Attribute Prototype Network を用いた音響イベントのゼロショット学習*

☆ LIN YIHAN, 高島遼一, 滝口哲也 (神戸大)

1 はじめに

音響イベント分類は，水が流れる音，足音，車の走行音といった音の種類を分類する技術で，高齢者や乳幼児の見守り [1, 2] や，機械の異常検知などの応用が期待されている．深層学習などの機械学習理論が大きく発展しており，環境音などあらゆる音の分析が可能になりつつある．一方，深層学習モデルを学習するためには多くのラベル付きデータが必要である．しかしこのタスクでは，例えば異常検知における異常データなど，イベントによっては学習データが入手困難という問題がある．一般的なクラス分類は学習されてないクラスを認識できない．この問題に対処する方法として，共通する情報を頼りに未知クラスを推定するゼロショット学習が研究されている．

ゼロショット学習は特に画像認識の分野において研究されており，Lampert らによる研究 [3] 以来様々な手法が提案されている [4, 5]．音声分野ではゼロショット学習の研究は少ないが，Huang らによる先行研究 [7, 8] ではクラスラベルの代わりに，Word2Vec [6] などによってクラスラベルの単語を埋め込んだベクトルを出力として音響埋め込みモデルを学習することで，ゼロショット学習を可能にしている．しかし，単語埋め込みによるクラスの表現方法は，各単語の意味的な近さは反映しているが，音の近さは反映していないため，音の分類においては不十分であると考えられる．

音の近さを反映したクラス表現を獲得するため，我々の先行研究 [9] では単語埋め込みの代わりに，音の高低などの属性情報を用いる手法を提案した．提案した属性情報はイベント間の音の近さを直接反映することが可能なため，従来の単語埋め込みを用いた手法よりも高いゼロショット分類性能が得られた．しかし，先行研究 [9] では，音響信号から属性情報を推定する精度が不十分であることが原因で，ゼロショット分類に失敗するケースも見られた．

本研究では，属性情報を用いるアプローチをベースに，定義した属性情報が音のスペクトログラム上のどこに反映されているかという局所情報を学習させることで，属性情報の学習を効果的に行い，それによりゼロショット分類性能を改善する手法を提案する．

2 属性情報を用いたゼロショット学習

本章では，以前我々が提案した手法 [9] について述べる．一般にゼロショット学習では，各クラスをクラスの意味空間へ変換し，クラスラベル空間ではなく意味空間上で分類を行う．画像のゼロショット学習を例とした場合,「シマウマ」というラベルに対して「馬の形」,「縞模様」といった意味空間へ変換することで，仮に「シマウマ」の画像そのものが学習データに存在しなくても，同じ意味空間の特徴を持つ画像（例えば「馬（馬の形）」や「トラ（縞模様）」）が学習データに存在していれば，それらから意味空間上の「シマウマ」を学習可能である

先行研究 [7, 8] では，意味空間として Word2Vec による単語埋め込み空間を使用していたが，前章で述べた通り，単語埋め込み空間は音の近さを反映していないという問題があった．そこで我々の先行研究 [9] では，音の特徴を直接表す意味空間として，音源の材質や音高といった属性情報を新たに定義し，それを意味空間として用いる手法を提案した．

属性は，例えば「音源は木材か否か」といった音源の材質に関する属性や,「高い音か否か」といった音高に関する属性,「同じパターンの音が繰り返されているか否か」など，計 16 種類の項目からなる．そして各音響イベントクラスの音に対して，16 種類の属性について当てはまっていれば 1，当てはまっていなければ 0 とすることで 16 次元のベクトルを作成し，それを各クラスの属性情報として定義した．

音響埋め込みモデルは，音のスペクトログラムを入力とし，その音が属しているクラスの属性情報ベクトルを出力するように，Binary cross entropy 基準によって学習する．

イベント分類時では，学習データに存在しない未知クラスの音を音響埋め込みモデルに入力することで，その未知クラスを説明する属性情報が出力される．未知クラスに関する正解の属性情報をあらかじめ辞書 (外部知識) として持っていることを前提とすると，音響埋め込みモデルの出力に対して，分類候補となる各クラスの属性情報とのユークリッド距離を計算し，距離が最小となるクラスを選出することで，未知クラスの分類を行う．

3 提案手法

先行研究 [9] では，属性情報を用いることで，単語埋め込みを用いるよりも高い精度でゼロショット分類が可能であることを確認した．しかし，音響イベントの種類によっては，属性情報の推定精度が不十分であることが原因で，イベント分類が高精度に行えないという課題があった．

