自然音の楽器性評価システム
Instrumentality Classification Evaluation System for Natural Sounds

オウ イクカン (WANG YUHUAN)*

法政大学 情報科研究科 情報科学専攻
yuhuan.wang.7f@stu.hosei.ac.jp

Abstract

Exploring the potential of natural sounds in music composition and performance often poses significant challenges. While creating melodies with glasses filled with varying amounts of water is both simple and captivating, replicating this flexibility in other natural sounds proves difficult. Most natural sounds lack the frequency adjustability found in glass-based instruments or exhibit acoustic characteristics that are either too homogeneous or excessively complex, making them unsuitable for direct instrumental use. To address these issues, this study proposes a deep learning–based method for evaluating natural sounds. First, a deep neural network is trained on a natural sound dataset to extract feature vectors that encapsulate their sonic characteristics. These vectors are then applied to an instrumental sound dataset, enabling the modeling of natural sound dimensions within the traditional instrumental domain. By employing a classification model to compare natural and instrumental sound ensembles, the method provides insights into the feasibility of utilizing natural sounds in music creation and performance. Illustrative comparisons between cowbell-like natural sounds and conventional instruments such as the marimba demonstrate the effectiveness of this approach and open up new avenues for incorporating natural sounds into creative musical practices.

Furthermore, to enhance the representation of pitch characteristics, this research explores a novel approach by using Mel-Frequency Cepstral Coefficients (MFCC) instead of mel spectrograms for training the natural sound model. Experimental results show that, while maintaining acceptable timbral characteristics, the new model significantly improves pitch representation. This advancement offers more refined technical support for research on the instrumentalization of natural sounds.

1　はじめに

近年、従来の楽器に代わり、自然音を直接楽器として活用した音楽演奏が注目を集めるようになっている。従来、音楽制作はピアノ、ドラム、バイオリンなど、工夫された設計と長い歴史に裏打ちされた既存の楽器を用いることが一般的であった。しかし、インターネットや SNS の普及により、誰もが手軽に映像や音声を発信できる時代となり、街中の雑踏、自然界の環境音、身近な物体が生み出す音など、日常のあらゆる音が新たな表現素材として再評価されるようになった。動画投稿サイトや SNS 上では、家庭内や野外の風景から録音された音が、従来の楽器演奏と同様、またはそれ以上の創造的なアプローチで提示され、視聴者の関心を一気に引きつけている。このような現象は、音楽制作の枠組み自体が広がり、従来の楽器演奏の枠にとどまらない、多様な音の表現手法が模索されていることを強く示している。

具体例としては、グラスの縁を指でこすって奏でる演奏や、水滴が落ちる音をリズム楽器として組み込む試みが挙げ

---

* 指導教員：伊藤克亘 教授

られる。これらは、単なる環境音としてではなく、その音自体に楽器としての機能や表現力を見出そうとするアプローチである。また、動物の鳴き声を楽曲の一部に取り入れ、従来の楽器の音色と組み合わせることで、全く新しいサウンドスケープが形成されるケースも見られる。これにより、自然界に存在する音が、単なる効果音として利用されるだけでなく、演奏の一要素として積極的に活用されるようになっている。こうした自然音を素材とした音楽は、従来の楽器が持つ固定的な音色や音域の制約を超え、常に変化し続ける多様な音響的魅力を提供するため、聴衆に新たな音楽体験と感動を与える可能性を秘めている。

また、自然音楽器の利用は、音楽理論や演奏技法に対しても革新的な視点をもたらす。既存の楽器は、その製造過程や設計段階で音楽的表現力が徹底的に考慮され、調律やダイナミクス、音域の制御が可能であるのに対し、自然音はそのままの状態では音楽的要件を十分に満たさない場合が多い。そのため、自然音を楽器として活用するには、従来の楽器とは異なる加工技術や音響処理、さらには演奏者自身の創意工夫が求められる。こうした挑戦は、自然音が持つ本来の多様な情報をいかに抽出し、音楽表現に転換するかという新たな研究課題となっており、音響学や信号処理、機械学習などの先進技術の融合が不可欠となる。

