人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-046-06 (11/09)

UAV 搭載マイクアレイを用いた 高雑音環境下における音イベント検出・識別の並列最適化

杉山　治[1*], 小島 諒介[1], 中臺 一博[1,3]

Osamu SUGIYAMA[1], Ryosuke KOJIMA[1], Kazuhiro NAKADAI[1,2]

1. 東京工業大学 大学院 情報理工学研究科, 2. （株）ホンダ・リサーチ・インスティチュート・ジャパン

1. Graduate School of Information Science and Engineering, Tokyo Institute of Technology,

2. Honda Research Institute Japan Co., Ltd.

sugiyama@kuhp.kyoto-u.ac.jp, kojima@cyb.mei.titech.ac.jp,
nakadai@jp.honda-ri.com

Abstract

無人航空機 (UAV) に搭載したマイクアレイは, 近くにノイズを発生するローターがあるため, 常に高雑音環境にさらされる. 本稿では, このような UAV に搭載したマイクアレイを用いて音源検出・音源識別をする際に現れる特有の課題に触れ, それらを解決するための並列最適化手法を提案する. 提案システムでは，定位と識別に異なるパラメータセットを用いた並列処理機構を持ち，さらに識別に用いる畳み込みニューラルネットワークのソフトマックス層から得られる確信度によって識別するフレームを取捨選択することで，定位と識別を同時最適化する．また，UAV 搭載マイクアレイによって収集した音声を用いた実験を通じて，提案システムの有効性を示した.

1　はじめに

本稿では，無人航空機 (Unmanned Aerial Vehicle, UAV) を用いた屋外音環境理解に取り組む．UAV に搭載したマイクアレイを用い収集した音を通じて「いつ」「どこで」「なに」といった 5W1H の情報を理解することができれば，例えば，災害時に人が立ち入ることが困難な場所においても UAV で空から被災者の探索を支援することができる．このように我々は屋外環境で音源を検出し, その位置や種類を推定する技術を「屋外音環境理解」とし，屋外音環境理解のための技術を確立するための研究を行っている．UAV に搭載したマイクアレイは，近くにローター音や風切り音などのノイズを発生するローターがあるため，常に高雑音環境にさらされる．本稿では，このような UAV に搭載したマイクアレイを用いて音源検出・音源識別をするときの特有の課題に触れ，それらを解決するためのフレームワークを設計・提案する．

*現在，京都大学医学部附属病院勤務

2　音源検出と音源識別を同時に行う際に現れる特有の課題

UAV にマイクアレイを搭載する場合の大きな課題の一つとして，UAV 自身のローター音や風切り音などのノイズに常にさらされることが挙げられる．このローター音は，UAV の飛行状態にある場合に強くなり，遠方から到来する本来識別したい音イベントの音声信号の大きさはローター音と比較して小さいことが多い．結果として，UAV に搭載したマイクアレイに入力される音声信号は，常に SN 比が悪い状態で収録されることとなる．さらに，そこに環境のノイズが加わるため，UAV に搭載したマイクアレイから収録される多チャネル音声信号を用いた音源定位と音源識別手法は高雑音環境下においてもロバストに動作する必要がある．このような高雑音環境下においては，音源検出と音源識別のパラメータ最適化を同時に行うことが難しい．音源定位をする場合，音源全体を漏れなく検出するために，音のパワーが強い部分だけでなく，その前後の SN 比が低い部分も検出する必要がある．対して，音源識別をする場合には，SN 比が良い部分のみを用いた方が識別精度が高くなる．したがって，音源定位・音源識別それぞれに特化した最適化を行う必要がある．