本研究では，属性情報をより正確に推定することで，音響イベントのゼロショット分類の性能改善を目的とする．そのため，我々はスペクトログラムの局所情報に着目した．例えば「高い音か否か」を表す属性

*Zero-shot sound event classification using Attribute Prototype Network. by Yihan Lin, Ryoichi Takashima, Tetsuya Takiguchi (Kobe University)

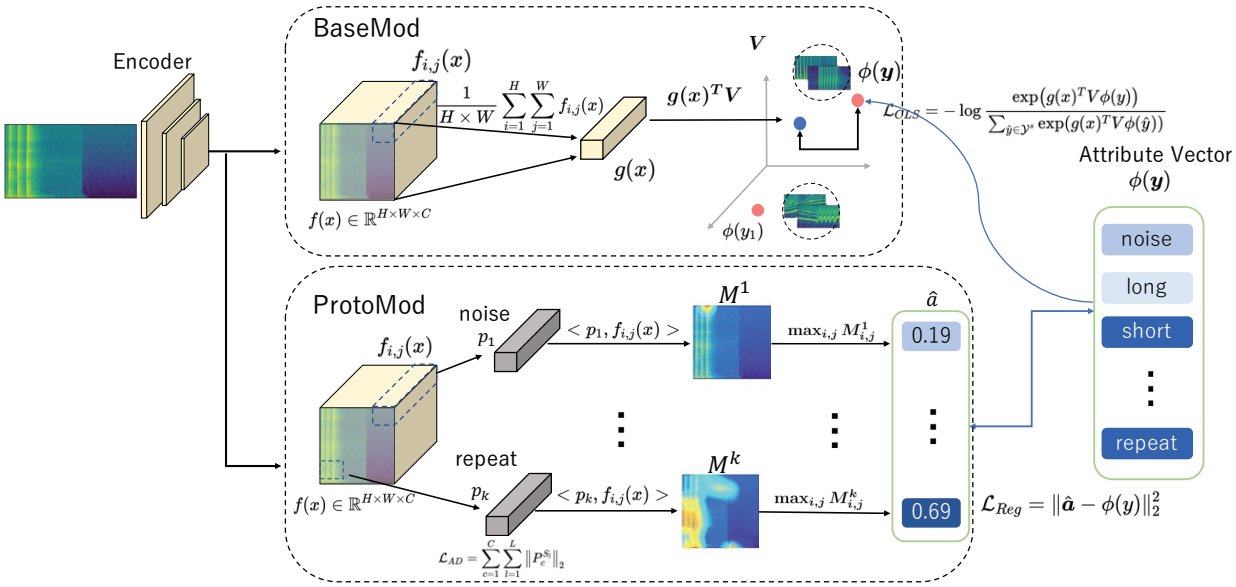

Fig. 1　Overview of the proposed method based on an attribute prototype network

は，スペクトログラム中の高周波成分が重要であり，「衝突音か否か」を表す属性は，衝突音が発生する時刻周辺の短時間領域が重要というように，属性の推定にとって重要な特徴は必ずしもスペクトログラム全体ではなく，局所的であると考えられる．そこで本研究では，属性とスペクトログラムの局所情報との関連を学習する手法として，Attribute Prototype Network (APN) [10] を使用した音響イベントのゼロショット学習手法を提案する．手法の概要を Fig. 1 に示す．APN モデルは画像認識の研究で提案されたモデルで，入力画像全体の特徴を考慮したモデルに加えて，属性と関連の強い局所特徴を考慮したモデルも導入している．文献 [10] では，画像の局所情報を考慮したモデルの導入により，属性情報の推定精度を改善し，その結果ゼロショット学習の性能が向上することを報告している．

APN モデルは，大きく分けて以下二つのモジュールがある．

Base Module (BaseMod)

BaseMod では，入力 x に対して全体特徴 $g(x)$ を計算し，全体特徴とクラス y 毎の属性ベクトル $\phi(y)$ との相関を計算することでイベント分類を行う．$g(x)$ を計算する Encoder の学習には，学習データに含まれるクラス（$\mathcal{Y}^s$）によって行われ，それにより，学習データに含まれない未知クラス（$\mathcal{Y}^u$）の分類を行う．