本研究では、上記の背景と課題に着目し、自然音と既存楽器との音響的特徴を比較することで、自然音が楽器としてどの程度の表現力を持つかを定量的に評価する手法を提案する。具体的には、最新の機械学習技術を活用し、自然音を楽器的に分類するシステムを構築することで、従来の楽器分類法では捉えきれなかった自然音の微細な音色や音高、ダイナミクスなどの情報を抽出し、それをもとに新たな楽器としての可能性を探ることを目的としている。このシステムは、自然音を単なる背景音や効果音として扱うのではなく、その中に秘められた音楽的価値を体系的に評価し、さらに新たな音楽表現の可能性を提示するためのものである。

2　関連研究

2.1　楽器特徴

音楽はメロディー（旋律）、リズム（律動）、ハーモニー（和声）の三要素で構成される。このうち、メロディーとリズムは音高（ピッチ）の有無によって区別される要素であり、梅本らの研究 [11] では、自然音を「音高を検出できるもの（メロディー）」と「音高を持たないもの（リズム）」に分類している。例えば、小鳥のさえずりや風鈴の音は、ある程度の音高を持つためメロディーとして扱うことができる。一方、波の音や雷鳴は、周期性のないノイズ成分が多く含まれるため、主にリズム要素として分類される。

この研究により、自然音をピアノやドラムなどの既存楽器で演奏される音楽に変換することはある程度可能であることが示された。しかしながら、自然音自身が楽器としての機能を果たすかどうかは、未だ十分に検証されていない。実際の楽器には、単にメロディーやリズムを奏でる機能だけでなく、独自の音色や演奏表現を持つため、単純な音高検出では楽器としての適用性を評価することは困難である。

楽器において最も大きな特徴は音色と音域である。楽器の音色について、従来、物理発声源に着目した研究 [4][13]、有限要素法でモデルを作る研究 [7]、楽器の倍音構造による多次元分布を提案する研究 [14] 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、ただの音高検出でも非常に困難である。例えば、図 1 のように、鳥の鳴き声はバイオリンのような規則的な倍音列がない。

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でトレーニングする CNN モデル [1] と Transformer モデル [10] を用いた研究がある。しかし、これらの研究は楽器を対象としたものでしか分類できず、自然音の複雑な特性に適用することは難しい。自然音には不規則な基本周波数、倍音、共鳴ピークなど、楽器よりも複雑な音楽的特徴があり、単純なピッチ検出でさえ非常に難しいからだ。

