位相スペクトルを利用した CNN に基づく環境音分類の検討

Study of Environmental Sound Classfication with Convolutional Neural Network using Phase spectrum

松原　拓未

Takumi Matsubara

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本研究では，環境音分類に用いる特徴量に位相スペクトルを追加することによる分類結果への影響について検討した．位相スペクトルを追加するために CNN の構造を用いた分類ネットワークを作成し，位相スペクトルを用いない手法との分類結果の比較を行った．実験結果から位相スペクトルを利用することで「虫」「バイク」「サイレン」の環境音ラベルにおいて分類性能の向上が見られた．

1　はじめに

環境音分類とは，収録された環境音から音源の種類を自動で分類することである．環境音分類の分野では Neural Network (NN) によるアプローチが行われている．再帰的な構造を持った Recurrent Neural Network (RNN) を用いた手法 [1] や Convolutional Neural Network (CNN) と RNN を組み合わせた手法 [2] などが提案されている．

環境音分類では，一般に入力特徴量として振幅スペクトルのみが用いられている．振幅スペクトルを得るために用いられるフーリエ変換の結果からは，周波数ごとの位相を表す位相スペクトルも求めることができるが，環境音分類では，利用されていない．そのため，振幅スペクトルのみを用いた環境音分類では音響信号の位相情報が欠落している．このことから位相スペクトルを追加の特徴量として用いることで，分類器へ与えられる音響信号の情報が増え分類の精度が向上すると考えた．

本研究では，CNN に基づく環境音分類に用いる特徴量に位相スペクトルを追加することによる分類結果に与える影響について検討する．

2　音響特徴量

本研究では音響特徴量として振幅スペクトログラムと位相スペクトログラムを使用する．振幅スペクトログラム，位相スペクトログラムはそれぞれ振幅パワースペクトルと位相スペクトルを時間軸に沿って並べることで行列として表現する特徴量である．時間ごとの振幅パワースペクトルと位相スペクトルは入力信号に短時間フーリエ変換を適応した結果である周波数スペクトルから求められる．周波数番号を k, 分析フレームの時刻を t, 短時間ごとの周波数スペクトル $X_k(t)$ とすると振幅パワースペクトルと位相スペクトルは式 (1) と式 (2) でそれぞれ求められる．

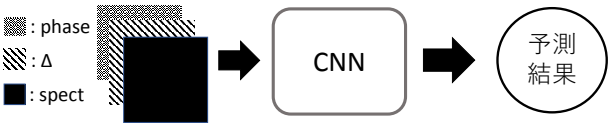


図 1: 環境音分類の概要

$$P_k(t) = |X_k(t)|^2 \tag{1}$$

$$\phi_k(t) = \arctan \frac{\Re X_k(t)}{\Im X_k(t)} \tag{2}$$

ここで，$\Re X_k(t)$ と $\Im X_k(t)$ はそれぞれ周波数番号 k における $X(t)$ の実部と虚部を表す．また，振幅スペクトログラムの時間微分を取ることで振幅スペクトログラムの動的特徴量を得ることができる．

3　CNN に基づく分類ネットワークの構成

本研究で用いる分類ネットワークの概要を図 1 に示す．分類ネットワークの構成は CNN の構造を利用したものであり，CNN は各入力特徴量を 1 つのチャネルとして複数の特徴量を同時に扱うことができる．入力された特徴量は各層で畳み込み (convolution) や プーリング (pooling) が行われる．畳み込みやプーリングは各入力特徴量から一定の領域ごとに値を取り出してそれぞれの処理を行うため，振幅スペクトログラムや位相スペクトログラムを入力として用いることで周波数方向と時間方向の情報を同時に扱うことができる．

分類ネットワークの構成は畳み込み層と最大プーリング層を交互に 3 層通過し，通過してきた出力ごとに平均プーリングを行うことでノード数を削減する．その後，2 層の全結合層を経て各ラベルごとの予測結果を出力する．収録された環境音に複数種類の音が含まれていることが考えられるので，あらかじめ設定した閾値を超えたものを予測されたラベルとして扱う．

4　評価実験

本実験は，入力特徴量の違いによる分類性能の比較を行う評価実験を行った．

4.1　使用データセット

本実験では Googl Nexus 7 を用いて Android 用アプリケーション「オトログマッパー」[4] により収録さ

れた環境音を用いる．収録条件はサンプリング周波数 32 kHz，標本化ビット数 16 bit，シングルチャネル，収録音の長さ 10 s で岡山大学周辺と岡山駅周辺で 14 名の協力者によって収録された．収録後に 14 種類の環境音ラベルが収録音に付与される．ラベル付与は 2 名のラベル付与者が分担して収録音を聴き行った．ラベルの付与された収録音の総数は 7107 であり，各ラベルごとにデータ数の偏りがある．

本実験では 14 種類のラベルから雑音ラベル 2 種類を除いた 12 種類の環境音を分類する．雑音ラベル 2 種類のみが付与された収録音は実験に使用しないため，実験に用いた収録音の総数は 6181 である．

