Vol.2020-MUS-128 No.12
2020/8/25

深層学習を用いた音をクエリとする類似効果音検索システムの基礎的検討

小宮 寛季[1,a)]　小坂 直敏[1,b)]

概要：効果音の検索を行う際，求める音を正確に言語化することは不可能であり，ゆえにテキストによる検索は困難である．そこで我々は CNN ベースの効果音分類モデルを用いて，クエリとなる音と検索対象の持つ特徴量をスペクトルから取り出し，それを基に効果音同士の類似度を算出し，3D 空間上に可視化することで，容易に効果音の検索が行えるシステムを提案する．システムの実装に先立って，50 クラスの効果音データセットをモデルに学習させたところ，精度は 83.40%となった。また，そのモデルにより音データから 128 次元の特徴ベクトルを抽出し，ユークリッド距離による類似度評価が行えることを確認した．

Basic Study of Searching Sound Effects by Audio Query Powered by Convolutional Neural Network

1. はじめに

音響効果（音効）は，映画・テレビ番組等の映像に，BGM や効果音を付加する工程である．この音効の制作において，ユーザは目的の音を膨大なデータベースから検索する必要がある．しかし，所望の効果音を手早く検索することは困難である．その理由としては，主に次の 2 つが挙げられる．

- 欲する音のイメージを言語化すること自体が，本質的に難しい．
- アノテーションの記法が統一されていない．外国製のデータベースを組み合わせる場合は尚更である．

つまり，テキストによる検索には限界があり，またユーザの側にも，言語化することが困難なイメージを，テキスト以外の形で検索したいという需要がある．

これに対して，相川ら [1] は，音を感性表現に基づく 8 次元の特徴ベクトルに変換し，そのベクトルの類似度を用いて音を検索する手法を提案した．ユーザは検索したい音の特徴をプルダウンメニューで選ぶことにより，平均して 60 %ほどの精度でイメージ通りの音を探すことができる．本研究では更に，この特徴ベクトルを，深層学習を用いて自動的に抽出することを目指す．

1 東京電機大学大学院
Graduate School of Tokyo Denki University
a) 19fmi15@ms.dendai.ac.jp
b) osaka@mail.dendai.ac.jp

また，画像処理の分野においては， 株式会社 ZOZO が提供している WEAR[2] など，畳み込みニューラルネットワーク（CNN）による画像分類モデルを特徴抽出器として用いることによる，画像をクエリとする類似画像検索が実用化されている．

さらに当研究室では，メルスペクトログラムを CNN に学習させる実験を行い，縦方向・横方向それぞれに一次元の畳み込みを行うことにより，より音データに適した分類モデルを構築できることを示している．[3]