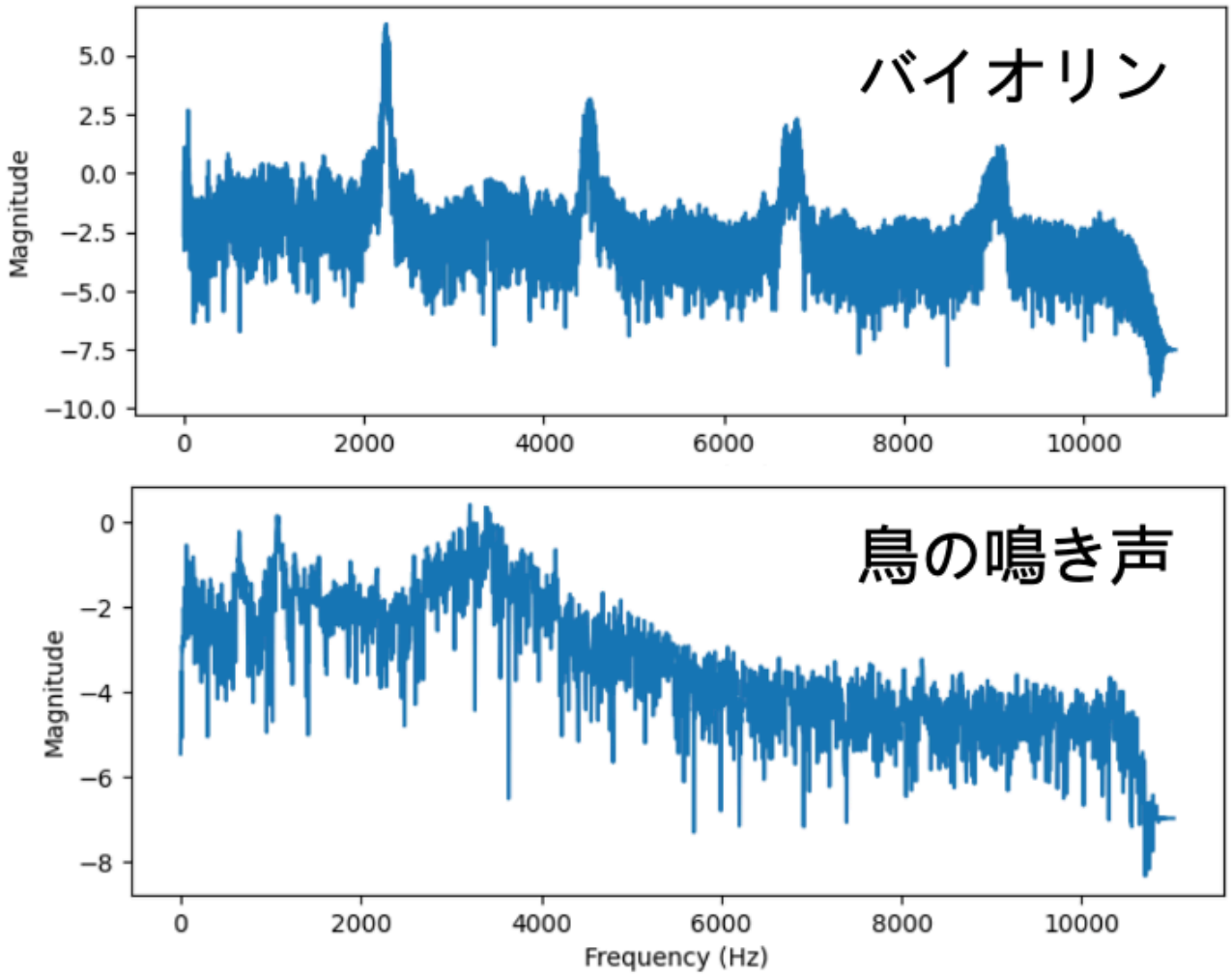

図 1. 楽器と非楽器のスペクトラム

既に純粋な 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 [6] もあるが、自然音と楽器の特徴両方同時に着目した研究はほとんど行われていない。本研究では、転移学習の手法を利用して、自然音モデルの特徴を楽器分類モデルをトレーニングする。自然音モデルについては、CNN モデルである「VGGish」[3] と純粋 Transformer モデルである「AST（Audio Spectrogram Transformer）」[6] を利用する。

2.2　楽器分類法

従来は楽器を打楽器、弦楽器、管楽器などに分類した西洋楽器分類法は一番知られている方法だが、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)[12] である。それは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。HS 分類法は、既存の楽器を何十種類を分けている。

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「水の流れ音」や「風の音」といった自然音は、この 5 つのカテゴリのどこにも明確に分類で

きない。水音は固体や膜の振動ではなく、気体や液体の運動によって発生するため、HS 分類の枠組みには収まらない。

さらに、HS 分類法は、楽器に対して非常に詳細な分類を提供するが、自然音に対してはその細分化が正しく機能しない。例えば、ピアノは「314.122」といった細かいカテゴリに分類されるが、自然音にはこのような詳細な分類が適用できない。自然音は楽器のように規則的な振動源を持たず、音響特性が時間とともに変化するため、HS 分類法の枠組みをそのまま適用することは適切ではない。

最後に最も重要なことは、HS 分類法は音色ではなく、発音原理に基づいて分類していることである。例えば、エレクトリックピアノは HS 分類法では「電鳴楽器」に分類されるが、実際にはアコースティックピアノとほぼ同じ音色を持つ。このように、発音原理が異なっても、音色が類似しているケースが多く存在する。楽器の音色は、発音メカニズムだけでなく、共鳴特性やスペクトル構造によって決まるため、音色を考慮しない分類は、自然音との比較において不十分である。

そこで本論文では、自然音の分類精度を向上させるために、HS 分類法を基準としつつ、楽器の音色と音域の特徴を考慮した新しいクラスタリング手法を提案する。具体的には、特徴空間における母集団中心点（Centroid）と特徴分布の平均距離の分散を用いたクラスタリングアルゴリズムを導入することで、従来の楽器分類手法では対応できなかった自然音の細分化を実現する。この方法については後述する。

2.3 AudioSet と VGGish

Google チームは 2017 年に「AudioSet」 [5] 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。AudioSet には、関連する機械学習モデルとして「VGGish」[3] が付属している。VGGish は、VGG（Visual Geometry Group）アーキテクチャをベースにした畳み込みニューラルネットワーク（CNN）モデルであり、AudioSet のメルスペクトログラムを用いて事前学習されている。

