アノテータごとのばらつきを考慮した音響イベント検出

古賀 直樹[1,2]，井本 圭佑[1,2]，坂東 宜昭[2]

[1]同志社大学，[2]産業技術総合研究所

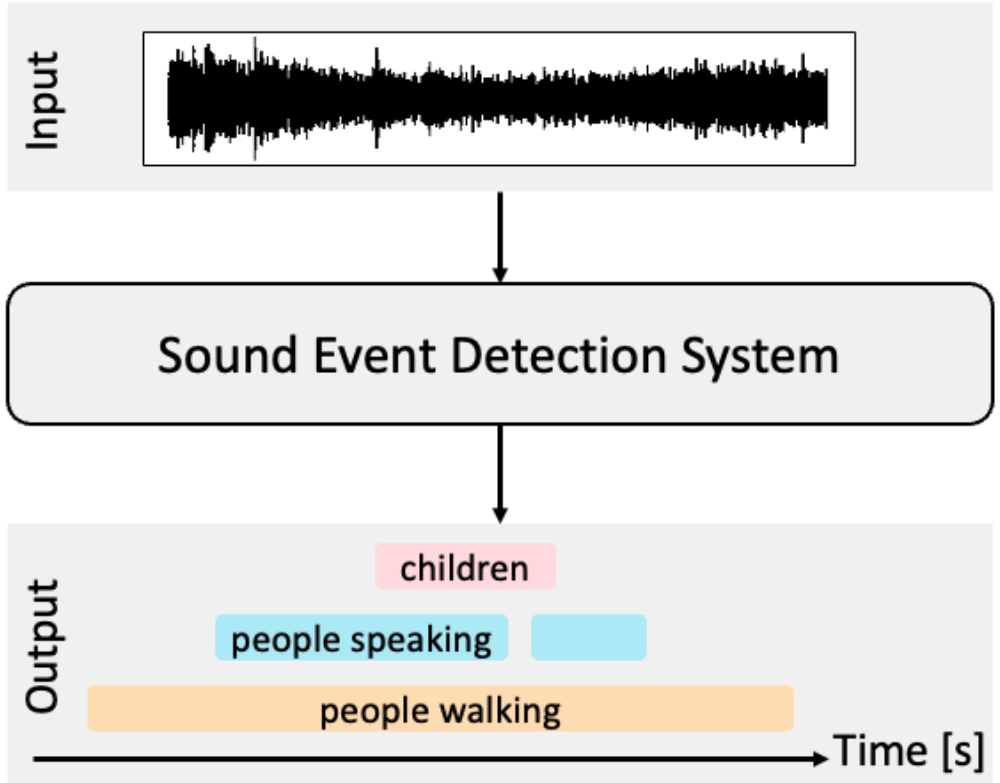
Imoto Lab

2024/03/15

音響イベント検出とは

音響イベントの種類と発生区間を推定するタスク

- 音響イベントの例：波の音，機械の作動音
- 応用例：住宅街の監視システム [1]，動画の自動タグ付け [2]