これまでも UAV 搭載のマイクアレイを用い，音源定位，分離，識別の研究を行ってきた [1, 2, 3]．音源検出と音源定位については，マイクアレイの音源定位手法である Multiple Signal Classification based on incremental Generalized Singular Value Decomposition with Correlation Matrix Scaling (iGSVD-MUSIC-CMS) 法 [4] を提案し，10-20 [m] 程度離れた場所の音源でも UAV 搭載マイクアレイから音源の定位および検出ができることを示した．この手法に加え，これまでに音源分離手法として，マイクアレイを用いた音源分離法として高性能であることが報告されている Geometric High-order Dicorrelation-based Source Separation with Adaptive Step-size control (GHDSS-AS) 法 [5] と，深層学習の一つである畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) [6] を組み合わせることで高雑音環境下でロバストに動作する音源識別手法を提案した [7]．CNN は元来，画像の識別に特化したニューラルネットワークであるが，入力として，縦軸を周波数成分，横軸を時間成分にとったスペクトログラムを画像的特徴量を用いることで，音イベント識別でも有効に作用する．本稿では，これまで個々に検証されてきた音源定位・検出・分離・識別を一つのフレームワークとして統合し，フレームワーク全体として高雑音環境下においても音源定位・識別性能を同時最適化する仕組みを提案する.

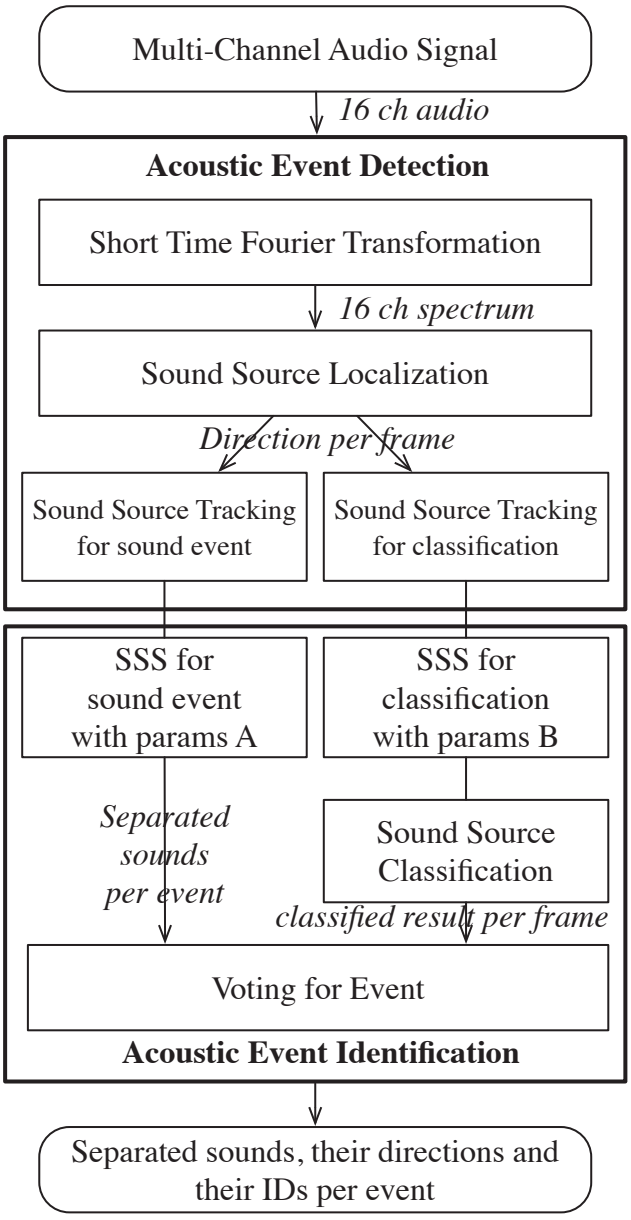


図 1: フレームワークの概要

3 高雑音環境下で音源検出と音源識別を同時最適化するフレームワーク

本稿で提案する高雑音環境下で音源検出と音源識別を行うフレームワークの概要を図 1 に示す．システムは大きく分けて，音イベント検出部 (Acoustic Event Detection, AED) と音イベント識別部 (Acoustic Event Identification, AEI) に分かれ，それぞれに音源定位・検出・分離・識別・統合の仕組みを内包する．