BaseMod の学習はラベル y とラベルに対応する属性ベクトル $\phi(y)$ を持つ学習データ x が与えられた時，以下の式 (1) で表される損失 $\mathcal{L}_{CLS}$ を最小化するように行われる．

$$\mathcal{L}_{CLS} = -\log \frac{\exp\left(g(x)^T V \phi(y)\right)}{\sum_{\hat{y} \in \mathcal{Y}^s} \exp\left(g(x)^T V \phi(\hat{y})\right)} \quad (1)$$

V は $g(x)$ と $\phi(y)$ の次元数を一致させるための，学習可能な射影行列であり，$g(x)^T V \phi(y)$ は全体特徴とクラス y の属性情報との相関を意味する．L_{CLS} では，正解クラスの属性情報との相関を高く，不正解クラスの属性情報との相関を低くするように学習を行う．

Prototype Module (ProtoMod)

BaseMod では全体特徴を出力するため，局所情報が失われている．そのため ProtoMod を用いることで，入力内の局所情報を考慮する．

ProtoMod では，属性 k ごとに，その局所特徴のプロトタイプ p_k が学習される．入力 x（本研究では短時間フレームごとに計算した対数メルフィルタバンク特徴の二次元行列を使用）に対して，メルフィルタバンクのビン i，フレーム j ごとに計算される局所特徴 $f_{i,j}(x)$ と属性ごとのプロトタイプ p_k との内積 $M^k_{i,j} = \langle p_k, f_{i,j}(x) \rangle$ によって局所特徴と属性との相関を表すヒートマップ $M^k_{i,j}$ を計算する．属性 k ごとに $M^k_{i,j}$ の最大値を並べたベクトルを $\hat{a}$ としたとき，$\hat{a}$ と属性ベクトル $\phi(y)$ が一致するように学習を行う．具体的には，以下の平均二乗誤差損失を最小化するように学習する．これにより，スペクトログラム上の局所領域と属性の相関が高くなるようにプロトタイプ p_k および Encoder が学習される．

$$\mathcal{L}_{Reg} = \|\hat{a} - \phi(y)\|_2^2 \quad (2)$$

例えば「高い音か否か」と「低い音か否か」は同じ「音高」に関する属性というように，属性には互いに関連あるものが存在する．そこで，属性のプロトタイプ p_k を学習する際に，関連のある属性同士はそれらのプロトタイプが相関を持つように学習する．具体的には以下の無相関損失を最小化する正則化項を加える．

$$\mathcal{L}_{AD} = \sum_{c=1}^{C} \sum_{l=1}^{L} \left\| P_c^{S_l} \right\|_2 \quad (3)$$

まず，属性を L 個のグループ $S_1, \ldots, S_l, \ldots, S_L$ に分割する．このとき，属性グループ S_l に含まれる属性のプロトタイプ $p_{k \in S_l}$ の c 次元目の値を並べたベクトルを $P_c^{S_l}$ とする．このとき，$P_c^{S_l}$ の l2 ノルムを最小

化することで，同一グループの属性は類似したプロトタイプになるように学習される．

モデルは以下の式 (4) を最小化しつつ，BaseMod と ProtoMod を最適化する．

$$\mathcal{L}_{APN} = \mathcal{L}_{CLS} + \lambda_1 \mathcal{L}_{Reg} + \lambda_2 \mathcal{L}_{AD} \qquad (4)$$

ここで，λ_1，λ_2 はハイパーパラメータである．

4 評価実験

4.1 実験条件

データセットとして RWCP-SSD [11] と ESC-50 [12] を使用した．RWCP-SSD は 105 種類の音響イベント (木板を叩く音，スプレーの噴射音，鈴の音など) が合計約 1 万ファイル含まれたデータセットである．一つの音声データには単一の音響イベントが含まれており，約 0.5~2 秒の長さになっている．ESC-50 は Freesound に登録されている音源から音響イベント分類に利用可能な音声を収集することで作成された，環境音・自然音のデータセットである．データセットの内容は大きな分類カテゴリーとして動物，自然 (音風景)，人間 (非言語音)，室内，室外の 5 つに分けられ，全 50 種類のクラスで，計 2,000 ファイルが用意されている．

音響特徴量として，全ての音声データはゼロ埋めにより 3 秒（298 フレーム）に長さを揃えた上で短時間毎に 40 次元の対数メルフィルタバンク特徴を計算することで得られた 298×40 の二次元特徴を使用した．

学習データと評価データの割り当てを Table 1 に示す．評価データは全て未知クラスによって構成され，分類候補もこの 4 クラスのみで構成される．すなわち 4 クラス分類タスクによる評価となる．