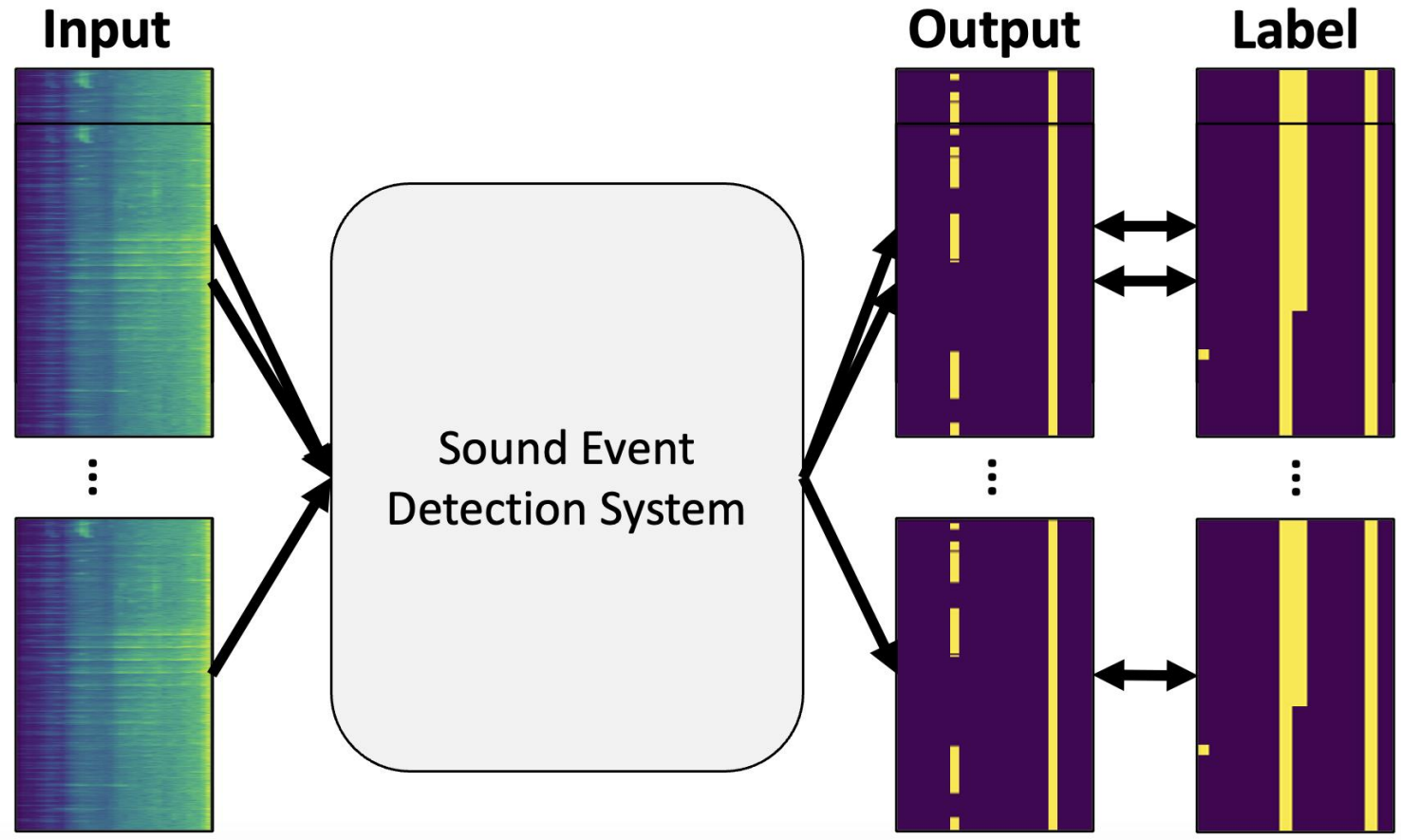


[1] C. Clavel, et al., “Events Detection for An Audio-based Surveillance System,” Proc. IEEE International Conference on Multimedia and Expo, pp. 1306–1309, 2005