音イベント検出部（図 1，AED) では，短時間フーリエ変換 (STFT) を用い，各チャネルの音声信号を周波数領域に変換する．次に iGSVD-MUSIC-CMS 法を用いて，雑音を抑圧しながら音源定位を行い，MUSIC 空間スペクトログラムを得る．得られた MUSIC 空間スペクトログラムを用いて，音イベントを追跡し，音源区間を音イベント識別部 (図 1，AEI) に出力する．音イベント識別部では，まず，得られた音源区間情報と 16 ch のスペクトラムを利用し，分離音を抽出する．次に，その分離音を CNN を用いて識別し，フレーム別の識別結果を得る．これらフレーム別の識別結果と音イベント検出部から得られた音源区間情報を組み合わせて，音イベント毎の識別結果を求める (図 1, Voting for Event)．識別部の構成については，3.1 節でより詳細に述べる．また，前述したとおり，高雑音環境下においては，音源検出と音源識別において，同じパラメータで性能を最適化することが困難であるため，図 1 中破線で囲んだ領域のように音源追跡と分離に関しては，2 つの異なるパラメータを用いて並列的に区間検出用，識別用のデータを切り出す．この処理については，3.2 節でより詳細に述べる．最後に，音イベントの区間情報と，識別結果をどのように投合するかについて，3.3 節で述べる．

3.1 音源分離と深層学習を組み合わせた雑音ロバストな識別

UAV 搭載のマイクアレイから得られる多チャネル音声信号は，ロータの雑音が常に入ってくるため，SN 比の悪い状態で収録される．このように SN 比の悪い音声信号をそのまま既存の GMM のような識別器にかけても，雑音の方が主要な成分となり，識別することができない．そこで，本稿では課題を分割し，予め音源分離を行うことで，雑音抑圧された分離音を抽出し，その分離音を識別するというアプローチをとる．ここで，音源分離手法としては，前述した GHDSS-AS 法を用いる．GHDSS-AS 法はビームフォーミングとブラインド音源分離の 2 つのコスト関数を組み合わせて分離を行うため，方向性のあるスパースな音声信号（つまり，本稿で対象とする人の発話や救助を求める人工的な音声）を効率よく分離できるものと考えられる．また，識別手法としては，CNN を用いた．CNN は画像の識別に特化した識別手法であり，本稿では，縦軸を周波数成分，横軸を時間成分にとったスペクトログラムを画像的特徴量とみなして，CNN に用いる．フィードフォワード型のフルコネクションのニューラルネットワークと異なり，CNN は全結合されていないため，音声信号の識別で重要であると考えられる時間成分を明に考慮できるものと考えられる．

使用した CNN の構成を図 2 に示す．入力はログフィルタバンク特徴量 20 次元を 20 フレーム分連ねた 20×20 のスペクトログラムとし，畳み込み層 (32 カーネル)，プーリング層（最大値プーリング）を 2 層ずつ連ねた後，出力層で統合し，識別学習を行った．学習には，識別用に調整された分離音を用い，人の発話や，救助を求めるときに人が出すと考えられるホイッスルなどの音に加え，災害現場で検出されそうな救急車の音声などを交え，8 クラスの識別タスクを学習した．最終層の softmax 層の出力は，識別学習だけでなく，3.3 節の音イベントの識別（Voting for Event）にも用いられる．

3.2 定位と識別を同時最適化するための区間検出の並列処理

高雑音環境下で音源検出と音源識別をする場合，区間検出において同じパラメータで性能を同時最適化することが困難である．そこで，本稿では，音源検出と音源識別においてパラメータの異なる 2 つの区間検出処理を並列で動かすことで検出と識別の同時最適化を図る．