Table 1 Class labels in the training and test data

Training Data : 32 classes (5,320 data in total)
cherry, wood, bank, bowl, candybwl, coffcan, colacan, metal, pan, trash-box, case, dice, bottle, china, cup, pump, spray, claps, alarm, dryer, tear, particle, bell, coin, tambourine, shaver, clock, kara, maracas, raining, water-fall, pouring-water

Test Data : 4 classes (160 data in total)
coughing, glass-breaking, door-knock, siren

本研究の実験で使用したクラスに対する属性情報の例を Fig. 2 に示す．2 章で述べた通り，先行研究 [9] では 16 種類の属性を定義していたが，本研究では先行研究で用いていた RWCP-SSD に加えて，先行研究で使用していない ESC-50 を使用しており，このデータセットに対しては先行研究で定義した属性の一部 (「大音量」,「小音量」,「大音量から小音量に変化」,「小音量から大音量に変化」の 4 属性）の定義が困難であったため，これらの属性を削除した．代わりに，音の動作を表す「落下音」と「衝突音」,「多数の音源」の 3 属性を追加し，合計 15 個の属性の属性情報を作成した．

Table 2 Recognition accuracies [%] of four unknown acoustic events

	coughing	glass breaking	door knock	siren	Average
BaseMod	0.0	0.0	0.0	100.0	25.0
+ProtoMod($\mathcal{L}_{Reg}$)	2.5	50.0	70.0	65.0	46.9
+$\mathcal{L}_{Ad}$	0.0	52.5	67.5	60.0	45.0
Prior work [9]	82.5	17.5	0	52.5	38.1

本実験では，Encoder (Fig. 1 参照) として事前学習無しの ResNet101 [13] を採用し，モメンタムを 0.9 とする SGD を使用して学習を行った．エポック数は 300，初期学習率は $1e^{-3}$ とし，10 エポック毎に 0.5 倍ずつ減少させた．式 (4) のハイパーパラメータである λ_1 および λ_2 はそれぞれ 1.0，0.05 とした．

提案された損失関数の有効性を検証するため，本研究は三つの条件で実験を行う:

- $\boldsymbol{\mathcal{L}_{CLS}}$

  ゼロショット学習を行う部分の損失 $\mathcal{L}_{CLS}$ だけを使用し，全体特徴のみを学習させる

- $\boldsymbol{\mathcal{L}_{CLS} + \lambda_1 \mathcal{L}_{Reg}}$

  各属性と局所情報との関連を付けさせる損失関数 $\mathcal{L}_{Reg}$ と $\mathcal{L}_{CLS}$ を用い，二つの損失を学習させる．

- $\boldsymbol{\mathcal{L}_{CLS} + \lambda_1 \mathcal{L}_{Reg} + \lambda_2 \mathcal{L}_{AD}}$

  関連する属性間の相関を強める損失関数 $\mathcal{L}_{AD}$ を加え，三つの損失関数を線形結合して学習させる．

4.2 実験結果

提案手法と従来手法 [9] の比較を Table 2 に示す．認識の結果，BaseMod のみ使用した場合，全てのデータが siren に分類されたため，siren の分類正解率は 100%となったが，他のクラスの正解率は 0 であった．一方，損失関数 $\mathcal{L}_{Reg}$ を追加した場合はほとんどのクラスにおいて分類正解率の向上が確認できたが，三つの損失全てを使用した場合において，正解率の向上は見られなかった．実際，学習過程において L_{AD} の値は下がっておらず，L_{AD} はモデルの学習に寄与できていなかったものと見られた．これについてはパラメータ調整など，今後の課題とする．

ProtoMod によって計算された，局所特徴と属性との相関を表すヒートマップ（$M_{i,j}$）の例を Fig. 3, Fig. 4 に示す．右はスペクトログラムにヒートマップを重ねた画像である．Fig. 3 はクラス “coughing” の音から計算された局所特徴と，属性 “繰り返す音” との相関ヒートマップである．クラス “coughing” は繰り返す音として属性を定義しており，このデータは咳を 6 回繰り返している音である．Fig. 3 によると，6 回の咳のうち，2 回から 5 回の咳に相当する部分が，“繰り返し” の属性と関連が高いと判断されていることが示された．また，Fig. 4 はクラス “door-knock”

	highfreq	lowfreq	midfreq	long	short	middle	wood	metal	plastic	ceramic	repeat	noise	fall	collision	sources
cup	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0
particle	0	0	1	1	0	0	0	1	0	0	1	1	1	0	1
pump	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0
spray	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0
claps	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0
clock	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0

