# RNN による実環境データからのマルチ音響イベント検出

Multi audio event detection from real life audio data by using RNN

鳥羽 隼司

Shunji Toba

岡山大学 阿部研究室

Abe Laboratory, Okayama University

**概要** 本報告では，実環境下で収録された環境音から，Recurrent Neural Network を用いて音響イベントを複数同時に検出する手法について述べる．提案する RNN は LSTM を用いることにより時系列の過去の情報を記憶し，また複数同時な音響イベント検出に対応する．RNN による提案手法と GMM による既存手法との比較の結果，提案手法は既存手法を検出性能で明確に上回らず，データ量の少なさが原因と考えられる．

## 1 はじめに

人の話し音や人工物，動物の鳴き声など我々が普段耳にする環境音には，有益な情報が多く含まれている．環境音を機械的に認識することにより，カメラによる視覚的な情報などからは得られない聴覚的な情報を取得し活用することが可能となる．この情報はホームセキュリティ，自動運転，オーディオデータを用いたライフログなど様々な活用方法が考えられる．

環境音は様々な音から成り立つため，環境音認識は単一の音のみではなく複数同時に行えることが望ましい．また，与えられたオーディオデータからは音の種類だけでなく，どの時点でその音が発生したかを知ることができれば，さらなる活用に繰り広げられる．

本研究では，実環境から収録された環境音を対象として，ある音響イベントの種類とその発生区間を同時に検出する音響イベント検出について検討する．ここでの音響イベントは複数同時に発生しうる，かつ発生区間はイベントごとに重複しうるものとし，音響イベント検出は複数イベント同時に検出するマルチラベル方式とする．音響イベント検出の実現には時系列データに有効とされる Recurrent Neural Network を用いる．

## 2 Recurrent Neural Network

Recurrent Neural Network（RNN）とは，時系列のデータを扱うためのニューラルネットワークの方式の一つであり，音声認識，自然言語処理の分野で用いられる．RNN では隠れ層に Long Short Term Memory を用いることでネットワークに過去の情報を記憶させられるため，時系列データを扱ううえで有効である．

### 2.1 Long Short Term Memory

Long Short Term Memory [1]（LSTM）とは隠れ層に用いるネットワーク方式の一つである．LSTM の内部には記憶装置が存在し，どのタイミングで隠れ層の状態を更新すべきかを学習することができる．

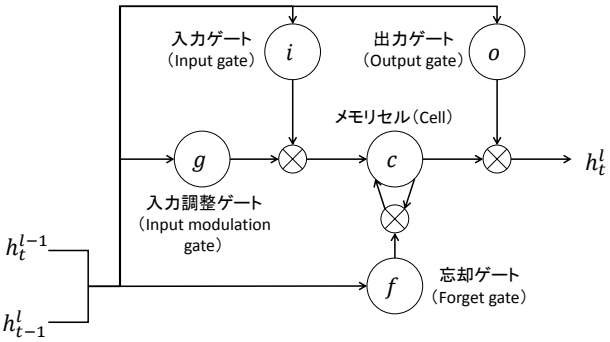


図 1: Long Short Term Memory の構成

LSTM の構成図を図 1 に示す．LSTM の構成要素としては入力ゲート $i$，入力調整ゲート $f$，出力ゲート $o$，忘却ゲート $g$ の各種ゲートと，過去の隠れ層情報の記憶部分であるメモリセル $c_t$ からなる．各構成要素は式で以下のように表される．

$$
\begin{aligned}
i &= \mathrm{sigm}(W_{i_1} h_t^{l-1} + W_{i_2} h_{t-1}^l + b_i) \\
f &= \mathrm{sigm}(W_{f_1} h_t^{l-1} + W_{f_2} h_{t-1}^l + b_f) \\
o &= \mathrm{sigm}(W_{o_1} h_t^{l-1} + W_{o_2} h_{t-1}^l + b_o) \\
g &= \tanh(W_{g_1} h_t^{l-1} + W_{g_2} h_{t-1}^l + b_g) \\
c_t &= f \odot c_{t-1} + i \odot g \\
h_t^l &= o \odot \tanh(c_t)
\end{aligned}
$$

$\mathrm{sigm}(x)$ はシグモイド関数，$\tanh(x)$ は双極正接関数を表す．各種ゲートは現在の時刻 $t$ の一つ前の層の出力 $h_t^{l-1}$ と時刻 $t-1$ の LSTM 層の出力 $h_{t-1}^l$ を受け取り，これらの情報が重み付けにより調整されてメモリセルと LSTM 層の出力の計算に用いられる．メモリセルは忘却ゲート $f$ で調整された過去のメモリセル $c_{t-1}$ と，入力ゲート $i$ と入力調整ゲート $g$ から決定される．$h_t^l$ は時刻 $t$ における LSTM 層の出力であり，出力ゲート $o$ とメモリセル $c_t$ により決定される．

## 3 マルチ音響イベント検出手法

### 3.1 音響的特徴量の抽出

音響的特徴量には Mel Frequency Cepstral Coefficient（MFCC）20 次元と，その時間微分 ΔMFCC 20 次元，さらにその時間微分 ΔΔMFCC 20 次元の合計 60 次元を使用する．その特徴量をオーディオデータから一定間隔のフレームごとに抽出し，時系列の特徴ベ

