自立生活支援のための音響イベント検出の連合学習

竹本志恩

2025 年 7 月 24 日

概要

少子高齢化に伴う医療・介護負担の増大に対し, 情報技術を用いて高齢者の自立生活を支援する AAL (Ambient Assisted Living) が注目されている．従来の AAL ではカメラやウェアラブル端末が主に用いられるが, プライバシー侵害や装着負担といった課題があった．本研究ではこれらの課題を解決するため, 安価かつプライバシー受容性の高い音響センサを用いた音響イベント検出 (Sound Event Detection, SED) に着目する．SED により生活音から利用者の行動を認識し, 異常検知や健康状態の把握を目指す．しかし, 機械学習モデルの学習にはプライバシー性の高い個人宅の音響データを要するため, データをサーバに集約する従来の中央集権的な手法は依然としてプライバシー上の懸念が残る．この課題に対し, 本研究ではデータを各家庭内に留めたまま分散的に学習を行う連合学習 (Federated Learning, FL) の適用を提案する．本稿では, 自立生活支援という文脈において, SED モデルを連合学習で構築する際の手法を検討し, 中央集権的手法に対する精度や, 最適なアルゴリズムについての問いを立て, その実験計画と評価方法を概説する．

1　緒言

1.1　背景: AAL と音響イベント検出

少子高齢化の進行により, 日本の医療および介護の負担は増大の一途をたどっている．この社会課題に対する一つの解決策として, Ambient Assisted Living (AAL) の研究が活発化している．AAL は, 情報通信技術を活用して高齢者などの自立した生活を支援し, 在宅介護における問題解決を目指すアプローチである [1, 2].

従来, AAL の実現にはカメラやウェアラブルデバイスが多く用いられてきた．しかし, カメラは設置コストが高いだけでなく, 常に監視されることによるプライバシー受容性に大きな課題を抱えている．また, ウェアラブルデバイスは充電の手間や装着し忘れ, 利用者への侵襲性（身体的・精神的負担）が問題となる．

本研究ではこれらの課題を解決する手段として, **音響イベント検出 (Sound Event Detection, SED)** に着目する．SED は, マイクなどの安価なセンサから得られる音響データに基づき, どのような事象がいつ発生したかを把握する技術である．カメラと比較してプライバシー侵害のリスクが低く, 音特有の異常兆候（咳, 叫び声, 転倒音など）を検出できる利点がある．機械学習を用いることで, 多様な環境や対象者に柔軟に対応可能であり, 「いつ, どのような行動があったか」という情報を高い説明性をもって提供できるモデルの構築が期待される．

1.2　音響イベント検出 (SED) の概要

音響イベント検出 (SED) は, 与えられた音響信号から音響イベントを検出するタスクである．具体的には, イベントのクラス（例: 「犬の鳴き声」）だけでなく, そのイベントがいつ始まり, いつ終わったかという時間情報も同時に予測する．

これは, 類似タスクである**音響シーン分類 (Acoustic Scene Classification, ASC)** とは異なる．ASC は, ある一定時間の音声クリップ全体に対して単一のラベル（例: 「公園」）を付与するのに対し, SED は音声中に発生する複数のイベントの持続時間を考慮する点でより詳細な情報を扱う（図 1 参照）．

2　研究課題と本研究の問い

AAL における SED の活用は有望であるが, その実現にはプライバシーに関する課題が残る．AAL は個人の生活空間という極めてプライバシー性の高いデータを扱うため, 音響データであってもその取り扱いには細心の注意が必要となる．従来の中央集権的な機械学習では, 各家庭から収集した音響データをサーバに集約してモデ

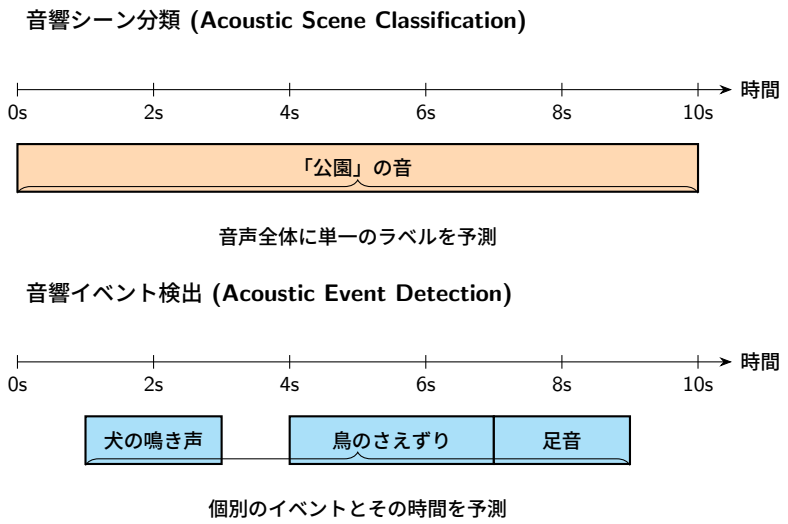

図 1　音響シーン分類と音響イベント検出の比較

ルを学習させる必要があり, プライバシー漏洩のリスクが懸念される.

本研究では, このプライバシー問題を解決するため, **連合学習 (Federated Learning, FL)** の導入を提案する. FL は, データをサーバに集約することなく, 各エッジデバイス（この場合は各家庭に設置された音響センサ）上でモデルを学習し, モデルの更新情報（重みパラメータなど）のみをサーバで集約・統合する分散学習手法である. これにより, プライバシーの根幹である生データを各家庭の外に出すことなく, 高精度なモデルの構築が期待できる. エッジデバイスでの処理に関する先行研究 [3] も存在するが, 本研究ではより柔軟な学習が可能な FL に着目する.

