LSTM を用いた環境音からのリアルタイム音響イベント検出の検討

Study of Sound Event Detection in Real Time from Environments Sounds Using LSTM

小原 俊一

Shunichi Kohara

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本研究では，リアルタイムでの動作を想定した音響イベント検出の方式を目指す．具体的には，Long short-term memory を用いた識別器に音データを 1 フレームごとに逐次入力し，識別器は，これに対する予測結果を逐次出力する．評価実験は，音響イベント検出とリアルタイム性の 2 種類の観点からおこなう．実験結果から，音響特徴量にメルスペクトログラムを用いる場合に，高い精度とリアルタイム性が実現されることが示された．

1 はじめに

私たちの日常生活で聞こえる音は，環境音と呼ばれ，環境音には車の走行音や人の話し声などが該当する．音響イベント検出とは，環境音データから音響イベントの種類とその発生区間を求めるタスクである．音響イベント検出の研究では Deep Neural Network (DNN) が利用されており，その中でも Convolutional Recurrent Neural Network (CRNN) を利用したアプローチが高い有効性を示している [1]．一方で，これらの研究では，検出精度の向上を主な目的としており，推定に未来の情報を用いる場合や，検出に必要な計算資源が十分に用意されていることを前提とするネットワークである場合など，方式のリアルタイム性を考慮したものは少ない．

リアルタイムで動作する音響イベント検出を想定した場合，システムは環境音を逐次入力，入力に対する識別結果を逐次出力する．リアルタイムに音響イベント検出がおこなえると，イヤホン装着者や聴覚障害者の街路歩行時に，背後から車などが接近する危険を即座に通知することができる．

本研究では，Long short-term memory (LSTM) [2] を用いて，リアルタイムでの動作を想定した音響イベント検出システムを目指す．LSTM は，RNN の持つ長期依存性の問題に対して提案されたニューラルネットワークであり，LSTM および RNN は過去の時系列情報を内包しうる特徴がある．また，入力となる各フレームごとに推定結果が得られるため，リアルタイム処理の音響イベント検出が可能となる．

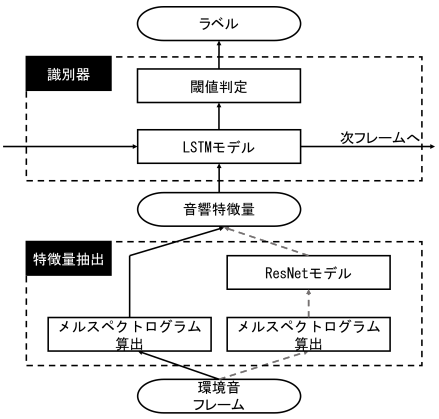

図 1: 提案方式の概要図

2 提案方式

2.1 提案方式の概要

提案方式の概要図を図 1 に示す．提案方式はフレームごとに環境音を逐次入力し，入力に対する音響イベント検出の予測結果を逐次出力する．なお，各フレームにおいて，識別器の LSTM モデルの出力は次フレームの LSTM モデルへ入力される．これにより，LSTM は過去の時系列情報を継承することが期待される．

2.2 特徴量抽出

本研究では，2 種類の音響特徴量を検討する．

1. メルスペクトログラム
2. 1. に ResNet モデル [3] を適用

2 つ目の特徴量について，音響イベント検出の特徴量として用いられるメルスペクトログラム [4] を，時間軸とメル周波数軸の 2 軸を持つ画像に見立て，ResNet モデルを適用することでより有用な特徴量抽出を期待する．

2.3 提案方式の実行時間

本研究では，提案方式の実行時間を測定する．実行時間は，図 1 の提案方式を以下の 4 つのモジュールに分割して測定する．

1. メルスペクトログラム算出部
2. ResNet モデル部

3. LSTM モデル部
4. 閾値判定部

3 評価実験

3.1 使用データセット

本実験で使用するデータセットは，URBAN-SED [5] である．データセットは 10 秒の環境音クリップ群で構成されており，各クリップに含まれる音響イベントの種類とその発生区間がラベル付けされている．クラス数は 10 クラスであり，各クリップ内で音響イベント同士のオーバラップが含まれる．クリップ数は 10000 個，総時間は約 28 時間であり，この中には約 50000 個の音響イベントが含まれる．

