畳み込みニューラルネットワークを用いた空間特徴抽出に基づく音響シーン識別の検討*

☆高橋玄, 山田武志, 牧野昭二 (筑波大)

1 はじめに

人間の行動や周囲の状況を自動認識しようとする取り組みがなされている．例えば，高齢者の見守りや動画への自動タグ付け，ライフログの収集などの応用システムが考えられている．これらのシステムを実現する要素技術の一つとして，環境音認識が注目されている．環境音認識は大きく二つに分けられる．一つは音響イベント検出，もう一つは音響シーン識別である．音響イベント検出は「いつ，何の音がしたか」を検出するものであり，音の種類としてはドアの開閉音，咳をする音，転倒音などの短い単発的な音が多い．一方，音響シーン識別は数秒から数十秒程度の長い音から録音された場所や状況を識別するものであり，音の種類としてはバスの中，オフィス，家などがある．本研究では音響シーン識別に焦点を当てる．

近年，音響シーン識別の研究が盛んに行われている．例えば環境音認識を対象とする DCASE 2017（Detection and Classification of Acoustic Scenes and Events 2017）Challenge が開催され，そこでは音響シーン識別のタスクが用意された [1]．DCASE 2017 では畳み込みニューラルネットワーク（Convolutional Neural Networks; CNN）を用いた手法がいくつか提案されたが（例えば [2][3]），中でも [3] では空間特徴を抽出することにより高い識別精度を達成した．この手法は，ステレオ入力信号に対して前処理を施し，左チャネル，右チャネルに加えて中央チャネル（左右チャネル加算），サイドチャネル（左右チャネル減算）等のそれぞれに対して独立に CNN を適用して識別を行う．

それに対して本稿では，前処理ではなく，CNN を用いて空間特徴を抽出する手法を提案する．これにより，識別に適した空間特徴を学習により獲得することが可能となる．提案手法では，空間特徴を抽出するために，時間・周波数領域に加えて時間・空間（左右チャネル）領域と周波数・空間領域に対して CNN を適用する．DCASE 2017 Challenge[1] における音響シーン識別のタスクを用いて提案手法における空間特徴抽出の有効性を評価する．

2 提案手法

2.1 提案手法の処理の流れ

Fig. 1 は提案手法の処理の流れである．提案手法では，DCASE 2017 に従いステレオ入力を想定している．また，入力信号の時間長は 10 s である．

まず左右チャネルそれぞれに対して 128 次元の対数メルフィルタバンク出力を計算する．ここでフレーム分析におけるフレーム長とフレームシフト長はそれぞれ 40 ms と 20 ms である．次に，この特徴量時系列を 10 フレームからなるブロックに分割して CNN に入力し，音響シーンごとに各ブロックの出力確率の対数和を求めることにより識別を行う．ここで，10 フレーム毎に分割するのは，時間的に大きく離れた特徴量は識別精度の改善にさほど寄与しないことが報告されていることによる [4].

次節では，CNN を用いた空間特徴抽出について述べる．

2.2 CNN を用いた空間特徴抽出

音響シーン識別では空間の情報を用いることが重要である．例えばモノラルチャネルからステレオチャネルに変えるだけで識別精度の改善に繋がる．しかし，従来音響シーン識別において，CNN は時間・周波数領域に対して適用されることが多かった．そこで提案手法では，時間・周波数領域に加えて，更に時間・空間領域，周波数・空間領域に対しても CNN を適用し，空間情報を含んだ特徴マップを抽出する．

Fig. 1 の上部，中部，下部はそれぞれ時間・周波数領域，時間・空間領域，周波数・空間領域に対する CNN の適用に対応している．また，青い四角は CNN を適用する際のフィルタを表している．例えば，時間・周波数領域に対する CNN では時間・周波数スペクトログラムを 1 枚の特徴マップとみなし，空間（ステレオ）を CNN のチャネルとみなしている．そのため，CNN のチャネル数は 2 になる．一方で，時間・空間領域に対する CNN では周波数を CNN のチャネルとみなすので，チャネル数はメルフィルタバンク出力の次元である 128 になる．時間・空間領域や周波数・空間領域に対する CNN の適用は，特徴マップ中のステレオチャネル間の空間情報を抽出することに

*Acoustic scene classification based on spatial feature extraction using convolutional neural networks. by Gen TAKAHASHI, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