VGGish の特徴は以下の通りであり、図 2 のように表す。

- **入力**: Mel スペクトログラム（96 × 64 の時間周波数表現）
- **ネットワーク構造**: VGG16 の畳み込み層をベースにした CNN
- **出力**: Softmax 層の前に **128 次元の特徴ベクトル**
- **適用範囲**: 楽器音や自然音を含む多種多様な音響データの分類

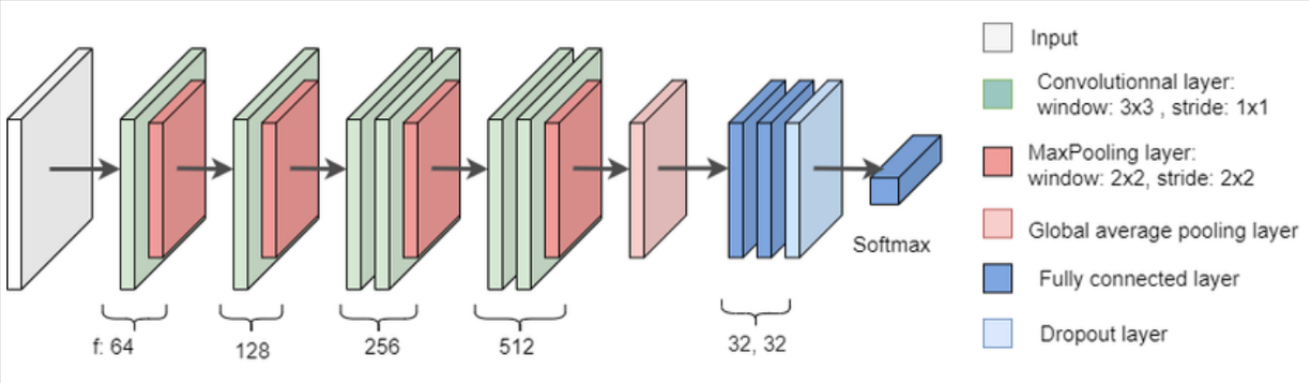

図 2. VGGish の構築

VGGish は、既存の多くのライブラリに統合されており、比較的容易に使用できるため、本研究ではこのモデルを活用し、提案手法の有効性を検証する。特に、VGGish を用いることで、楽器音と自然音の特徴空間を統一的に解析し、自然音の分類精度を向上させることが期待される。

2.4 Audio Spectrogram Transformer

「AST」（Audio Spectrogram Transformer）[6] は、近年注目を集めている Attention モジュールと Transformer モデルを用いた音響処理モデルである。従来の CNN ベースの音響分類モデルに対し、AST は自己注意機構（Self-Attention）を活用し、長時間依存関係をより効果的に捉えることができる。

近年、画像認識領域において、Transformer ベースのモデルが CNN を超える性能を示すことが検証されている [2]。特に、Dosovitskiy らによって提案された Vision Transformer（ViT）[2] は、画像を固定サイズのパッチに分割し、それぞれをトークンとして処理することで、CNN よりも優れた特徴抽出を実現した。このアプローチに基づき、Yuan らは音のスペクトログラムと画像の構造の類似性に着目し、ViT のパラメータを転移学習（Transfer Learning）して AudioSet 上で再学習させた [6]。

AST のモデル構造は、ViT のアーキテクチャを音響分類に適用したものであり、図 3 のように構築し、主に以下の特徴を持つ。

- 入力データ: - 任意長さの 16kHz 音声のメルスペクトログラム- 入力は 128 × T の 2D 特徴マップとして表現
- パッチ分割: - **16 × 16 のパッチ** に分割（時間・周波数方向）- マイヤーのスペクトラムを **オーバーラップ率 6** で分割
- Transformer エンコーダ: - 各パッチを線形埋め込み（Linear Embedding）で変換- マルチヘッドアテンション（Multi-Head Attention）を適用
- 出力ベクトル: - **AudioSet の 527 クラス** に対応する分類出力

AST の最大の利点は、CNN ベースの音響モデルと比較して、よりグローバルな特徴を学習できる点である。従来の CNN は局所的な特徴抽出に優れるが、長時間依存関係を捉えるのが難しい。一方、AST は自己注意機構により、時間・周波数方向のグローバルな相関を学習できる。

また、AST は AudioSet の大規模データセットで事前学習されており、転移学習に適している。これにより、データ量が限られた環境音や特殊音源の分類にも応用可能である。

