様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 １ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１１８８０

研究課題名（和文）空間アテンション機構に基づく新しい音響シーン識別手法の確立

研究課題名（英文）Acoustic scene classification based on spatial attention mechanism

研究代表者
山田　武志（YAMADA, Takeshi）

筑波大学・システム情報系・教授

研究者番号：２０３１２８２９

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では、ビームフォーマを前処理として用いる音響シーン識別手法の性能を改善するために、音響シーンに存在する複数の音の中から注目すべき音（識別に有用な音）に自動的に焦点を当てる空間アテンション機構という新しいアイデアを導入した。その実現のために、複数の空間フィルタ出力への自動重み付けに基づく識別手法、及びその拡張である空間フィルタの自動推定に基づく識別手法を提案し、実験によりその有効性を示した。

研究成果の学術的意義や社会的意義
本研究成果の学術的独自性と創造性は、空間アテンション機構という新しいアイデアを実現した点にある。これにより、目的音方向などの事前情報を必要とせず、注目すべき音がどの音なのかを自動的に見つけると共に、それを強調するための空間フィルタを自動推定することが可能となった。これは信号処理技術と識別技術の有機的な統合によって成し得たものであり、音響シーン識別のみならず、雑音下音声認識などの他の様々なタスクへの展開が期待できる。

研究成果の概要（英文）：In order to improve the performance of acoustic scene classification that uses a beamformer as preprocessing, this study introduced a new idea of a spatial attention mechanism that automatically focuses on the sound of interest (useful for classification) among multiple sounds present in the acoustic scene. To realize this idea, we proposed a classification method based on automatic weighting of multiple spatial filter outputs and, as its extension, a classification method based on automatic estimation of spatial filters, and demonstrated their effectiveness through experiments.

研究分野：音声・音響情報処理

キーワード：音響シーン識別　空間アテンション機構　ビームフォーマ　空間フィルタ　ニューラルネットワーク　損失関数

１．研究開始当初の背景

数秒、または数十秒程度の音響信号から、それが録音された場所や状況を識別するタスクを音響シーン識別という。自動運転車の周囲状況把握や高齢者の見守り、動画の自動タグ付け、ライフログの収集など、様々な応用が期待できることから、近年、音響シーン識別の研究が世界的に活発化している。中でも複数のマイクで録音したマルチチャネル音響信号を入力とし、特定の方向から到来する音を強調するビームフォーマを前処理として適用する手法に注目が集まっている。これは事前に注目すべき音（識別に有用な音）を取り出すことができれば、識別精度向上が期待できるからである。

一般にビームフォーマの空間フィルタ（指向特性）を形成するためには、目的音方向などの事前情報が必要となる。しかし、個々の音響シーンにおいて注目すべき音がそもそもどの音であり、またその音がどの方向にあるのかは自明ではないため、前処理としてビームフォーマを適用することは本質的に難しい。例えば人の話し声は、レストランにおいては識別の重要な手がかりとなり得るが、図書館においてはイレギュラーな音なのでむしろ注目すべきではないであろう。以上から、注目すべき音を自動的に判断して強調するような自律型ビームフォーマの確立が急務である。

２．研究の目的

本研究の目的は、ビームフォーマと識別器の融合による新しい音響シーン識別手法を確立することであり、そのためにアテンション機構を有するニューラルネットワークに着目する。これは、入力中のより重要な部分に自動的に焦点を当てる（重み付けする）ことを可能とし、機械翻訳や音声認識の性能向上に大きく寄与することが知られている。本研究ではこれを応用し、音響シーンに存在する複数の音の中から注目すべき音に自動的に焦点を当てる機能（本研究ではこれを空間アテンション機構と呼ぶ）をニューラルネットワークにより実現する。これにより、目的音方向などの事前情報を必要とせず、識別に適した指向特性を入力信号から自動的に形成することが可能となる。空間アテンション機構と識別器は同じ損失関数のもとで End-to-End に学習を行うことが可能であり、またどの音に焦点を当てたのかを容易に分析できるようになる。

