音響イベント検出と位置推定における転移学習の効果の検証 *

☆陳軼夫，山田武志，牧野昭二（筑波大）

1　はじめに

近年，周囲の音環境から抽出した情報を高齢者の見守りや動画のタグ付けなどに活用するシステムへの関心が高まっている．このようなシステムでは，音響イベント検出（SED：sound event detection），すなわち周囲に存在する音響イベントの名称と開始･終了時間を検出する技術が重要な役割を担う．最近ではSEDと音響イベント方向推定（SEL:sound event localization）を同時に行うSELD（SEL＋SED）タスクが注目されており，例えばDCASE（Detection and Classification of Acoustic Scenes and Event）Challenge 2020ではTask 3として設定された [1].

このように関連性の高い2つのタスクを同時に取り扱う場合には，検出用モデルの学習にマルチタスク学習を適用することが考えられる [2]．これにより，個々のタスクに専用のモデルを学習する場合と比べて，検出精度と汎化性能が向上することが知られている．それに対して，Caoらは転移学習を導入し，DCASE 2019 Task 3において第2位の検出精度を達成した [3]．転移学習では，特定のタスクに対する学習済みモデルを他のタスクに転用することにより，限られたデータから高精度なモデルを得ることができる [4]．しかし，Caoらの手法では，2つのタスクのどちらを転移元とするのか，またネットワークのどの部分を転移対象とするのかについて，十分な議論がなされていなかった．

そこで本稿では，SELDタスクに対して広く用いられている，CNN（convolutional neural network）とBGRU（bidirectional gated recurrent unit）を組み合わせたモデルを対象とし，転移学習の効果的な適用方法を検証する．

2　SELDタスクへの転移学習の適用

2.1　ネットワーク構造

本稿では，Fig. 1に示すようなCNNとBGRUを組み合わせたモデルを用いてSELDを行う．ここで，CNNとBGRUは，SELタスクとSEDタスクに対して同一のネットワーク構造を有する．

まず，CNNでは，GCC-PHAT（generalized cross correlation phase transform）[5]により求めた空間スペクトログラムとmbe（mel band energy）分析により求めた時間周波数スペクトログラムを入力として，主にスペクトル構造を分析する．次に，BGRUでは，CNNの出力特徴量を入力として，主に時間構造を分析する．最後にdence層などを経て各タスクに対応する出力を得る．

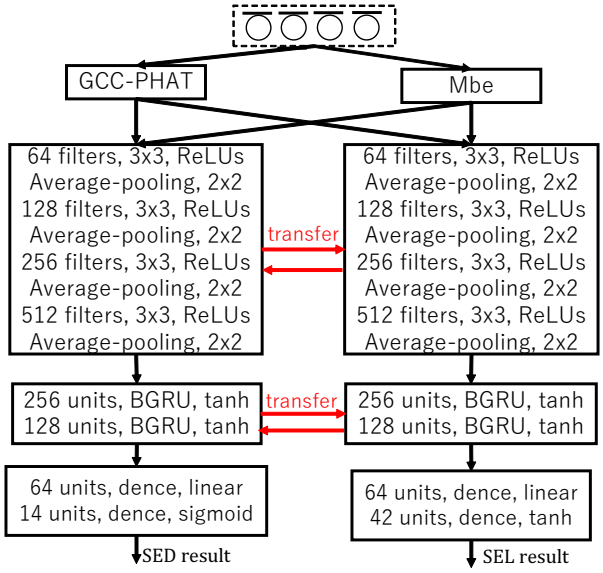

Fig. 1　Network configuration for SELD.

2.2　転移学習の適用

上述したように，転移学習では，特定のタスクに対する学習済みモデルを他のタスクに転用することにより，限られたデータから高精度なモデルを得ることができる．しかし，そのためには，転移元と転移先のタスクの関連性や転移の対象を適切に見定める必要がある．

そこで本稿では，SEDからSELへの転移学習とSELからSEDへの転移学習のそれぞれに対して，CNNのみ，BGRUのみ，両方を転移する場合の検出精度を比較する．具体的な組み合わせは次の通りである．

- SED → SEL（CNN）
- SED → SEL（BGRU）
- SED → SEL（CNN & BGRU）
- SEL → SED（CNN）
- SEL → SED（BGRU）
- SEL → SED（CNN & BGRU）

左は転移元のタスク，右は転移先のタスク，括弧内は転移対象のモデルを表す．なお，本稿では，転移したモデルを含めて転移先のネットワーク全体を学習する．

---

*Investigation on the effect of transfer learning in sound event localization and detection. by Yifu CHEN, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

Table 2　Experimental results.