本研究では、AST を活用することで、楽器音と自然音の関係性をより精度高く解析し、VGGish と同じく、自然音モデルとして使用する。

また、モデルの音高表現を向上させるために、本研究では、AST のアーキテクチャを参考にし、「MFCC-AST」モデルを提案する。具体的には後述する。

3 提案手法

簡単というと、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[3] と AST[6](Audio Spectrogram Transformer) モデルを使用する。既存の楽器データセットから、モデルが推定した自然音クラスの分類確率とする特徴ベクトルである「楽器の自然音特徴」を用いて、もう一つの分類モデルを学習する。分類モデルは MLP(Multi-layer Perceptro) で構築、特徴から楽器に分類する。対象となる自然音があれば、分類モデルの出力である「Logits」は「自然音の楽器性特徴」として、その自然音をある程度評価することができる。全体的には、図 4 のようになる。

本研究において、図 4 中の青色の線はデータを用いたモデルの学習プロセスを示し、赤色の線は特定の自然音をテストする際の適用プロセスを表している。本システムにはモデル 1 とモデル 2 の 2 つの異なるモデルが含まれる。

モデル 1 は自然音モデルであり、大規模な自然音データセットを用いて学習を行い、自然音の音響情報を分類する役

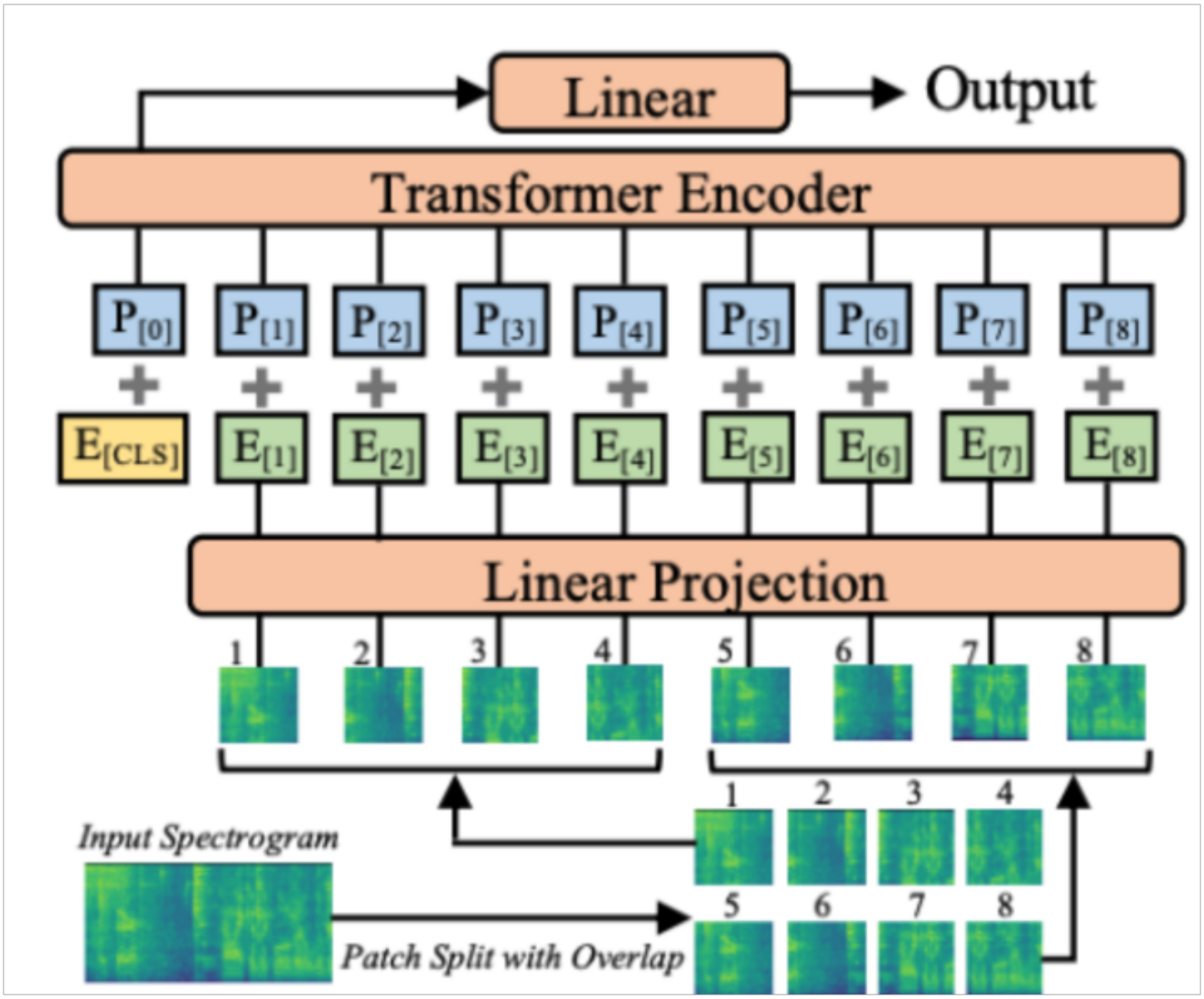

図 3. Audio Spectrogram Transformer（AST）の構造