３．研究の方法

本研究では次の 2 つの手法を提案する。

(1) 複数の空間フィルタ出力への自動重み付けに基づく識別手法

提案手法(1)では、異なる空間フィルタ（指向特性）を複数個用意し、各空間フィルタ出力の重要度をアテンション機構によって推定する。これは、様々な方向から到来する音の中から注目すべき音を自動的に判断することに相当する。

図 1 に提案手法(1)の概要を示す。

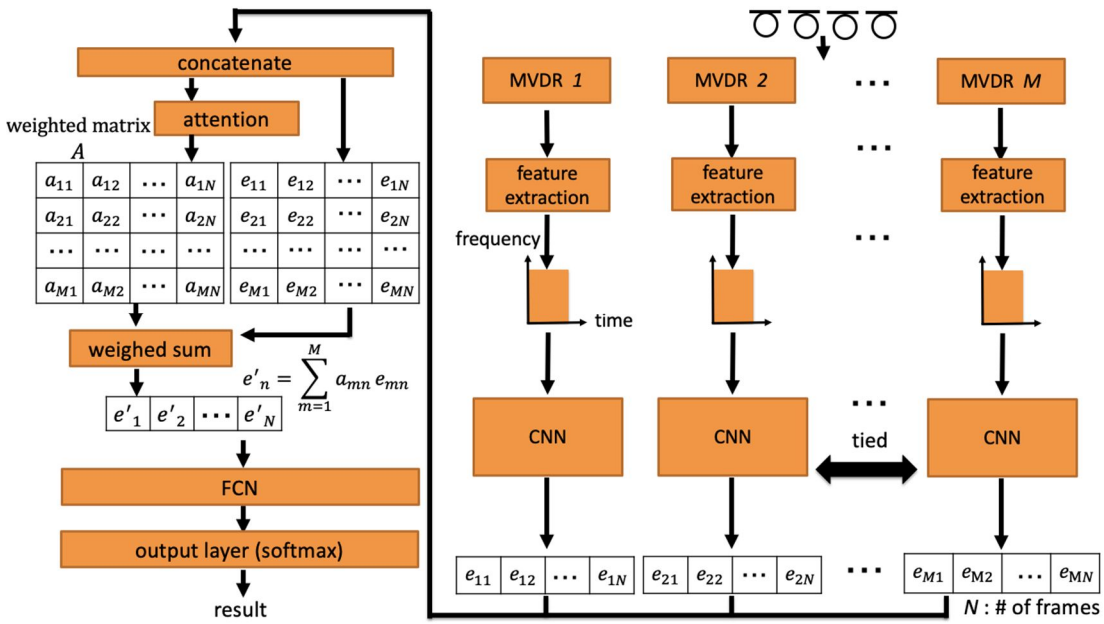

図 1：提案手法(1)の概要

まず、M個の MVDR（minimum variance distortionless response）ビームフォーマを用いて、そ

れぞれに割り当てた目的方向からの強調音を得る。次に各強調音から音響特徴量（対数メルスペクトログラム）を求めて CNN（convolutional neural network）に入力する。その後、アテンション機構を用いて各 CNN 出力の重要度（重み）を推定し、各 CNN 出力の重み付き和を求める。最後にこれを全結合層に入力して識別結果を得る。ここで、図中のNは時間フレーム数である。また、e_{mn}はm番目の MVDR ビームフォーマにおけるn番目の時間フレームの CNN 出力、a_{mn}はそれに対する重みである。提案手法(1)の学習は、識別タスクの学習によく用いられるクロスエントロピー損失を用いて行う。

(2) 空間フィルタの自動推定に基づく識別手法

提案手法(2)では、空間フィルタそのものを推定し、その空間フィルタの出力を用いて識別する。これは、音響シーンに存在する複数の音にどのように焦点を当てるか（どのような指向特性を形成するか）を自動推定するという意味で空間アテンション機構に他ならず、提案手法(1)の拡張とみなすことができる。