[2] Y. Ohishi et al., “Bayesian Semi-supervised Audio Event Transcription Based on Markov Indian Buffet Process,” Proc. ICASSP, pp. 3163–3167, 2013

音響イベント検出モデルの学習

教師あり学習

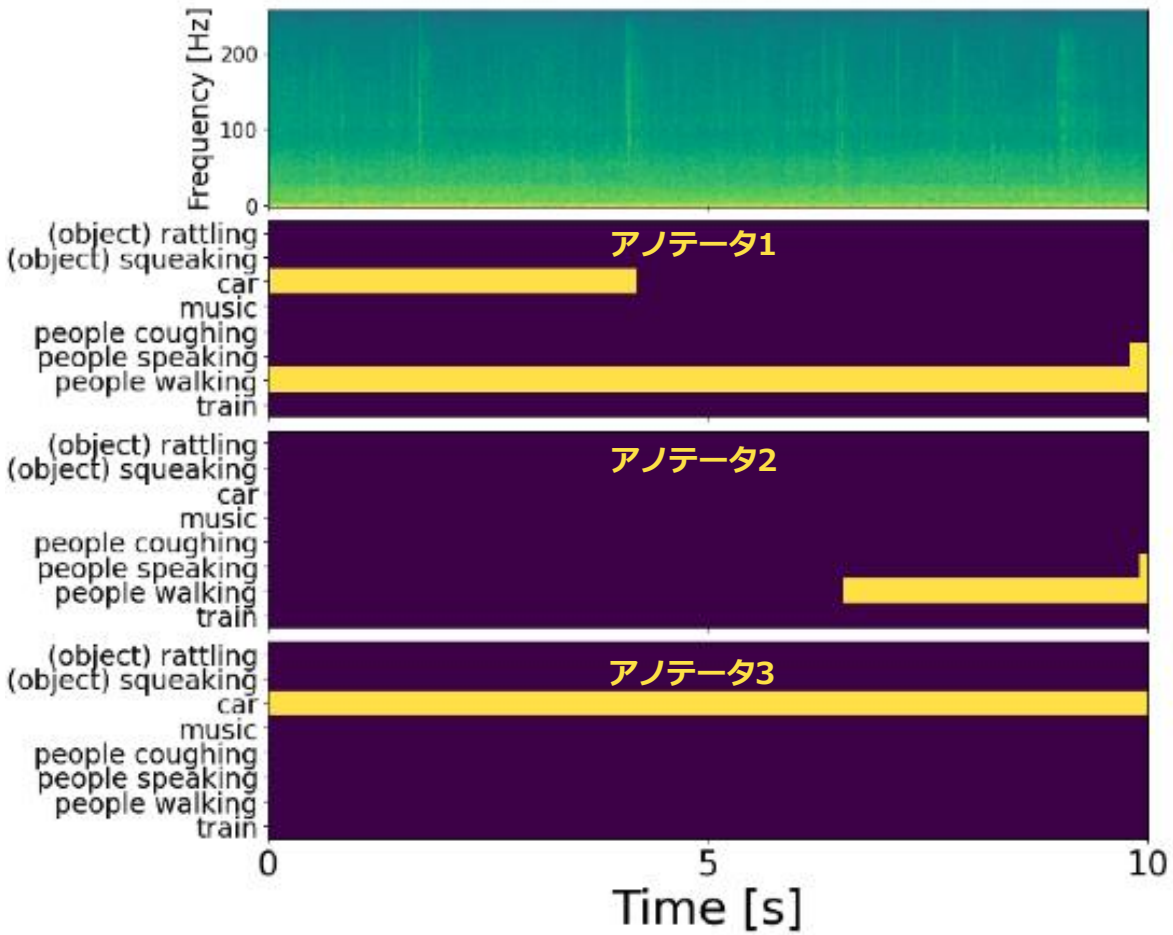
- モデルの学習に大量の音響信号とラベルが必要
  - ラベルを付与する人のことをアノテータと呼ぶ



