CE-009

音色に着目した自然音の楽器分類評価システム *

オウ イクカン † 伊藤 克亘 ‡
Wang Yuhuan Katunobu Itou

1 まえがき

ここ数年、非楽器で音楽を演奏する映像が流行している。ピアノやドラム、バイオリンといった既存の楽器の代わりに、グラスや水滴、動物の鳴き声といった自然音を楽器として使う音楽である。このような演奏は非常な魅力がある。しかし、新しい自然音楽器を見つけるには多くの課題がある。そのひとつは、ほとんどの自然音はグラスのようにピッチを自由に調整できない。例えば、テーブルをたたく音はピッチが均一で、打楽器のそれに近い。二つ目は、自然音が持つ複雑で幅広い音色は、そのまま使うと逆効果になることが多い。例えば、グラスの演奏が流行している一方で、プラスチックのコップは長い音域があるが、音色はグラスほどクリアではなく、人間には受け入れられないことが多い。自然音に近い既存の楽器が見つかれば、その楽器を真似て演奏したり、楽曲に追加してアレンジしたりすることで、聴く人をリフレッシュさせることができる。

本論文では、対象となる自然音と既存の楽器との類似度を算出し、その類似度を用いて、自然音の楽器としての価値を評価できるシステムを提案する。同時に、類似楽器が見つかれば、その楽器の演奏手法を模倣して新しい楽器を作ることも可能である。本研究では、自然音に含まれる複雑な音色の特徴を抽出するために VGGish[5] と AST[7] モデルを使用し、抽出した特徴を用いて、各楽器種類の次元における自然音の大きさを代表する分類確率を持つ楽器音データセットで第 2 の分類モデルを学習した。この確率を特徴量として用いることは、対象となる自然音をある程度評価できることを意味する。

2 関連研究

2.1 楽器分類法

古来、楽器を分類する方法はたくさんある。最も広く知られているのは、西洋楽器の分類、つまり打楽器、弦楽器、管楽器などである。最も広く使われているのは、実は発音原理に基づく HS 分類（ザックス＝ホルンボステル分類法）である [3]。この分類法は楽器の発音原理に基づいており、既存の楽器を大きく 5 つに分類している：体鳴楽器、膜鳴楽器、弦鳴楽器、気鳴楽器、電鳴楽器である。例えば、よくピアノと呼ばれるものは、この分類法では「314.122」に分類され、3 は弦鳴楽器を表し、それ以降の数字は平板と共鳴箱に基づく細分化である。

しかし、本研究では、自然音を対象とした分類法が必要であり、HS 分類は非常によく機能するが、本研究ではいくつかの問題がある。例えば、自然音である水の流れ音は、このシステムでは 5 つの種類のどれにも分類できない。

さらに、5 つの種類は幅広い自然音に対して明らかに不十分であるにもかかわらず、HS 分類は「314.122」のように楽器に対してのみ細分化されており、この細分化は自然音に対しては正しく機能しない。

最後に最も重要なことは、この分類は結局のところ、楽器の音色ではなく、楽器の発音原理に基づいているということである。例えば、全く異なる発音原理を持つ電鳴楽器といったエレクトリックピアノの音色は、弦鳴楽器といったピアノとほとんど同じである。

そこで本論文では、音色細分のために、HS 分類を基準にして、特徴の母集団中心点と平均距離のクラスタリング法を提案する。この方法については後述する。

2.2 音色特徴

様々な楽器の音色に関する研究は、物理発声源に着目した研究 [6][9]、有限要素法でモデルを作る研究 [8]、楽器の倍音構造による多次元分布を提案する研究 [10] などがあるが、自然音と楽器の音色の両方に着目した研究は行われていない。しかし、自然音と楽器の両方に着目した研究は非常に少ない。なぜなら、楽器と比べて、自然音の基本周波数、倍音などの音楽特徴はより複雑で、共鳴体も多くて、ただの音高検出でも非常に困難である。

