音響イベントのゼロショット学習における属性情報の拡張*

☆ LIN Yi-Han, 高島遼一, 滝口哲也 (神戸大)

1 はじめに

音響イベント分類 (SEC) 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．

この問題に対処する方法として，少量のデータで音響イベントを分類する few-shot 学習が提案されている [3]．しかし，few-shot 学習を含む従来のイベント分類手法は，あらかじめ定義されたクラスに対してのみ分類を行うものであった．そのため，従来手法では学習データが一切存在しない未知のイベントを分類することはできない．

学習データが一切存在しないクラスを分類する方法として，ゼロショット学習 (ZSL) の研究が行われている．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 [4] 以来様々な手法が提案されている [5, 6]．画像認識における ZSL の代表的なアプローチとして，クラスの外観を表す視覚属性（例：クラス “シマウマ” は馬の形，白黒，ストライプなど）を用いて，視覚属性空間で画像を分類する方法がある [7, 8]．クラス名の情報の代わりに属性情報を使うことで，学習データの存在しないクラスを識別可能としている．

画像認識分野と比べて，音響イベント分類における ZSL の研究は少ない．先行研究 [9, 10] は意味的埋め込みを用いたゼロショット SEC の手法を提案している．それらの手法では，クラスラベルの代わりに，Word2Vec [11] 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．

我々の先行研究 [12] では，画像認識における視覚属性のように，クラスの音を直接表すことができる属性情報ベクトルを提案した．属性情報を用いることで，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．さらに，我々の研究 [13] 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) [14] 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し，性能向上を確認した．しかし，先行研究 [12, 13] では定義した属性の種類が少ないために，クラスを分類するための属性情報の不足により分類性能が低下するという課題があった．

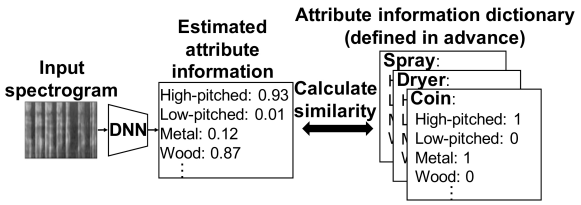

Fig. 1 Overview of sound event classification based on sound attribute vectors.

本研究では，属性情報を用いるアプローチをベースに，オノマトペを用いて属性情報を拡張することで，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

Fig. 1 に属性情報ベースの音響イベント分類システムの概要を示す．本システムでは，一般的な音響イベント分類で用いられるクラスラベル空間ではなく，音の属性で定義される意味空間において分類が行われる．音の属性は音のイベントを表現するものであり，当てはまっていれば「1」，当てはまっていなければ「0」を割り当てる 2 値ベクトルで表現される．例えば，クラス「コインの落下」に対して,「高音」「金属」「衝突」の属性は 1 に設定され,「低音」「木」は 0 に設定される．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類似度を計算することで，未知クラスの分類を行う．このように，学習データが無いイベントであっても，そのイベントの属性情報があらかじめ定義されていれば，識別することができる．我々の先行研究 [13] では，Attribute Prototype Network (APN) 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した．APN は，Xu ら [14] がゼロショット画像認識のために提案したネットワークであり，これを音響イベント分類に適用したものである．

Fig. 2 に APN モデルの概要を示す．APN モデルは，大きく分けて以下二つのモジュールがある．

*Extending attribute information for zero-shot learning of sound events. by Yihan Lin, Ryoichi Takashima, Tetsuya Takiguchi (Kobe University)

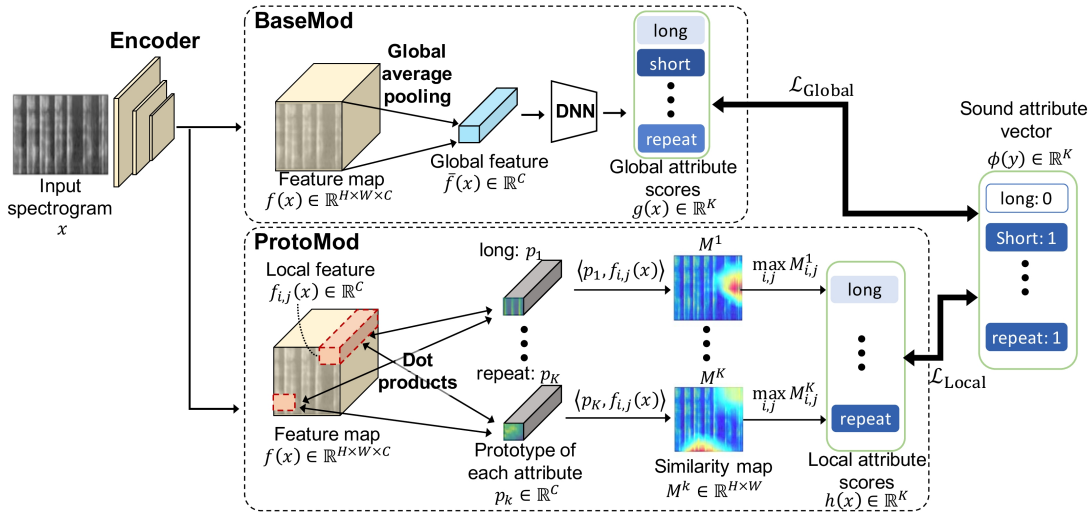

Fig. 2　Overview of the SEC system based on an attribute prototype network.

