Vol.2021-MUS-130 No.17
Vol.2021-EC-59 No.17
2021/3/16

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

小宮 寛季[1,a)]　小坂 直敏[1,b)]

概要：環境音の検索を行う際，求める音を正確に言語化することは不可能であり，ゆえにテキストのみによる検索は困難である．そこで我々は CNN ベースの環境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，それを基に環境音同士の類似度を算出し，3D 空間上に可視化することにより，容易に環境音の検索が行えるシステムを提案する．音データから提案モデルを用いて 128 次元の特徴ベクトルを抽出し，主成分分析を用いて 20 次元に圧縮した後，ユークリッド距離/コサイン距離を基準として類似音検索を行った．15 種類の環境音をクエリとして 2339 個の音から上位 10 個の類似音検索を行ったところ，クエリと類似した音が提示された例は全体の 79.3%であった．さらに 27699 個の音を検索対象として用いて，既存手法との比較を行ったところ，提案手法はより正確な類似音検索が行えると示された．

1. はじめに

近年，動画投稿サービスの台頭により，個人が動画や BGM，効果音を扱う機会が増加している．また，ゲーム制作や音楽制作においても楽音以外に環境音を音素材として使用することも多い．効果音を動画に付与する際，あるいは制作で音素材を選ぶ際には，目的の音を膨大なデータベースから探し出さなければならない．ゆえに現在，効果音等で用いるための環境音検索の需要が高まっている．しかし，所望の音を手早く検索することは困難である．理由としては，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なおさらである．

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．

この問題に対して相川らは，音を感性表現に基づく 8 次元の心理値ベクトルに変換し，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 [1]．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%ほどの精度で所望の音を探すことができる．

具体的には，1 分程度の楽曲 88 曲に対し「楽しさ」「悲しさ」「恐怖」「落ち着き」「怒り」「不気味」「明るさ」「爽やかさ」の 8 つの特徴語を用いた各 5 段階の評価を，14 名の被験者を集めて回答させた．収集した回答を，平均せずに分布をそのまま用いることにより，楽曲の心理値を $8 \times 5 = 40$ 次元のベクトルとして表現した．

検索の際には，各特徴語に対する 5 段階の値を入力して，対象の楽曲との類似度を計算する．心理値の類似度はコサイン距離を用いて算出している．また，提示された候補を基準として，これらの最初の印象との変化分をユーザが入力することにより，心理値ベクトルを更新して，より所望の楽曲に近い候補を選ぶことができる．しかし，楽曲から心理値ベクトルを割り出すためには，逐一アンケート調査を行う必要があり，新しくデータを追加することが難しい．

これに対して Qi らは，音響特徴量と音色との関係をモデル化することにより，音響学や音響特徴量の知識がなくとも，直観的かつ効率的に音の検索が行える，音による音データベースの検索システムを提案した [2]．まず，検索クエリとなる音と，データベース内の音それぞれについて，波形およびスペクトルから，64 次元の特徴量を抽出する．特徴量は，エンベロープ（5 次元），周波数構造（15 次元），倍音構造（44 次元）の三種類（計 64 次元）からなる．

検索時には，その特徴量を基に，類似度が高い順に候補を 5 つ表示する．ユーザが候補の中から音を選ぶと，システムはそれをクエリとしてさらに検索を行う．その際，前のクエリとユーザが選んだ候補の特徴空間内での距離が，

[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

他の候補と比べて最も近くなるように，特徴量の重みの調整を行う．また，候補を提示する際には，一度に 5 つの音すべてを確認できるように，再生タイミングと左右定位をずらして同時に音を再生している．

さらに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[3] 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 に学習させることで環境音分類モデルを構築する実験を行い，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モデルを構築できることを示している [4]．

2. 本研究の位置づけ

本研究では，当研究室で開発していた環境音管理ツールである電子音色辞書 [5] を基にしてシステムを開発する．電子音色辞書の主な機能は，ブラウザ機能・3D 音色表示機能・マイリスト機能の 3 つである．特にブラウザ機能は，以下の 3 つの検索手法を有する．

- カテゴリ検索（ディレクトリを上から辿る）
- キーワード検索（レーベンシュタイン距離を用いる）
- 類似音検索（音響特徴量を用いる）

本システムの検索の特徴は，聴感上の類似音の検索であり，音源の生成の立場での検索ではない，という点である．これはテキスト検索においては，擬音語での検索により可能である．

3D 音色表示機能は，検索結果やマイリスト内の音群を，1 つの音を 1 つの球として 3 次元空間上に表示する機能である．次元軸には，3 次元に圧縮した MFCC，もしくは平均音圧レベル・再生時間・減衰度・周波数重心の中から 3 つの音響特徴量を選んで設定する．マイリスト機能は，ブラウザ機能や 3D 音色表示機能で見つけた所望の音を，ユーザの操作により動的にグルーピングするための機能である．

システムに登録した環境音には，ラベルとして，擬音語・巨視的音色・キーワード・音源分類の情報をテキスト形式で付与できる．ただし巨視的音色を除き，これらは音を追加するたびにユーザが手で入力しなければならない．

そこで本研究では，ラベルを入力せずとも検索が行えるように，またテキストを介さずに検索が行えるように，深層学習モデルを応用した特徴抽出器を組み込んで，電子音色辞書を新たに開発する．今回は初期段階として，類似音検索と 3D 音色表示のみを実装する．音データから特徴ベクトルを抽出する深層学習モデルの実装には TensorFlow，ファイルリストと 3D 音色表示の GUI の実装には PySimpleGUI と Matplotlib を用いる．