4.2 実験条件

実験に用いる音響特徴量はメル周波数振幅スペクトログラムとその動的特徴量，メル周波数位相スペクトログラムであり，分類ネットワークに入力をする際に各特徴量を最大 1，最小 0 となるように正規化している．特徴量抽出の条件はフレーム長 40 ms，フレームシフト 20 ms でメルフィルタバンクのフィルタ数は 50 である．分類に使用する特徴量の組み合わせは提案手法である「全ての音響特徴量 (3ch)」と従来手法の「メル周波数振幅スペクトログラムとその動的特徴量 (2ch_d)」，位相のみを振幅スペクトルと組み合わせた場合を考慮した「メル周波数振幅スペクトログラムとメル周波数位相スペクトログラム (2ch_p)」の 3 パターンで実験を行った．評価は 10-fold cross validation を用いて行い，評価尺度は F-score (F)，Error Rate (ER) を用いる．各尺度の計算方法は以下の通りである．

$$ER = \frac{FP + FN}{N} \qquad F = \frac{2P \cdot R}{P + R} \tag{3}$$

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \tag{4}$$

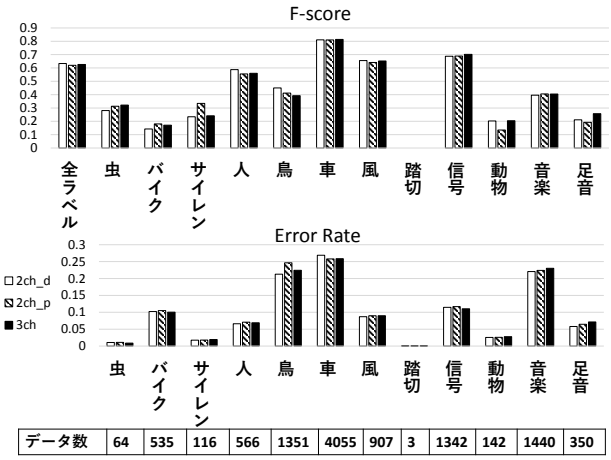
N は総収録音数，TP は正例に対して正の予測をした予測の正解数，FP は負例に対して正の予測をした数，FN は正例に対して負の予測結果をした数を示す．

環境音ラベルの予測結果を決定する閾値は F-score をできるだけ大きく，Error Rate をできるだけ小さくするように設定する必要がある．そのため各ラベルごとに閾値を [0.01，0.99] の値域で 0.01 刻みに変化させ式 (5) の Score を最大化する値を閾値と設定した．これにより式 (5) の条件で最も良い分類結果を得ることができる．

$$\text{Score} = F + (1 - ER) \tag{5}$$

4.3 実験結果

ラベル全体と各ラベルの分類結果を図 2 に示す．結果の値は各 fold の結果を平均して得られたものである．図よりラベル全体の結果では F-score と Error Rate には大きな変化が見られなかった．



データ数	64	535	116	566	1351	4055	907	3	1342	142	1440	350

図 2: 実験結果とラベルごとのデータ数

各ラベルごとの結果では，虫，バイク，サイレンのラベルで位相スペクトルを利用した 3ch と 2ch_p の F-score が 2ch_d の F-score より高い値を示している．特にサイレンの音では 2ch_p の条件で分類した場合に他の条件と比べて，F-score が 0.1 ポイント上昇しており，サイレンのような周期的な音を分類する際に位相スペクトルが有効だと考えられる．

5 まとめ

本報告では，位相スペクトルを用いた CNN に基づく環境音分類の検討について述べた．位相スペクトルを用いて環境音分類を行うことによりラベル全体での F-score の向上は見られなかった．しかし，バイクやサイレンのラベルの音では F-score の上昇が確認された．このことから位相スペクトルを追加することで分類性能が向上する環境音があることが分かった．今後の課題はラベル全体での F-score を向上させるために位相を追加することによる分類性能の低下を防ぐことである．

参考文献

[1] M. Zohrer, F. Pernkopf, “Virtual Adversarial Training and Data Augmentation for Acoustic Event Detection with Gated Recurrent Neural,” in *Proc. INTERSPEECH*, pp. 493-497, 2017.

[2] J. Guo, N. Xu, L. Li, A. Alwan, “Attention based CLDNNs for short-duration acoustic scene classification,” in *Proc. INTERSPEECH*, pp. 469-473, 2017.

[3] 李権俊, 滝口哲也, 有木康雄 “深層学習による位相情報を考慮した音声合成の検討,” 日本音響学会講演論文集， pp. 281–184，2-Q-19，2017.

[4] S. Hara, S. Kobayashi, and M. abe, “Sound collection systems using a crowdsourcing approach to construct sound map based on subjective evaluation,” IEEE International Conference on Multimedia Expo Workshops (ICMEW), pp. 1–6, 2016