音響イベント検出における BLSTM-CTC を用いた弱ラベル学習法の有効性評価 *

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

1　はじめに

音環境理解による防犯監視システムや動画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，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 を用いる手法は少量のデータで推定モデルを学習することができる．しかし，線形処理であるために推定精度のさらなる改善は難しい．一方，NN を用いる手法は非線形処理であるため，より精密なモデル化が可能である． 特に BLSTM (Bidirectional Long Short-term Memory) を用いた手法 [4] は DCASE (Detection and Classification of Acoustic Scenes and Events) 2016 [5] における SED のタスクでトップレベルの性能を発揮した．しかし，これらの手法は推定モデルの学習に大量のデータが必要である．BLSTM を用いる手法における推定モデルの学習にはイベント音の種類，開始時刻，終了時刻の正解ラベル (強ラベル) が必要であるが，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多大なコストがかかるために極めて困難である．この問題を解決する一つの方法は，低コストでラベル付けが可能なイベント音の種類のみを含む正解ラベル (弱ラベル) を用いて学習を行うことである．

これまでに我々は，弱ラベルを用いて BLSTM を学習する手法 (BLSTM-CTC) を提案した [6]．これは，強ラベルを用いた学習の際に実行する正確な誤差計算の代わりに CTC (Connectionist Temporal Classification) [7] の損失計算を適用することにより，弱ラベルの使用を可能にした手法である．前報では，1 つのイベント音が単独で発生している条件下で有効性を示した [6]．本稿では，DCASE2016 Task 2 で提供されているデータセットを用いて複数のイベント音が同時に発生する条件下で有効性を検証する．

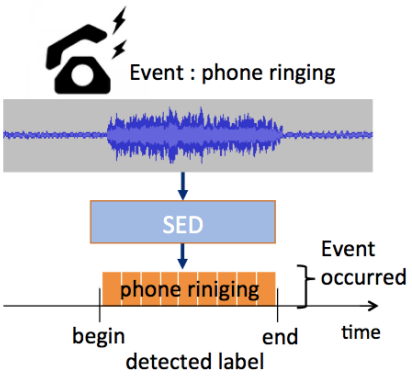

Fig. 1　SED の概要

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，出力層を時間フレームごとに並べた NN である．RNN では，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．この処理によって，前フレームとの相関の情報を反映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が大きく，長時間離れたフレームとの相関の情報が利用できない．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)[8] 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 を用いた SED 手法 [4] が提案され，高い推定性能を示している．

---

*Evaluation of 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)

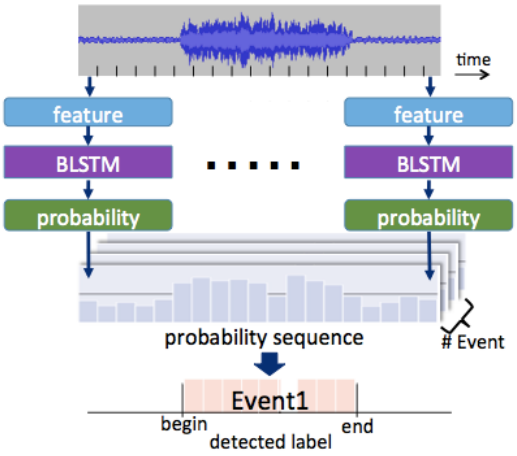

Fig. 2　BLSTM を用いた SED の概要

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強ラベルを用意する必要がある．このような強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での音響信号データの聴取，波形の確認などにより，イベント音の種類，開始時刻，終了時刻をラベル付けする必要がある．高い推定精度を得るためには学習に有効な大量のデータが必要になるが，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．そこで低コストでラベル付けが可能な弱ラベルを用いて学習を行う手法が必要とされている．

3　提案手法

3.1　BLSTM を用いた SED の概要

Fig. 2 に提案手法における BLSTM を用いた SED について示す．まず，音響信号に対して 25 ms のフレーム長，40%のオーバーラップで特徴抽出を行い，その特徴量を入力とする．特徴量の情報を BLSTM の隠れ層に伝播させていき，各フレームにおけるイベント音ごとの発生確率を BLSTM から出力する．これらの処理を全フレームで行い，得られた確率列に対し閾値計算を行うことにより，イベント音の種類，開始・終了時刻のラベルを推定する．

3.2　提案手法における BLSTM の学習方法

提案手法ではまず少量の強ラベル付きの学習データで BLSTM を初期学習し，次に初期学習済みの BLSTM を大量の弱ラベル付きの学習データで追加学習する．

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では，

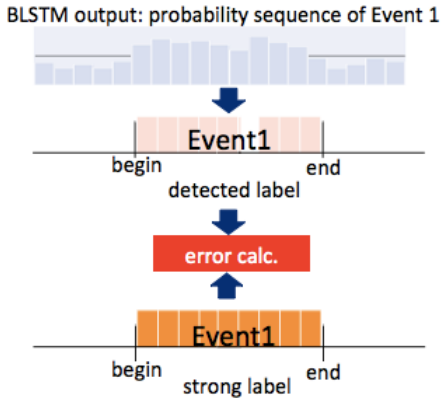

Fig. 3　強ラベルを用いた誤差計算

Fig. 3 に示すように推定されたラベルと強ラベルの誤差計算を行う．誤差計算は softmax-cross entropy によって行われ，その誤差が小さくなるように BLSTM のパラメータの更新を行う．