図 2 に提案手法(2)の概要を示す。

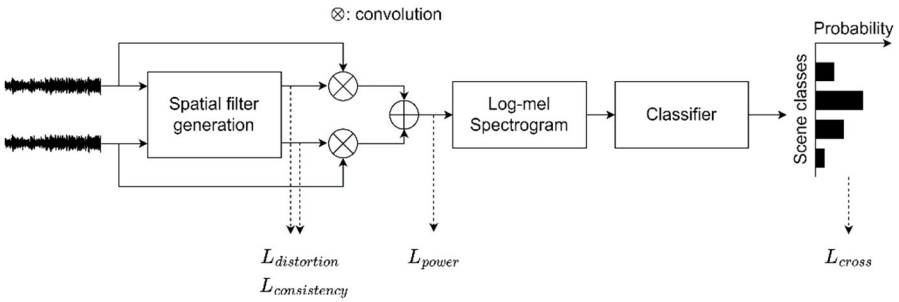

図 2：提案手法(2)の概要

まず、空間フィルタ生成器を用いて注目すべき音を強調するための空間フィルタを推定する。次に、この空間フィルタを入力信号に適用して強調信号を得る。その後、この強調信号から音響特徴量（対数メルスペクトログラム）を求め、識別器に入力して識別結果を得る。ここで、空間フィルタ生成器は LSTM（long short-term memory）、識別器は CNN で構成されており、全体を End-to-End に学習する。学習の際の損失関数は次の通りである。

$$L = \alpha_1 L_{directivity} + \alpha_2 L_{power} + \alpha_3 L_{consistency} + \alpha_4 L_{cross}$$

ここで、α_1～α_4は重み係数、L_{cross}はクロスエントロピー損失である。また、$L_{directivity}$、L_{power}、$L_{consistency}$は指向特性の適合性を評価するために新たに考案した損失である。

4．研究成果

提案手法(1)と(2)の有効性評価の結果を述べる。

(1) 複数の空間フィルタ出力への自動重み付けに基づく識別手法

提案手法(1)の有効性を評価するために、DCASE2018 タスク 5 のデータセットをもとに目的シーンと妨害シーンを混合した音響データを作成し、識別実験を行った。マイクアレイと音響シーンの配置を図 3 に示す。図中の緑の円は 4 つのマイクアレイを表し、各マイクアレイは 4 個のマイクで構成される。また、青い円は目的シーンが生起する場所、赤い円は妨害シーンが生起する場所をそれぞれ表している。本実験では、目的シーンを料理、皿洗い、食事、仕事、会話の 5 種類、妨害シーンをテレビと掃除の 2 種類とした。同じマイクアレイを用いて録音された目的シーンと妨害シーンの音響データをランダムに一つずつ選択して混合した。混合後の音響データは計 32,444 個である。また、図 1 におけるMは 3、各 MVDR ビームフォーマの目的方向は 30°、90°、150°とし、識別器の学習には混合した音響データのみを用いた。