Learning method	Development dataset				Evaluation dataset			
	SED task		SEL task		SED task		SEL task	
	$ER_{\leq 20^\circ}$	$F_{\leq 20^\circ}$	LE_{CD}	LR_{CD}	$ER_{\leq 20^\circ}$	$F_{\leq 20^\circ}$	LE_{CD}	LR_{CD}
No transfer	0.66	52.7%	25.6°	59.2%	0.60	55.0%	23.8°	63.2%
SED → SEL（CNN）	0.64	53.3%	24.2°	63.4%	0.58	54.8%	21.3°	65.2%
SED → SEL（BGRU）	0.65	53.1%	16.2°	73.2%	0.59	55.6%	15.3°	75.3%
SED → SEL（CNN & BGRU）	0.64	52.9%	15.5°	74.1%	0.59	55.2%	14.4°	76.3%
SEL → SED（CNN）	0.78	35.2%	25.1°	61.4%	0.71	44.2%	23.5°	63.1%
SEL → SED（BGRU）	0.58	55.4%	25.3°	60.3%	0.55	56.3%	23.4°	62.9%
SEL → SED（CNN & BGRU）	0.63	50.1%	24.9°	60.9%	0.57	55.8%	23.5°	62.7%

Table 1　Input features and network configurations.

Input features	GCC-PHAT	mbe
Input size (frame×order×ch)	256 × 12 × 6	256 × 40 × 4
Frame length	40 ms	
Shift length	20 ms	
Audio block size	256 frames	
Activation function	CNN: ReLU BGRU: tanh	
Optimization method	Adam	
Epoch	300	
Learning rate	0.001	

3　実験

3.1　実験条件

本実験では，DCASE 2020 Task 3[1] のデータセットのうち，4 チャネルのマイクロホンアレーで測定した室内インパルス応答を各種の音響イベントに畳み込んで生成されたデータを用いた．開発用データセットと評価用データセットの総時間長はそれぞれ 600 分，200 分である．検出対象となる音響イベントは，アラームや乳児の泣き声などの計 14 種類である．なお，サンプリング周波数は 24kHz，量子化ビット数は 24 である．入力特徴量とネットワークの詳細を Table 1 に示しておく．

本実験では，DCASE 2020 Task 3 と同じ評価尺度を用いて検出精度を評価する．SED タスクに対しては，20° の誤差範囲を許容した $ER_{\leq 20^\circ}$（error rate）と $F_{\leq 20^\circ}$（F 値）により評価する．また，SEL タスクに対しては，LE_{CD}（localization error）と LR_{CD}（localization recall）により評価する．完全に正しく検出できた場合，$ER_{\leq 20^\circ}$ は 0，$F_{\leq 20^\circ}$ は 100%となり，また LE_{CD} は 0°，LR_{CD} は 100%となる．

3.2　実験結果と考察

Table 2 に各学習手法の検出精度を示す．ここで，表中の No transfer は各タスクを独立に学習した場合である．

SED タスクの検出精度を見ると，SEL → SED（BGRU）が最も高い（赤字）．また，SEL タスクの検出精度については，SED → SEL（CNN & BGRU）が最も高く，SED → SEL（BGRU）がそれに迫っていることが分かる（緑字）．このことから，SED タスク，SEL タスク共に BGRU を転移すること，すなわち時間構造の分析を行うモデルをあらかじめ片方のタスクで学習することが効果的であることが分かる．一方，SED タスクにおいて，SEL → SED（CNN）の検出精度は No transfer よりも大きく低下していることが分かる（青字）．CNN は主にスペクトル構造の分析を担うが，SEL と SED とでは抽出すべき特徴が大きく異なるため，このような性能低下を招いたと考えられる．

4　おわりに

本稿では，SELD タスクに対して広く用いられている，CNN と BGRU を組み合わせたモデルを対象とし，転移学習の効果的な適用方法を検証した．その結果，SEL タスク，SED タスク共に BGRU を転移することが効果的であること，また，SED タスクにおいては CNN を転移すると性能低下を招くことが明らかとなった．

謝辞　本研究は JSPS 科研費 20K11880，19H04131 の助成を受けた．

参考文献

[1] http://dcase.community/challenge2020/task-sound-event-localization-and-detection.
[2] S. Adavanne et al. , “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” IEEE Journal of Selected Topics in Signal Processing, Vol. 13, Issue 1, pp. 34–48, 2019.
[3] Y. Cao et al., “Polyphonic sound event detection and localization using a two-stage strategy,” Proc. DCASE 2019 Workshop, pp. 30–34, 2019.
[4] S. j. Pan et al., “A survey on transfer leatning,” IEEE Transactions on Knowledge and Data Engineering, Vol. 22, Issue 10, pp. 1345–1359, 2010.
[5] M. Azaria et al., “Time delay estimation by generalized cross correlation methods,” IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 32, Issue 2, pp. 280–285, 1984.