弱ラベル付き学習データを用いた追加学習

弱ラベル付き学習データを用いた追加学習では．Fig. 4 に示すように BLSTM の出力である確率列と弱ラベルの損失計算を行う．CTC の損失計算により，その損失が小さくなるように BLSTM のパラメータの更新を行う．以下で CTC の損失計算について説明する．

Fig. 4 において，横軸は時間を示し，縦軸は各フレームごとのイベント発生の有無を表す状態の遷移を示す．状態は「occurred」と「not occurred」と「blank」の 3 種類がある (図では「blank」の状態を省略して示している)．弱ラベルと一致する各状態遷移の確率が高くなるほど損失が小さくなり，低くなるほど損失が大きくなる．図 4 では，弱ラベルが表す音響信号内で発生しているイベント音の種類が「Event1」であることを示しているため，「Event1」が音響信号内で 1 回発生することが期待されている．一方，「Event1」に対する状態遷移は「not occurred」から「occurred」へ遷移し，再び「not occurred」の状態に戻っていることから，イベントが音響信号内で 1 回発生したことを表している．そのため，図 4 では損失が小さい状況を示している．これらの処理により正解として与えられた弱ラベルに対し，BLSTM の出力が適正になるように BLSTM の学習を行う．

4　提案手法の有効性の検証

4.1　実験条件

提案手法の有効性を示すため，評価実験を行った．実験には DCASE2016 Task 2 で提供されている評価

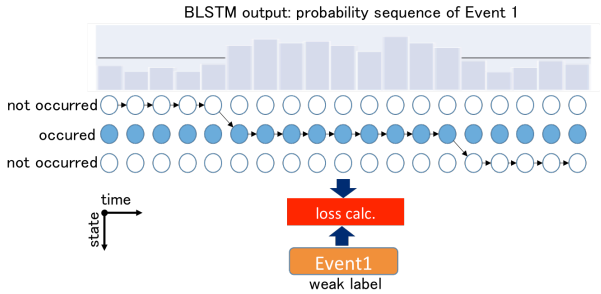

Fig. 4　弱ラベルを用いた損失計算

Table 1　実験条件 (BLSTM)

Learning rate	強ラベル学習: 0.0005 弱ラベル学習： 0.00001
Gradient clipping norm	強ラベル学習: 5 弱ラベル学習： 1
Batch size	強ラベル学習: 50 弱ラベル学習： 1
Epoch	強ラベル学習: 20 弱ラベル学習： 5
Hidden layer size	400
# of hidden layers	2

用，開発用データセットを使用した．開発用データセットには音響イベントごとに 20 個のクリーンなサンプルが用意されている．このクリーンなサンプルのみでは学習を行う上で十分ではないので，データ拡張によりサンプルから学習用データを生成した．データの生成は以下の 3 ステップで行う．1) 開発用データセットから 5 秒の背景雑音をランダムに切り出す，2) クリーンなサンプルをランダムに 2 つ選択する．3) 選択したサンプルを 5 秒の背景雑音に所定の SNR (−6，0，6 dB) で，ランダムな位置に重畳する．以上の処理によって，5 秒間の背景雑音内に 2 つのイベント音が発生してる音響信号を作成した．実験ではこの音響信号を使用する．その他の実験条件を Table 1 と Table 2 に示しておく．

実験では，以下の 4 つの手法において比較した．

(i) **Strongly/all**: 48400 個のデータ (全 20 サンプルから生成) で BLSTM を強ラベル学習したもの．

(ii) **Strongly/small**: 12400 個のデータ (5 サンプルから生成) で BLSTM を強ラベル学習したもの．これを提案手法の初期学習とする．

(iii) **(Proposed method) strongly/small + weakly/all**: (ii) で学習済みの BLSTM に対し，48400 個のデータからランダムに選択した 2200 個のデータ (全 20 サンプルから生成) で弱ラベルを用いて追加学習したもの．

Table 2　実験条件 (audio data)

Sampling rate	44100 Hz
SNR	−6, 0, 6 dB
feature	39 Mel-filter bank outputs
Frame size	25 ms
# of event class	11
# of learning data for initial learn	12600 (=11 sound events × 5 samples × 220 patterns)
# of learning data for additional learn	2200 (=11 sound events × 20 samples × 10 patterns)

(iv) **Strongly/small + strongly/all**: (ii) で学習済みの BLSTM に対し，(iii) と同じ 2200 個 (全 20 サンプルから生成) のデータで強ラベルを用いて追加学習したもの．