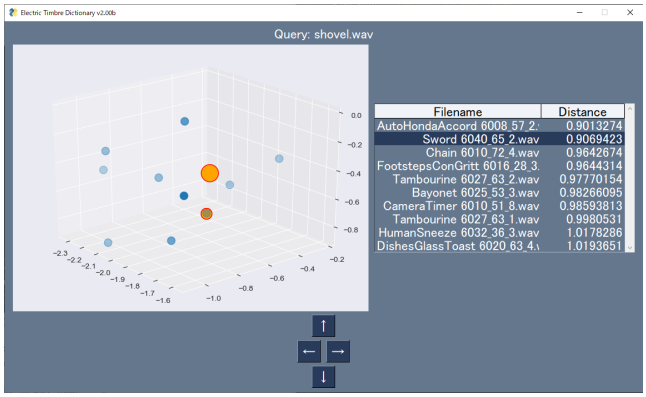

図 1　3D timbre browser

2.1 分類と検索の相違点

一般にユーザは，唯一の音をイメージできていてこれを検索するのではなく，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適な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

深層学習モデルを用いた音の類似度計算と，それを基にした環境音検索システムについて以下に記す．まず 3.1 節でシステムの基本となる音のリスト表示について述べ，3.2 節で特徴抽出器，3.3 節で 3D 音色表示について述べる．

3.1 リスト表示

PySimpleGUI を用いて環境音のリスト表示画面を実装する．ユーザはファイルもしくはフォルダを選ぶことにより，音ファイルをリストに追加することができる．この際システムは音データから特徴ベクトルを抽出し，内部にそのデータを保持する．

3.2 特徴抽出器

画像処理の分野においては，大規模なデータセットを学習させた分類モデルを用意し，データを与え推論させた際の隠れ層の重みを取り出して，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．また，ターゲットとなるデータの特徴が，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は，改めて学習を行わずとも，モデルを特徴抽出器として転用できる [6]．

しかし音声処理分野においては，大規模なデータセットを学習させ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．ここでは画像処理分野における先行研究を参考に，CNN をベースとしたモデルを実装する．このモデルを特徴抽出器として

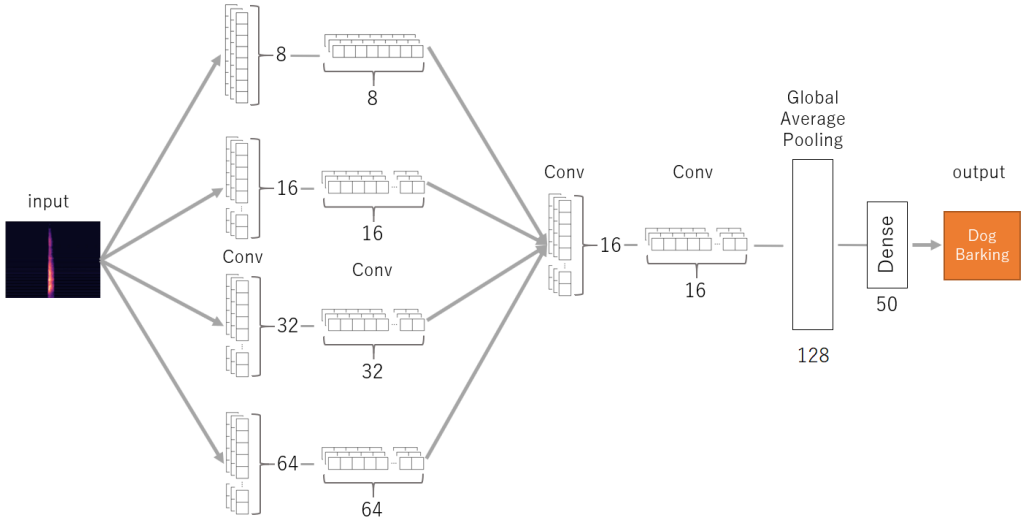

図 2 Model architecture

用いることにより，音データから 128 次元の特徴ベクトルを抽出する．その後，類似度計算の計算量削減や，後述する類似度の音色空間上での可視化のために，主成分分析を用いて特徴ベクトルを 20 次元に圧縮する．

3.3 類似度の可視化

当研究における，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 20 次元の多次元空間（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，ユーザは各次元軸に対して，20 次元の中から自由に特徴量を割り当てることができる．

4. 実装

開発言語として Python を用いて，特徴抽出器とシステムの GUI を実装した．類似度計算を実装するまでの流れは以下の通りである．

(1) 50 クラスの環境音分類モデルを学習させる．
(2) 学習済モデルを特徴抽出器として用いて，音から 128 次元の特徴ベクトルを抽出する．
(3) 主成分分析を用いて，特徴ベクトルを 20 次元に圧縮する．
(4) 圧縮した特徴ベクトル（20 次元）を元に，音の類似度を計算する．
(5) 算出した類似度をもとに，音を 3D 空間上に表示する．

まず 4.1 節で (1) のモデルの学習について述べ，続く 4.2 節で (2) ～ (4) の処理について述べる．最後に 4.3 節で，システム全体の実装と (5) の処理の実装について記す．

4.1 環境音分類モデルの構築

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環境音のデータセットを与え，環境音分類モデルを学習させる実験を行った．

4.1.1 モデルの構造

図 2 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り，Shibui が提唱した環境音分類モデル [7] を基に，計算量削減のために，分割後の縦・横の畳み込み層の数を 4 層から 2 層に省略した構造とした．

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，環境音の 50 クラスの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．

また宮田らは，このモデルにおいて縦方向の畳み込み層を追加することで，より周波数軸を重視した環境音分類モデルが構築できることを示している [4]．

4.1.2 データセット

50 種類の環境音が各クラス 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[8] 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，環境音と