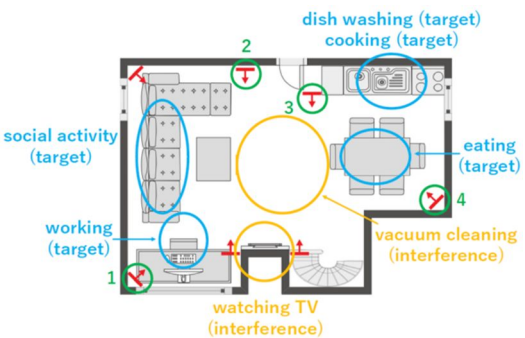

図 3：マイクアレイと音響シーンの配置

識別実験の結果（F 値）を以下に示す。

- single channel　64.61%
- proposed（correct weight）　83.46%
- proposed　76.18%

ここで、single channel は 4 チャネルマイクアレイの 1 チャネルのみを用いて識別した場合であり、混合データをそのまま識別することに相当する。また、proposed（correct weight）は、提案手法(1)において正しい重み行列を与えた場合であり、提案手法(1)により得られる上限性能に相当する。まず、single channel の F 値は 64.61%であった。これに対して、proposed（correct weight）の F 値は 83.46%であり、目的シーンの方向の強調音を重視することにより、識別精度を大きく改善したことが確認できる。proposed の F 値は 76.18%であり、proposed（correct weight）には及ばないものの、single channel と比べると約 12%の改善が得られた。また、図 4 にアテンション機構により推定した重みの例を示す。これは、マイクアレイ 3 において食事とテレビの音が混合した場合に推定された重みを示している。横軸は時間フレームの番号、縦軸は MVDR ビームフォーマの番号と目的方向である。この図から、目的シーンは 30 ° 方向の領域にあると概ね正しく推定できていることが確認できる。一方、局所的に 90 ° や 150 ° の重みが大きくなっている部分があり、これが proposed（correct weight）と proposed の識別精度の差の原因であると考えられる。さらに M を 11 に増やした実験を行い、同様の結果が得られることを確認した。

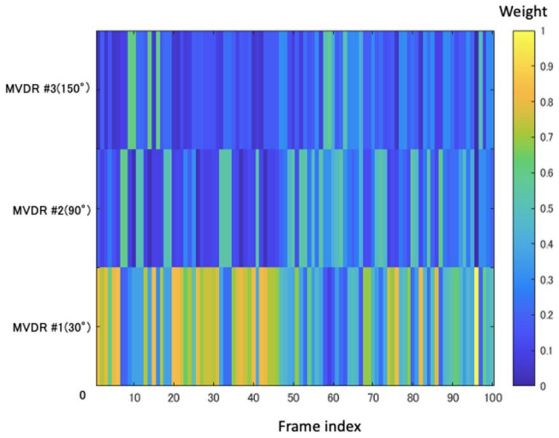

図 4：推定した重みの例

(2) 空間フィルタの自動推定に基づく識別手法

注目すべき音とその到来方向が未知という条件の下で、注目すべき音がどの音なのかを学習し、またその音が到来する方向の音を強調する空間フィルタを自動生成できるかを検証する。注目すべき音を強調しているかを容易に評価できるようにするため、騒音下で男性が話しているシーンと、騒音下で女性が話しているシーンの 2 つのシーンを識別する実験を行った。入力信号は、音声とピンクノイズが異なる方向から同時に到来する状況を想定して生成した。学習の際には混合音のみを用い、また話者の方向は未知とするので、ネットワークは音声が識別に有用であることを学習し、また音声を強調（ピンクノイズを抑制）する空間フィルタを自動生成することが求められる。