本実験では DCASE 2016 Task 2 の 120 秒のデータを使用した．評価時は，各データを 24 分割した 5 秒長のデータを用いた [9]．

4.2　実験結果と考察

まず，手法 (i) と提案手法 (iii) の性能と，DCASE2016 に提出された他の手法の性能を比較した結果を Fig. 5 に示す．図の縦軸は F 値，横軸は各手法をランキング順に並べたものである．手法 (i) は提案手法における BLSTM を用いた SED における上限性能を示している．図から手法 (i) が 5 位相当，提案手法 (iii) が 6 位相当であることがわかる．手法 (i) において，提案手法における BLSTM を用いた SED は Fig .5 の 3,4 位の手法を簡略化したものであるため，この結果から妥当な検出性能が得られていることがわかる．

次にラベル付けコストと検出性能の関係について考察を行う．弱ラベルは一度データを聞けばラベル付けすることが可能であるため，イベントがデータ内にランダムに配置されている際，平均するとデータの長さの半分の時間でラベル付けすることができる．一方強ラベルの場合，イベントの開始・終了時刻を正確にラベル付けする必要があり，一つのデータを何度も聞く必要があり，ラベル付けに弱ラベルより長い時間がかかる．実際に強ラベル付けを行った際，データ長の 10 倍の時間がかかったことが報告されている [10][11]．そのため，ここでは時間ベースで弱ラベルは強ラベルの 20 分の 1 のコストでラベル付けできると考える．

Fig. 6 に，縦軸をラベル付けコストとし，横軸を性

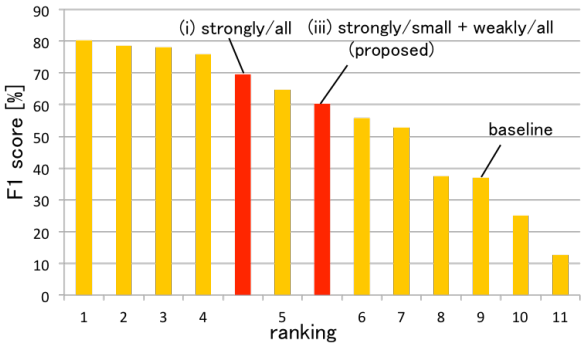

Fig. 5　DCASE2016 に提出された他手法と提案手法の比較

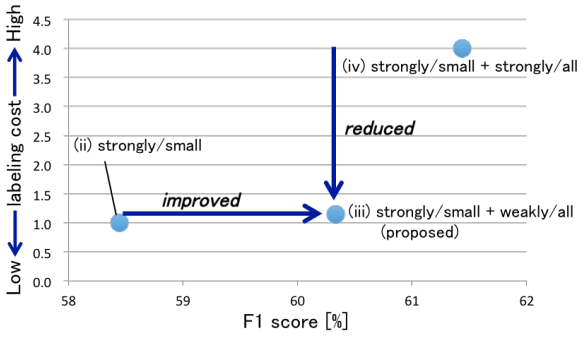

Fig. 6　F 値とラベル付けコストの関係

能 (F 値) として，提案手法 (iii) を手法 (ii)，(iv) と比較した結果を示す．これより，提案手法 (iii) はラベル付けコストのわずかな増加で，手法 (ii) より 1.9% の F 値向上を達成したことがわかる．また，手法 (iv) と比較すると，提案手法 (iii) は F 値が 1.3%劣るものの，ラベル付けのコストを大幅に $(95\% = 1 - 0.15/3)$ 削減していることが確認できる．

5　おわりに

本稿では，BLSTM-CTC を用いた弱ラベル学習法の有効性を，DCASE2016 Task 2 で提供されているデータセットを用いて複数のイベント音が同時に発生する条件下で検証した．実験結果から，提案手法による追加学習により，初期学習時に比べ，F 値が 1.9% 向上することが確認できた．また，提案手法と強ラベルを用いて追加学習を行った手法を比較すると，F 値が 1.3%劣るものの，ラベル付けコストが 95%削減できたことが確認できた．以上のことから，提案手法が有効であることを示した．

参考文献

[1] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, “Sound event detection in multi-source environments using source separation,” Proc. Workshop on machine listening in Multi-source 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,” Proc. IEEE International Joint Conference on Neural Networks, pp. 1-7, 2015.

[4] 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.

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

[6] 松吉大輝, 小松達也, 近藤玲史, 山田武志, 牧野昭二, ”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,” 日本音響学会講演論文集, pp. 63–66, March 2018.

[7] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” Proc. International Conference on Machine Learning, 2006.

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

[9] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” Applied Sciences, pp.1–17, 2016.

[10] T. Komatsu and R. Kondo, “Detection of anomaly acoustic scenes based on a temporal dissimilarity model,” Proc. Internal Conference on Acoustics, Speech Signal Processing pp.376–380, 2017.

[11] T. Komatsu, Tani, Takahiro Toizumi, Narisetty Chaitanya, Masanori Kato, Yumi Arai, Osamu Hoshuyama, Yuzo Senda and Reishi Kondo, “An acoustic monitoring system and its field trials,” Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–6, 2017.