# 音響イベント検出における BLSTM-CTC を用いた弱ラベル学習の検討 *

☆松吉大輝 (筑波大), 小松達也 (NEC), 近藤玲史 (NEC), 山田武志 (筑波大), 牧野昭二 (筑波大)

## 1 はじめに

防犯監視システムや音環境理解による動画の自動タグ付けシステムにおいて, 音響イベント検出 (SED: Sound Event Detection) は重要な役割を担う. SED は防犯監視システムにおいては異常音の検出, 音環境理解による動画の自動タグ付けシステムでは, タグ付けする音の検出に使用される. SED のタスクは音響信号データ内で発生しているイベント音の種類, 開始時刻, 終了時刻を推定することである. 例として Fig. 1 ではイベント音（電話の音）が入っている音響信号データを SED システムに入力したとき, 電話の音がいつからいつまで発生しているかを表すラベルを出力していることを示している.

従来, SED 手法として主に NMF(Non-negative Matrix Factorization) を用いる手法 [1][2] と NN(Neural Network)[3] を用いる手法が提案されている. NMF を用いる手法は線形処理のため, 少量の学習データで推定モデルを学習することができる. しかし, 推定モデルの表現力が十分でないため, 複雑なタスクにおいて高精度に検出を行うことは困難である. 一方, NN を用いる手法は非線形処理のため表現力が高く, 複雑なタスクに対応可能である. しかし, 推定モデルの学習にはイベントの種類, 開始時刻, 終了時刻の正解ラベルが必要となり, 大量の音響信号データに対し人手でラベル付けすることは極めて困難である.

本稿では, NN を用いた SED 手法における学習データのラベル付けのコスト削減を目的とし, BLSTM-CTC を用いた弱ラベル学習を提案し, その有効性を示す.

## 2 従来の NN を用いた SED 手法

### 2.1 RNN を用いた SED

SED の推定対象はイベント音の種類, 開始時刻, 終了時刻であり, 時間情報を含んでいるため, 時系列データに対応した RNN(Recurrent Neural Network) がよく用いられる. RNN は入力層, 隠れ層, 出力層を時間

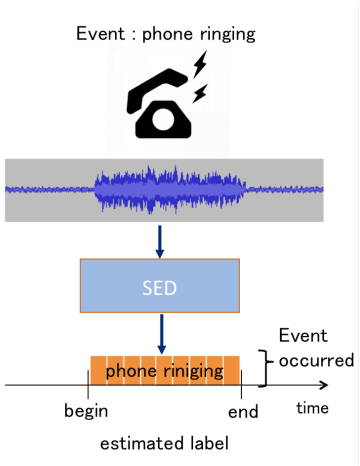


Fig. 1 SED の概要

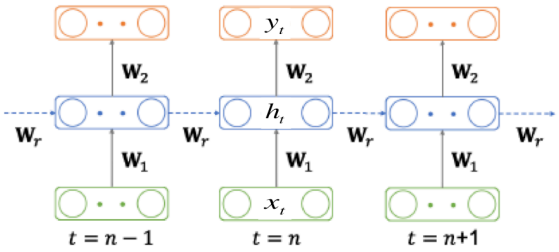


Fig. 2 RNN の例

フレームごとに並べた NN である. Fig. 2 の RNN において, ある時間フレーム $t = n$ の入力と $t = n - 1$ の隠れ層の情報を用いて $t = n$ の隠れ層を決定し, その隠れ層から出力を決定する. これらの処理を次式に示す.

$$h_t = f(W_1 x_t + W_r h_{t-1} + b_1), \tag{1}$$

$$y_t = g(W_2 h_t + b_2) \tag{2}$$

この処理によって, 前フレームとの相関の情報を加味した出力をすることができる.

### 2.2 BLSTM を用いた SED

2.1 節で述べたシンプルなモデルの RNN では, 直前のフレームの影響力が大きく, 長い時系列データに対して, 長時間離れたフレームとの相関の情報が

* Weakly-Labeled Learning Using BLSTM-CTC for Sound Event Detection. by Taiki Matsuyoshi (University of Tsukuba), Tatsuya Komatsu, Reishi Kondo (NEC), Takeshi Yamada, Shoji Makino (University of Tsukuba)

