5S-05

視聴覚自己教師あり学習に基づく音響イベント検出

大田 竹蔵[1,2]　坂東 宜昭[2]　井本桂右[3,2]　大西 正輝[2,1]

[1]筑波大学　[2]産業技術総合研究所　[3]同志社大学

1. はじめに

どの種類の音響イベントがいつ発生しているかを推定する音響イベント検出は，見守りシステムやスマートスピーカなど様々な応用における基盤技術である．音響イベント検出では，深層学習に基づく手法が高い性能を達成しており，畳み込みニューラルネットワーク (CNN) [1]，再帰型 NN (RNN) [2]，畳み込み再帰型 NN (CRNN) [3] に基づく手法が提案されている．これらの学習には，混合音に含まれる音響イベントの種類を表すラベルと開始・終了時刻が付与された教師データが用いられる．

音響イベント検出の学習における課題の 1 つに，アノテーション済み学習データの構築に要する膨大なコストが上げられる．特に持続音などでは開始・終了時刻の判断が難しい場合も多く，音響タギングなど隣接タスクと比べても多くの時間を要する．本問題を解決するため，開始・終了時刻を殆ど必要としない，弱教師あり学習 [4] に基づく枠組みが研究されている．この枠組みでは，ネットワーク構造の帰納バイアスを活用することで，音響イベントのラベル情報のみからその発生時刻を学習させる．弱教師あり学習に基づく枠組みは，アノテーションコストを大幅に削減でき広く研究されているが，本研究ではさらに少量の教師情報から学習できる枠組みを目指す．

人手による教師データを用いない枠組みとして，自己教師あり学習（SSL）が研究されている．SSL は，教師ラベルのない大量のデータからデータ自身を教師として学習する枠組みであり，音声認識では高い性能が報告されている [5]．さらに，音響信号だけでなく動画情報も用いる視聴覚 SSL（AV-SSL）は，音響イベントの特徴について音響信号単体では得づらい情報を視覚情報から補うことができる．音響イベントは映像上の物体の変化や移動に伴って発されることも多く，音色が酷似した音響イベントの弁別に有効であると期待できる．視聴覚対応学習（AVC）や視聴覚同期学習などが提案されており，音響シーン分類で高い性能を達成している [6,7]．

本研究では AVC により得られる音響特徴抽出器を音響イベント検出へ応用する (図 1)．従来の AVC では，音響イベントの識別に主眼が置かれており，発生・終了時刻の推定が必要な音響イベント検出には不向きだった．そこで，弱ラベル学習における帰納バイアスを活用して，従来の AVC に応用する．動画配信サービスで収集された映像データセットを用いて提案法を評価した．

2. AV-SSL に基づく音響イベント検出

教師ラベルが無くとも音響特徴抽出器を学習できる AVC に基づき，音響イベント検出を事前学習する．AVC では，音響信号と画像のそれぞれの特徴を抽出する 2 つの抽出器と，得られた特徴量を統合する出力層を学習する (図 1)．提案法は，AVC の音響特徴抽出器の時間統合層を工夫することで，帰納バイアスに基づき音響イベントの発生時刻が推定されるよう学習する．

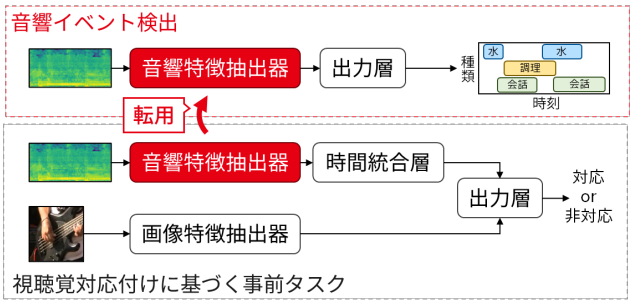

図 1: AV-SSL による音響イベント検出の事前学習

2.1 音響・画像特徴抽出器

音響特徴抽出器 $f^{(a)}$ は，入力音響信号の対数メルスペクトログラム $\mathbf{A} \in \mathbb{R}^{F\times T}$ を入力として，一定の時間間隔で特徴量 $\tilde{\mathbf{z}}_k^{(a)} \triangleq [\tilde{z}_{k1}^{(a)}, \ldots, \tilde{z}_{kD}^{(a)}]^\mathsf{T} \in \mathbb{R}^D$ を出力する．

$$\{\tilde{\mathbf{z}}_1^{(a)}, \ldots, \tilde{\mathbf{z}}_T^{(a)}\} = f^{(a)}(\mathbf{A}) \tag{1}$$