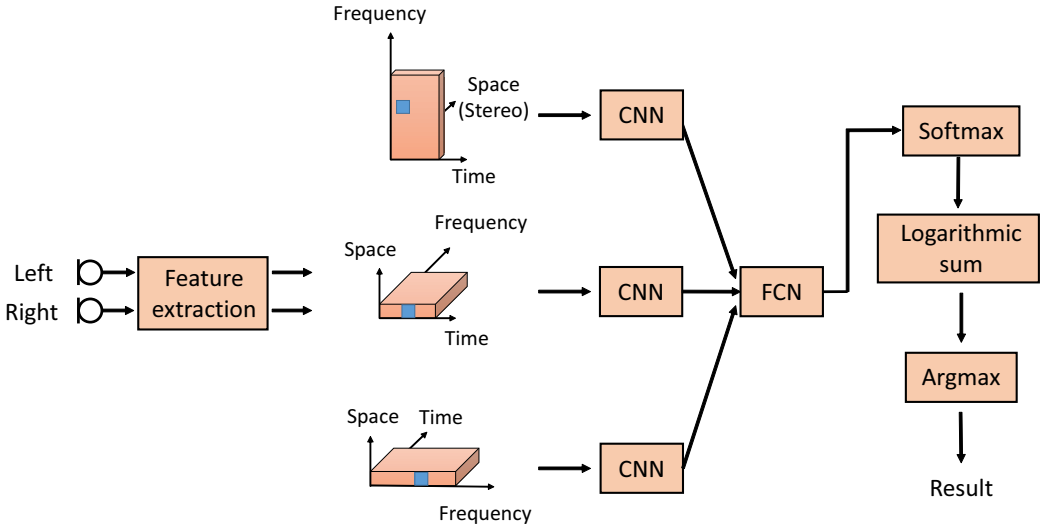

Fig. 1　Process flow of the proposed method.

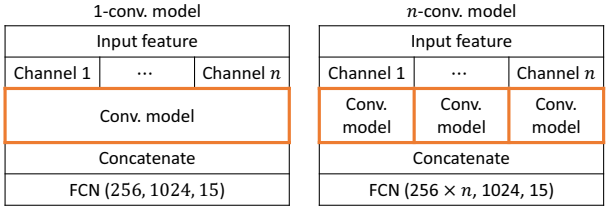

Fig. 2　Overview of the CNN of the proposed method.

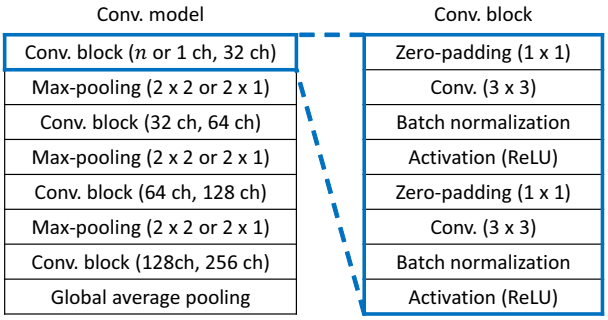

Fig. 3　Conv. model and conv. block.

相当している．

2.3　ネットワーク構造

提案手法では，文献 [3] の手法をもとに 8 層の畳み込み層を持つ CNN を用いる．Fig. 2 は提案手法で用いる CNN の概要を示している．まず，図の左側の 1-conv. model について説明する．1-conv. model では，n チャネルの特徴マップを入力とし，$1 \times 1 \times 256$ の特徴マップを出力するような Conv. model を適用する．Conv. model に関しては後述する．次に Conv. model の出力を連結し，256 次元の特徴量とする．最後に 1024 次元，15 次元の全結合ネットワーク（Fully Connected Neural Network; FCN）を適用し，出力確率を求める．次に n-conv. model について説明する．n-conv. model では，チャネルごとに独立に，1 チャネルの特徴マップを入力とし，$1 \times 1 \times 256$ の特徴マップを出力するような Conv. model を適用する．このため，各 Conv. model の出力を連結した特徴量の次元は $256 \times n$ となる．以降は 1-conv. model と同様である．n-conv. model は CNN の適用の際にチャネルを独立に入力するため，チャネルごとの特徴を抽出することができる．一方で，CNN 内のパラメータ数が 1-conv. model と比べて約 n 倍になる．