Fig. 2　Examples of attribute information

の音から計算された局所情報と，属性 “低音” との相関ヒートマップである．“door-knock” は低音のイベントとして定義されており，Fig. 4 によると，比較的低周波帯域かつ音の存在する場所が，“低音” の属性と関連が高いと判断されていることが示された．

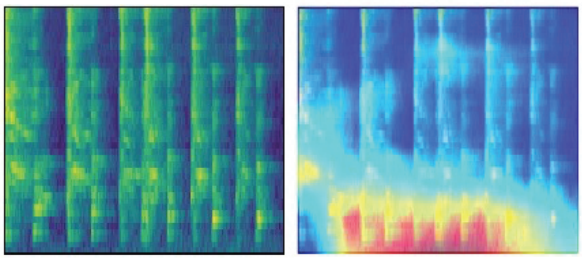

Fig. 3　A sample of the heatmap (left) for an attribute “repeat” calculated from a spectrogram

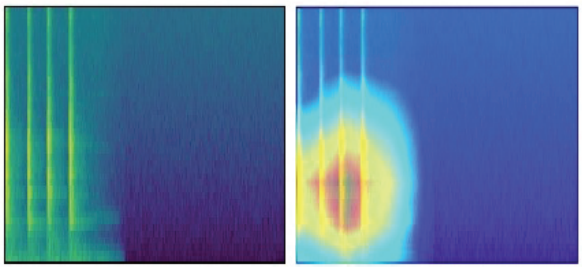

Fig. 4　A sample of the heatmap (left) for an attribute “lowfreq” calculated from a spectrogram (right) of “door-knock”

5　おわりに

本研究では，APN モデルと属性情報を用いた，音響イベント分類のゼロショット学習手法を提案した．従来手法では属性の推定精度が不十分という課題に対して，本研究では APN モデルを用いることで，スペクトログラムの全体特徴だけでなく，属性と関連の高い局所情報を考慮して学習を行う手法を検討した．実験の結果，推定された属性と局所情報との関連をつけることで，ゼロショット分類性能が向上することが確認できた．また，属性の相関ヒートマップから，属性と関連の高い局所情報をスペクトログラム上で確認できたことから，局所情報の有効性を確認した．今後はクラス分類性能を高めるための属性情報の改良法を模索する．

謝辞　本研究の一部は，JSPS 科研費 JP22K12168 の支援を受けたものである．

参考文献

[1] P. Guyot et al., “Water sound recognition based on physical models,” Proc. ICASSP, pp. 793-797, 2013.
[2] Y.-T. Peng et al., “Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models,” Proc. ICME, pp. 1218–1221, 2012.
[3] C. H. Lampert et al., “Learning to detect unseen object classes by between-class attribute transfer,” Proc. ICCV, 2009.
[4] W. Wang et al., “A survey of zero-shot learning: Settings, methods, and applications,” ACM Transactions on Intelligent Systems and Technology (TIST), no. 13, pp. 1-37, 2019.
[5] Y. Xian et al., “Latent embeddings for zero-shot classification,” Proc. CVPR, pp. 69-77, 2016.
[6] T. Mikolov et al., “Distributed representations of words and phrases and their compositionality,” Proc. NIPS, pp. 3111-3119, 2013.
[7] H. Xie et al., “Zero-shot audio classification via semantic embeddings,” IEEE/ACM Trans. Audio, Speech and Lang, no. 29, pp. 1233-1242, 2021.
[8] H. Xie et al., “Zero-Shot Audio Classification with Factored Linear and Nonlinear Acoustic-Semantic Projections,” Proc. ICASSP, pp. 326-330, 2021.
[9] Lin Yihan 他, “属性情報を用いた音響イベントのゼロショット学習”, 情報処理学会, 2022.
[10] W. Xu et al., “Attribute prototype network for zero-shot learning,” Proc. NIPS, no. 33, pp. 21969-21980, 2020.
[11] S. Nakamura et al., “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” Proc. LREC2000, pp. 965-968, May. 2000.
[12] K. J. Piczak “ESC: dataset for environmental sound classification,” Proc. ACM international conference on Multimedia (MM ’15), pp. 1015-1018, 2015.
[13] K. He et al., “Deep residual learning for image recognition,” Proc. CVPR, pp. 770-778, 2016.