音源定位と音源識別で最適な音源区間の概要を図 3 に示す．図 3 に示す通り，音源区間検出では，図中の a) Th, b1) pre-margin, b2) post-margin の各パラメータを用いて，音源区間を検出する．Th は音源と雑音を分ける閾値を，pre-margin は音の開始地点のパワーが低い部分が何ミリ秒続くかを，post-margin は音の終了地点のパワーが低い部分が何ミリ秒続くかをそれぞれ表す．音源定位に最適な区間検出を考える場合，音イベントの開始・終了時点の音源のパワーは低いため，SN 比の低い区間を含んで検出する必要がある．したがって，検出音源の性質に合わせて，pre-margin, post-margin を適切に設定する必要があり，また音源と雑音をわける閾値 Th も音源全体を検出できるように低めに設定する必要がある（図 2，下部の閾値設定参照）．一方，音源識別に最適な区間検出を考える場合，SN 比が大きい音イベントの特徴がよく現れたフレームを用いて識別したほう

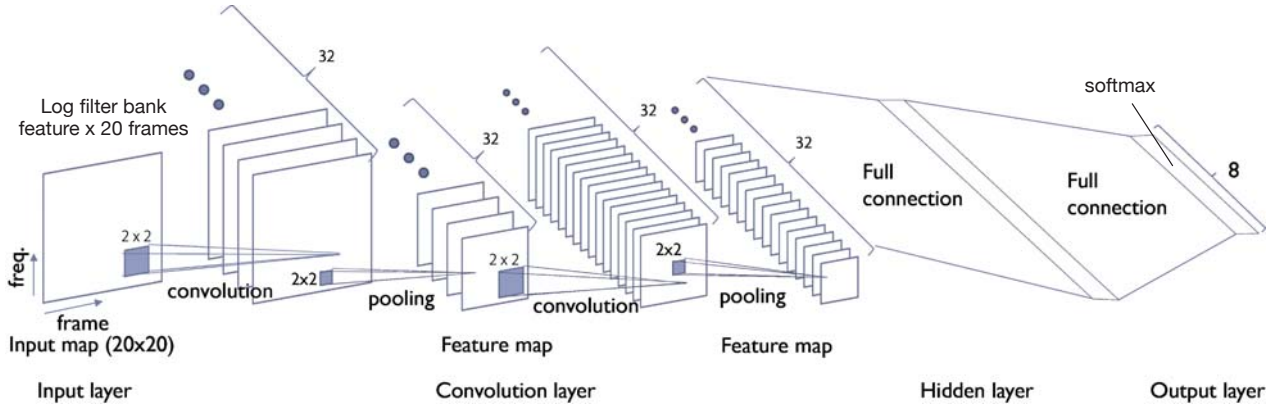


図 2: 識別学習に用いた CNN

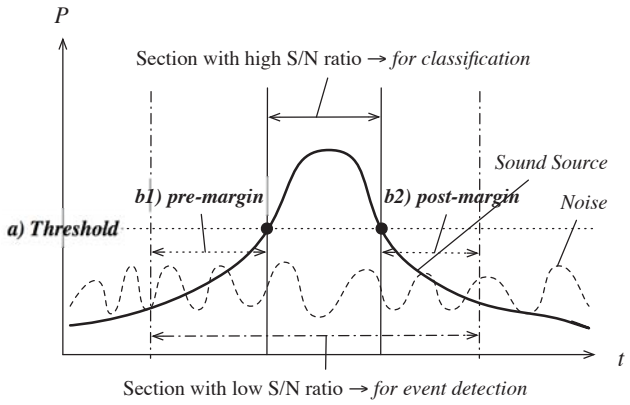


図 3: 音源定位と音源識別で最適な音源区間

図 4: 実験で試用した UAV とマイクアレイ

が良いため，pre-margin, post-margin はなるべく小さい方がよく，閾値 Th も高めに設定する必要がある（図 3，上部の閾値設定参照).

本稿では，これらの 2 つの条件には矛盾があるため同じパラメータを用いた最適化は不可能であると考え，定位用・識別用の 2 つのパラメータセット Th, post-margin, pre-margin によって，並列に定位用・識別用の区間検出を行う.