音響イベント検出の課題

ラベルの時間情報のばらつき

- アノテータによって時間情報のばらつきが生まれる [3]



[3] A. Mesaros et al., “Sound event detection: A tutorial,” Proc. IEEE Signal Processing Magazine, vol. 38, pp. 67–83, September 2021

本研究の目的と方針

目的

- 音響イベント検出の課題を解決する
  - ラベル内で時間情報のばらつきが生まれる
    →**モデルの検出性能が低下してしまう**

方針

ラベルの時間情報のばらつきを考慮しモデルの
検出性能を向上させる機構の導入

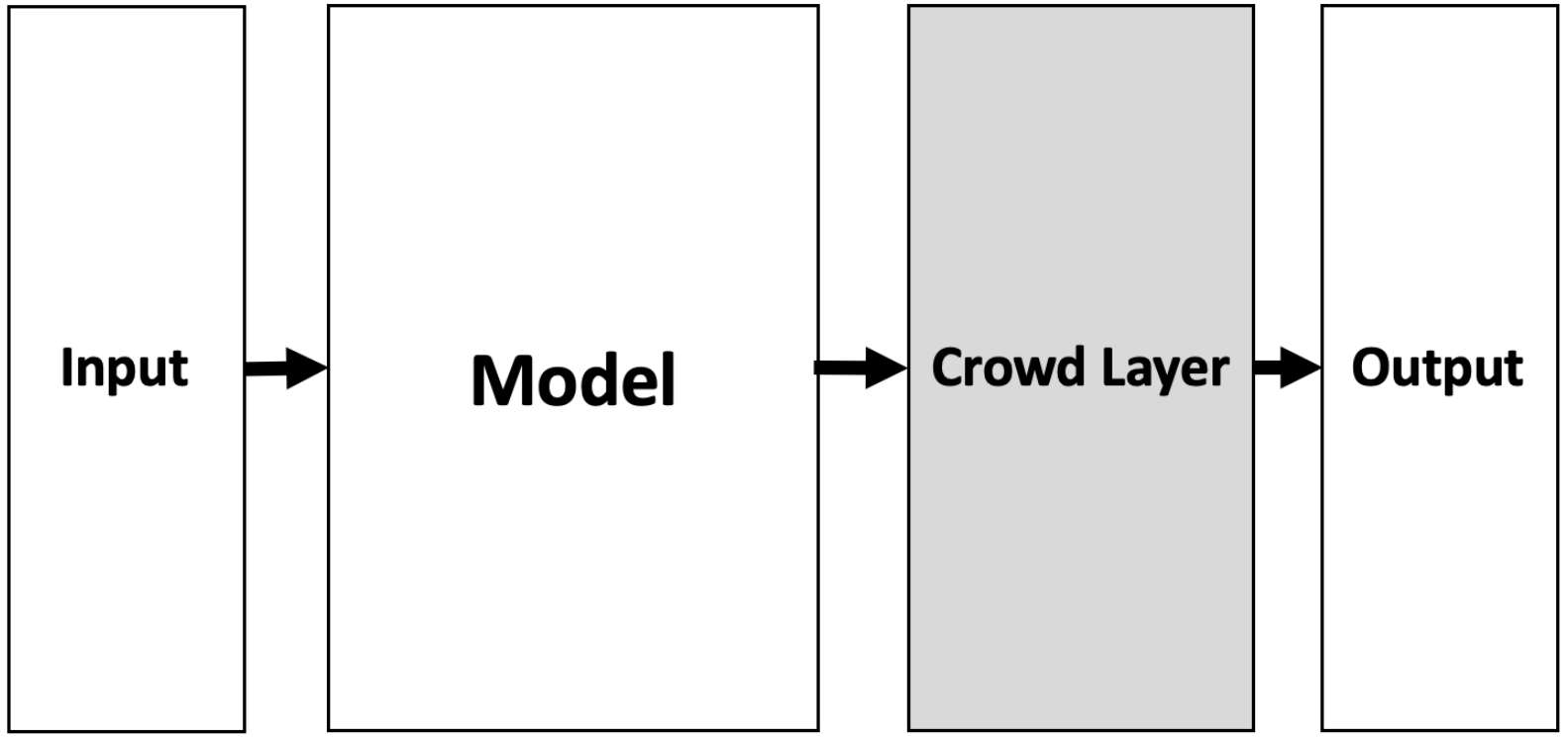
- アノテータ情報を把握するCrowd layer [4] を導入する

[4] F. Rodrigues, “Deep Learning from Crowds,” Proc. AAAI, Vol. 32, pp. 1611–1618, 2018

提案手法

Crowd layer [4] を用いた音響イベント検出モデル

- ネットワークの出力層の一種
  - CRNN (Convolutional Recurrent Neural Network) などの従来モデルと組み合わせて用いる

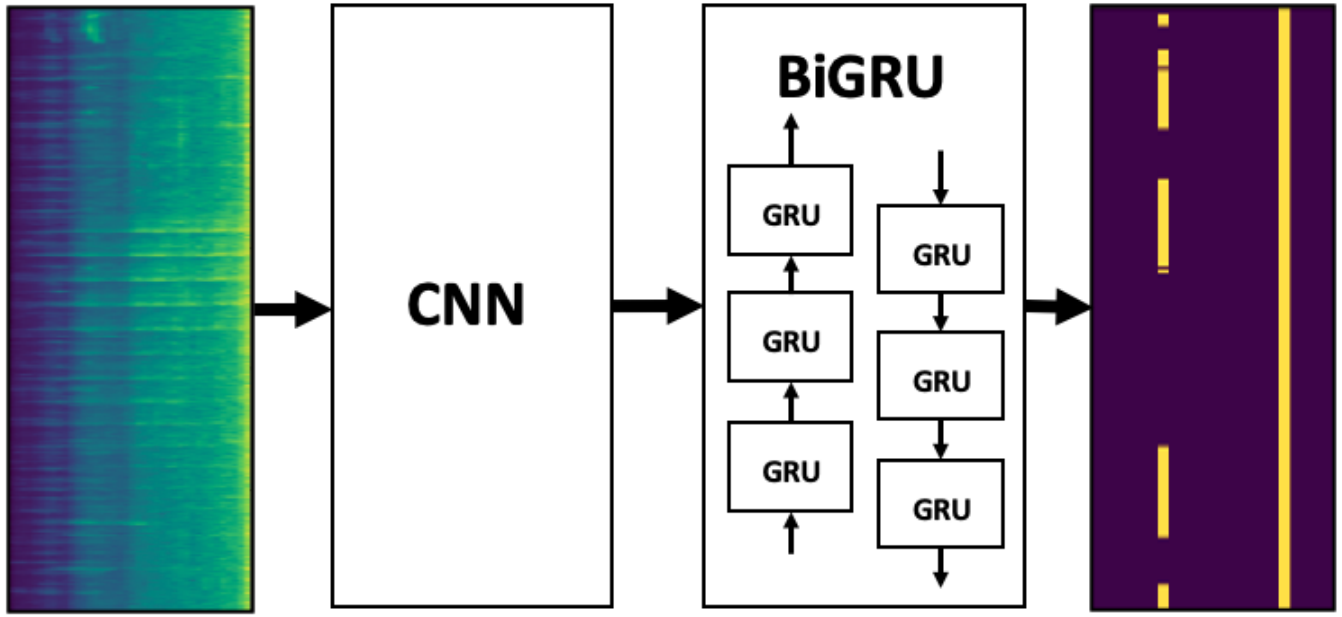


[4] F. Rodrigues et al., “Deep learning from crowds,” Proc. AAAI, vol. 32, pp. 1611–1618, 2018