しかし近年、ディープラーニングの進展により、このような複雑の問題を AI で解決する可能になった。既に楽器のメルスペクトログラムと CNN を用いた研究がよくある。そして最近、純粋な Transformer モデルや CNN-attetion モデルが AT タスク（Audio Tagging）に有効であることが実証された研究 [7] もある。十分なデータがあれば、ディープニューラルネットワークは複雑な音に含まれる情報を自動的に学習することができる。本研究では、CNN モデルである VGGish と純粋 Transformer モデル AST を特徴抽出モデルとして利用する。

2.3 データセット

Google チームは 2017 年に AudioSet[1] と呼ばれる大規模なサウンドデータセットをリリースした。このデータセットには 200 万以上の 10 秒間の音声と 500 以上のタグが含まれている。AudioSet のタグは、Google チームが YouTube 動画の音声に手動でタグを追加したものである。様々な楽器音や音楽だけでなく、鳥の鳴き声や水の流れ音、車のエンジン音などの自然音も含まれて、自然

* Instrument classification evaluation system natural sound tones focusing on tone color
WANG YUHUAN (Grad. School of CIS, Hosei Univ.) et al.

† 法政大学大学院 情報科学研究科
‡ 法政大学情報科学部

音の割合がこれまでで最も多いデータセットである。また、AudioSet の付属といった VGGish[5] は、VGG アーキテクチャに基づく CNN モデルで、AudioSet データセットの Mel スペクトログラムを使用して事前学習され、Softmax 層の前に 128 次元の特徴ベクトルを出力できる。VGGish モデルは多くのライブラリに含まれているため、より簡単に使用できるので本論文では、VGGish を使用して提案手法の有効性を検証する。

AudioSet に加え、本研究では RWC2003[2] というデータセットも使用する。RWC2003 には、人間の歌声を含む 50 の楽器があり、それぞれ異なるピッチで異なる演奏方法で、データはすべて単音である。楽器の単音データセットについては、単音検出にエネルギーウィンドウ法を用い、ランダム組み合わせ法とランダム削減法を用いて、各種類のデータ量を約 5000 個に制御する。

本手法の有効性を検証するため、グラス音、水流音などの自然音データを採録し、それぞれ約 4 分間収集した。グラス音は、異なる形状のグラスに異なる量の水を入れ、プラスチック、木や鉄の棒で叩いた音である。水流音は、水量や条件の異なる蛇口やシャワーヘッドの音である。各種類のサンプリングは約 1 秒間、75 %のオーバーラップで行われ、各データ量は約 1000 個である。

2.4 AST

AST モデル [7] とは、「Audio Spectrogram Transformer」である。近年様々な領域で活躍している Attention モジュールと Transformer モデルが画面認識で CNN モデルより高い精度があることを検証した [4]。そのため、Yuan たちは音のスペクトログラムと画像の類似性に着目して、ViT いわゆる「Vision Transformer」モデルのパラメーターを転移学習して、AudioSet で再トレーニングした。AST の構造は、1 のように、入力は任意長さの 16kHz オーディオのメルスペクトログラム、出力は AudioSet のタグとなる 527 種類の特徴ベクトル。

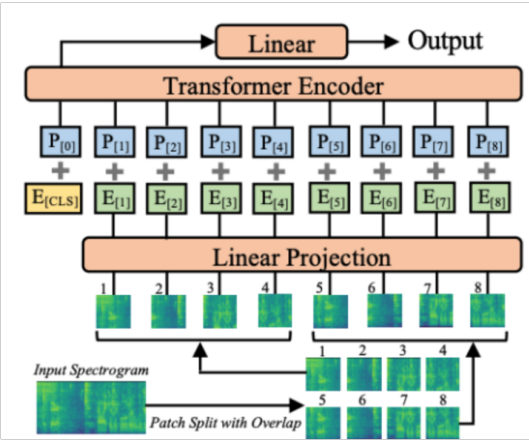

図 1. AST の構造

3 データセット

3.1 AudioSet

自然音を研究しようとする際に最も難しい点は、データセットの取得の問題である。自然音には多種多様な音が含まれているため、すべての種類の自然音を含むデータセットを用意することは不可能だ。しかし、グーグルのチームは 2017 年に AudioSet[1] と呼ばれるデータセットをリリースした。AudioSet は、有名な動画サイト YouTube の動画から、さまざまなラベルを付けてサウンド事件をダウンロードできる。合計で、200 万以上の 10 秒間のサウンドセグメントがあり、500 以上のラベルがある。

