環境音分析

井本桂右（同志社大学）
keisuke.imoto@ieee.org

1. 環境音分析とは

環境音分析とは，音声や楽音に限らないあらゆる種類の音(注1)を分析し，音の種類や発生時刻，発生場所を推定したり，音が発生した状況を説明する技術である．環境音分析技術を活用することで，補聴器の高度化やメディアコンテンツへの自動タグ付与，高齢者や乳幼児の見守りシステム，工場における機器の自動監視，自動運転など様々なサービスへの応用が期待できる．環境音分析の実現方法としては音声や画像などのメディア分析手法と類似した部分が多く，機械学習，とりわけニューラルネットワークによる手法を中心として研究が進められている．

2. 環境音分析で扱われる主なタスク

音響シーン分類は環境音分析の中で最も取組み事例が多いタスクである．図1に示すように，音響シーン分類は，あらかじめ決められたクラスの中から，入力された音を最もよく表す音響シーンを一つ推定する．ここで，音響シーンとは音が収録された場所や状況，周囲にいる人の行動を表す．具体的に，音響シーン分類は以下のように定式化される．

$$\hat{s} = \arg\max_{s} y_s = \arg\max_{s} f(\mathbf{X}, \boldsymbol{\theta}) \qquad (1)$$

ただし，s 及び y_s，f，$\mathbf{X}$，$\boldsymbol{\theta}$ はそれぞれ，音響シーンのクラスインデックス，音響シーン s の予測確率，モデル，音響特徴量，モデルパラメータを表す．

環境音分析の中でも音響シーン分類は簡易な問題設定と言える．そのため，他の環境音分析タスクと比較して技術が成熟しており，今日では実践的な課題を中心に研究が進められている．例えば，学習データと異なる収録機器でシステムを利用した際に，分類性能が低下する問題への取組みや，携帯型デバイス向けの軽量な機械学習モデルの検討である．

音響イベント検出は，入力された音に含まれる環境音の種類とその発生時刻を推定するタスクである．図2に音響イベント検出の概略を示す．具体的には，短区間の時間フレーム，音響イベントごとに音の発生の有無を推定する問題として以下のように定式化される．

$$\mathcal{L} = I[y_{t,c} \geq \phi] \in \{0, 1\}^{T \times C} \qquad (2)$$

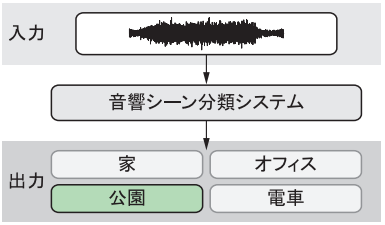

図1　音響シーン分類システム　　あらかじめ決められたクラスの中から，入力された環境音を最もよく表す音響シーンを一つ推定する．

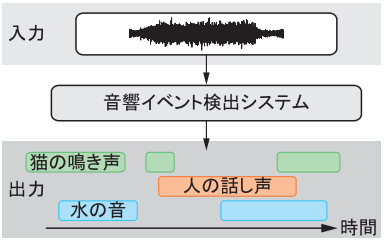

図2　音響イベント検出システム　　入力された環境音に含まれる音の種類とその発生時刻を推定する．

ただし，$I[\cdot]$，$y_{t,c}$，ϕ，T，C はそれぞれ，指示関数，音響イベント c の時間フレーム t における予測確率，検出しきい値，分析する音信号の時間フレーム数，音響イベントのクラス数を表す．検出アルゴリズムとしては，畳込みリカレントニューラルネットワークや Transformer，Conformer など，時系列型のニューラルネットワークを用いることで高い性能を示すことが報告されている．

機械学習に基づく音響イベント検出では，環境音の種類とその開始時刻，終了時刻をラベル情報として持つ学習データセットを事前に用意する必要がある．しかしながら，環境音に対して音の開始時刻，終了時刻を付与する作業は非常に手間が掛かる．その理由として，環境音は複数の音が同じ時刻に発生しやすいことや，音源が遠方に存在し，音の開始時刻や終了時刻を判断しづらい状況が発生しやすいことが挙げられる．音響イベント検出のデータセットにおける音の開始・終了時刻の付与作業の課題を解決する方法として，弱ラベルデータ（Weakly-labeled Data）を用いた音響イベント検出(1)の研究も取り組まれている．音響イベント検出における弱ラベルデータとは，音の開始時刻や終了時刻の情報を持たず，音信号に含まれる環境音の種類のみを情報として持つラベルデータを指す．弱ラベルデータを用いた音響イベント