割を担う。本研究では、このモデルとして VGGish および Audio Spectrogram Transformer（AST）の 2 つを採用した。いずれも AudioSet という大規模な自然音データセットで事前学習されたモデルであり、出力は 527 種類の自然音ラベル（AudioSet のカテゴリ）に対応している。学習が完了すると、モデル 1 は自然音の音響特性を含む特徴ベクトルを出力できるようになる。

VGGish では、分類層の前にある埋め込み層（embedding layer）の出力をそのまま使用し、特徴ベクトルの次元数は 128 である。一方、AST では分類層の出力結果を直接使用し、特徴ベクトルの次元数は 527 となる。これは、AST の構造（図 3 参照）に起因するものであり、Transformer エンコーダの最終層（head 層）が 768 次元の特徴ベクトルを出力し、その後、1 層の MLP による分類層を通過する。分類層のパラメータには多くの情報が含まれているため、この層を削除することが難しい。

図 4 に示すように、自然音データセットを用いてモデル 1 の学習を完了した後、楽器単音データセットを処理し、すべての楽器単音データを対応する自然音の特徴ベクトルに変換する。図 5 に示すように、AudioSet には様々な種類の自然音が含まれているだけでなく、楽器音や音楽演奏もデータセット内に存在する。そのため、モデル 1 による処理後の楽器単音データの特徴ベクトルには、楽器そのものの特性に加え、より精細な自然音の特徴も含まれることになる。

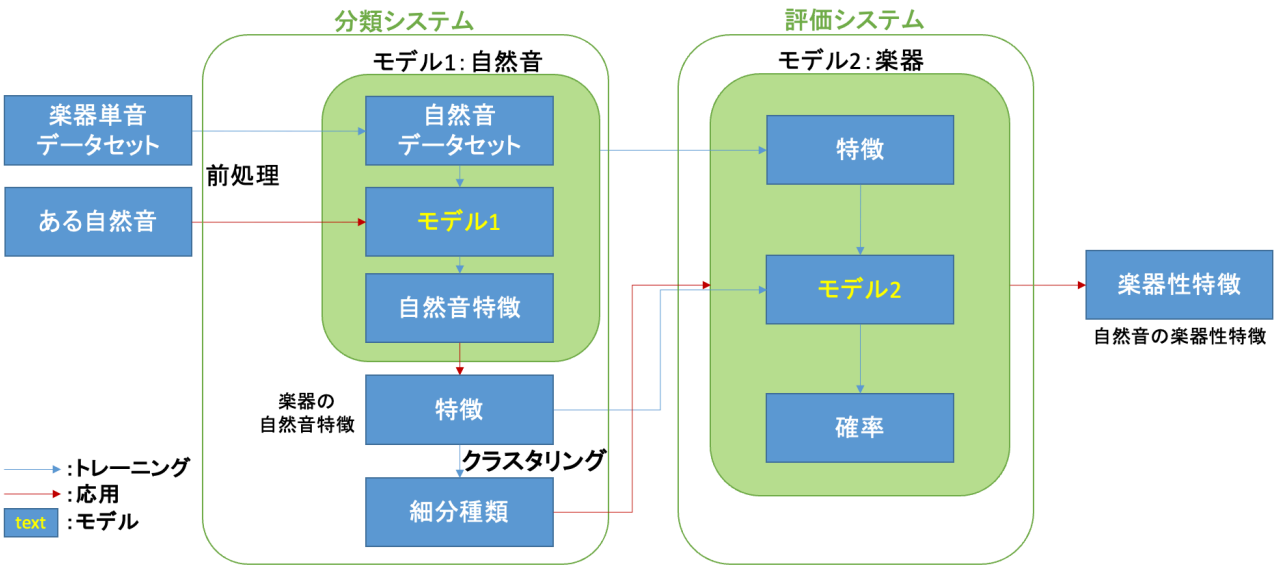

図 4. システムの全体

これは、広範な特徴空間から、より限定された特徴空間への転移学習（transfer learning）に相当し、後続の楽器分類モデルに自然音の情報を適用できるようにするための重要な処理である。