このデータセットはもともとサウンドイベント検出の分野で使用されていたため、ほとんどのサウンドセグメントには複数のラベルが含まれている。複数のラベルが付けられたデータは本研究のモデルの精度に影響を与えるため、本研究では単一ラベルのイベント、約 80 万のセグメントのみが収集された。

3.2 楽器単音データセット

AudioSet データセットにはすでに楽器音と音楽イベントが含まれているが、楽器単音の特徴をより正確に調べるため、本研究では楽器単音のみを含むデータセット RWC2003[2] も使用する。このデータセットは、ボーカルを含む 50 の楽器で構成され、各楽器に 2〜3 個があり、さまざまな演奏方法があり、各楽器のすべての単音と部分ハーモニーが含まれている。

4 提案手法

提案手法では、まず、自然音と楽器音の両方を含むデータセット AudioSet から特徴ベクトルを抽出できるモデル 1 をトレーニングする。次に、このモデルを用いて、楽器単音のみを含む RWC2003 から「楽器の自然音特徴ベクトル」を抽出・分類し、各種類の推定確率を出力できるもう一つのモデルを学習する。最後に、学習したモデル 2 を用いて、対象となる自然音の種類ごとの確率を推定し、対象となる自然音が楽器として使用できるかどうかを評価する。

データセットは生のオーディオであるため、いくつかの前処理を行う必要がある。次に、現在利用可能な分類法（HS 分類法等）では、自然音に対してうまく表現できないため、より細かい分類法を行う必要がある。そこで、本研究では HS 分類法をベースとした母集団の特徴に基づくクラスタリング法を提案する。

4.1 前処理

RWC2003 楽器単音データセットに対して、本稿では以下の前処理を行う。まず，完全な単音データをエネルギー窓法によって検出する．この方法は，式 1 のように，音声データ上で窓をスライドさせることで，窓内のエネルギー値を計算する．

$$E(n) = \sum_{i=1}^{N} x^2(i) \quad (1)$$

開始閾値と終了閾値を設定し、窓が開始閾値より大きいエネルギーを検出したときに開始時刻をカウントし、窓が終了閾値より小さいエネルギーを検出したときに終了時刻をカウントする。

これにより、長いデータに含まれるすべての単音の開始時刻と終了時刻のインデックスを一度に検出することができ、単音をカットすることができる。データ損失率最低限にするために、色々試した結果は、最大振幅を 1.0 にして、開始閾値を 0.1 にして、結束閾値を 0.005 にして、持続長さ最小値 0.3 秒、最大値 3 秒に設定する。

RWC2003 に含まれるデータ量は楽器によって異なり、例えばドラムは音域が狭く、100 単音以下のデータしか含まれないことが多い。そこで本研究では、ランダム組み合わせ法を用いてデータセットを拡張する。それは、変更なし・増加・減少 3 種類の音量変換、変更なし・左シフト・右シフト 3 種類の位相変換で、9 種類のデータを処理できる。そして、同じ種類に含まれるすべてのデータから、1 対 1 の組み合わせがいくつかある。それぞれの組み合わせに対して、最大 81 の可能性がある。ランダムで 81 以下の値を設定すれば、違うサイズのデータが得られる。しかし、この方法で特定な数を設定できないため、得られたデータを平滑のランダムで特定な数で選択する必要がある。RWC2003 でデータ量が一番多いの楽器は 3300 くらいがあるため、以上の方法で全部の楽器を 5000 くらいに処理する。