Vol.2021-MUS-130 No.17
Vol.2021-EC-59 No.17
2021/3/16

表 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

してよく用いられる音が含まれているため，環境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．

4.1.3　実験条件

データセットは fft の次数 1024，フレームシフト 256，出力の次元数 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データの拡張を行い，学習用データを 5 倍の 8000 に増量した．これらの処理には，Python 用の音声信号処理ライブラリである Librosa を用いた．

モデルは TensorFlow を用いて実装し，最適化関数には AMSGrad[9] 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．

4.1.4　結果と考察

学習の結果， 評価データでの 50 クラスの分類の精度は 83.40%となった．これは，Shibui が提唱した元のモデルと同等の性能である．また，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．

4.2　特徴抽出器の構築

前節で構築したモデルを用いて，音データから 128 次元の特徴ベクトルを抽出し，主成分分析により 20 次元に圧縮した後に，それらを用いて，クエリと探査空間内の特徴ベクトルの距離による類似音検索を行い，その特性と有用性を検証した．

計算量削減のため，特徴ベクトルを圧縮した．まず ESC-50 から各カテゴリの音を 1 個ずつ用意して，特徴ベクトルを抽出した後，それらをまとめた 50×128 の行列に対して主成分分析を行い，128×20 の変換行列を算出した．

そして変換行列をクエリとターゲットの特徴ベクトルに適用することで，特徴ベクトルを 20 次元に圧縮した．この 20 次元のベクトルの，元の 128 次元のベクトルに対する寄与率は 83%となった．

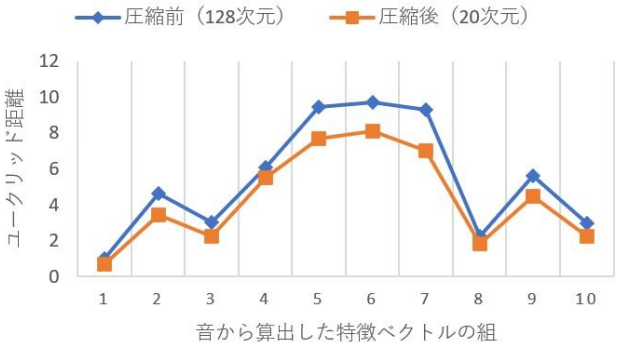

図 3　Comparison of distance of features, before and after PCA

無作為に選んだ 10 組の音に対して，128 次元の場合と 20 次元の場合のユークリッド距離を比較したところ，図 3 に示すように，値の大小関係に大きな変化は見られず，音の特徴が保たれていることが確認された．

4.3　システム全体の実装

使用時のシナリオとしては，まずユーザは環境音データベース，およびクエリ用の環境音をシステム内の検索対象リストに登録する．その際にシステムは予め特徴ベクトルを抽出し，各環境音のパスと関連付ける．この際特徴ベクトルは，4.2 節の処理で ESC-50 のデータから算出した変換行列を用いて，128 次元から 20 次元に圧縮される．

検索の際には，ユーザは検索対象リストの音の中からクエリを選び，システムはそれに対応する特徴ベクトルを用いて，クエリ以外の音の特徴ベクトルに対して類似度計算を行う．その計算結果を元に，システムはクエリと類似する音を類似度の高い順に 10 個抽出し，3D 空間上に可視化する．

音ファイルのパスと特徴ベクトルの組は，システム外部に書き出し可能であり，システムを起動するたびに自動的に読み込まれる．また，システムに未登録の音ファイルからの特徴抽出と，それをクエリとした類似音検索を同時に行うことも可能である．

表 2　Details of query sounds and summary of search results (Exp. 1)

No.	クエリ音の内容	同	異	誤	距離の範囲	検索結果の例
1	赤ん坊の泣き声 A	1	4	5	1.520〜2.025	フルートのトリル
2	赤ん坊の泣き声 B	1	3	6	1.624〜2.456	スティールパン，カラス
3	自転車のベル	0	6	4	1.744〜2.174	トライアングル，鈴
4	子供の声の騒音（22 秒）	3	5	2	1.148〜3.650	騒音，ワライカワセミ
5	子供の声の騒音（4 秒，F.O.）	3	5	2	2.438〜3.735	フクロウ，フレクサトーン
6	柱時計の針の音（5 回）	2	7	1	1.584〜5.128	足音，手拍子
7	ドアノック（2 回）	1	8	1	1.554〜2.030	打撃音，朗読音声
8	フルートのタンギング（1 回）	0	10	0	0.271〜0.307	ごく短い減衰音，水滴
9	フルートのタンギング（7 回）	6	2	2	0.883〜1.024	フルートの奏法各種，鶴
10	鳥の鳴き声（＋森の環境音）	8	0	2	2.078〜2.958	鳥の鳴き声各種
11	銅板の打撃音	1	8	1	1.709〜1.887	スティールパン
12	黒電話のベル音（1 回）	0	6	4	1.381〜1.805	鈴，コオロギ
13	黒電話のベル音（2 回）	1	9	0	0.326〜3.155	スティールパンのトリル
14	黒電話のベル音（3 回）	1	9	0	0.326〜3.090	スティールパンのトリル
15	足音（＋ノイズ）	8	1	1	2.895〜4.003	足音各種

5. 実験

提案手法による類似環境音検索の有用性を確認するため，2 種類の実験を行った．本稿における特徴抽出器は，学習時に用いたものと同じ形式の音データしか扱えないため，用いる音データのサンプリング周波数はすべて 44.1 kHz に変換し，5 秒より長い音は，音量が最大のフレームの後方 5 秒間を切り出して，5 秒以下の音は，末尾に無音を挿入して使用した．