（注1）　環境音の定義は様々あるが，本稿では「音声や楽音に限らないあらゆる種類の音」を環境音と称する．

本会ハンドブック「知識の森」
https://www.ieice-hbkb.org/portal/doc_index.html

図３　音響キャプション生成システム　入力された環境音を説明する文を生成する．

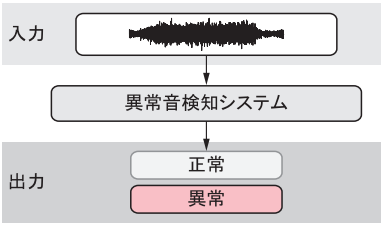

図４　異常音検知システム　入力された環境音から異常状態か否かを判断する．

検出では，モデル学習は弱ラベルデータを用いて行うが，システムの利用時には通常の音響イベント検出同様，環境音の種類とその発生時刻を推定する．そのため，通常の音響イベント検出と比較して，難易度が高い問題設定となっている．

音響イベント検出では「いつ」「何の音が発生したか」を分析するが，それに加えて「どこで音が発生したか」も同時に分析する Sound Event Localization and Detection（SELD）[2] という環境音分析タスクもある．環境音分析のコンペティションである Detection and Classification of Acoustics Scenes and Events（DCASE）Challenge では，2019 年から 2023 年にわたり，環境音の種類とその発生時刻，音の発生方向（水平角及び仰角）を同時に推定するタスクが開催され，人気を博している．

音響キャプション生成は，「何の音が」「なぜ」「どのように生成されたか」を説明するタスクである[注2]．図３に示すように，音響キャプション生成では，環境音の特徴量系列 $\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$ を入力として，音を説明する単語系列 $\mathbf{W}=(w_1, w_2, \cdots, w_N)$ を生成する．これを系列の変換問題とみなすと，音響キャプション生成は機械翻訳や声質変換と同様に，Encoder-Decoder 型ニューラルネットワークなどの系列変換モデルで実現することができる．このときの音響キャプション生成システムの定式化を以下に示す．

$$\boldsymbol{v}=\mathcal{E}(\mathbf{X}) \tag{3}$$

$$p(w_n|\mathbf{X}, \mathbf{w}_{n-1})=\mathcal{D}(\boldsymbol{v}, \mathbf{w}_{n-1}) \tag{4}$$

$$\overline{w}_n=\arg\max_{w_n} p(w_n|\mathbf{X}, \mathbf{w}_{n-1}) \tag{5}$$

ただし，$\boldsymbol{v}$ はエンコーダによる音響特徴量の埋込みを，$\mathcal{E}$ 及び $\mathcal{D}$ はそれぞれ，エンコーダとデコーダを表す．また，$\mathbf{w}_{n-1}=(w_1, w_2, \cdots, w_{n-1})$ とする．

異常音検知も研究事例が多い環境音分析タスクである．図４に示すように，異常音検知は，入力された環境音に基づき，何らかの異常が発生しているかどうかを判断する．例えば，セキュリティ用途ならば環境音に含まれる銃声や悲鳴を検出し，製品検査であれば製品の稼動音から不良品を検知する．ただし，検知される異常は必ずしも，正常時の音に対して異常な音が加わった状況とは限らないことに注意が必要である．例えば，正常時であれば含まれるはずの成分が入力音に含まれていないことを手掛かりに異常か否かを判断することも，異常音検知に含まれる．これは，我々が検知したい異常状態と「異常音」の定義が常に一致しているとは限らないためであり，異常音検知に取り組む場合は注意すべき事項である．したがって，異常音検知と呼ぶよりも「環境音からの異常検知」と呼ぶ方が，異常音検知の定義を正確に表していると言える．