提案手法と組み合わせる従来モデル

CNN-BiGRU

- 以下の2つのモデルを組み合わせたモデル
  - CNN (Convolutional neural network)
  - GRU (Gated recurrent unit) [5]
- 音響イベント検出で高い検出性能を達成することが報告されている [6]

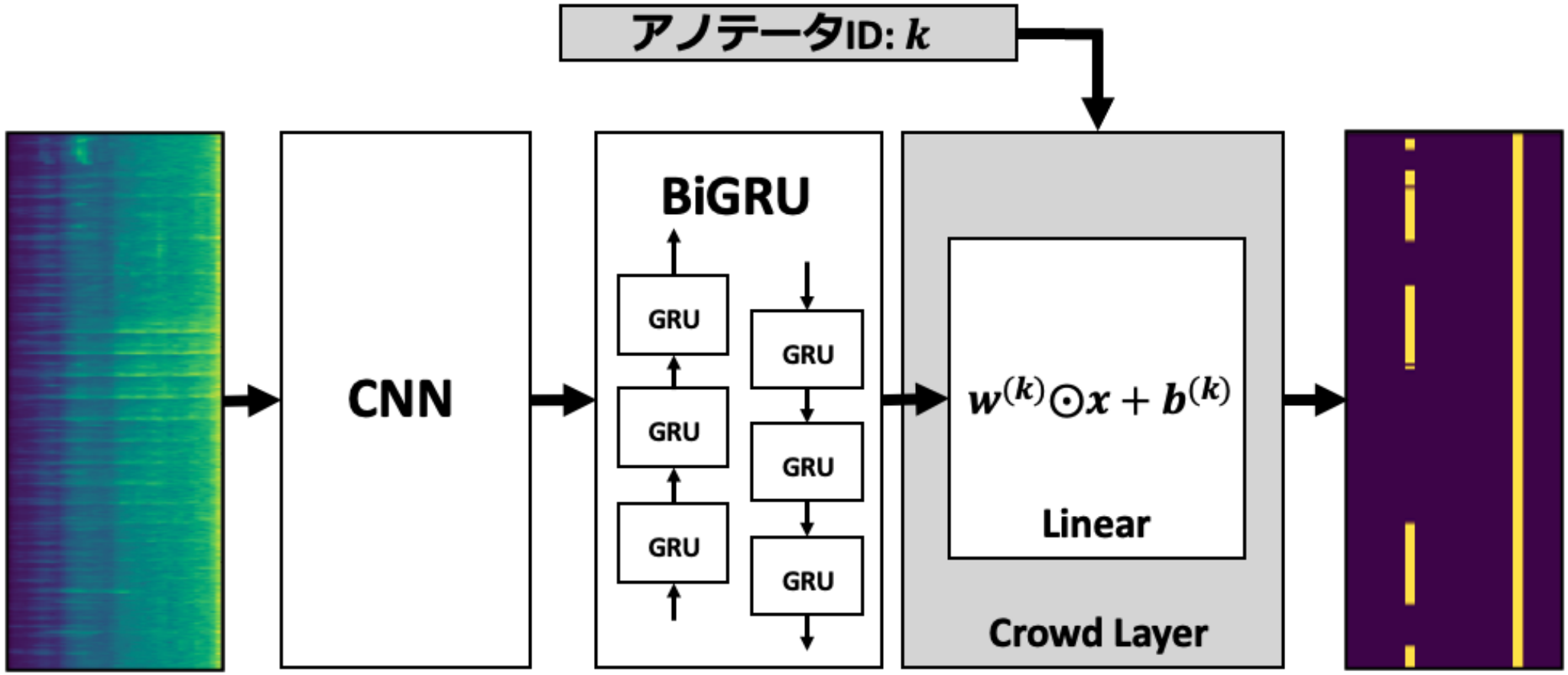


[5] K. Cho et al, “Learning phrase representations using rnn encoder–decoder for statistical machine translation,” Proc. EMNLP, pp.1724–1734, October, 2014
[6] S. Adavanne et al, “A report on sound event detection with different binaural features,” Proc. DCASE, pp. 12–16, November, 2017