Base Module (BaseMod)

BaseMod は，各属性のスコアを推論するために，グローバル特徴を計算する．BaseMod は入力特徴マップ $f(x)$ に対して，時間-周波数空間軸上のグローバル平均プーリング（すなわち，$\bar{f}(x) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} f_{i,j}(x)$）を行って，グローバル特徴 $\bar{f}(x) \in \mathbb{R}^{C}$ を算出する．ここで，$f_{i,j}(x) \in \mathbb{R}^{C}$ は特徴量 $f(x)$ の位置 (i,j) における値である．次に，グローバルの属性スコア $g(x) \in \mathbb{R}^{K}$（K は属性数）を DNN モデルによって推定する（すなわち，$g(x) = \mathrm{DNN}(\bar{f}(x))$）．BaseMod とエンコーダの学習には softmax を最小化するように行われる．

$$\mathcal{L}_{\mathrm{Global}} = -\log \frac{\exp\left(\phi'(y)^T \mathrm{Tanh}(g(x))\right)}{\sum_{y'\in\mathcal{Y}^{\mathrm{seen}}} \exp\left(\phi'(y')^T \mathrm{Tanh}(g(x))\right)} \quad (1)$$

ここで，$\mathcal{Y}^{\mathrm{seen}}$ は学習データに含まれるクラスの集合である．$\phi(y)$ はイベントクラス y に対してあらかじめ定義されている属性ベクトルである．

Prototype Module (ProtoMod)

BaseMod のみを用いると，スペクトログラムの局所情報が失われる．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 は，各属性に対して学習可能なパラメータ p_k を持ち，これが各属性に対応する局所特徴の代表的なパターンを表現する．ある属性 k について，p_k と位置 (i,j) の局所特徴 $f_{i,j}(x) \in \mathbb{R}^{C}$ との内積を計算し，類似度を算出する．すべての位置について類似度を計算すると，類似度マップ $M^k_{i,j} = \langle p_k, f_{i,j}(x)\rangle$ が求まる．ここで，類似度マップの最大値 $max_{i,j} M^k_{i,j}$ を属性のスコアと定義する．全ての属性スコアを $h(x) = [\max_{i,j} M^1_{i,j}, \ldots, \max_{i,j} M^K_{i,j}]$ として定義し，平均二乗誤差損失を用いて，推定された属性スコアが学習クラスの属性情報と一致するように局所特徴を学習させる．

$$\mathcal{L}_{\mathrm{Local}} = \|h(x) - \phi(y)\|_2^2 \quad (2)$$

BaseMod と ProtoMod の両者を考慮した総合損失関数は，以下のように定義される．

$$\mathcal{L} = \mathcal{L}_{\mathrm{Global}} + \lambda \mathcal{L}_{\mathrm{Local}} \quad (3)$$

ここで，λ は $\mathcal{L}_{\mathrm{Local}}$ の重みである．

3　提案手法

先行研究 [13] では，用いた属性の種類が少ないため，異なるイベントクラスであっても，似たような属性ベクトルで定義されてしまうケースが存在した．このようなイベント同士は分類が困難であるため，認識率が低下する原因となっていた．

そこで本研究では，オノマトペを用いて属性の種類を拡張することで，異なるクラスで属性ベクトルが類似する問題を解消し，それにより分類性能を向上させることを提案する．オノマトペは音の特徴を自然言語を使用して表現したものであり，音高や音色の違いを表現できるため，音響イベントを表す属性情報として利用可能と考えられる．例えば，イベント「ガラス瓶を叩く」音のオノマトペは「ピン」，クラス「木を叩く」音のオノマトペは「トン」というように，叩く対象によって音の違いが文字として表される．本研究ではオノマトペに含まれる音素を属性情報とし，既存の属性情報に追加することを提案する．

本研究では RWCP-SSD 環境音データセット [15] を用いてイベント分類実験を行う．RWCP-SSD に収録された音についてオノマトペを定義したデータセットとして RWCP-SSD-Onomatopoeia [16] が存在する．このデータセットには Table 1 で示されるような「オノマトペの ID，オノマトペ，信頼度スコアを付与した作業者 ID，信頼度スコア」が順に定義されており，各音声ファイルにつき複数のオノマトペのラベル