異常音検知の実現方法としては，音響特徴量 $\mathbf{X}$ に基づいて異常スコアを計算し，スコアがしきい値よりも大きければ異常と判定し，そうでなければ正常と判定するシステムを構築する方法が一般的である．つまり，パラメータ $\boldsymbol{\theta}$ を有する異常スコア計算器 $\mathcal{A}_{\boldsymbol{\theta}}$ を用いて，異常音検知は以下のように定式化される．

$$\text{Decision}=\begin{cases}\text{Anomaly} & \text{if } \mathcal{A}_{\boldsymbol{\theta}}(\mathbf{X})>\phi \\ \text{Normal} & \text{otherwise}\end{cases} \tag{6}$$

ただし，ϕ は異常スコアに対する判定しきい値である．異常音検知では，ニューラルネットワークに基づく教師なし学習手法が盛んに研究されている．これは，正常音と比べて異常音が発生することは極めてまれであり，多量のデータを入手することが現実的ではないためである．具体的なアルゴリズムとしては，AutoEncoder の再構成誤差に基づく手法が広く用いられている．

その他の環境音分析タスクとして，分析対象の音に含まれる全ての音響イベントを予測する音響タグ付け[3] や，複数の環境音・音声分析タスクに有効な汎用音響信号表現学習[4,5] などがある．また，関連する話題として，混ざり合った複数の環境音を個々の音源に分ける環境音分離[6] や，音の説明文・オノマトペを入力として環境音を生成する環境音合成[7,8] も急速に広まってきている．

3. 環境音分析を始めるために

環境音分析の学術団体である DCASE Community の活動を知ることは，環境音の研究開発を始める上で非常に役に立つ．というのも，DCASE は，環境音分析に関するワークショップやコンペティションの開催を通して得られた，データセット，ベースラインシステム，技術レポート，メタ分析結果，評価指標を無償公開している．これらの研究資源は DCASE の Web ページ[注3] からアクセス可能である．また，環境音分析の解説として文献（９），（10）がある．併せて参照してほしい．

（注２）　データセットによっては，音の説明文として「いつ」「どこで音が発生したか」が付与されている場合もある．このようなデータセットを機械学習に基づく音響キャプション生成に利用した場合，より詳細で多様な説明文の生成が可能となる．

（注３）　https://dcase.community/ (2023 年 4 月 17 日現在)

文　　献

（1） A. Kumar and B. Raj, “Audio event detection using weakly labeled data,” Proc. ACM Int. Conf. Multimedia, pp. 1038-1047, 2016.

（2） S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” IEEE J. Sel. Top. Signal Process., vol. 13, no. 1, pp. 34-48, 2018.

（3） E. Fonseca, M. Plakal, F. Font, D.P.W. Ellis, X. Favory, J. Pons, and X. Serra, “General-purpose tagging of freesound audio with audioset labels : Task description, dataset, and baseline,” Proc. Detection and Classification of Acoustic Scenes and Events, pp. 69-73, 2018.

（4） A. Saeed, D. Grangier, and N. Zeghidour, “Contrastive learning of general-purpose audio representations,” Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 3875-3879, 2021.

（5） D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, “BYOL for audio : Self-supervised learning for general-purpose audio representation,” Proc. Int. Joint Conf. Neural Networks, pp. 1-8, 2021.

（6） I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J.L. Roux, and J.R. Hershey, “Universal sound separation,” Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 175-179, 2019.

（7） F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, “AudioGen : Textually guided audio generation,” Proc. Int. Conf. Learning Representations, 2023.

（8） Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, “Onoma-to-wave : Environmental sound synthesis from onomatopoeic words,” APSIPA Trans. Signal and Information Processing, vol. 11, no. 1, e13, 2022.

（9） 井本桂右，川口洋平，“環境音分析・異常音検知の研究動向，”信学 Fundam. Review, vol. 15, no. 4, pp. 268-280, 2022.

（10） 井本桂右，“ドメイン知識を利用した環境音分析，”信学誌 , vol. 105, no. 12, pp. 1434-1440, 2022.

（2023 年 4 月 15 日受付）