提案手法

Crowd layerの実装

- アノテータkのラベルではk番目の重み・バイアスのみ使用

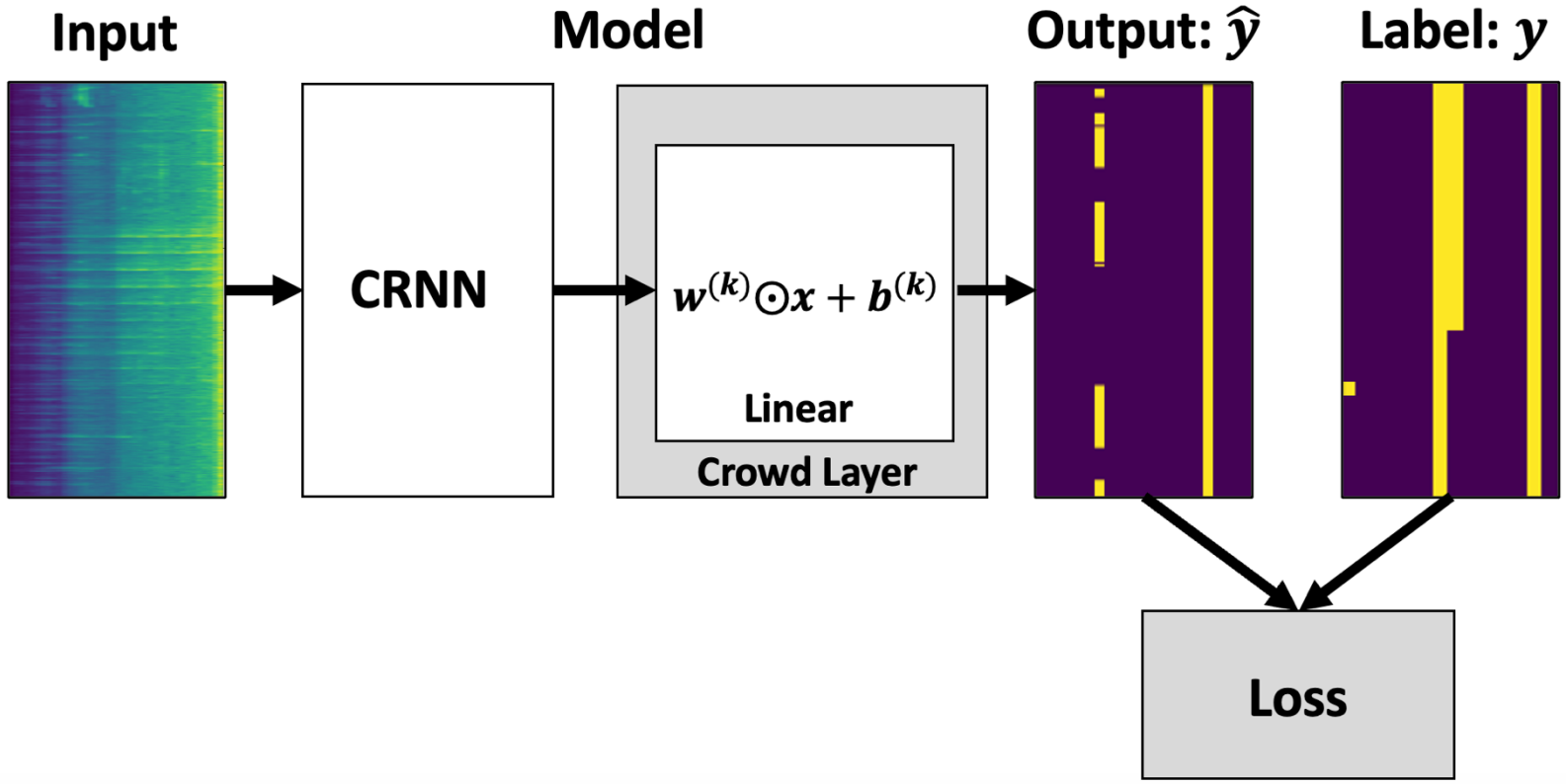


提案手法

モデルの学習

- BCE (Binary Cross Entropy) で損失を計算

$$L = -\sum y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})$$



実験目的・条件

実験目的

- Crowd layerを組み込んだモデルの従来法に対する検出性能向上を確認する

実験条件

表 1: 実験条件

Model	CNN-BiGRU (Crowd layer なし)
	CNN-BiGRU + CL (Crowd layer あり)
Dataset	TUT Sound Events 2016 [7]
	TUT Sound Events 2017 [8]

[7] A. Mesaros et al., “TUT Database for Acoustic Scene Classification and Sound Event Detection,” Proc. EUSIPCO, pp. 1128–1132, 2016
[8] A. Mesaros et al., “DCASE 2017 challenge setup: tasks, datasets and baseline system,” Proc. Workshop of DCASE, 2017