よって本研究では， CNN ベースの深層学習モデルを用いて，テキストラベルではなく，音データそのものを検索クエリとして類似効果音の検索を行うシステムを提案する．

2. システムの機能概要

当研究室で開発していたツールである電子音色辞書 [4] を新たに実装しなおし，以下 5 つの機能を実装する．

(1) 効果音のリスト表示
(2) タグの自動付与
(3) タグによる検索
(4) プレイリスト機能
(5) 効果音の類似度の空間上での可視化

この中で，(2) は今回新たに実装する機能である．GUI の開発には GTK+3 を，類似度を算出する深層学習モデルの実装には TensorFlow を用いる．

ⓒ 2020 Information Processing Society of Japan

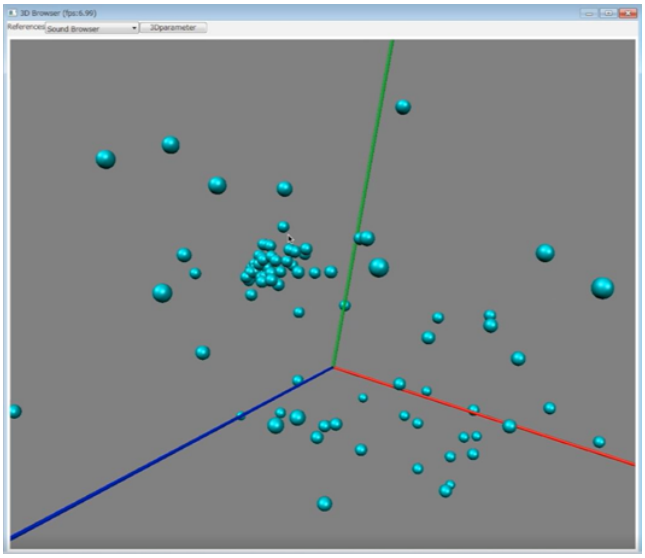

図 1　音色空間ブラウザ
Fig. 1　3D Timbre Browser

効果音の類似度の可視化のイメージを図 1 に示す．特徴量の次元圧縮を行った上で，音色の特徴を表すたかだか 10 次元程度の多次元空間（音色空間）の中に，一つの音を一つの粒子として表示している．聴感上似ている音同士は，空間内で近い位置に配置される．空間の x/y/z 軸はそれぞれ別の特徴量を表しており，ユーザは多次元の中から，任意の次元軸に対して自由に特徴量を割り当てることができる．

2.1　分類と検索の相違点

検索においてユーザは，クエリに類似している上位 N 個の音について，どの程度近いのかをシステムの音色空間の表示を見ながら検討する．これを実現するためには，分類・認識問題のように最尤のカテゴリを表示するのではなく，上位 N 個の候補データを適切に音色空間上に配置する必要性があり，そのために，カテゴリ認識のみではなく，音の類似度を抽出する必要がある．

3.　提案手法

深層学習による音の類似度計算，およびそれを基にした効果音検索システムを実現するための手法について以下に述べる．

3.1　リスト表示

GTK+3 を用いて実装した，効果音のリスト表示画面のプロトタイプを図 2 に示す．ユーザは左上のボタンからファイルもしくはフォルダを選ぶことで，音ファイルをリストに追加することができる．現段階ではまだ実装できていないが，このときシステムは音の特徴量を抽出し，それに基づいて自動的にタグを付与する．

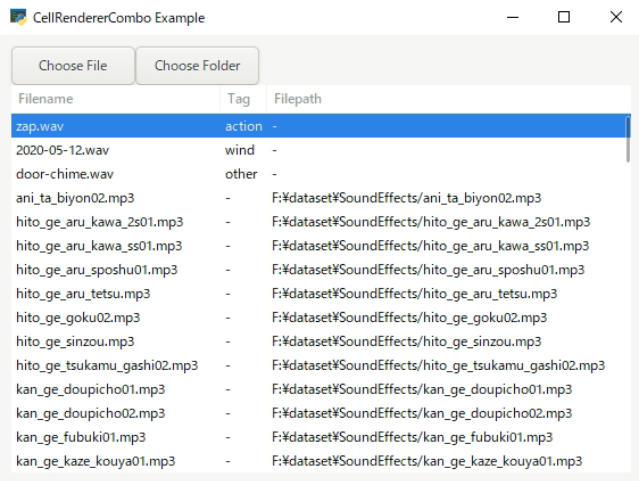

図 2　効果音のリスト表示のプロトタイプ
Fig. 2　A prototype of timble list window

3.2　SSDH モデル

画像処理の分野においては，深層学習を利用してデータの類似度を算出する際，分類モデルの推論時の隠れ層の重みを取り出し，それを特徴ベクトルとして類似度計算を行うことが多い．

Yang ら [5] は，CNN ベースの画像分類モデルに，隠れ層の出力が 0 または 1 の 2 値となるように損失関数を追加することで，高速に特徴ベクトル同士の距離計算を行える Supervised Semantics-preserving Deep Hashing(SSDH) モデルを提案した．また，モデルの学習時には，後述する転移学習を用いている．このモデルを参考に，効果音の分類モデルを構築して類似度計算を行う．

3.3　転移学習

分類モデルを学習させる際，各クラスの効果音を大量に含むデータセットと違い，本番環境においては，特定のクラスに属するデータを大量に用意することは難しい．そのような場合，画像処理の分野においては，転移学習という手法をとる．

これは，高性能なモデルに多クラス・多量のデータセットを学習させたものを事前に用意しておき，その学習済みモデルの，出力層及びその直前の層以外の重みを固定した状態で，再び少量のデータを学習させるという手法である．十分に学習した分類モデルの浅い層は，汎用的な特徴抽出器の役割を果たすため，その部分を再利用して学習させることにより，学習データが少量であっても分類モデルを構築することができる．

