I-005

音響イベントを構成する音響イベントの情報量に関する考察
Study of the information entropy of Acoustic Events Comprising an Acoustic Event

山城 遼* 荒井 秀一*
Haruka Yamashiro Shuichi Arai

1 はじめに

人間が生活する環境は様々な音で溢れており, 環境や状況を理解するための多くの情報を音から得ている. この人間の音から環境を理解する能力をコンピュータでも実現しようとして, 古くから様々な研究が行われてきた [1]. そして, 近年では環境音分析のコンペティションである DCASE Challenge が開催されることによって, 音響シーン推定, 音響イベント検出, 音響タグ付けといった環境音分析も活発な研究が行われている [2][3].

2 従来手法

2.1 音響シーン分類

音響シーン分類は環境音分析技術の 1 つであり, 入力された音データを予め定義されている音響シーンに推定する技術である. DCASE Challenge が初めて開催された年から現在に至るまで, 音響シーン分類のタスクは毎年存在し, 多くの研究が行われている分野である.

2.2 Pretrained Audio Neural Networks(PANNs)

環境音分析の手法は様々なものが提案されており,RNN や CNN などのニューラルネットワークを用いた手法も多く存在する. そのニューラルネットワークを用いた手法の 1 つに,Pretrained Audio Neural Networks(PANNs)[4] がある. PANNs は CNN や ResNet で構成されたアーキテクチャを持ち, 音響イベント 527 クラスで構成された 5800 時間の大規模オーディオデータセットである AudioSet[5] で学習されたモデルである. PANNs は音響タグ付けや音響イベント検出に用いることができるだけでなく, 転移学習させることによって更に多くのタスクに利用することができる.

3 研究目的

近年の DCASE のタスクの音響シーン分類の学習と推論には,1 秒の音データが用いられる. 音響シーンは様々な音響イベントが複雑に組み合わさることで構成されており, 音響シーンに含まれる音響イベントの種類だけでなく, 音響イベントの鳴り方, 観測頻度も音響シーン分類を行う際の重要な手がかりになると考える. しかし, 近年の DCASE のタスクの音響シーン分類の学習と推論には 1 秒の音データが用いられている. よって,DCASE のタスクで行われている音響シーン分類は, これらの情報が含まれていないクラス分類する手がかりのない音データを用いて行われており, 出来るはずがなく, やるべきではないと考える.

音響シーン分類は, 人間が実際に音響シーンを推定する際と同じように, 一定以上の長さの音を聞き, 音響シーンを構成している音響イベントを明らかにした上で, 音響イベントの鳴り方や観測頻度といった特徴の類似性や相違性を解析して行うべきだと考え, 音響シーン分類を音響シーンを構成している音響イベントから行うことを研究目的としている.

4 提案

本稿では, 音響シーンを構成している音響イベントから音響シーンの解析を行うために, PANNs を音響イベント検出機として用い, 得られた出力から音響イベントベクトルを定義する. また, 作成した音響イベントベクトルのベクトル間距離に基づいた音響シーンの類似性の比較尺度を定義する.

4.1 音響イベントベクトル

音響イベント検出機として PANNs を用いると, フレーム毎に 527 クラスの音響イベントの推定確率を求めるため, 入力した音はフレーム数×音響イベント数の行列として出力される. この行列はフレーム方向に情報を持っているため, 音響イベントがどのフレームで観測されたか, 音響イベントがどの順番で鳴ったか, 音響イベントが何フレーム鳴り続けたか, などの時間にも意味を持つことになる. そのため,PANNs の出力行列をそのまま特徴行列として用いてしまうと, 観測した音響イベントの種類, 音響イベントの鳴り方に加えて, 観測した音響イベントがどのフレームにあるかといった情報も特徴になる. だが, 音響シーンは音楽とは異なり, 音響イベントが観測される順番や音響イベントが観測される時間は決まっていない. そのため, 音データのどのフレームで観測した音響イベントであっても, 同じ特徴を持った音響イベントを観測したのであれば, 音響イベントベクトルの値は同じになるようにしたいと考えた.

そこで, 時間域での情報を排除し, 音響イベントの持つ情報のみから特徴を定義するため, PANNs の出力する行列から, 音響イベント毎に出力した推定確率の最大値をその音響イベントの値として, 1 ×音響イベント数のベクトルを定義した. 図 1 に音データの入力から音響イベントベクトル生成までのフローチャートを示す.

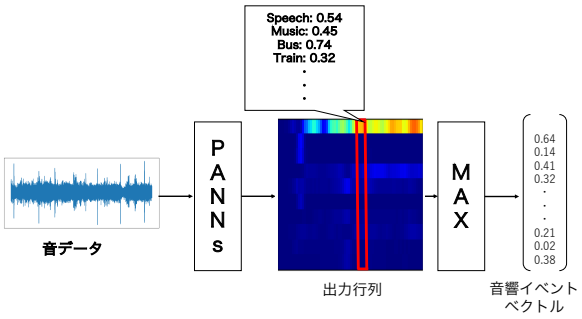

図 1 音響イベントベクトル生成のフローチャート

* 東京都市大学

4.2 音響イベントベクトルの類似尺度

音響イベントベクトルは, 音響シーンを構成している音響イベントの持つ特徴の値を並べて作成している. そのため, 音響イベントの持つ特徴を 1 つずつ比較することで, 音響シーンの類似度を比較できると考えた.