実験結果

提案手法

- 提案手法の**有効性が示唆された**

表 2: TUT Sound Events 2016 (20 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	33.33±1.49	54.17±1.82	41.23±0.99	10.81±0.63	74.55±0.96	634.30±286.56
Crowd layer あり	Train	34.53±1.22	52.72±3.72	41.66±1.10	11.15±0.65	74.57±2.18	332.93±158.65
Crowd layer なし	Eval	60.48±0.70	78.02±0.86	68.14±0.20	35.08±0.29	46.27±0.40	78.94±0.46
Crowd layer あり	Eval	**61.08±0.69**	**78.08±0.58**	**68.54±0.25**	**35.26±0.42**	**45.89±0.39**	**78.51±0.21**

表 3: TUT Sound Events 2017 (10 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	41.93±0.75	55.20±1.40	47.65±0.76	14.11±0.70	68.35±1.09	100.66±1.78
Crowd layer あり	Train	42.20±0.58	54.87±1.72	47.69±0.76	14.09±0.19	68.99±1.13	100.96±2.24
Crowd layer なし	Eval	65.19±0.51	**67.70±0.22**	66.42±0.17	**34.81±0.42**	51.50±0.21	82.90±0.11
Crowd layer あり	Eval	**65.40±0.39**	67.67±0.10	**66.51±0.18**	34.78±0.87	**51.36±0.21**	**82.89±0.17**

- 検出性能の向上は全て**1%以内だった**

考察

提案手法

- 検出性能の向上は**1%以内だった**
  - 処理が線形変換→パラメータで表現できることが限られるから

表 2: TUT Sound Events 2016 (20 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	33.33±1.49	54.17±1.82	41.23±0.99	10.81±0.63	74.55±0.96	634.30±286.56
Crowd layer あり	Train	34.53±1.22	52.72±3.72	41.66±1.10	11.15±0.65	74.57±2.18	332.93±158.65
Crowd layer なし	Eval	60.48±0.70	78.02±0.86	68.14±0.20	35.08±0.29	46.27±0.40	78.94±0.46
Crowd layer あり	Eval	**61.08±0.69**	**78.08±0.58**	**68.54±0.25**	**35.26±0.42**	**45.89±0.39**	**78.51±0.21**

表 3: TUT Sound Events 2017 (10 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	41.93±0.75	55.20±1.40	47.65±0.76	14.11±0.70	68.35±1.09	100.66±1.78
Crowd layer あり	Train	42.20±0.58	54.87±1.72	47.69±0.76	14.09±0.19	68.99±1.13	100.96±2.24
Crowd layer なし	Eval	65.19±0.51	**67.70±0.22**	66.42±0.17	**34.81±0.42**	51.50±0.21	82.90±0.11
Crowd layer あり	Eval	**65.40±0.39**	67.67±0.10	**66.51±0.18**	34.78±0.87	**51.36±0.21**	**82.89±0.17**

実験結果

■ データセット

- macro ERの平均，標準偏差が**大きい**

表 2: TUT Sound Events 2016 (20 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	33.33±1.49	54.17±1.82	41.23±0.99	10.81±0.63	74.55±0.96	634.30±286.56
Crowd layer あり	Train	34.53±1.22	52.72±3.72	41.66±1.10	11.15±0.65	74.57±2.18	332.93±158.65
Crowd layer なし	Eval	60.48±0.70	78.02±0.86	68.14±0.20	35.08±0.29	46.27±0.40	78.94±0.46
Crowd layer あり	Eval	**61.08±0.69**	**78.08±0.58**	**68.54±0.25**	**35.26±0.42**	**45.89±0.39**	**78.51±0.21**

表 3: TUT Sound Events 2017 (10 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	41.93±0.75	55.20±1.40	47.65±0.76	14.11±0.70	68.35±1.09	100.66±1.78
Crowd layer あり	Train	42.20±0.58	54.87±1.72	47.69±0.76	14.09±0.19	68.99±1.13	100.96±2.24
Crowd layer なし	Eval	65.19±0.51	**67.70±0.22**	66.42±0.17	**34.81±0.42**	51.50±0.21	82.90±0.11
Crowd layer あり	Eval	**65.40±0.39**	67.67±0.10	**66.51±0.18**	34.78±0.87	**51.36±0.21**	**82.89±0.17**

考察

データセット

- 特定の音響イベントのラベルが激しくばらついている可能性

表 2: TUT Sound Events 2016 (20 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	33.33±1.49	54.17±1.82	41.23±0.99	10.81±0.63	74.55±0.96	634.30±286.56
Crowd layer あり	Train	34.53±1.22	52.72±3.72	41.66±1.10	11.15±0.65	74.57±2.18	332.93±158.65
Crowd layer なし	Eval	60.48±0.70	78.02±0.86	68.14±0.20	35.08±0.29	46.27±0.40	78.94±0.46
Crowd layer あり	Eval	**61.08±0.69**	**78.08±0.58**	**68.54±0.25**	**35.26±0.42**	**45.89±0.39**	**78.51±0.21**