ここで，F と T はメルスペクトログラムの周波数ビン数と時間フレーム数を表し，$k = 1, \ldots, K$ は特徴量の時間インデックスを表す．得られた特徴量系列は，時間方向の最大値プーリングを用いて音響信号全体の埋め込み $\mathbf{z}^{(a)} \triangleq [z_1^{(a)}, \ldots, z_D^{(a)}] \in \mathbb{R}^D$ に変換する．

$$z_d^{(a)} = \max_k \tilde{z}_{kd}^{(a)} \tag{2}$$

最大値プーリングを用いることで，音響イベントの発生区間を学習できるようになる．音響特徴抽出器 $f^{(a)}$ は，具体的には文献 [8] の CNN10 として構成する．

画像抽出器 $f^{(v)}$ は幅 W で高さ H の入力画像 $\mathbf{V} \in \mathbb{R}_+^{W\times H}$ から埋め込みベクトル $\mathbf{z}^{(v)} \in \mathbb{R}^D$ を出力する．

$$\mathbf{z}^{(v)} = f^{(v)}(\mathbf{V}) \tag{3}$$

画像特徴抽出器 f_v は，具体的には文献 [6] の 8 層の VGG 型の CNN を用いる．

2.2 視聴覚対応学習

AVC では，2 つの特徴抽出器 $f^{(a)}$ と $f^{(v)}$ を，一つの動画内の画像フレームと音響信号に共通に含まれる情報をそれぞれ抽出するよう学習する．具体的には，正例として同一の動画から得た画像と音響信号の組を，負例として異なる動画から得た組を準備し，2 つの特徴抽出器で得られた特徴量から正例か否かを予測するよう学習する．

この学習では，2 層の全結合層から成る出力層 g を用いて，2 つの特徴量抽出器の出力 $\mathbf{z}^{(a)}$ と $\mathbf{z}^{(v)}$ から，入力が正例である確率 $\hat{y} \in [0, 1]$ を予測させる．

$$\hat{y} = g(\mathbf{z}_a, \mathbf{z}_v) \tag{4}$$

Audio-Visual Self-Supervised Learning for Sound Event Detection: *T. Ohta, Y. Bando, K. Imoto, M. Onishi*