処理した結果は、92 種類の 467,042 のデータである。

4.2 楽器細分法

前章で記述した HS 分類法には限界があるため、本稿は、HS 分類を基準にして、特徴の母集団中心点と平均距離のクラスタリング法を提案する。

上記の前処理の後、処理された RWC2003 のデータを特徴モデルに通すと、各 1 秒間のデータに対して 128 次元の埋め込み特徴量が得られる。ここで VGGish と AST 二つのモデルを使う。VGGish は固定長の入力オーディオの必要があるため、データを 1 秒にカットする。VGGish の出力値は 128 次元の 0 から 255 までの整数なので、正則化は必要ない。一方で、AST は固定長の必要がなくて、出力値は AudioSet の各種類の 0 から 1 までの正規化した値である。

例えば、気鳴楽器の場合、2 のようになる。こちらの x 軸は中心点が原点からの距離、y 軸はデータが中心点からの距離の平均値である。つまり、x 軸は音色の情報を含んで、y 軸は音域の情報を含んで、両方が正規化したので、比較できる。

同じ楽器の音色と音域二つに着目するために、母集団の中心点から原点の距離を計算して、次元 1 とする。そして、毎データから中心点の距離を計算し、平均距離を次元 2 とする。算出した二次元の母集団特徴を利用し、HS 種類から新しい種類にクラスタリングできる。細分した結果は、92 種類から 17 種類にクラスタリングできで、「体 1-5」「膜高」「膜低」「弦 1-4」「気 1-4」「電」「ボーカル」である。

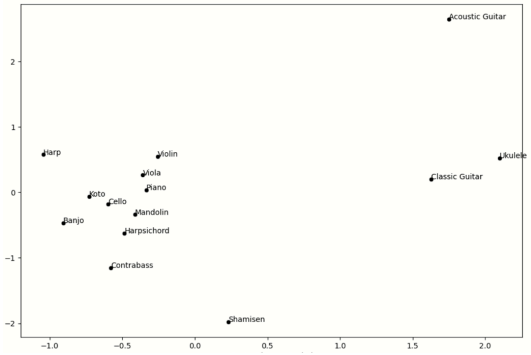

図 2. 気鳴楽器の場合

表 1. VGGish と AST の精度

Model	Test Accuracy
VGGish	0.7595
AST	0.9726

上記の分類方法でより正確な楽器種を得た後、その楽器種をラベルとして RWC2003 データセットを再ラベル化した。

4.3 評価手法

本研究では、自然音の楽器性能を評価するために以下の方法を用いた。特徴モデルから得られた特徴ベクトルで、もう一つのモデルをトレーニングする。

評価対象となる自然音は、上記のプロセスを経て特徴ベクトルが得られ、対象となる自然音の特徴ベクトルと既存の楽器の特徴ベクトルとの距離を計算することで、その類似度を知ることができる。この類似度によって、その自然音の楽器としての性能を評価することができる。同時に、類似楽器の演奏方法を模倣したり、新しい楽器を開発することも可能である。

5 評価試験

本研究は、特徴抽出モデルが VGGish と AST である二つの場合、1 次元 CNN モデル、純粋完全接続モデル、残差接続モデル三つのモデルを構築し、細分した 17 種類に分類する。トレーニングの結果は、表 1 と表 2 になる。

VGGish が精度低い原因の一つは、VGGish の入力は固定長である。そのため、データの拡張は困難で、音の情報を損失する場合も多い。そしてもう一つは、VGGish はより古い CNN モデルである、自然音データセットの情報を取得する能力が低い。

三つの試験は全部 epoch が 1,000、学習率 0.0001、Patch サイズが 128 である。結果を見ると、1D 畳み込み層を使うと、精度が高い一方、パラメーター数が多いの

表 2. 三つのモデルの精度とパラメーター数

Model	Test Accuracy	Param#
Conv1D	0.9810	540k
FC	0.9763	180k
RFC	0.9686	100k

表 3. The structure of the FC model.