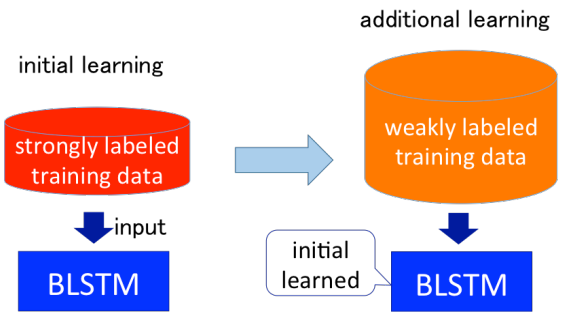


Fig. 3　学習の流れ

利用できない．そこで長い時系列データに対応可能になるよう, RNN にメモリセルなどの機能を追加した LSTM(Long Short-Term Memory)[4] が提案されている．これにさらに時間的に前向きの隠れ層の情報だけでなく, 後ろ向きの隠れ層の情報も利用する BLSTM(Bi-directional LSTM) を用いた手法 [5] が提案され, 高い推定性能を示している．本研究では BLSTM を用いた SED 手法に注目する．

### 2.3　BLSTM を用いた SED の課題

BLSTM を用いた SED の推定モデルを学習するためには大量の正解ラベルを用意する必要がある．このような正解ラベルを作成するためには，1 つ 1 つの音響信号データに対し人手による音響信号データの聴取, 波形の確認などをすることで, イベントの種類, 開始時刻, 終了時刻をラベル付けする必要がある．NN では, 高い推定精度を出すためには学習に有効な大量のデータ（数百〜数千個）が必要になるが, これに対応する正解ラベルを作成するのは極めて困難と言える．

そこで本研究では正解ラベルの作成のコストを小さくするために, 発生しているイベントの種類のみのラベルで学習することを目的とする．ここで, 従来のイベントの種類, 開始終了時刻を含んだラベルを強ラベルとし, 発生しているイベントの種類のみのラベルを弱ラベルとする．

## 3　提案手法

### 3.1　BLSTM を用いた SED における学習の流れ

提案手法では Fig. 3 に示すように, まず少量の強ラベル付きの学習データで BLSTM を学習し, その BLSTM を大量の弱ラベル付きの学習データで追加学習する．弱ラベル学習前後の BLSTM での推定精度を比較することで, 弱ラベル学習による推定性能の向上を確認することができる．

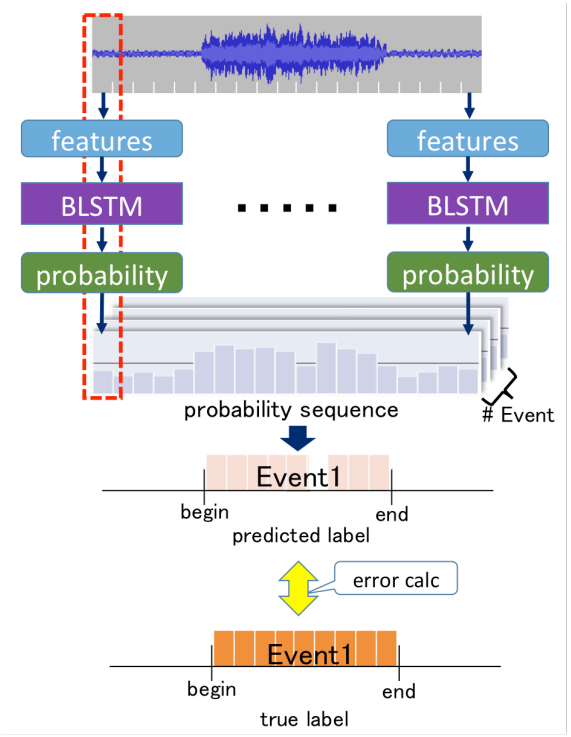


Fig. 4　強ラベル学習の流れ

### 3.2　強ラベル付き学習データを用いた初期学習

強ラベル付き学習データを用いた初期学習の処理フローを Fig. 4 に示す．まず音響信号データに対し, 時間フレームごとに特徴抽出し, それを BLSTM の入力とする．BLSTM 内の隠れ層に入力の特徴の情報を伝播し, 各フレームごとにイベントの存在確率を出力する．以上の処理によってフレーム数と同じ大きさのイベント存在確率列が得られる．そのイベント存在確率列から, 閾値計算を行うことでラベルを推定する．推定されたラベルと正解ラベル (強ラベル) を比較し, その誤差を小さくするように BLSTM の内部パラメータを更新していく．