ただし，モデルを特徴抽出器として利用する場合には，ターゲットとなるデータの特徴が，元のモデルの学習に用いられたデータセットと類似している場合に限り，転移学習を行わずとも学習済みモデルを特徴抽出器として転用できる．[6]

ⓒ 2020 Information Processing Society of Japan

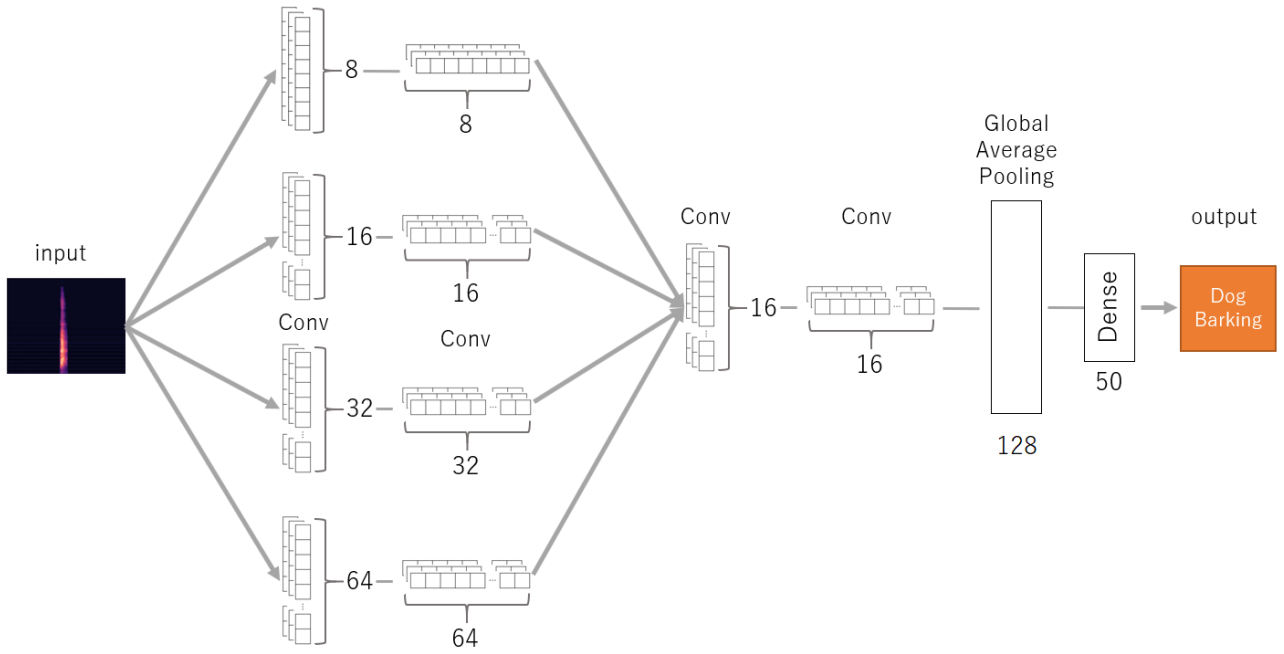

図 3 モデルの構造
Fig. 3 Model Architecture

画像処理分野では，20,000 以上のクラスからなる大規模な画像データセットである ImageNet[7] を学習させた高性能なモデルが多数提供されており，それを流用して転移学習を行うことが多い．音声処理分野においても AudioSet[8] という 527 クラスからなる音のデータセットが公開されているが，それを用いた学習済みモデルは提供されていない．そのため，本研究にあたって新たにモデルを構築する．

4. 実験：分類モデルの構築

汎用的な特徴抽出器を構築するために，CNN ベースのモデルに中規模な効果音のデータセットを与え，効果音分類モデルを学習させる実験を行う．

4.1 データセット

50 種類の効果音が各クラス 40 個ずつ，計 2000 個のファイルが収録されたデータセットである ESC-50[9] を用いる．各データはサンプリング周波数 44.1 kHz，量子化 16 bit，長さ 5 秒の wav ファイルである．

表 1 に ESC-50 に含まれる音のカテゴリ一覧を示す．このデータセットには，減衰音から持続音まで多岐に渡る音が収録されており，またドアが軋む音（Door, wood creaks）やガラスの割れる音（Glass breaking）などの，効果音としてよく用いられる音が含まれているため，効果音の特徴抽出器を構築するにあたって有用であると考えられる．