無響室環境と残響環境で推定された空間フィルタの指向特性の例をそれぞれ図 5 と図 6 に示す。

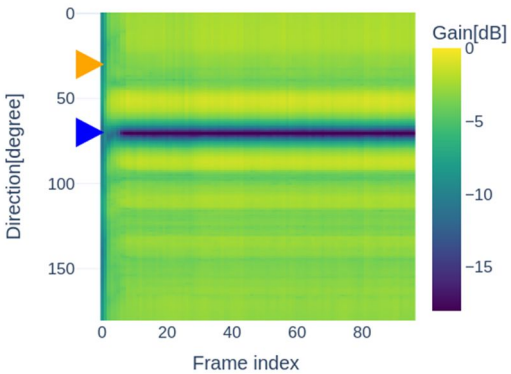

図 5：指向特性の例（無響室環境）

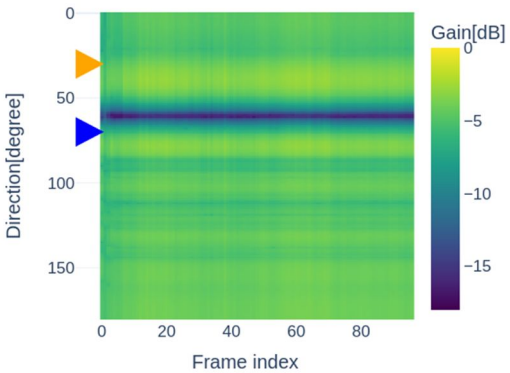

図 6：指向特性の例（残響環境）

ここで、縦軸は方向、横軸は時間フレーム番号であり、左端にある矢印は青がノイズ方向、橙が音声方向を表している。また、カラーマップは色が青くなるほど抑圧量が大きいことを表す（0 dB は抑圧なし）。これらの図から、ノイズ方向に明確な null（指向特性の死角）が形成されていることを確認できる。これは、ネットワークが注目すべき音が音声であることを学習し、その妨げになっているノイズを抑制するフィルタを自動生成したことを意味している。また、識別精度を以下に示す。

- classifier only　　76.2%（無響室環境）、65.4%（残響環境）
- proposed　　92.8%（無響室環境）、57.4%（残響環境）

無響室環境においては、識別器のみの手法と比較して提案手法(2)による大幅な精度向上を確認

できる。これは、提案手法(2)が注目すべき音を強調するフィルタを生成したことにより、後続の識別器における識別が容易になったからであると考えられる。一方、残響環境においては、識別器のみの手法に及ばないことが判明した。今後その原因を調査する必要がある。

(3) まとめ

本研究では、目的音方向などの事前情報を必要とせず、注目すべき音がどの音なのかを自動的に見つけると共に、それを強調するための空間フィルタを自動推定する手法を提案し、実験によりその有効性を示した。本研究成果の学術的独自性と創造性は、空間アテンション機構という新しいアイデアを実現した点にある。これにより、目的音方向などの事前情報を必要とせず、注目すべき音がどの音なのかを自動的に見つけると共に、それを強調するための空間フィルタを自動推定することが可能となった。これは信号処理技術と識別技術の有機的な統合によって成し得たものであり、音響シーン識別のみならず、雑音下音声認識などの他の様々なタスクへの展開が期待できる。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Kaneko Yuki、Yamada Takeshi、Makino Shoji	４．巻 25
２．論文標題 Monitoring of Domestic Activities Using Multiple Beamformers and Attention Mechanism	５．発行年 2021年
３．雑誌名 Journal of Signal Processing	６．最初と最後の頁 239～243
掲載論文のDOI（デジタルオブジェクト識別子） 10.2299/jsp.25.239	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Sota Ichikawa, Takeshi Yamada, Shoji Makino
２．発表標題 Neural beamformer with automatic detection of notable sounds for acoustic scene classification
３．学会等名 APSIPA ASC (Asia-Pacific Signal and Information Processing Association Annual Summit and Conference) 2022（国際学会）
４．発表年 2022年

１．発表者名 市川創大，山田武志，牧野昭二
２．発表標題 音響シーン識別のための注目すべき音を自動検出するニューラルビームフォーマの検討
３．学会等名 音学シンポジウム2022
４．発表年 2022年

１．発表者名 Kazuya Ouma, Takeshi Yamada, Shoji Makino
２．発表標題 Semi-supervised learning using weakly labeled data generated by GAN in sound event detection
３．学会等名 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2022 (NCSP'22)（国際学会）
４．発表年 2022年

１．発表者名
山田友紀，山田武志，牧野昭二

２．発表標題
Wave-U-Netと識別器のエンドツーエンド学習による音響シーン識別の検討

３．学会等名
日本音響学会春季研究発表会

４．発表年
2022年

１．発表者名
合馬一弥，山田武志，牧野昭二

２．発表標題
音響イベント検出におけるGANを用いた弱ラベルデータ生成による半教師あり学習

３．学会等名
日本音響学会秋季研究発表会

４．発表年
2021年

１．発表者名
Yuki Kaneko, Takeshi Yamada, Shoji Makino

２．発表標題
Monitoring of domestic activities using multiple beamformers and attention mechanism

３．学会等名
RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2021 (NCSP'21)（国際学会）

４．発表年
2021年

１．発表者名
陳軼夫，山田武志，牧野昭二

２．発表標題
音響イベント検出と位置推定における転移学習の効果の検証

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関