まず 5.1 節では，提案手法を用いて 2339 個の環境音素材集に対して類似音検索を行う．続く 5.2 節では，27699 個の大規模な環境音素材集を用いて，既存手法での検索結果との比較を行う．

5.1 実験 1：小規模な環境音素材集

表 2 に示す 15 種類の音をクエリとして，複数の素材集を組み合わせた 2339 個の wav ファイルからなる環境音データベースに対して，単純なユークリッド距離を用いて，類似度の高い順に 10 個の音を求めた．

表 2 に実験結果の概略を示す．同じカテゴリの音が見つかった件数を「同」の列，違うカテゴリだが耳で聴いて類似性があるものの件数を「異」の列，関連性がみられないものの件数を「誤」の列に記している．

5.1.1 全体の評価

明らかにクエリと関連性がみられないものを除くと，類似した音をシステムが提示した例は 150 件中 119 件で，全体の 79.3%であった．これは，あくまでも予備実験であるため，さらに外部の被験者を募った実験を行う必要がある．またクエリによっては，同カテゴリの音が検索対象に含まれていないものもあった．

5.1.2 距離の特性

8 のような短かい音に関しては距離が小さくなり，逆に 6 のような長い音や，10 と 15 のようなノイズの多い音に関しては距離が大きくなる傾向があった．

5.1.3 特徴量の優先順位

クエリと音色構造の類似した音の方が，時間構造が類似した音よりも多く上位に表れた．ただし 8 のような極端に短い音の場合は，時間構造を優先した結果が得られた．

5.1.4 長い音をトリミングしてクエリとした場合

13 の音を繰り返しただけの音である 14 の検索結果はほとんど変化しなかった．しかし 5 に関しては，4 の音と違いフェードアウトしているため，より減衰系の音が多く検索結果に表示された．

5.1.5 異カテゴリの類似音

検索結果には，カテゴリや音源という観点で類似している音以外にも，様々な音が表示された．中でも特に特徴的な例として，2 からはカラスの鳴き声，3 からは神楽鈴の音，4 からはワライカワセミの鳴き声，6 からは手拍子の音，7 からはオランダ語の朗読音声，8 からは水滴の音，9 からは鶴の鳴き声，12 からはコオロギの鳴き声などが提示されている．

カテゴリや音源はクエリと異なるものの，これらはエンベロープやピッチがクエリと類似した音であった．このことから，人が聴いて「似ている」と思える音を検索できており，人間の聴覚特性を反映した類似度計算が実現できていると考えられる．

5.2 実験 2：大規模な環境音素材集

さらに詳しく当システムの特性を検証するため，5.1 の実験と同じクエリを用いて，プロユースの環境音素材集である Sound Ideas 社の General Series 6000 および Extension

表 3 Details of query sounds and summary of search results (Exp. 2)

No.	クエリ音の内容	類似音の件数				距離の平均			
		$E\alpha$	$C\alpha$	$E\beta$	$C\beta$	$E\alpha$	$C\alpha$	$E\beta$	$C\beta$
1	赤ん坊の泣き声 A	4	3	1	1	1.545	0.093	31.771	2.506×10^{-3}
2	赤ん坊の泣き声 B	8	7	2	4	1.645	0.110	36.158	4.188×10^{-3}
3	自転車のベル	8	7	1	1	1.767	0.099	26.170	1.912×10^{-3}
4	子供の声の騒音（22 秒）	5	3	6	4	2.901	0.246	43.140	3.409×10^{-3}
5	子供の声の騒音（4 秒，F.O.）	3	3	5	4	2.593	0.188	53.616	7.204×10^{-3}
6	柱時計の針の音（5 回）	8	8	2	3	3.654	0.163	51.292	2.451×10^{-3}
7	ドアノック（2 回）	7	8	1	4	1.226	0.037	23.359	7.307×10^{-4}
8	フルートのタンギング（1 回）	10	10	10	10	0.327	1.936×10^{-3}	9.002	7.822×10^{-4}
9	フルートのタンギング（7 回）	6	8	1	3	0.906	0.018	24.087	1.371×10^{-3}
10	鳥の鳴き声（＋森の環境音）	4	4	4	3	2.293	0.244	104.036	2.009×10^{-2}
11	銅板の打撃音	6	5	1	5	1.594	0.081	46.423	2.113×10^{-3}
12	黒電話のベル音（1 回）	7	6	3	3	1.457	0.079	37.472	3.846×10^{-3}
13	黒電話のベル音（2 回）	7	5	3	4	2.522	0.189	59.939	3.292×10^{-3}
14	黒電話のベル音（3 回）	7	5	2	3	2.464	0.187	61.854	3.176×10^{-3}
15	足音（＋ノイズ音）	7	7	1	2	3.397	0.226	71.376	3.537×10^{-3}

表 4 Summary of search results (Exp. 2)

距離指標＼特徴量	CNN (α)	MFCC (β)
ユークリッド距離 (E)	64.7%	28.7%
コサイン距離 (C)	59.3%	36.0%

表 5 Average of sigma of 10 results (Exp. 2)

距離指標＼特徴量	CNN (α)	MFCC (β)
ユークリッド距離 (E)	0.109	3.502
コサイン距離 (C)	0.017	6.758×10^{-4}

1 ～ 10（計 27699 個の wav ファイル，約 160GB）を検索対象に，環境音を検索する実験を行った．この環境音素材集には，15 種のクエリそれぞれと同じカテゴリ（または音源）の環境音が含まれていることを事前に確認している．

5.2.1 実験条件

比較対象の音響特徴量として MFCC を基にした特徴量を用いた．提案手法と同じく，音データのサンプリング周波数はすべて 44.1 kHz に変換し，5 秒より長い音は音量が最大のフレームの後方 5 秒間を切り出して，5 秒以下の音は末尾に無音を挿入して使用した．