### 3.3　弱ラベル付き学習データを用いた追加学習

#### 3.3.1　Connectionist Temporal Classification

2.3 節で述べた弱ラベルでの学習を行うために, 音声認識の分野で弱ラベル学習において成果を上げている CTC(Connectionist Temporal Classification)[7] という手法を利用する．CTC は, 入力と出力の系列長の異なる時に用いられる損失関数であり, RNN の出力に適用可能である．CTC の処理について Fig. 5 を用いて説明する．横軸は時間を示し, 縦軸は状態の遷移を示す．状態は「Event1 未発生」と「Event1 発生」の 2 種類とし, ある時間フレームにおいて, 3.2 節で述べた BLSTM で出力される Event1 の発生確率が低い場合「Event1 未発生」の状態になり, 高い場合は

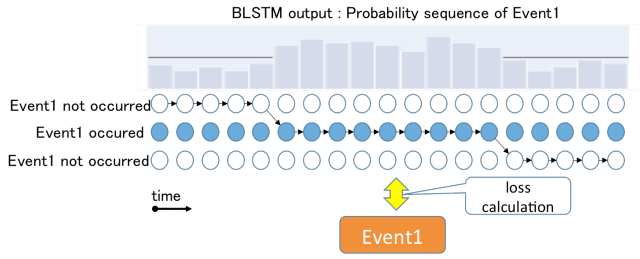


Fig. 5　CTC の処理

「Event 発生」の状態になる．Fig. 5 ではイベントが未発生, 発生, 未発生の順になっており, 音響データ内で Event1 が 1 回発生している確率が高いことを示している．正解として与えられた弱ラベルが示す種類のイベントが音響データ内で 1 回発生する確率が高いとき小さい損失を算出し, 発生する確率が低いときに大きい損失を算出することで BLSTM の内部パラメータを更新する．Fig .5 では, Event1 が 1 回発生している確率が高いため, 弱ラベルが示す種類が Event1 のとき, 小さい損失となる．以上のような CTC の処理を利用することで, 弱ラベルによる BLSTM の内部パラメータの学習を行うことができる．

### 3.3.2　BLSTM-CTC の学習の流れ

提案手法の処理フローを Fig. 6 に示す．まず学習時について述べる．3.2 節で述べた強ラベルを用いた BLSTM の学習と同様, 時間フレームごとに抽出した特徴量を BLSTM の入力とし, イベントの発生確率列を出力する．そのイベント発生確率列を CTC の入力とし, 弱ラベルを用いて損失計算をし, BLSTM の内部パラメータを更新する．推定時は学習済みの BLSTM の出力であるイベント発生確率列を閾値計算し, 推定したラベルを出力する．

## 4　提案手法の有効性の検証

### 4.1　実験の概要

CTC を用いた弱ラベル学習の効果検証をするための実験を行った．イベントの種類は clear throat の 1 つのみとし, イベント開始時間, 終了時間の推定を行う．実験条件を Table1,2 に示す．イベントデータには DCASE2016[6] の Task2 のデータセット収録の clear throat を用い, 5 秒の背景雑音に対し 1 秒から 3 秒程度のイベントデータを重畳することで 85 個の学習データを生成した．そのうち 5 個を強ラベル, 80 個を弱ラベルの正解ラベルとし,

(i) 強ラベル付き学習データ 5 個のみで学習した BLSTM

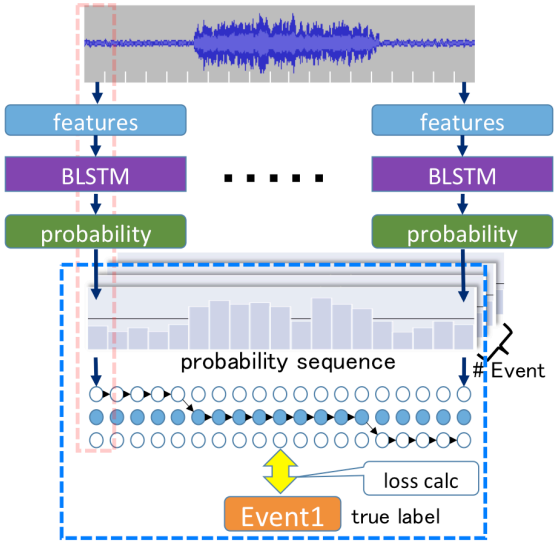


Fig. 6　弱ラベル学習の流れ

Table 1　実験条件 (NN)

| | |
|---|---|
| Learning rate | 0.0005 |
| Gradient clipping norm | 5 |
| Batch size | 5 |
| Epoch | initial leaning : 30<br>additional learning : 5 |
| Hidden layer size | 100 |