Table 1　Samples in RWCP-SSD-Onomatopoeia

bell1 000.acc	0271_7,ch i r i N r i r i N,0326,5 0271_7,ch i r i N r i r i N,0299,3 ...
bell1 001.acc	0900_1,ch i r i N q r i r i q,0408,4 ... 0808_2,p i r i p i r i:,0617,1
...	...
bell1 049.acc	... 1004,1004_2,r i N ch i N r i N,4 1004,1004_3,r i N ch i r i N,2

が付与されている．

本研究で提案するオノマトペを用いた属性情報ベクトルでは，音素ごとに，そのイベント音のオノマトペでよく使われる音素であれば 1，そうでなければ 0 の値を取るバイナリベクトルをイベントクラスごとに定義する．まず Table 1 の情報を用いて，イベントクラスごとに各音素の信頼度スコアの合計値を計算する．例えば表中の 1 行目では，使用されている音素 /ch/，/i/，/r/，/N/に対して信頼度スコアを 5 とし，2 行目では同じ音素の信頼度スコアを 3 とする．このように各行で登場する音素の信頼度スコアを得た後，音素ごとに信頼度スコアを総和する．その後，総和値の最大値で割ることで最大値を 1 に正規化する．正規化した値が 0.5 以上であれば 1，0.5 未満であれば 0 として，最終的なバイナリベクトルを作成する．

4　評価実験

4.1　実験条件

3 章で述べた通り，本研究では環境音データセット RWCP-SSD およびオノマトペデータセット RWCP-SSD-Onomatopoeia を使用して実験を行った．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．RWCP-SSD-Onomatopoeia は RWCP-SSD に含まれる 105 種類の環境音に対して，計 155,568 個のオノマトペをカタカナと英文で付与したデータセットである．

テストデータとして，bowl (金属製のボウルを金属棒で叩く音)，clock2 (電子目覚まし時計の音)，kara (ガラガラを振る音)，maracas (マラカスを振る音)，ring (ハンドベルの音)，tambouri (タンバリンを振る音) の 6 クラス，計 543 環境音データを未知クラスとして選定する．学習データとしては，前述の 6 クラスを除く 62 クラス，計 5,647 環境音データを既知クラスとして用いた．

音響特徴量として，全ての音声データはゼロ埋めにより 1 秒（100 フレーム）に長さを揃えた上で短時間毎に 80 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 80×100 の二次元特徴を使用した．

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. 3 に示す．3 章で述べた通り，先行研究 [13] では 15 種類の属性を定義していたが，本研究ではさらにオノマトペ情報を追加して使用する．本来音素は 39 種類存在していたが，3 章で述べた方法で音素毎の属性値をバイナリ値にした結果，全てのクラスで値が「0」となる，つまりオノマトペとして全く使用されていないとみなされる音素が複数あった．これらの音素を削除した結果，使用した音素は 11 種類となり，従来の 15 種類の属性情報と合わせて 26 種類の属性情報を定義した．

本実験では，先行研究 [13] で使用した APN モデルを踏襲する．Encoder としては VGGish [17] を採用し，モデルを一から学習させ，学習済みのモデルは使用しなかった．グローバル平均プーリング層の次にある DNN は，ReLU 活性化を持つ 4,096 ノードの線形層からなる 2 つの中間層と，26 ノードの出力層から構成されている．

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

	bowl	clock2	kara	maracas	ring	tambouri	Average
previous [13]	83.9	15.4	21.0	99.3	64.1	55.1	56.6
proposed	99.5	0.0	34.5	99.3	86.1	91.5	67.1

4.2　実験結果

提案手法と従来手法 [13] の比較を Table 2 に示す．認識の結果，オノマトペを属性情報に導入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，クラス clock2 について，全てのデータが誤認識され，clock2 の分類正解率は 0 であった．

Table 3 は属性ベクトルの推定性能を示す．ここでは，Sigmoid($g(x)$) を用いて，閾値を 0.5 として各属性のバイナリ値を検出し，検出結果を事前に定義された属性情報 $\phi(y)$ と比較して評価した．従来法は 15 種類の属性情報の検出精度，提案法ではさらにオノマトペを追加した 26 種類の属性情報の検出精度を表しているが，これらの検出精度はほぼ同等であった．これらのことから，オノマトペの情報は従来の属性情報と同程度に検出できてることが伺える．