MFCC のパラメータは，4.1.3 で扱ったモデルの学習データと同じ条件である fft の次数 1024，フレームシフト 256，メルスペクトルの次元数 128 に揃え，最終的な出力の次元数を 10 に設定した．これにより，5 秒の音から 862 フレーム分のデータが得られ，10×862 の行列が出力された．この行列を基に各次元における全フレームの平均（周波数構造の特徴）と，フレーム方向の標準偏差（時間構造の特徴）を算出して 10 次元のベクトルを 2 種類求め，これを組にして，最終的に 1 つの音データにつき 20 次元のベクトルを出力した．

そして，5.1 の実験と同じクエリを用いて，提案手法を使う場合（α）と MFCC を使う場合（β）のそれぞれに対して，ユークリッド距離（E）と，符号を反転し 1 を足したコサイン距離（C）を指標に用いた計 4 種の条件で検索を行い，距離の短い順に 10 個の音を求めた．

5.2.2 結果

表 3 に，ユークリッド距離・提案手法（$E\alpha$），コサイン距離・提案手法（$C\alpha$），ユークリッド距離・MFCC（$E\beta$），コサイン距離・MFCC（$C\beta$）の各条件における実験結果の概略を示す．表の左側にクエリ音の内容，中央に類似音の件数，右側にクエリとの距離の平均を記している．

また，巻末の表 6・7・8 に左から順に各条件における詳細な検索結果を示す．聴感上クエリと関連のある項目には○を付加した．

加えて，検索結果のうちクエリと類似した音の割合を表 4 にまとめて記し，各条件の検索結果における距離の分散を平均したものを表 5 に記す．

結果から，提案手法は MFCC の時間平均と分散を用いる手法よりも高い精度で類似音を検索出来ていると言える．

5.2.3 詳細な比較

各クエリにおける検索結果を比較すると，クエリ 4・5 以外では（α）の方が類似音を提示した件数が多い．またクエリ 8 では，4 条件すべてにおいてクエリと類似した「ごく短い減衰音」が検索結果に表れている．

5.2.4 クエリを加工することの影響

クエリ 4 では（β）と（α）どちらもクエリと同じカテゴリの音（歓声）を検索できているが，クエリ 5 では（α）の検索結果にはそのような音が出ていない．これはクエリ音が約 1.5 秒かけてフェードアウトするように加工されたことで，本来の特徴を損なったためだと考えられる．

5.2.5 コサイン距離を用いる事の影響

（α）ではユークリッド距離を，（β）ではコサイン距離を用いた方が精度が高かった．またクエリ 10 の（α）において，コサイン距離を用いた方が，クエリとの関連性が高い「カエルの鳴き声」を抽出できている．またクエリ 15 の（α）においても，コサイン距離を用いた方が，クエリと同じ「足音」を多く検索出来ている．このことから，ノイズの多い音をクエリとする場合は，コサイン距離を用いた方がより正確な結果を得られると考えられる．

5.2.6 検索が失敗した例について

赤ん坊の泣き声（1，2）で検索を行った際，本来であれば同じ「赤ん坊の泣き声」が見つかるはずだが，達成できていない．これは検索対象の音から特徴ベクトルを抽出する際，音量が最大のフレームの後方 5 秒間を切り出したために，音の立ち上がり（アタック）の部分が欠落しており，正しく特徴を抽出できていないためだと考えられる．

また鳥の鳴き声＋森の環境音（10）では単体の「鳥の鳴き声」を検索することはできていない．これはクエリの鳴き声の音色と，検索対象に含まれている鳴き声の音色が異なることと，クエリの音色の時間変化が激しいことが原因と考えられる．

同様に黒電話のベル音（12，13，14）を元に検索した際，（α）では同じ「黒電話のベル音」を抽出できていない．（β）では 1 件だけではあるが，（E）（C）の両条件ともに黒電話のベル音を検索出来ている．これはクエリに用いた日本式の黒電話と，Series 6000 に収録されている欧米式の黒電話のベルのピッチが異なるためであると考えられる．これらを解決するためには，音色やピッチよりもカテゴリを重視するような特徴抽出器の構築が必要である．

6. まとめ

深層学習モデルを特徴抽出器として用いて，環境音データベースから音をクエリとして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，50 クラスのデータセットを用いて CNN ベースの環境音分類モデルを構築する実験を行ったところ，精度は 83.40%となった．

また，学習済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し，主成分分析を用いて 20 次元に圧縮したうえで，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．加えて，15 種類の環境音をクエリとして 2339 個の音から上位 10 件の類似音検索を行ったところ，クエリと類似した音が提示された例は全体の 79.3%であった．

さらに 27699 個の音を検索対象として用いて，既存手法を想定した MFCC の時間平均と分散を用いる手法と比較を行ったところ，既存手法は最大で 36.0%，提案手法は 64.7%の割合でクエリと類似した音が検索結果として得られ，提案手法はより正確な類似音検索が行えると示された．

以降は，当システムの有用性について，何を基準として評価するかを検討する必要がある．物理評価に重きを置く場合，先行研究や電子音色辞書で用いられていた，多くの音響特徴量との比較が必要である．また主観評価に重きを置く場合は，何をもって正しい検索結果とするかの基準の設定が重要である．

そして今回，検索結果が正解であるか（クエリと類似しているか）どうかを筆者の耳による聞こえで判定している．より正確な評価のためには，周波数構造・時間構造などに基づいた，人間の耳に依らない評価尺度を定義し，それを用いて結果を評価する必要がある．