3.3 確信度が高いフレームを用いた区間識別の最適化

定位用・識別用，2 つのパラメータセットを用いた並列区間検出は最終的に図 1 の音イベント識別部 (Voting for Event) で統合される．この際，音イベント識別のための CNN は SN 比が高い音声信号に合わせて学習しているため，検知したイベントから抽出された音声特徴量全てを対象とするのではなく，図 3 で示した音の特徴を良く表現しているフレームを選択し識別を行うことが望ましい．しかしながら，観測時にはその音声信号の SN 比は未知であり，また，図 3 のモデルで示すように観測区間の中心が必ずしも SN 比が良い区間とも限らない．そこで，本研究では，SN 比の代わりに CNN の最終層であるソフトマックス層の出力のうち，argmax で選ばれたノードの値を確信度とみなし，この確信度が一定値以上のフレームのみを用いて，識別を行うこととする．CNN の畳み込み層のカーネルが各音イベントの特徴を正しく学習しているのであれば，ソフトマックス層から得られる確信度は音の特徴をよく表すフレームで高くなるものと考えられる．本研究では，実験を通して，この仮説を検証した（5.3 節参照）.

4 UAV 搭載マイクアレイによる音環境理解システム

提案したフレームワークに基づき，UAV 搭載のマイクアレイを用いた音環境理解システムを構築した．図 4 に示すように，UAV には，Asctech 社のクアドロコプタ Pelican を用いた．この機体の外周上に等間隔にマイクロホンを 16 個配置しマイクアレイとした．各マイクロホンは，黒い毛状の風防で覆うことで風切り音の低減を図った．これらのマイクロホンの信号は 16 [bit], 16 [kHz] で同期収録される．マイクアレイで収録した音声信号は，ロボット聴覚のオープンソースソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [8] を用いて，音源定位，音源追跡，音源分離を行い，得られた分離音を入力として Python の深層学習パッケージである chainer[1] で実装した CNN で識別学習を行った.

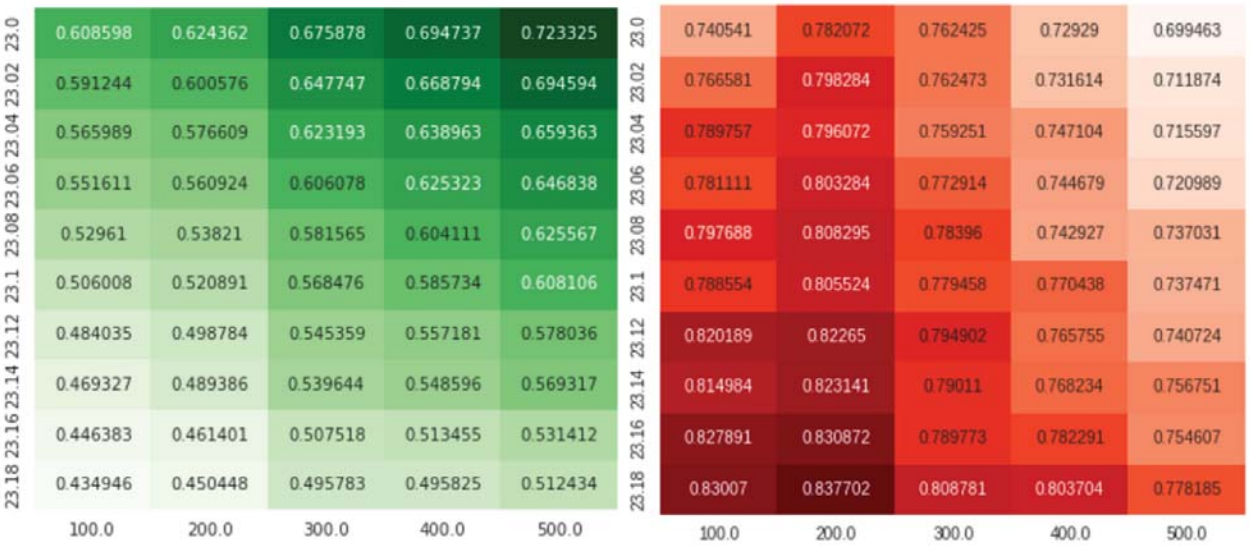
5 実験

本実験の目的は，提案したフレームワークの有効性を検証することである．本稿では，5.2 節で、提案した並列処理が有効に動作しているかどうかの検討，5.3 節で，CNN のソフトマックス層から得られた確信度を用いた音イベント識別において，識別率が向上するかどうかの検討をそれぞれ詳細に述べる.