また,SED は DCASE (Detection and Classification of Acoustic Scenes and Events) コンペティションの一部として取り組まれており, 近年の研究では半教師あり学習や事前学習済みモデルの活用が進んでいる [4, 5, 6]. それらの研究を踏まえ, 出来るだけ高精度かつエッジでの動作を見据えた SED モデルを構築し, 連合学習を適用することを目指す.

以上の背景から, 本研究では以下の問いを立てる.

1. 中央集権的な学習手法と比較して, 連合学習は SED タスクにおいてどの程度の精度を維持できるか？
2. 複数の連合学習アルゴリズムのうち, 家庭内の音響データという不均一な（Non-IID）データ環境に対して, どのアルゴリズムが最も適切か？

将来的には, この連合学習によって構築された SED モデルを手がかりとした異常検知システムの実現を展望する.

3　実験計画

本研究の問いに答えるため, 以下の計画で実験を進める.

- **7 月: 準備期間**
  - 評価指標（F1 スコア, PSDS など）の意味を詳細に理解.
  - 具体的な目標精度を設定.
- **7-8 月: モデルアーキテクチャと学習戦略の検討**
  - 適切なモデルアーキテクチャの比較・検討 [7, 8] を行う. 事前学習済みモデル, CNN, RNN の最適な構成を模索する.
  - 適切な学習戦略を決定する. データの前処理・後処理, アンサンブル学習の有無などを検討し, 特に半教師あり学習の手法として Mean-Teacher または FixMatch の導入を検討する (DCASE 2024 を参考).
- **9 月: ベースラインモデルの性能評価**
  - データの前処理, 後処理, およびモデルの各構成要素が精度に与える影響を詳細に調査する.
- **9-10 月: 連合学習モデルの実装と比較評価**
  - 複数の連合学習アルゴリズムを実装し, その精度を中央集権的手法と比較する. ベースライ

ンとして FedAVG [9] を用い, 比較対象として FedProx [10] や SCAFFOLD [11] などを実装し評価する.
  - 各 FL アルゴリズムのハイパーパラメータ調整を行う.
- **11 月: 考察と論文執筆**
  - 実験結果を分析・考察し, 論文としてまとめる.

4 評価方法

実験における評価は以下の枠組みで行う.

- **モデル共通の前提**
  - **基本モデル**: DCASE 2024 のベースラインモデルを基礎とする.
  - **データセット**: DESED, MAESTRO などの公開データセットを使用する.
- **モデルの比較対象**
  - 中央集権的に学習させたベースラインモデルや, 関連研究における SOTA (State-of-the-Art) モデルを比較対象とする.
- **評価指標**
  - DCASE 2024 で用いられる Supplementary metrics を参照し, イベントベースの各種 **F1 スコア**と, **PSDS (Polyphonic Sound Detection Score)** 1 および 2 を主たる評価指標として使用する.
- **精度の基準**
  - 連合学習を適用した際の精度を, 中央集権的な手法で学習させた場合の精度と比較する. 理想的には同等の精度を目指すが, プライバシー保護という利点を考慮し, 従来手法からわずかに劣る程度の精度を目標とする.

5 現在の進捗

本稿執筆時点での進捗は以下の通りである.

- **各種サーベイの実施**: 連合学習, 音響イベント検出, 半教師あり学習に関する技術調査, および自立生活支援という研究目的の理解を進めた. また, DCASE Task4 (2018-2024) の大まかな内容を把握した.
- **実験評価計画タスクの進捗**
  - 研究計画書: ほぼ完了.
  - 各評価指標, 目標精度: 指標は把握済み.
  - モデルアーキテクチャ: 使用するモデルは概ね把握済み.
  - 学習戦略: 半教師あり学習手法は把握済み. 全体の戦略は関連論文を参考に検討中.
  - 精度影響の調査: 未着手.
  - 連合学習: FedAVG と FedProx は追試済み. 他のアルゴリズムは未調査.

References

[1] S. Blackman et al. “Ambient assisted living technologies for aging well: a scoping review”. In: *Journal of Intelligent Systems* 25.1 (2016), pp. 55–69.

[2] R. Stodczyk and F.-H. Uhp. “Ambient assisted living an overview of current applications, end-users and acceptance”. In: *Biomedical Journal of Scientific & Technical Research* 30.3 (2020), pp. 23374–23384.

[3] R. M. Alsina-Pagès et al. “homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring”. In: *Sensors* 17.4 (2017), p. 854.

[4] S. Cornell et al. “DCASE 2024 task 4: Sound event detection with heterogeneous data and missing labels”. In: *arXiv preprint arXiv:2406.08056* (2024).

[5] H. Yue et al. “Local and global features fusion for sound event detection with heterogeneous training dataset and potentially missing labels”. In: *Detection and Classification of Acoustic Scenes and Events 2024* (2024).

[6] S. W. Son et al. “Sound event detection based on auxiliary decoder and maximum probability aggregation for DCASE challenge 2024 task 4”. In: *arXiv preprint arXiv:2406.12721* (2024).

[7] Y. Li et al. “A hybrid system of sound event detection transformer and frame-wise model for dcase 2022 task 4”. In: *arXiv preprint arXiv:2210.09529* (2022).

[8] F. Schmid et al. “Multi-iteration multi-stage fine-tuning of transformers for sound event detection with heterogeneous datasets”. In: *arXiv preprint arXiv:2407.12997* (2024).

[9] B. McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[10] T. Li et al. “Federated optimization in heterogeneous networks”. In: *Proceedings of Machine learning and systems* 2 (2020), pp. 429–450.

[11] S. P. Karimireddy et al. “Scaffold: Stochastic controlled averaging for federated learning”. In: *International conference on machine learning*. PMLR. 2020, pp. 5132–5143.