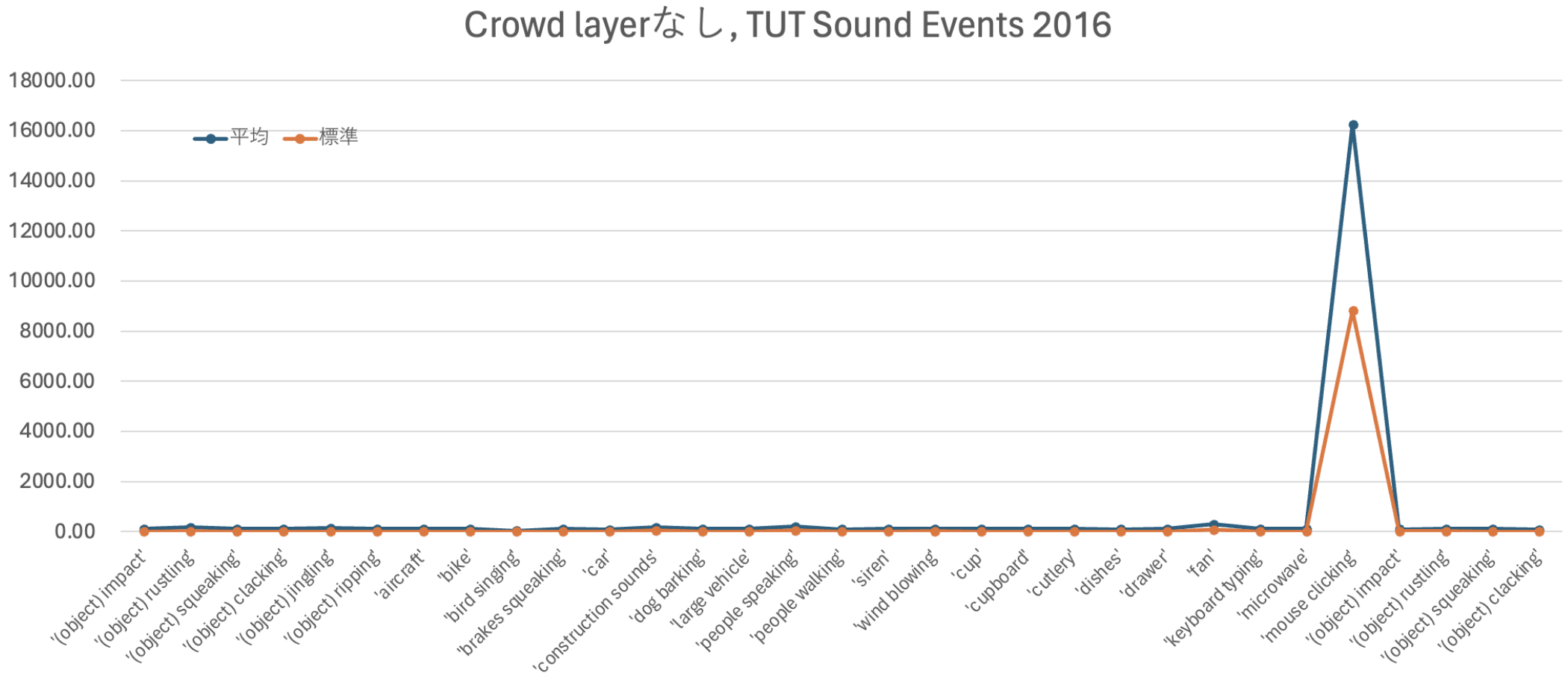
表 3: TUT Sound Events 2017 (10 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	41.93±0.75	55.20±1.40	47.65±0.76	14.11±0.70	68.35±1.09	100.66±1.78
Crowd layer あり	Train	42.20±0.58	54.87±1.72	47.69±0.76	14.09±0.19	68.99±1.13	100.96±2.24
Crowd layer なし	Eval	65.19±0.51	**67.70±0.22**	66.42±0.17	**34.81±0.42**	51.50±0.21	82.90±0.11
Crowd layer あり	Eval	**65.40±0.39**	67.67±0.10	**66.51±0.18**	34.78±0.87	**51.36±0.21**	**82.89±0.17**

考察

データセット

- 実際に特定の音響イベントラベルのERにばらつき



まとめ

本研究の目的

- 音響イベント検出の課題を解決する

実験結果

- アノテータごとの時間情報のばらつきを把握し検出性能が向上した
- 検出性能の向上の程度には改善の余地がある

今後の展望

- 特定の音響イベントラベルのばらつきを抑えたデータセットの作成
- 音響信号の数を増やして実験を行う