5.1 データ収集

実験には，RWCP 実環境音声・音響データベースと電子協騒音データベースに含まれる計 8 種類の音源を用い，音声 1 種類 (男性の呼び声) と非音声 7 種類（携帯, 救急車, 拍手, 目覚まし時

[1] http://chainer.org/



a) Recall rate of Sound Source Localization

b) Precision rate of Sound Source Identification

図 5: 定位再現率と識別適合率のヒートマップ

計, シンバル, ホーン, カラスの鳴声) から構成される．音源はスピーカから 3.0-4.0 秒音源を，その後に無音区間を 3.0 - 4.0 秒それぞれ出力し，それを 1 単位として 15 回繰り返して収録を行った．収録は，屋外にて UAV をプロペラを回転させた状態で固定，UAV の中心から 45 [°] 方向，距離 3.0 [m] 地点に音源を設置し収録した. 音源定位は, 各周波数ビンでは, 同時に存在する音源数は高々ひとつであるという仮定の下, iGSVD-MUSIC-CMS を実行した (ただし, 周波数方向に統合を行ったブロードバンド空間スペクトルを用いること, および音源追跡時に一定の音源生存期間 (500 [ms]) が仮定されることから，検出結果には同時に複数の音源が含まれる可能性がある). 音源識別は, 全データの 8 割を学習に，残りの 2 割を評価に用いて，8 種類の音源を識別する 8 クラス識別タスクを行い，K-分割交差検証 (K=5) を行った.

5.2 並列区間検出の性能評価

並列区間検出の性能評価では，3.2 節で述べた Th, pre-margin, post-margin を変化させ，フレームベースの定位再現率，フレームベースの識別適合率から最適なパラメータ群を探索的に求めた．Th は, 事前検討において音源定位時に最も定位数が変化した 23.0 から 23.18 までを 0.02 刻みで変化させた．一方で，pre-margin, post-margin に関しては，100 [ms] 刻みで，100 [ms] から 500 [ms] まで変化させ，定位再現率，識別適合率の推移を検証した．そして，a) 定位用のパラメータを用いた場合，b) 提案システムで 2 つの区間検出を組み合わせた場合の識別適合率を比較し，提案手法の有効性を検討した.

実験結果を図 5 に示す．図 5 は，縦軸を閾値 Th，横軸を pre-margin, post-margin とした時の定位再現率，識別適合率の推移をヒートマップとして表したものである．ヒートマップは各値の推移を見るために，数値の高い部分ほど色が濃く，数値の低い部分ほど薄くなるように図示したもので，直感的に値の推移を読み取ることができる．図を見ると，定位再現率は，Th が低く，pre-margin, post-margin が高くなるにつれて向上するのに対し，識別適合率は，Th が高く，pre-margin, post-margin 共に低くなるにつれて向上することが示された．このことから，定位と識別を単独で最適化するパラメータセットは真逆の設定になることが示された.

次に，2 つのパラメータセットを組み合わせてシステム全体の識別性能を最大化したときの識別率の比較を図 6 に示す.