Layer	Input	Output
Linear-1	527	128
ReLU-2	128	128
Linear-3	128	64
ReLU-4	64	64
Linear-5	64	32
ReLU-6	32	32
Dropout-7	32	32
Linear-8	32	17

で、無効的なパラメーターがよくあるとわかる。残差接続はパラーメーター数を減少できる一方、精度の損失が高いとわかる。これは、モデルの訓練に使われたデータセットがもともと AST で処理されたもので、オーバーフィッティングが少ないからである。そこで本研究では、最も単純な完全連結ネットワークを選択する。モデル構造は表 3 である。

初期学習率 0.001 で 100 エポック学習し、30 エポックごとに学習率を 0.1 に調整した結果を 3 に示す。

最後は精度が 0.9938 に達成して、パラメーター数が 99,676 である。提案手法で一番高い精度を達成する。

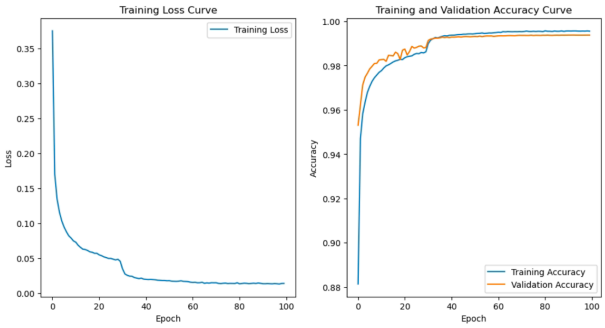

図 3. トレーニング結果

本研究では、「鳥の鳴き声」、「車のエンジ音」、「カウベル」、「グラスを叩く音」、「水流」五つの自然音データを使用する。提案手法で試した結果は、表 4 のようになる。

実際は、鳥の鳴き声は、「気 2」に一番近い（0.284971）だけではなく、「ボーカル」にも近い（0.248289）であり、他の自然音でも見られる。鳥の鳴き声が人間の歌や気鳴楽器に近いことは常識に合致しており、提案手法の有効性を証明している。

表 4. 自然音のテスト

Natural Sounds	Closest category	Percentage
Bird	Ki2	0.284971
Car	vocal	0.230435
Cowbell	Gen1	0.226562
Glass	Tai5	0.298551
Water	LowMaku	0.205572

6 おわりに

以上より、本稿では、音色に着目して、自然音の楽器性を評価できる評価システムを提案した。母集団の中心点距離と平均距離を用いるクラスタリング法で、人間の歌を含む 92 の楽器から、17 種類に細分した。そして、自然音と既存の楽器との類似度を出せるモデルをトレーニングした。この手法の有効性を検証した。

しかし、この手法はいくつかの問題点もある。ひとつは、出せる類似度ベクトルは、多種類に似ている場合は、どちらが良いと判断できない。そして、目標となる自然音は楽器としての価値がない場合は、閾値を設定する必要があるため、閾値は何を設定すれば良いかと微妙的に難しい。また、自然音を応用して、楽曲のアレンジから既存の楽器を置き換えることは、楽器分離、ピッチ変換、自動採譜とか、他の多くの技術が関わっており、それらを活用するには限界や課題がある。

参考文献

[1] Audioset. http://research.google.com/audioset/ index.html. [Online]. [Accessed Dec 26, 2023].

[2] Rwc 研究用音楽データベース：楽器音. https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-i-j.html. [Online]. [Accessed Aug 7, 2023].

[3] 「楽器」. https://ja.wikipedia.org/wiki/楽器. [Online]. [Accessed Aug 7, 2023].

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[5] S. H. et al. Cnn architectures for large-scale audio classification. *ICASSP*, pages 131–135, 2017.

[6] A. French. In vino veritas:a study of wineglass acoustics. *American Association of Physics Teachers*, 51(8):688–694, August 1983.

[7] Y. Gong, Y.-A. Chung, and J. Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021.

[8] T. N. Keiichi Oku, Atsushi Yarai. A new tuning method for glass harp based on a vibration analysis that uses a finite element method. *J. Acoust. Soc. Jpn.*, 21(2):97–104, 2000.

[9] 中. 勲. 発音機構のシミュレーション. *Acoustical Society of Japan*, 37(2):65–75, 8 1981.

[10] 奥. 博. 北原 鉄朗, 後藤 真孝. 音高による音色変化に着目した楽器音の音源同定：f0 依存多次元正規分布に基づく識別手法. **情報処理学会論文誌**, 44(10):2448–2458, Oct 2003.