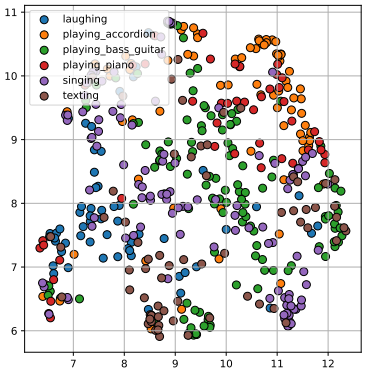

図 2: 次元削減した埋め込みベクトル

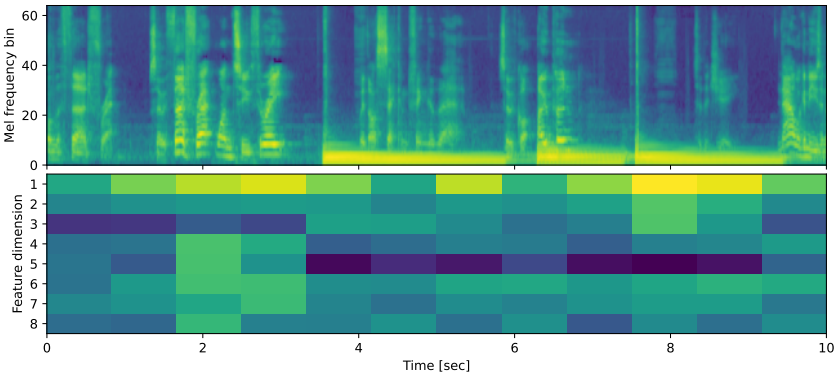

図 3: 入力信号のメルスペクトログラムと得られた埋め込みベクトル列

正解ラベル $y \in \{0, 1\}$ を用いて，以下の交差エントロピーを損失関数として DNN を学習する．

$$\mathcal{L} = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \tag{5}$$

ここで，y は，入力の画像フレームと音響信号が同一の動画から得たもの (1) か否 (0) かを表す．

3. 評価実験

提案法により，音響特徴抽出器が音響イベントの種類と発生区間を学習できたかを評価した．

3.1 データセット

動画配信サービス YouTube の動画から構成される Kinetics 400 データセット [9] を用いた．本データセットは，人間の行動に関する動画データセットで，道具の使用や楽器演奏，人間間の相互作用に関する，400 種類のラベルが付与された動画が収録されている．人物が映っている動画が多いため，音と画像の共起性が高く AVC の学習に適している．約 30 万個の動画が収録されており，その中から 10 秒の動画を抜粋して実験に用いた．約 20 万個を学習用データに，約 2 万個を検証用データに，約 3 万個を評価用データに利用した．

3.2 実験条件

入力音響信号は 10 秒間の動画からランダムに 1 秒間の区間を切り出して生成し，入力画像は同じ区間のフレーム 1 枚をランダムに選択して生成した．DNN の更新ごとに切り出す区間をランダムに変更することで，学習データの多様性を確保した．また，学習を安定させるためにミニバッチ内の正例と負例の比は常に 1:1 として学習をした．入力音響信号は 32 kHz にリサンプリングし，メルスペクトログラムに変換した．短時間フーリエ変換は窓長を 1024 サンプル，ホップ長を 320 サンプルとし，メルビン数を 64 とした．入力画像は，256×256 の元画像をデータ拡張として 224×224 にランダムにクロップして画像特徴抽出器に入力した．ただし，推論時は同じ大きさで中心をクロップした．埋め込みは $D = 512$ 次元とした．学習には学習率 1.0×10^{-4} の Adam を用いた．バッチサイズは 256 とし，学習エポック数は 400 とした．また，重み 1.0×10^{-5} の荷重減衰を適用した．

3.3 実験結果

時間統合層が出力する 512 次元の埋め込みベクトルを，UMAP [10] を用いて 2 次元に次元削減して可視化した結果を図 2 に示す．texting や playing_accordion, singing など，いくつかの音響イベントは，埋め込みがクラスタを形成していることが変わる．本結果より，AVC により学習した音響特徴抽出器が，音響イベントの種類が判別できる埋め込み空間を学習できていることを示している．

入力信号のメルスペクトログラムと得られた埋め込みの出力のうち値の大きい順に 8 次元を抜粋したものを図 3 に示す．この信号では，クリップ全体を通して音声が発されながら，4 秒付近から楽音が生じている．楽音の発生に対応し，3 と 5 次元目の値が大きく変わっており，これらの次元が楽音に関する特徴を表していると考えられる．以上より，本手法で学習した音響特徴抽出器は，音響イベントの発生区間を学習できていることが分かる．

4. おわりに

弱教師あり学習における帰納バイアスに基づいた視聴覚自己教師あり学習による，時系列方向の情報を含む特徴表現の学習手法を提案した．埋め込みベクトルの可視化結果から，音響イベントの種類，発生区間の情報を学習していることが確認した．今後は，本手法により獲得した特徴表現の音響イベント検出での性能を評価する．

謝辞: 本研究の一部は，NEDO の支援を受けた．

参考文献

[1] N. Takahashi *et al.* Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition. In *Proc. Interspeech*, 2982–2986, 2016.
[2] P. Giambattista *et al.* Recurrent neural networks for polyphonic sound event detection in real life recordings. In *Proc. IEEE ICASSP*, 6440–6444, 2016.
[3] C. Emre *et al.* Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM TASLP*, 25(6):1291–1303, 2017.
[4] K. Anurag *et al.* Audio event detection using weakly labeled data. In *Proc. ACMMM*, 1038–1047, 2016.
[5] B. Alexei *et al.* wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*, 12449–12460, 2020.
[6] Relja A *et al.* Look, listen and learn. In *Proc. ICCV*, 609–617, 2017.
[7] K. Bruno *et al.* Cooperative learning of audio and video models from self-supervised synchronization. In *Proc. NeurIPS*, 7774–7785, 2018.
[8] K. Qiuqiang *et al.* PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM TASLP*, 28:2880–2894, 2020.
[9] K. Will *et al.* The kinetics human action video dataset. *arXiv*, 2017.
[10] M. Leland *et al.* UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2018.