4.2 モデルの構造

図 3 にモデルの構造を示す．入力はメルスペクトログラム，出力は識別結果のラベルである．

入力を 8，16，32，64 とフィルタサイズが異なる 4 つの並列な畳み込み（Conv）層に分割し，縦方向，横方向それぞれに一次元の畳み込みを行って足し合わせ，更にフィルタサイズ 16 で縦と横に畳み込みを行い，128 次元の Global Average Pooling 層に通し，50 次元の Dense 層を通して，識別結果を出力する．

各 Conv 層のフィルタの数は 32 枚であり，Conv 層の活性化関数は ReLU，出力の活性化関数は Softmax である．また，特徴ベクトルの抽出のために，Global Average Pooling 層からバイパスを設けている．これにより，学習後のモデルを用いて音データから 128 次元の特徴ベクトルを抽出することができる．

4.3 実験条件

データセットは nfft = 1024，hop length = 256，mels = 128 のメルスペクトログラムに変換し，学習用 1600，評価用 400 に分割した．更にノイズ付加，時間シフト，時間伸縮およびそれらの組み合わせにより Augumentation を行い，学習用データを 5 倍の 8000 に増量した．最適化関数には AMSGrad[10] を用いて，パラメータは学習率 10^{-5}，学習率の減衰率 10^{-6}，バッチサイズ 32，エポック数 1000 とし，4 台の GTX-1080Ti を搭載した環境で実験を行った．

4.4 結果と考察

学習の結果， 評価データでの精度は 83.40 %となった．TensorBoard を用いて表示した学習曲線を図 4 に示す．縦軸は精度，横軸はエポック数であり，橙色の線が学習データ，青色の線が評価データでの精度を表している．

表 1　ESC-50 に含まれる音データのカテゴリ一覧
Table 1　Content categories of ESC-50

Animals	Natural & water	Human (non-speech)	Interior	Exterior
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footstep	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insect (Flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clockt tick	Fireworks
Crow	Thunderstorm	Drinking	Glass breaking	Hand saw

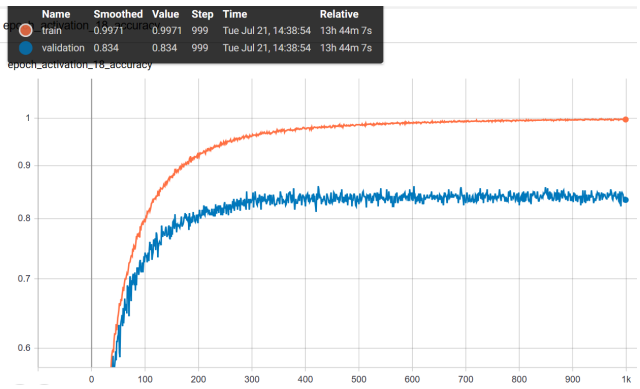

図 4　モデルの学習曲線
Fig. 4　Learning curve of the model

ESC-50 を提供している Piczak が行った実験によると，データセット内の音を人間に分類させた場合の正解率は 81.30 %であるため，今回学習させたモデルの精度は妥当であると言える．

5.　実験：類似度計算

ここでは，ESC-50 とターゲットとなる効果音データに関連があると想定して，転移学習を行わずにモデルをそのまま特徴抽出器として用いる．

前章で構築したモデルを用いて，学習時に用いたものとは異なる猫の鳴き声を入力し，特徴ベクトルを抽出させたところ，平均して 0.003〜2.149 の範囲の値を持つ，128 次元の密なベクトルを抽出できることが確認された．この特徴ベクトルから類似度計算を行い，その有用性を検証する．

5.1　実験条件

以下に示す 10 種類の音を入力として，各々の類似度を算出する．ベクトルの距離指標としては，単純なユークリッド距離を用いる．フォーマットは ESC-50 のデータと同じ，44.1 kHz, 16 bit, 5 秒の wav ファイルに揃えている．

(1) 猫の鳴き声 1
(2) 猫の鳴き声 2
(3) 牛の鳴き声
(4) 車のエンジン音
(5) ドアを閉める音
(6) 炎のエフェクト
(7) 旗が風に揺れる音
(8) 足音（コンクリート）
(9) 笑い声
(10)剣の打ち合い