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-SLP-065，No. 11，pp. 5-10，2007.
[2] H. Qi, P. Hartono, et. al, ”Sound Database Retrieved by Sound”, Acoustical Science and Technology. 23, 6, 2002.
[3] 株式会社 ZOZO，“ファッションコーディネート WEAR”，https://wear.jp/news/imagesearch/，2020/7/22 参照
[4] 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.
[5] 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”，情報処理学会研究報告，Vol. 2014-MUS-103，No. 57，pp. 1-6，2014.
[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”，情報処理学会研究報告，Vol. 2017-IOT-37，No. 4，pp. 1-8，2017.
[7] W. Shibui,
“Audio classification usingKeras with ESC-50 dataset.”, https://github.com/shibuiwilliam/audio_classification_keras, 2019/12/10 参照
[8] K. J. Piczak.,“Esc: Dataset for environmental sound classification.”, Proceedings of the 23rd ACM International Conference on Multimedia, 2015.
[9] S. J. Reddi，S. Kale，and S. Kumar.，“On the convergence of adam and beyond.”, International Conference on Learning Representations, 2018.

表 6　The detail of search results, Query 1-5(Exp. 2)

Eα	Cα	Eβ	Cβ
クエリ 1：赤ん坊の泣き声 A			
○おもちゃのタイヤが軋む音	○男性の叫び声	カウベル連打	テープレコーダの操作音
○男性の叫び声	バスのクラクション	テープレコーダの操作音	カウベル連打
猫の鳴き声「ナー」	猫の鳴き声「ナー」	落石の音	落石の音
跳弾の音	○女性の歌声	レーザーの環境音	レーザーの環境音
○鉄の門が軋む音	男性の歌声	薬瓶を振る音	紙を丸める音
トラックのクラクション	金属の打撃音	金属の打撃音	テープレコーダの操作音
○女性の悲鳴	バスのクラクション	○金具が軋む音	○金具が軋む音
軽乗用車のクラクション	○女性の悲鳴	氷を踏む音	釘を打つ音
パネルの操作音「ピッピッ」	軽乗用車のクラクション	自転車のホイールの音	氷を踏む音
ホイッスルの下降音	トラックのクラクション	ドアを開ける音	梯子を立て掛ける音
クエリ 2：赤ん坊の泣き声 B			
○赤ん坊が唇を鳴らす音	○赤ん坊が唇を鳴らす音	石を彫る音	お盆が落下する音
○ヴァイオリンの上昇音	ホイッスルの上昇音	お盆が落下する音	石を彫る音
○うめき声	○ヴァイオリンの上昇音	○土を掘る音	○着水音
跳弾の音	○男性の叫び声	ムチの音	消火器の音
ホイッスルの下降音	○うめき声	炎の環境音	○土を掘る音
○犬のうめき声	バスのクラクション	金属の打撃音	ムチの音
○おもちゃのタイヤが軋む音	○男性の叫び声	お盆が落下する音	○土を掘る音
○男性の叫び声	○犬のうめき声	○鉄の門が軋む音	ホッケーの打撃音
○男性の叫び声	短い汽笛	小銭を自販機に入れる音	○馬のいななき
○短いサイレン	○女性のうめき声	木片が落下する音	薬瓶を振る音
クエリ 3：自転車のベル			
○携帯電話の着信音	爆弾のアラーム	ベルトを締める音	アスファルトの上の着地音
爆弾のアラーム	○携帯電話の着信音	○鎖の音	落ち葉の上の着地音
○卓上ベル	○卓上ベルの音	ゴミ袋を投げる音	ベルトを締める音
○携帯電話の着信音	○携帯電話の着信音	土の上の着地音	落ち葉の上の着地音
○携帯電話の着信音	○携帯電話の着信音	砂利の上の着地音	落ち葉の上の着地音
○アメリカの学校のベル	金属片の落下音	ゴミ袋を投げる音	○カメラのフラッシュ
○アメリカの学校のベル	○アメリカの学校のベル	ワイヤーを引く音	アスファルトの上の着地音
○携帯電話の着信音	○携帯電話の着信音	スプレーの音	コンクリート床の着地音
ブラインドを閉める音	ブラインドを巻き上げる音	車のドアを閉める音	コンクリート床の着地音
○卓上ベル（2 回）	○アメリカの学校のベル	木の床の着地音	コンクリート床の着地音
クエリ 4：子供の声の騒音（22 秒）			
怪獣の鳴き声	○ドアが軋む音	氷を擦る音	氷を擦る音
怪獣の鳴き声	○サッカーの環境音	○女性の話し声	男性の叫び声
お盆が落下する音	○観客席の環境音	男性のうめき声	男性の叫び声
○歓声	シンバルの音	○ブーイングの環境音	○歓声
○歓声	シンバルの音	○子供の声の騒音	男性の叫び声
レーシングカーの通過音	レーシングカーの通過音	○歓声	○女性の話し声
○歓声（屋内）	怪獣の鳴き声	○歓声	落石の音
○観客席の環境音	怪獣の鳴き声	落石の音	梯子を上る音
○ブーイングの環境音	木材が軋む音	大勢の手拍子	○拍手と指笛
ブラインドを巻き上げる音	無線機のノイズ	○大勢の拍手と指笛	○歓声
クエリ 5：子供の声の騒音（4 秒，F.O.）			
材木が軋む音	木材が軋む音	急ブレーキと衝突音	○映写機の駆動音
怪獣の鳴き声	ゴムボールが軋む音	トランペットの演奏	トランペットの演奏
怪獣の鳴き声	怪獣の鳴き声	鉄の門を揺する音	ノコギリの音
○シャッターを開ける音	○シャッターを開ける音	急ブレーキと衝突音	○大勢の拍手
○紙を破る音	電撃の環境音	○ミキサーの音	金属を擦る音
ゴムボールが軋む音	電撃の環境音	○大勢の拍手	ノコギリの音
電撃の環境音	○紙を破る音	○大勢の拍手	落石の音
木の床が軋む音	○水をかき回す音	○大勢の拍手	ドアが軋む音
材木が軋む音	怪獣の鳴き声	○大勢の笑い声	○大勢の拍手
○電動ドライバーの音	木材が軋む音	ドアが軋む音	○ミキサーの音