次に Conv. model の説明をする．Fig. 3 は Conv. model とその中で使われている Conv. block を表している．Conv. model は畳み込みを行う Conv. block と Max プーリングを繰り返し行う．Conv. block の括弧の中のコンマで区切られた 2 つの数字は左が畳み込み層の入力チャネル数で，右が出力チャネル数である．最初の Conv. block の入力チャネル数は，1-conv. model の場合は n，n-conv. model の場合は 1 となる．また，プーリング層の括弧の中の数字はプーリングを行う際のウィンドウサイズである．ここで，プーリング層におけるウィンドウサイズは時間・周波数領域の場合は 2×2 である．一方，時間・空間領域，周波数・空間領域の場合，空間方向の次元は 2 次元しかなく，プーリングをすると 1 次元になってしまうので，プーリング層のウィンドウサイズは 2×1 とする．最後のプーリング層では特徴マップ全体に対して Average プーリングを行

Table 1　Overview of the development dataset.

# of scenes	15
# of sound data	4680 (=15 scenes × 312 data)
data length	10 s
# of channels	2 (left and right)
sampling frequency	44.1 kHz
quantization bits	24 bit

う．よって，Average プーリングを行ったあとの特徴マップの次元は $1 \times 1 \times 256$（チャネル数）となる．

最後に Conv. block の説明をする．Conv. block ではまず，入力となる特徴量に対して幅 1 のゼロパディングを行う．次にサイズが 3×3 のフィルタを用いて畳み込みを行う．その後，バッチ正規化 [5] を行い，最後に活性化関数として ReLU 関数を適用する．これを 2 回繰り返したものが Conv. block である．

本稿ではそれぞれの領域に対して 1-conv. model と n-conv. model を用いて識別を行い，比較検討する．

3　提案手法の有効性の評価

3.1　実験条件

本章では，DCASE 2017 Challenge の開発用データセットを用いて，提案手法における空間特徴抽出の有効性を評価する．Table 1 は開発用データセットの概要を示している．音響シーンは 15 種類である（Table 4 参照）．各音響シーンには 312 個の音響データがあり，それぞれ 10 s のステレオ信号である．サンプリング周波数は 44.1 kHz，量子化ビットは 24 ビットである．

Table 2 は特徴量と CNN の条件を示している．特徴量は 128 次元の対数メルフィルタバンク出力で，フレーム分析におけるフレーム長とフレームシフト長はそれぞれ 40 ms と 20 ms である．フレーム長とフレームシフト長に関しては DCASE 2017 Challenge のベースラインで用いられているものと同じである．特徴量時系列を分割する際のブロックサイズは 10 フレームである．識別器の学習における最適化手法には Adam[6] を用い，学習時のエポック数は 20 とした．

識別器の評価は 4-fold クロスバリデーション（データオープン，音響シーンクローズド）による識別精度の平均で行った．4-fold クロスバリデーションにおけるデータの分割方法は DCASE 2017 Challenge で指定されたものと同様である．なお，提案手法の実装には Chainer[7] を用いた．

Table 2　Condition of the acoustic features and the CNN.

feature	128th-order log mel filter bank outputs
frame length	40 ms
frame shift length	20 ms
audio block size	10 frames (without overlap)
domain	time-frequency, time-space, frequency-space combination of 3 domains
# of conv. layers	8
optimization method	Adam
epoch	20

Table 3　Classification accuracies of proposed methods.

domain \ model	1-conv.	n-conv.
time-frequency domain	81.95	80.01
time-space domain	78.81	80.84
frequency-space domain	82.20	83.55
combination of 3 domains	81.69	84.14

本実験では，

- 時間・周波数領域のみの場合
- 時間・空間領域のみの場合
- 周波数・空間領域のみの場合
- 上記 3 つを組み合わせた場合

の 4 通りのそれぞれに対して，1-conv. model と n-conv. model の 2 通りのモデルを適用して識別を行う．

3.2　実験結果

Table 3 は領域・モデルの各組における識別精度を示している． 表の行は CNN を適用した領域で，列は CNN のモデルである．3 種類の領域ごとの識別精度を比較すると，周波数・空間領域（3 行目）を用いたときに識別精度が最も高い．これは周波数・空間領域に対して CNN を適用することによって，時間・周波数領域（1 行目）を用いたときよりもよい空間特徴を抽出できていることを意味している．一方で，同じように空間を含む時間・空間領域（2 行目）における識別精度では大きな改善は得られなかった．次に

Table 4　Classification accuracies for each sound scene of proposed methods.