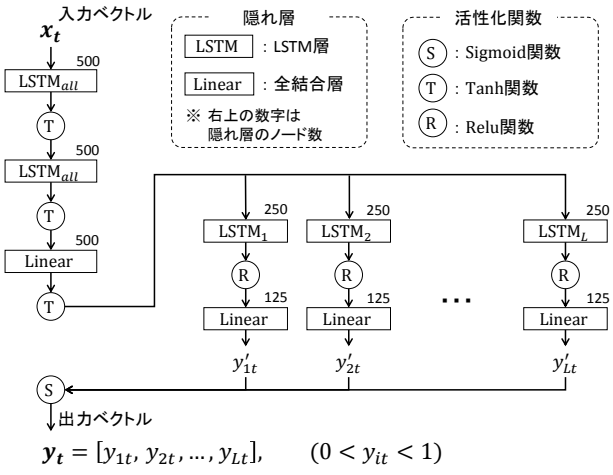


図 2: マルチ音響イベント検出 RNN の構成

クトルとして RNN の入力に用いる．フレーム分割条件はフレーム長 40 ms，フレームシフト 20 ms としている．

### 3.2 RNN による音響イベント検出器

マルチラベル対応の音響イベント検出器の実装には Chainer[*1] を使用した．実装した RNN は時間 $t$ の音響的特徴量 $x_t$ を入力として，その時刻のラベル情報予測 $y_t$ を出力する．ネットワークの学習時には本来のラベル情報と $y_t$ の誤差が最小となるよう学習を繰り返す．

RNN のネットワークの構成を図 2 に示す．ネットワークの入力 $x_t$ は $\mathrm{LSTM}_{all}$ 層 2 層と全結合層を通過したのち，各ラベルそれぞれの予測 $y'_{it}$ $(i = 1, 2, \ldots, L)$ のための $\mathrm{LSTM}_i$ 層と全結合層を通過する．そして，各ラベルについての予測 $y'_{it}$ をひとまとめにしてシグモイド関数を適用したものを最終的な出力 $y_t$ とする．

このネットワークの狙いとしては，最初の 2 層の $\mathrm{LSTM}_{all}$ 層で全ラベルについて認識に有効な情報を記憶させ，その後の $\mathrm{LSTM}_i$ 層で各ラベルそれぞれについて有効な情報を記憶させることを目的としている．また，最後にシグモイド関数を適用することで，最終的な出力 $y_t$ の各要素の値を $(0, 1)$ としラベルの類似度と見立てられるため，複数同時なラベルの認識が可能となる．

## 4 評価実験

今回実装した音響イベント検出器の評価のため，既存手法との比較による評価実験を行った．比較する既存手法は文献 [2] で示されている GMM（Gaussian Mixture Model）による手法とする．

### 4.1 データセット

今回実験には TUT Sound Events 2016 [2] データセットを用いた．データセットには Home（室内）と Residential area（室外）の 2 つの場面で収録されたオーディオデータと，人手による音響イベントのラベル付けデータが付与されている．各オーディオデータは 1 ファイルあたり 3〜5 分程度で，トータルの時間では Home が 36 分 16 秒，Residential area が 42 分であり，サンプリング周波数は 44.1 kHz，量子化ビット数は 24 bit である．音響イベントラベルは Home では全 11 種類，Residential area では全 7 種類が付与されている．

[*1] Chainer: http://docs.chainer.org

表 1: GMM と RNN の各手法の評価結果

| | GMM | | RNN | |
|---|---|---|---|---|
| place | ER | F[%] | ER | F[%] |
| Home | 0.95 | 18.1 | 1.26 | 28.7 |
| Residential area | 0.83 | 35.2 | 0.95 | 28.7 |
| total | 0.89 | 26.6 | 1.11 | 28.7 |

### 4.2 評価結果

Home，Residential area それぞれについて 4-fold cross validation で検出モデルの学習と評価を行った．評価には文献 [2] で示されている検出の誤り率を表す Error Rate と，検出の適合性と再現性を表す F-score を用いる．

評価結果を表 1 に示す．評価結果より，F-score では Home において RNN が 10 % 以上の差をつけて GMM を上回ることができた．一方で Error Rate では RNN は GMM に合計値で劣り，F-score でも RNN は合計では勝るが Residential area では GMM に劣る．総合的に見て，RNN が GMM に性能において必ずしも勝るとはいえない結果となった．

RNN での検出性能が不十分であった理由として，学習に用いるデータ不足が考えられる．今回使用したデータセットは総時間が合計で 72 分程度と，音響イベント検出というタスクに対して RNN が十分に性能を発揮するにはデータが不足しており，統計的な手法の GMM が有利であったと考えられる．

## 5 まとめ

本報告では，実環境で収録された環境音からマルチ音響イベント検出を行う方法について述べた．音響イベント検出の実現には RNN を用い，既存手法 GMM との比較による評価実験を行った．評価実験の結果，RNN が GMM を性能で明確に上回ることは示せなかった．今後はデータ量の多いデータセットで RNN の検出性能を確認する予定である．

## 参考文献

[1] W. Zaremba, I. Sutskever, and O. Vinyals, “Recurrent neural network regularization”, arXiv preprint arXiv:1409.2329, 2014.

[2] A. Mesaros, T. Heittola, and T. Virtanen , “TUT Database for Acoustic Scene Classification and Sound Event Detection”, 24rd European Signal Processing Conference, 2016.