この内，(1) と (2) の音色は聴感上ほぼ同一である．また (6) は人工的に作られた効果音であり，ESC-50 の中には該当するクラスが存在しない．(10) も同じく，学習データのカテゴリには当てはまらない．時間構造の観点からみると，(1) (2) (5) (9) (10) は発音区間の短い減衰音という共通点がある．

5.2　結果と考察

類似度（ベクトル間の距離）の算出結果を 2 に示す．() 内の数字は，前述の 10 種類の音と対応しており，重複する組み合わせは省略している．表中の値はベクトル同士の距離であり，小さいほど類似している事を表す．本来は 32 bit 浮動小数点の数として算出されるが，紙面の都合上，少数第 4 位以下を切り捨てて表示している．

おおよそ 1.0〜10.0 の範囲に値が収まっており，(1) と (2) の距離が小さくなっていることから，正しく音色の違いを表現できていると言える．また，時間構造が類似している (1) (2) (5) (9) (10) の距離が 2.0〜3.0 程度と短いため，時間構造も考慮した類似度計算を行えていると考えられる．さらに，学習データのカテゴリに当てはまらない (6) についても，問題なく距離を算出できている。

今後は特徴量の次元圧縮を行った上で，さらに別の音データや距離指標を用いて特徴ベクトルの類似度を算出し，その有用性を検証する．またその結果に応じて，モデルの表現力を向上させるために大規模なデータセットを学習させることや，類似度の計算量削減のために SSDH モデル [5] を参考に損失関数を追加することを検討する．

表 2　類似度の算出結果
Table 2　Similarities of sample data

-	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	0	0.962	4.627	8.345	2.988	6.079	9.462	9.261	2.211	3.124
(2)	-	0	4.660	8.456	2.756	6.113	9.595	9.506	2.103	2.904
(3)	-	-	0	8.071	5.174	5.558	9.273	9.078	4.361	5.418
(4)	-	-	-	0	9.211	7.736	9.791	9.517	8.398	9.508
(5)	-	-	-	-	0	5.969	10.096	9.974	2.208	2.123
(6)	-	-	-	-	-	0	8.320	9.499	6.028	6.420
(7)	-	-	-	-	-	-	0	6.057	9.816	10.512
(8)	-	-	-	-	-	-	-	0	9.402	10.464
(9)	-	-	-	-	-	-	-	-	0	2.500
(10)	-	-	-	-	-	-	-	-	-	0

6. まとめ

大規模な効果音データベースから, 深層学習モデルによる特徴抽出器を用いて，音をクエリとして効果音を検索する方法を検討した．実装に先立って，類似画像検索を参考に，CNN ベースの効果音分類モデルを構築する実験を行った．50 クラスの効果音データセットを学習させたところ，精度は 83.40 %となった．また，そのモデルを特徴抽出器として用いることで，128 次元の特徴ベクトルを抽出できることを確認し，ユークリッド距離で類似度を評価可能であることを確認した．今後はモデルを改良しつつ，特徴ベクトルの次元数を圧縮する処理を実装し，その有用性の検証を行う．

参考文献

[1] 相川清明, 谷島加奈子. ベクトル空間法を用いた相対的感性表現による音検索. 情報処理学会研究報告, Vol. 2007, No. 11, pp. 5–10, feb 2007.
[2] 株式会社 ZOZO. ファッションコーディネート WEAR. https://wear.jp/news/imagesearch/ 2020 年 7 月 22 日参照
[3] 宮田康弘, 小坂直敏. CNN を用いた環境音認識. 東京電機大学未来科学部情報メディア学科 平成 30 年度卒業論文, 2019.
[4] 山田祐雅, 小坂直敏. C++ によるクロスプラットフォーム化した電子音色辞書の構築. 研究報告音楽情報科学 (MUS), Vol. 2014, No. 57, pp. 1–6, 2014.
[5] H. Yang, K. Lin, and C. Chen. Supervised learning of semantics-preserving hash via deep convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 2, pp. 437–451, 2018.
[6] 三宅悠介, 松本亮介, 力武健次, 栗林健太郎. 特徴抽出器の学習と購買履歴を必要としない類似画像による関連商品検索システム. Technical Report 4, may 2017.
[7] J. DENG. Imagenet : A large-scale hierarchical image database. *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition 2009*, 2009.
[8] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.
[9] Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, p. 10151018, 2015.
[10] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

ⓒ 2020 Information Processing Society of Japan