図 4 は ProtoMod で計算したテストデータの類似度マップ M^k を示している．未知クラス「clock2」に対して，音素の属性 “i” (上) を推定する際，ProtoMod は電子目覚まし時計が音を発する瞬間に注目する傾向がある．既知クラス「bells5(自転車ベルを鳴らす)」に対して，音素の属性 “撥音 N” (下) を推定する場合，ProtoMod はベルが鳴らされた瞬間を注目するだけではなく，音が伸ばす時間にも注目する傾向があった．類似度マップの妥当性に関する考察は今後の課題とする．

class	highfreq	lowfreq	midfreq	long	short	middle	wood	metal	plastic	ceramic	repeat	noise	fall	collision	many	a	o	i	k	r	N	q	ch	p	i:	a:
ring	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0
coins1	0	0	1	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0
bells5	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0
clap1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
dice2	0	0	1	0	1	0	0	0	1	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0

Fig. 3　Example of binary attribute information of some classes in our experiments.

Table 3　Attribute detection performance.

	Precision	Recall	F1score
previous [13]	0.62	0.59	0.60
proposed	0.62	**0.61**	**0.61**

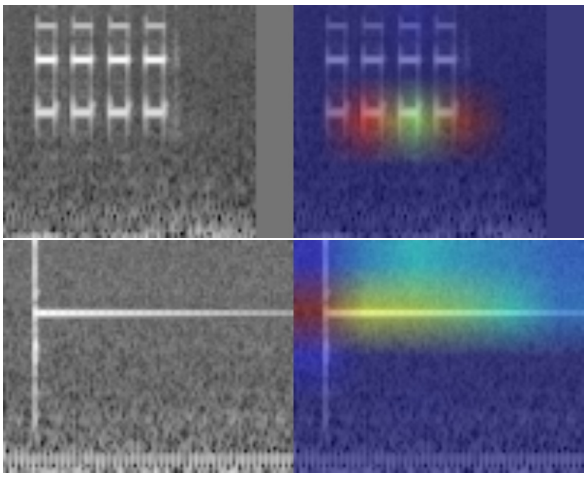

Fig. 4　Input mel-spectrograms (left) and their similarity map M^k for attributes k calculated in ProtoMod (right); top: unseen event "clock2" and its map for "/i/", bottom: seen event "bells5" and its map for "/N/".

5　おわりに

本研究では，APN モデルと属性情報に基づく，属性情報をさらに拡張するため，環境音についてのオノマトペを導入する手法を提案した．従来手法では分類に必要な属性が不十分という課題に対して，本研究ではオノマトペを用いることで，クラスの中で出現率が高い音素を活用し，環境音を詳しく表現する手法を検討した．実験の結果，属性情報の種類が増加し，かつそれらが従来の属性情報と同程度の精度で検出できることで，ゼロショット分類性能が向上することが確認できた．今後は属性情報をさらに正確に認識するため音響埋め込みモデルなどの改良について検討する．

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. Chou et al., "Learning to match transient sound events using attentional similarity for few-shot sound recognition," Proc. ICASSP, pp. 26-30, 2019.

[4] S. Y. Lampert et al., "Learning to detect unseen object classes by between-class attribute transfer," Proc. ICCV, 2009.

[5] 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.

[6] Y. Xian et al., "Latent embeddings for zero-shot classification," Proc. CVPR, pp. 69-77, 2016.

[7] Y. Xian et al., "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," IEEE transactions on pattern analysis and machine intelligence, vol.41, no. 9, pp. 2251-2265, 2018.

[8] L. Chen et al., "Zero-shot visual recognition using semantics-preserving adversarial embedding networks," Proc. CVPR, pp. 1043-1052, 2018.

[9] H. Xie et al., "Zero-shot audio classification via semantic embeddings," IEEE/ACM Trans. Audio, Speech and Lang, no. 29, pp. 1233-1242, 2021.

[10] H. Xie et al., "Zero-Shot Audio Classification with Factored Linear and Nonlinear Acoustic-Semantic Projections," Proc. ICASSP, pp. 326-330, 2021.

[11] T. Mikolov et al., "Distributed representations of words and phrases and their compositionality," Proc. NIPS, pp. 3111-3119, 2013.

[12] Y. Lin et al., "Binary attribute embeddings for zero-shot sound event classification," Proc. GCCE, pp. 13-14, 2022.

[13] Lin Yihan 他, "Attribute Prototype Network を用いた音響イベントのゼロショット学習", 日本音響学会, 2022.

[14] W. Xu et al., "Attribute prototype network for zero-shot learning," Proc. NIPS, no. 33, pp. 21969-21980, 2020.

[15] 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.

[16] Y. Okamoto et al., "RWCP-SSD-Onomatopoeia: Onomatopoeic Word Dataset for Environmental Sound Synthesis," Proc. DCASE, pp. 125-129, 2020.

[17] S. Hershey et al., "CNN architectures for large-scale audio classification," Proc. ICASSP, pp. 131-135, 2017.