(ii) 強ラベル付き学習データに加え, BLSTM-CTC により弱ラベル付き学習データ 80 個を用いて追加学習を行った BLSTM

の SED の性能比較を行った．評価データには, 学習データ同様, 5 秒の背景雑音に対しイベントデータを重畳することで, 40 個の評価データを生成した．

評価尺度は時間フレーム単位での Recall と Precision とする．Recall は実際にイベント音が存在しているもののうち、イベント音があると推定した割合, Precision はイベント音があると推定したデータのうち実際にイベント音が存在しているものの割合である．

### 4.2　実験結果と考察

実験結果を Fig. 7,8 に示す．これにより, CTC による弱ラベルを用いた追加学習によって Recall においてほぼ同等の精度を示し, Precision において 5.5% の精度の向上を示し, 弱ラベルによる追加学習の効果が確認できる．Precision が向上したことから, 誤推定してしまう割合を削減できたことがわかる．今後は推定したラベルに対し, 事後処理を適用することによる Recall の精度向上及び, 複数種類のイベント推定など

Table 2　実験条件 (音響データ)

| | |
|---|---|
| Sampling rate | 44,100Hz |
| SNR | 6dB |
| feature | 39 Mel-filter bank |
| Frame size | 25ms |
| event name | clear throat |
| train data of initial learning | 5 |
| train data of additional learning | 80 |
| test data | 20 |
| length of data | 5s |

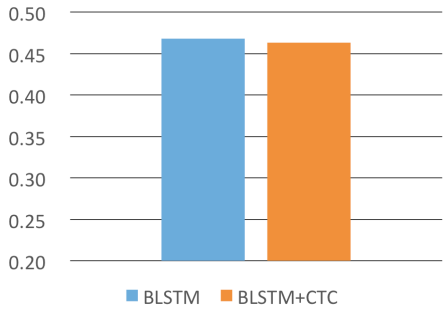


Fig. 7　実験結果・Recall

のより難しいタスクを用いて提案手法による有効性の評価を行う予定である.

## 5　おわりに

本稿では, 強ラベル付きデータのラベル付けのコスト削減を行うことを目的とし, BLSTM を用いた SED 手法において CTC を利用することで, 弱ラベル学習を行う手法を提案した. 提案手法により, 弱ラベルの学習データを用いた追加学習によって, Precsion において 5.5%の性能向上を得られ, CTC を用いた弱ラベル学習が有効であることを確認した.

## 参考文献


[1] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, “Sound event detection in multisource environments using source separation,” Workshop on machine listening in Multisource Environments, pp. 36-40, 2011.

[2] T. Komatsu, T. Toizumi, R. Kondo, Y. Senda, “Accouctic Event Detection Method Using Semi-supervised Non-negative Matrix Factorization with a Mixture of Local Dictionaries,” Detection and Classification of Acoustic Scenes and Events 2016.

[3] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, “Polyphonic sound event detection using multi label deep neural networks,”IEEE IJCNN, pp. 1-7, 2015.

[4] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, No. 9, Vol. 8 pp. 1735-1780, 1997.

[5] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Roux, K. Takeda, “Bidirectional LSTM-HMM Hybrid System For PolyPhonic Sound Event Detection,”Detection and Classification of Acoustic Scenes and Events 2016.

[6] DCASE2016, http://www.cs.tut.fi/sgn/arg/dcase2016/.

[7] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks,” IDSIA, 2006.

[8] Y. Wang, F. Metze, “A First Attempt at Polyphonic Sound Event Detection Using Connectionist Temporal Classification,” In Proc. ICASSP, New Orleans, LA; U.S.A., March 2017. IEEE.

[9] J. Schluter, “Learning to Pinpoint Singing Voice From Weakly Labeled Examples,”Proceedings of the 17th ISMIR Conference, pp. 44-50, August 2016.

[10] Chainer, http://chainer.org/.

[11] Theano, http://deeplearning.net/software/theano/.


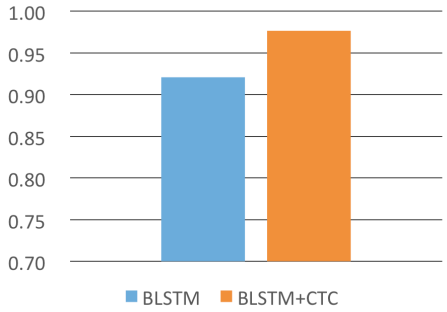


Fig. 8　実験結果・Precision