domain	time-frequency		time-space		frequency-space		combination	
model	1-conv.	n-conv.	1-conv.	n-conv.	1-conv.	n-conv.	1-conv.	n-conv.
bus	89.74%	79.49%	76.92%	91.99%	86.86%	91.35%	84.94%	83.97%
café/restaurant	62.18%	51.28%	42.31%	38.46%	47.44%	37.18%	61.86%	47.44%
car	92.63%	97.12%	86.22%	88.14%	95.19%	97.44%	90.06%	96.79%
city_center	95.83%	85.26%	90.38%	96.15%	85.90%	95.19%	86.86%	82.37%
forest_path	95.19%	93.27%	78.85%	75.96%	96.79%	92.63%	92.95%	93.91%
grocery_store	70.83%	60.58%	77.24%	88.14%	77.56%	91.67%	70.51%	84.62%
home	75.79%	80.82%	78.62%	77.36%	75.16%	83.33%	77.04%	77.99%
beach	85.26%	70.51%	80.77%	80.77%	87.18%	89.42%	85.90%	83.01%
library	86.22%	77.56%	82.37%	77.56%	83.65%	83.01%	80.13%	88.14%
metro_station	95.83%	95.51%	97.12%	98.72%	98.08%	96.79%	93.59%	97.12%
office	91.03%	99.68%	98.08%	98.40%	91.99%	97.12%	99.04%	98.72%
residential_area	70.19%	86.22%	80.13%	76.60%	78.21%	78.85%	83.01%	84.62%
train	68.59%	70.51%	59.94%	65.71%	75.00%	69.23%	71.47%	78.85%
tram	82.37%	89.10%	85.58%	90.06%	86.86%	88.78%	92.63%	88.78%
park	67.63%	63.14%	67.63%	68.59%	67.31%	61.22%	55.45%	75.96%

モデルごとの識別精度を比較すると，空間を含む領域では改善が見られたが，時間・周波数領域では改善しなかった．これは時間・周波数領域でチャネルごとに独立に CNN を適用すると，計算の過程で空間の情報が失われてしまうためだと考えられる．最も識別精度がよかったのは 3 種類の領域を組み合わせ，n-conv. model を適用したときで，時間・周波数領域に 1-conv. model を適用したときと比較すると 2.19% 改善した．また，DCASE 2017 Challenge のベースラインの識別精度と比較すると，9.34%改善した．なお，ベースラインは 40 次元のメルフィルタバンク出力を入力特徴量とし，隠れ層が 2 層 50 ユニットの FCN である．

最後に，Table 4 は領域・モデルの各組における音響シーン別の識別精度である．オレンジ色のセルは各シーンにおける識別精度が最も高かった領域・モデルの組のセルである．時間・周波数領域と周波数・空間領域の列を比べると，オレンジ色のセルが周波数・空間領域の方が多い．特に grocery store や home などでは大きく識別精度が改善されているのが分かる．また，領域を組み合わせた場合は grocery store, home, residential area, train, park のように識別精度の低かったシーンが改善されている．一方で cafe / restaurant などでは大きく識別精度が下がっていたので，どのようなシーンで識別精度が下がるのかを調査し，それらを改善していくことが必要である．

4　おわりに

本稿では CNN を用いた空間特徴量抽出手法を提案した．DCASE 2017 Challenge の開発用データセットを用いて実験を行い，提案手法における空間特徴抽出の有効性を評価した．周波数・空間領域に対して n-conv. model の CNN を適用することにより，時間・周波数領域を用いたときよりも識別精度が向上することを確認した．更に，それぞれの領域を組み合わせることで，時間・周波数領域のみを用いた場合と比較して 2.19%の改善を確認した．

参考文献

[1] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj and T. Virtanen, “DCASE 2017 Challenge setup: Tasks, datasets and baseline system,” DCASE 2017 Challenge.

[2] S. Mun, S. Park, D. Han and H. Ko, “Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane”, DCASE 2017 Challenge.

[3] Y. Han and J. Park, “Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification,” DCASE 2017 Challenge.

[4] G. Takahashi, T. Yamada, N. Ono and S. Makino, “Performance evaluation of acoustic scene classification using DNN-GMM and frame-concatenated acoustic features,” Proc. APSIPA 2017, Paper ID 219, December 2017.

[5] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” ArXiv e-prints, February 2015.

[6] D. P. Kingma, J. L. Ba, “Adam: A method for stochastic optimization,” In Proceedings of International Conference on Learning Representations, 2015.

[7] `https://chainer.org/`