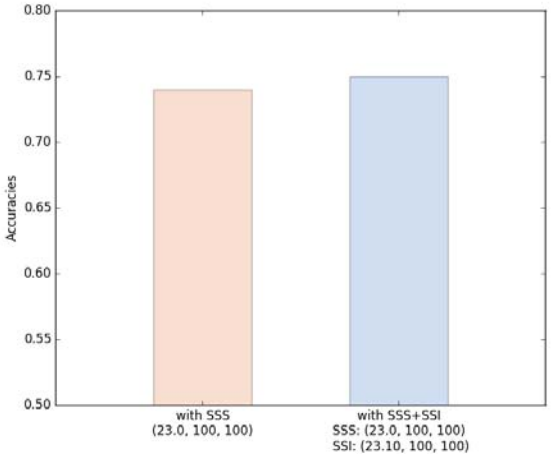


図 6: 識別性能の比較

パラメータセットは，Th を 23.0 から 23.18 まで 0.02 ずつ，post-margin, pre-margin を 100 から 500 [msec] まで変化させて，音源定位と音源識別をおこなうのに最適な組み合わせを探索し，定位：Th, post-margin, pre-margin $= 23.0, 100, 100$，識別：Th, post-margin, pre-margin $= 23.10, 100, 100$ を得た.

図 6 はで左から求めたパラメータセットのうち定位用のパラメータセットで学習した学習器を用いた識別適合率，定位と識別を異なるパラメータセットで同時最適化したときの識別適合率をそれぞれ表す．図 6 を見ると分かる通り，同時最適化した識別適合率は，定位用のパラメータセットを用いた識別適合率よりわずかながら優れた識別性能を示すことがわかる．したがって，本稿で提案する定位と識別の同時最適化機構が機能することが示唆された．一方で，思うよりも識別性能が延びていないことから，単純に定位された区間のフレームを全て用いた場合，含まれる雑音成分によって，識別性能の向上が阻害されていることが予想される．次節でより効率的な音イベント識別について詳細に述べる.