3.2 音響イベント検出に関する評価実験

本実験では，音響イベント検出の評価として，入力特徴量 1 フレームに対するマルチラベル分類の評価をおこなう．音響イベント検出におけるマルチラベル分類では，環境音フレームに対して，同時に 2 つ以上の音響イベントが割り当てられうる [6]．評価指標には，F1-score を用いる．

音響イベント検出に関する F1-score は，音響特徴量にメルスペクトログラムを用いた場合に 0.423，メルスペクトログラムに ResNet モデルを適用した場合に 0.353 であった．特徴量抽出器に ResNet モデルを追加した場合，有用な特徴量が抽出できなかったことがわかる．これは，ResNet モデルへメルスペクトログラムを入力する際に，アスペクト比を等しくするため，リアルタイムのフレームと適当な過去のフレームをまとめて入力したことで，リアルタイムのフレームの情報が削減されたからと考えられる．

3.3 リアルタイム性に関する評価実験

本実験では，リアルタイム性の評価として，実行時間を測定する．実行時間は，提案方式をスマートフォンや同等のサイズのデバイスで実現できる性能で動作させることを想定し，Raspberry Pi 3 Model B で測定する．Raspberry Pi 3 Model B は，1.2GHz のクアッドコア CPU と 1GB のメモリを搭載した小型の計算機である．評価指標には，リアルタイムファクター (RTF) を用いる．RTF は $\mathrm{RTF} = \mathrm{T_p}/\mathrm{T_f}$ で算出される．ここで，$\mathrm{T_p}$ と $\mathrm{T_f}$ は，それぞれ実行時間と入力の時間長である．RTF が 1 を下回る場合，リアルタイム動作を実現できると考えられる．

各モジュールの実行時間を表 1 に示す．入力特徴量 1 フレームの時間長は，約 23.2ms である．特徴量をメルスペクトログラムとした場合，提案方式はメルスペクトログラム算出部と LSTM モデル部，閾値判定部で構成され，RTF は 0.49 となり，リアルタイム動作を実現できるといえる．ResNet モデル部は他モジュールと比較して，大幅に実行時間が長くなった．これは．ResNet を構成する CNN のパラメータが非常に多く，実験に使用した Raspberry Pi にこのパラメータ群を高速に処理できる性能がなかったためと考えられる．

表 1: Raspberry Pi 3 Model B での各モジュールの実行時間 (1 フレーム: 23.2 ms)

	平均 (ms)	標準偏差
メルスペクトログラム算出部	8.738	2.325
ResNet モデル部	5591.678	151.687
LSTM モデル部	1.730	0.092
閾値判定部	0.829	0.048

4 まとめ

本報告では，LSTM を用いてリアルタイムでの動作を想定した音響イベント検出システム方式について述べた．提案方式の評価に，音響イベント検出とリアルタイムに着目した実験をおこなった．実験結果から，音響特徴量にメルスペクトログラムを用いる場合に，高い精度とリアルタイム性が実現されることが示された．今後の課題は，精度向上のための音響特徴量やネットワーク構成の検討が挙げられる．

参考文献

[1] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural networks for polyphonic sound event detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291-1303, 2017.

[2] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc.f the IEEE Conference on CVPR, pp. 700–778, June 2016.

[4] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A.P. Shah, “Large-scale weakly labeled semi-supervised sound event detection in domestic environments,” Proc. DCASE2018, pp. 19–23, November 2018.

[5] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J.P. Bello, “Scaper: A library for soundscape synthesis and augmentation,” 2017 IEEE WASPAA, pp. 344–348, 2017.

[6] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” Applied Sciences, vol. 6, no. 6, 2016.