そこで, 音響イベントベクトルの類似度をユークリッド距離を用いて定義する. 比較を行う音響イベントベクトルを X=$[x_1,\ldots,x_M]^{\mathrm{T}}$ と Y=$[y_1,\ldots,y_M]^{\mathrm{T}}$ とすると, 式 1 によって計算出来る. 音響イベントベクトルが類似しているほど,0 に近づき, 逆に類似していなければ値は大きくなる.

$$\sin_{XY} = \|X - Y\|_2 = \left(\sum_{k=1}^{M}(x_k - y_k)^2\right)^{\frac{1}{2}} \tag{1}$$

5 実験と考察

5.1 データセット

音響イベントベクトルは, 一定以上の長さの音を聞き, 音響シーンを構成している音響イベントを明らかにした上で, 音響シーン解析を行うために定義した. よって, 音響イベントベクトルを作るために使用するデータセットは, ある程度の長さの音データであること, 1 つの音データが多くの音響イベントで構成されていることの 2 つの条件を満たす必要がある. だが, この 2 つの条件を満たすデータセットは公開されているものはない. そこで [6] に記載されている東京都 23 区にある 60 箇所の観光地の環境音を各場所で 1 時間録音し, 計 60 時間のデータセットを作成した. このデータセットにおいて, 観光地 1 つ 1 つがクラスとなり, 合計 60 クラスから構成される.

5.2 音響イベントベクトルに関する考察

本稿で定義した音響イベントベクトルは, 観測したデータに含まれる音響イベントは考慮しているが, 音響イベントの観測頻度や鳴り方を考慮できていない. 今後, 音響イベントベクトルが, 何の音がどれくらい鳴っているかを特徴として表すことができるようにするため, 観測頻度や鳴り方を考慮するための方法を検討する必要がある. この音響イベントの観測頻度や鳴り方を考慮していない点に加えて, 音響イベントの持つ音響シーンを分類するための手がかりとなる情報量を同一に扱っている点も問題であると考える.

PANNs は 527 クラスの音響イベントを検出することが可能であるが, 多くの音響シーンで観測される音響イベントと僅かなクラスでのみ観測される音響イベントがある. 僅かな音響シーンでのみ観測される音響イベントは, 多くの音響シーンで観測される音響イベントと比較して, 音響シーン分類をする際の分類クラスの候補を絞るための情報を多く持っている. そのため, 情報を多く持つ音響イベントを観測した際と観測しなかった際で, 音響イベントベクトルの特徴量に大きな差が出るように定義すれば, 音響シーン分類をする際に情報量の大きな音響イベントが与える影響を大きく出来ると考えた.

そこで,527 クラスの音響イベントがどの音響シーンで観測されており, 作成したデータセット 60 クラスの音響シーンを解析する際は, どの音響イベントがクラス分類先を絞る情報を多く持つかを調べるために, 図 2 のヒートマップを作成した.

図 2 の縦軸は音響シーンを表す 60 クラスが並べられており, 横軸は PANNs の出力する音響イベント 527 クラスが並べられている. また, 推定確率は, その音響シーンの 1 時間の中で観測された最大の値を用いており, 確率が 1 に近いほど赤に近づき, 確率が 0 に近いほど白に近づく.

図 2 より speech,music,vehicle, といった音響イベントは多くの音響シーンで観測されているため, クラス分類する際の手がかりとなる情報は少なく, 逆に chime,Violin fiddle といった音響イベントは僅かな音響シーンでのみ観測されているため, クラス分類する際の手がかりとなる情報を多く持っていると考える. 今後, これらの音響イベントの持つ情報量を定式化する方法と音響イベントベクトルに組み込む方法を検討する必要がある.

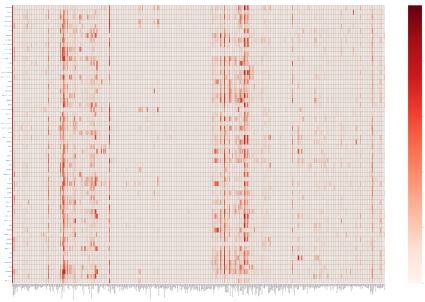

図 2 各音響シーンが出力する各音響イベントの最大推定確率

6 おわりに

本稿で, 音響シーンを構成している音響イベントを明らかにした上で, 音響イベントの鳴り方や観測頻度といった特徴の類似性や相違性を解析して音響シーン分類を行うため, PANNs を用いて音響イベントベクトルの定義と音響イベントベクトルを用いた音響シーンの類似度の評価尺度を定義した.

今後, 本稿で行った考察をもとに音響イベントの観測頻度や鳴り方, 音響イベントが持つ情報量を音響イベントベクトルに組み入れられるよう検討を進めていく.

参考文献

[1] 井本桂右, 川口洋平. 環境音分析・異常音検知の研究動向. 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review Vol.15 No.4 pp.268-280 2022 年 4 月

[2] 井本桂右. DCASE Challenge: 環境音分析・理解のための総合的コンペティション. 日本音響学会誌 79 巻 9 号 (2023), pp.470-476

[3] 井本桂右. 環境音分析の研究動向. 日本音響学会 75 巻 9 号 (2019), pp.512-518

[4] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang and Mark D. Plumbley. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880-2894, 2020.

[5] J. F. Gemmeke et al. Audio Set: An ontology and human-labeled dataset for audio events. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 776-780.

[6] 公益財団法人東京観光財団. 東京トラベルガイド, 2023.