを意図している．しかし，本実験で用いたコーパスである DCASE2016 の Acoustic scene classification 用のデータセットは，音を収録した場所（公園，レストラン，電車等）の識別を行うベンチマークデータセットであり，収録されている音は様々な雑音を元々含んでいる．実験ではこの収録音を教師データとして用いたため，学習された音源分離器はクアドロコプタ由来の音以外を除去せず，むしろその他の雑音を積極的に残していたと考えられる．残りの雑音の抑圧は識別器側の共有されていない部分のみを用いて学習することになるため，識別精度が大きく向上しなかったと考えている．

4 結論

本稿では，マイクロホンアレイを搭載したクアドロコプタによる災害地での要救助者の捜索を目的とした，低 SN

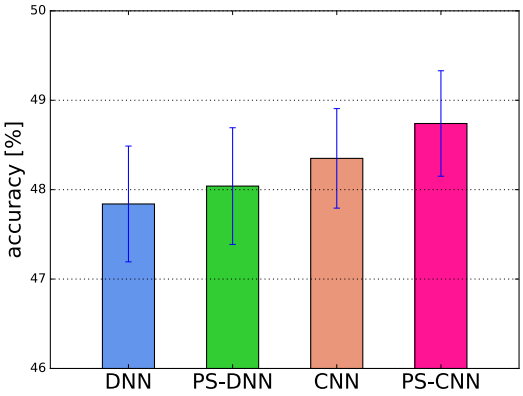

図 9: Trained with 25% annotated data

比環境下での音源同定器の学習手法について述べた．多チャネル音響信号を入力とする音源同定器に，音源分離の処理を積極的に学習させる手法を提案した．提案手法は一般的な DNN, CNN と比べて若干高い同定精度を実現した．今後は，別のデータセットを用いた提案手法の有効性の検証を行う予定である．

謝辞

本研究は JSPS 科研費 24220006, 16H02884, 16K00294 および，JST ImPACT タフロボティクスチャレンジの助成を受けた．

参考文献

[Abadi 15] Abadi, M., et al.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, http://tensorflow.org/ (2015)

[Caruana 97] Caruana, R., et al.: Multitask learning, Machine Learning, vol.28, no. 1, pp. 41-75 (1997)

[Mesaros 16] Mesaros, A., et al.: TUT database for acoustic scene classification and sound event detection, 24th Acoustic Scene Classification Workshop 2016 European Signal Processing Conference (EUSIPCO) (2016)

[Hannun 14] Hannun, A., et al.: Deepspeech: Scaling up end-to-end speech recognition, arXiv preprint arXiv:1412.5567 (2014)

[Lawrence 97] Lawrence, S., et al.: Face recognition: A convolutional neural-network approach, IEEE Transactions on Neural Networks, vol. 8, no. 1, pp. 98-113 (1997)

[Nakadai 10] Nakadai, K., et al.: Design and Implementation of Robot Audition System "HARK", Advanced Robotics, vol. 24, pp. 739-761 (2010)

[Nakajima 10] Nakajima, H., et al.: Correlation matrix estimation by an optimally controlled recursive average method and its application to blind source separation, Acoustical Science and Technology, vol. 31, no. 3, pp. 205212 (2010)

[Ohata 14] Ohata, T., et al.: Inprovement in outdoor sound source detection using a quadrotor-embedded microphone array, IEEE/RSJ International Conference on Intelligent Robots and Systems (2014).

[Uemura 15] Uemura, S., et al.: Outdoor acoustic event identification using sound source separation and deep learning with a quadrotor-embedded microphone array, The 6th International Conference on Advanced Mechatronics (2015)