表 7 The detail of search results, Query 6-10(Exp. 2)

Eα	Cα	Eβ	Cβ
クエリ 6：柱時計の針の音（5 回）			
○置時計の針の音	○置時計の針の音	テープレコーダの操作音	アメリカの学校のベル
○置時計の針の音	○置時計の針の音	○車のバック音	テープレコーダの操作音
○木槌で釘を打つ音	レコードプレーヤの操作音	テープレコーダの操作音	ノート PC のノイズ
レコードプレーヤの操作音	○車のウインカーの音	車のドアを閉める音	○置時計の針の音
○車のウインカーの音	○釘を打つ音	アメリカの学校のベル	風鈴の音
本のページをめくる音	本のページをめくる音	テープレコーダの操作音	機織り機の音
○車のウインカーの音	○車のウインカーの音	○心電図の音	皿を置く音
○金属版の上の足音	○金属版の上の足音	テープレコーダの操作音	○心電図の音
○置時計の針の音	○置時計の針の音	車のエアコンの駆動音	○車のバック音
○置時計の針の音	○置時計の針の音	車のドアを閉める音	テープレコーダの操作音
クエリ 7：ドアノック（2 回）			
BBQ コンロを設置する音	BBQ コンロを設置する音	スイッチを入れる音	車のドアを閉める音
○皿を置く音	○皿を置く音	車の窓を閉める音	車のドアを開ける音
○ドアノック（5 回）	○ドアノック（5 回）	ロッカーを閉める音	水を飲む音
○ドアノック（3 回）	○ドアノック（3 回）	鍵を閉める音	ダンクシュートの音
○ドアノック（5 回）	○ドアノック（5 回）	電話のノイズ	車のドアを開ける音
○ドアノック（4 回）	○ドアノック（4 回）	電話のノイズ	○ドアノック（2 回）
車のドアを閉める音	車のドアを開ける音	○車のドアを開く音	○ビリヤードの玉が跳ねる音
○皿を置く音	○木片の落下音	鍵を閉める音	車のドアを閉める音
○木片の落下音	○皿を置く音	鍵を閉める音	○棚を開ける音
車のドアを閉める音	○ドアを開ける音	鍵を閉める音	○棚を閉める音
クエリ 8：フルートのタンギング（1 回）			
○携帯電話のキー音	○携帯電話のキー音	○スイッチを入れる音	○布を掴む音
○金属の打撃音	○金属の打撃音	○スタンプを押す音	○雪を掘る音
○電話の切断音	○携帯電話のキー音	○キータイプ音	○雪を掘る音
○扇風機のスイッチ	○携帯電話のキー音	○スイッチを入れる音	○スーツケースを開ける音
○携帯電話のキー音	○携帯電話のキー音	○スイッチを入れる音	○金属の打撃音
○携帯電話のキー音	○扇風機のスイッチ	○スーツケースを閉める音	○シートベルトを締める音
○金属の打撃音	○携帯電話のキー音	○グラスを叩く音	○スイッチの音
○金属の打撃音	○ボールをキャッチする音	○スイッチを入れる音	○スーツケースを開ける音
○車のドアを閉める音	○金属の打撃音	○受話器を取る音	○スイッチの音
○ボールをキャッチする音	○金属の打撃音	○受話器を取る音	○缶を開ける音
クエリ 9：フルートのタンギング（7 回）			
○エアホーン「パフパフ」	○エアホーン「パフパフ」	○電動ドリルの音	缶を投げる音
○エアホーン「パフパフ」	○エアホーン「パフパフ」	缶を投げる音	○工具箱の音
○エアホーン「パフパフ」	○エアホーン「パフパフ」	斧を叩く音	○電動ドリルの音
○エアホーン「パフパフ」	○エアホーン「パフパフ」	弓を構える音	くしゃみの音
缶を叩く音	○エアホーン「パフ」	テープレコーダの操作音	石の打撃音
車のクラクション	車のクラクション	テープレコーダの操作音	金属の打撃音
○エアホーン「パフパフ」	○エアホーン「パフパフ」	レバーが軋む音	猫の鳴き声
レコードのスクラッチ音	○エアホーン「パフパフ」	車のクラクション	ポストに投函する音
車の窓を開ける音	レーダーの音	金属の打撃音	くしゃみの音
○ドアノック（7 回）	○エアホーン「パフパフ」	金属の打撃音	○工具箱の音
クエリ 10：鳥の鳴き声（＋森の環境音）			
○女性の悲鳴	○カエルの鳴き声	グロッケンの環境音	グロッケンの環境音
ハウリング音	○カエルの鳴き声	グロッケンの環境音	グロッケンの環境音
○女性の悲鳴	○女性の悲鳴	電動ドリルの音	電動ドリルの音
携帯電話の着信音	携帯電話の着信音	電動ドリルの音	携帯電話の着信音
携帯電話の着信音	ハウリング音	魚が跳ねる音	電動ドリルの音
携帯電話の着信音	携帯電話の着信音	○おもちゃのタイヤが軋む音	魚が跳ねる音
携帯電話の着信音	携帯電話の着信音	○おもちゃのタイヤが軋む音	○おもちゃのタイヤが軋む音
ハウリング音	携帯電話の着信音	電動ドリルの音	○ゼンマイの音
○女性の悲鳴	携帯電話の着信音	○ゼンマイの音	○おもちゃのタイヤが軋む音
○女性の悲鳴	○女性の悲鳴	○女性の悲鳴	電動ドリルの音