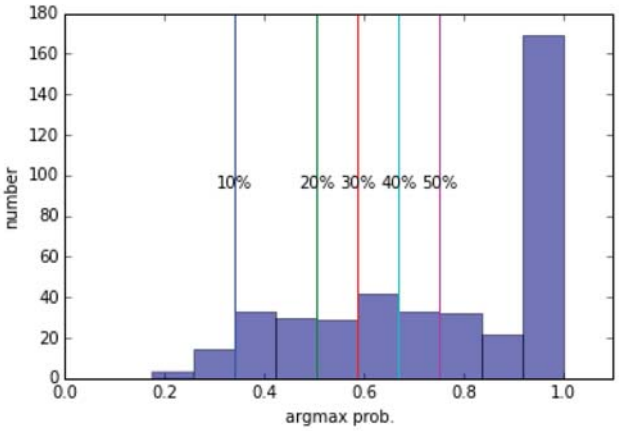


図 7: 確信度のヒストグラム

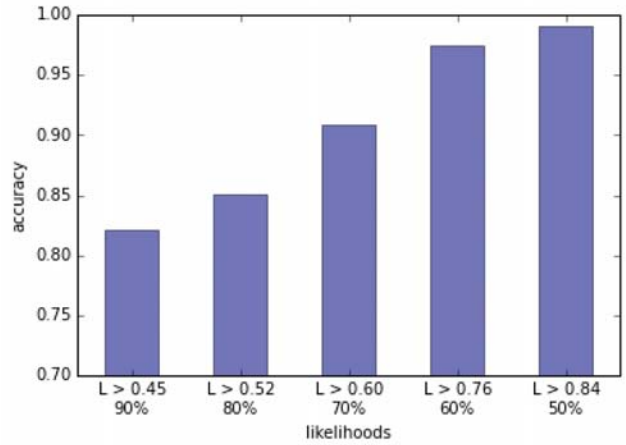


図 8: 確信度に基づいたフレームの取捨選択を行ったときの正答率の推移

5.3 確信度を用いた最良識別区間の選択と識別

次に上述の検討から得られた 2 つのパラメータセットを用いて学習した CNN と定位・区間検出した音イベントのデータセットを用いて，CNN のソフトマックス層の中で選択されたノードの値を確信度とみなしたとき，確信度の高さを閾値として最良の識別区間を選択できるのかを検証した．図 7 にデータセットを CNN に入力することで得られた確信度のヒストグラムを示す．図 7 を見るとわかるように，50%以上のデータの確信度が 0.80 を超える値を示しており，これらの区間のみを用いることで，よりよい識別性能を得ることができると考えられる．本実験では，選択区間が全体のデータセットの 90%, 80%, 70%, 60%, 50%となるように確信度の閾値（順番に 0.45, 0.52, 0.60, 0.76, 0.84）を設定し，それぞれの閾値で CNN の識別性能がどのように変化するのかを調べた．図 8 に，確信度の閾値と得られた識別性能の推移を示す．図 8 を見るとわかるように，確信度の閾値を上げていくことで，識別性能も向上していくことが示せた．したがって，本研究の仮説が検証できたものと考えられる．なお、識別性能が最も高くなるのは，確信度の閾値が 0.84 のときだが，この際，50%以上のフレームが信頼できないフレームとして棄却されるため、区間が短い音イベントなどの識別が困難になる可能性がある．確信度の閾値は，どの程度の精度の音源定位・区間検出を行いたいのかを考慮に入れつつ，調整する必要がある．

6 おわりに

本稿では，UAV 搭載マイクアレイを用いた音イベント検出・識別のためのフレームワークを設計・実装した．UAV に搭載されたマイクアレイはローター音や風切り音が混入するため，常に高雑音環境下での音源の検出と識別を行う必要がある．このような環境下でも，高精度な音イベント検出と識別を同時におこなうため，本稿では，音源分離と音源識別を組み合わせた識別手法と，音源定位から識別までを組み合わせた統一的なフレームワークを提案した．識別手法としては，雑音ロバストな定位手法である GHDSS-AS 法と CNN を組み合わせ，統合フレームワークにおいては，区間検出のパラメータとしてイベント検出と識別時に異なる値を用い，並列に処理する機構を導入した．UAV 屋外飛行実験で収集した実録音のデータを用いて実験した結果，仮説どおり，定位と識別には異なる閾値設定が必要であること，これらの異なる閾値設定を用いて，並列に定位と識別用の区間検出を行うことで，識別適合率を向上することができることを示した．また，同時に学習した CNN のソフトマックス層の出力から得られる確信度を用いることで，音イベントの特徴を表すフレームを取捨選択することができ，識別性能の向上に寄与できることを示した．

今後の課題としては様々な環境音下で学習データを作成し，雑音ロバスト性の向上を図ること，オンタイム処理を行うシステムに組み込むことなどが挙げられる．

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うけた．

参考文献

[1] H. Nakajima *et al.*, Blind Source Separation with Prameter-Free Adaptive Step-Size Method for Robot Audition, *IEEE Trans. ASLP*, 18(6), pp. 1476-1484.

[2] 上村 他，クアドロコプタ搭載マイクロホンアレイを用いた音源分離と深層学習による音源識別，第 33 回日本ロボット学会学術講演会, 2015.

[3] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai. Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter. In *IEEE/RSJ IROS*, pp. 3288–3293, 2012.

[4] 大畑他，相関行列スケーリングを用いた iGSVD-MUSIC 法による屋外環境音源探索の向上第 32 回日本ロボット学会学術講演会, 2014

[5] 中村 他，Latent Dirichlet Allocation と Nested Pitman-Yor Process に基づく雑音に頑健な音響イベント同定の検討，第 31 回日本ロボット学会学術講演会, 2013.

[6] S. Lawrence *et al.*, (1997). Face recognition: A convolutional neural-network approach. Neural Networks, IEEE Transactions on, 8(1), 98-113.

[7] 上村 他，クアドロコプタ搭載マイクロホンアレイを用いた深層学習による音声識別，第 15 回計測自動制御学会システムインテグレーション部門講演会, 2014.

[8] K. Nakadai *et al.* Design and Implementation of Robot Audition System “HARK”, *Advanced Robotics*, vol.24, pp.739-761 (2010).