表 8　The detail of search results, Query 11-15(Exp. 2)

Eα	Cα	Eβ	Cβ
クエリ 11：銅板の打撃音			
怪獣の鳴き声	怪獣の鳴き声	鍵を閉める音	○落石の音
怪獣の鳴き声	怪獣の鳴き声	車のドアを閉める音	銃の発砲音（4 回）
○ボタン音	○ボタン音	○銃の発砲音	ドアノック（4 回）
○木管楽器の低音	○木管楽器の低音	金属製のドアを閉める音	○銃の発砲音
怪獣の鳴き声	○トラックのクラクション	金属製のドアを閉める音	銃の発砲音（2 回）
○金属製のドアが軋む音	○トラックのクラクション	車のドアを閉める音	○落石の音
○金属板の打撃音	怪獣の鳴き声	スライドドアを閉める音	○バスのドアを閉める音
○金属版の上を引きずる音	○金属製のドアが軋む音	スライドドアを閉める音	○ショットガンの発砲音
○金属性のドアが軋む音	木製のドアがきしむ音	車のドアを閉める音	マシンガンの発砲音
タイプライターの改行音	ギターをでたらめに弾く音	車のドアを閉める音	マシンガンの発砲音
クエリ 12：黒電話のベル音（1 回）			
○固定電話の着信音	○金属片が落下する音	泥の上の足音	銃のリロードの音
○鍵の落下音	○金属片が落下する音	○金属片が落下する音	木を切り倒す音
○固定電話の着信音	○固定電話の着信音	○鎖の音	泥の上の足音
ホイッスルの音	○携帯電話の着信音	砂利の上の足音	スクリーンを垂らす音
○携帯電話の着信音	○爆弾のアラーム	風切り音	砂利の上の着地音
ゲップの音	電話の不通音	スクリーンを垂らす音	○鎖の音
○馬のいななき	ホイッスルの下降音	汽笛の音	○黒電話のダイヤルを回す音
車のクラクション	○エレベーターのベル	木片が落下する音	炎の環境音
○時計のベル	フレクサトーンの下降音	ライフルの操作音	砂利の上の着地音
○携帯電話の着信音	ドアが軋む音	○金属片が落下する音	○金属片が落下する音
クエリ 13：黒電話のベル音（2 回）			
○グロッケンの演奏	○携帯電話の着信音	○黒電話のベル音	○黒電話のベル音
○グロッケンの演奏	○携帯電話の着信音	○壁掛け電話のベル音	○壁掛け電話のベル音
○モーターの駆動音	○グロッケンの演奏	すり鉢の音	カセットテープの落下音
無線機のノイズ	○グロッケンの演奏	○固定電話の着信音	紙を広げる音
柱時計の時報	トランペットの演奏	タイプライターの操作音	泥の上の足音
○携帯電話の着信音	携帯電話のキー音	公衆電話の操作音	すり鉢の音
○火災報知機のベル	携帯電話のキー音	野菜を切る音	○固定電話の着信音
○金属を削る音	ミュートトランペットの演奏	公衆電話の操作音	○黒電話のベル音
柱時計の時報	柱時計の時報	ライフルの操作音	ライフルの操作音
○観客席の環境音	○モーターの駆動音	銃が落下する音	釘を打つ音
クエリ 14：黒電話のベル音（3 回）			
○グロッケンの演奏	○携帯電話の着信音	○黒電話のベル音	○黒電話のベル音
○グロッケンの演奏	○携帯電話の着信音	すり鉢の音	○壁掛け電話のベル音
○モーターの駆動音	○グロッケンの演奏	ライフルの操作音	カセットテープの落下音
○火災報知機のベル	○グロッケンの演奏	野菜を切る音	すり鉢の音
柱時計の時報	携帯電話のキー音	野菜を切る音	フェンシングの剣の音
無線機のノイズ	携帯電話のキー音	釘を打つ音	紙を広げる音
○金属を削る音	トランペットの演奏	テープレコーダの操作音	釘を打つ音
○置時計の時報	柱時計の時報	タイプライターの操作音	公衆電話の操作音
柱時計の時報	○モーターの駆動音	フェンシングの剣の音	○固定電話の着信音
○観客席の環境音	柱時計の時報	電灯のひもを引く音	ヘルメットの装着音
クエリ 15：足音（＋ノイズ音）			
○斧で木を切る音	○斧で木を切る音	アメリカの学校のベル	アメリカの学校のベル
○レコードプレーヤの操作音	○レコードプレーヤの操作音	冷蔵庫のドアを開ける音	○テープレコーダのノイズ
ドアの開閉音	ドアの開閉音	鍵を閉める音	肉を食べる音
○木靴の足音	○木製の床の足音	ブラウン管モニタの起動音	電話の呼び出し音
○車のウインカー	○タップダンスの音	テープレコーダの操作音	○レコードプレーヤの操作音
電動ドリルの音	○車のウインカーの音	鳩時計の時報	車のドアを閉める音
○小銭の音	○釘を打つ音	○トレーニング器具の音	車のドアを閉める音
○機織り機の音	○金属板の上の足音	車のドアの開閉音	機織り機の音
梯子を立てかける音	梯子を立てかける音	ゴングの音	草野球の環境音
○置時計の針の音	台車を動かす音	鳩時計の時報	車のエアコンの音