仮に、楽器の音声データそのものや、その音響特徴（メルスペクトログラム、MFCC、音高、倍音成分など）を直接用いて楽器分類モデルを学習した場合、モデルが未学習の自然音データに遭遇した際には予測が困難になる可能性が高い。しかし、自然音モデルで処理した後の特徴ベクトルを用いることで、モデルは楽器本来の特徴とは無関係な自然音の特性も学習できる。これにより、未知の自然音に対しても適切に分類を行うことが可能となる。

楽器単音データセットの自然音特徴を取得した後、次にモデル 2（楽器分類モデル）の学習を行う。VGGish および AST の出力特徴はどちらもベクトル形式であるため、モデル 2 には 全結合ニューラルネットワーク（Fully Connected Network, FCN）または一次元畳み込みニューラルネットワーク（1D-CNN）のいずれかを選択できる。実験の結果、1D-CNN は全結合ネットワークに比べて若干分類精度が向上するものの、パラメータ数が大幅に増加することが確認された。これは、特徴ベクトル内の各成分間には時系列的な依存関係がほとんど存在しないため、1D-CNN の適用による大幅な性能向上は期待できないためである。そのため、最終的には全結合ニューラルネットワークを採用した。

モデル 2 の学習が完了した後、自然音のテストを行う際には、対象とする自然音の複数のデータを取得し、モデル 1 およびモデル 2 の処理を順次適用する。まず、自然音モデルを通じて各データの特徴ベクトルを抽出し、その後、楽器分類モデルによって各データが推定される楽器カテゴリを取得する。最終的に、各カテゴリの出現割合を算出し、この割合をその自然音の楽器性（instrumentality）を示す指標として利用することができる。

また、現在利用可能な分類法 (HS 分類法 [12] 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。既存の HS 種類（体・膜・弦・気・電）と人間の歌声を、17 種類に細分した。

しかし、細分した種類が実と大きな差がある問題がある。それは、自然音モデルが学習した音高特徴は足りない。この問題を解決するため、本研究では、「MFCC-AST」というモデルをトレーニングして、AST の精度を低下させない同時に、音高表現を向上させる。MFCC-AST の音高表現を評価できる「ピアノの 88 音分類モデル」の学習結果を観察する方法も提案した。

本研究の提案手法では、以下のデータセットを利用している。

3.1 データセット

本研究では、自然音データセットと楽器音データセットの 2 つのデータセットを使用する。自然音データセットには、AudioSet[5] と ESC-50[9] を使用し、ガラスと水の流れの音データを追加して収録する。楽器音データセットには、RWC2003[8] を使用する。

3.1.1 AudioSet

Google チームは 2017 年に「AudioSet」[5] 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「2,084,320 個の 10 秒間の音声クリップ」が含まれており、「527 種類の音響カテゴリ」がラベル付けされている。AudioSet のタグは、Google チームが YouTube の動画音声から抽出し、手動または半自動的にアノテーションを付与したものである。

このデータセットの特徴として、従来の音楽関連データセットと比較して、自然音の割合が極めて多い点が挙げられる。具体的には、図 5AudioSet には以下のような多様な音声データが含まれている。

- **楽器音**（ピアノ、バイオリン、ギター、ドラムなど）
- **自然音**（鳥の鳴き声、風の音、水の流れ、雷鳴など）
- **人工音**（車のエンジン音、機械の作動音、都市の環境音など）
- **ヒューマンボイス**（話し声、笑い声、叫び声など）

このような包括的な音響データを持つ AudioSet は、自然音と楽器音の関係性を解析するための基盤データセットとして非常に有用であり、本研究の提案手法の適用可能性を検証するために活用する。

3.1.2 ESC-50

本研究では、AudioSet に加えて、テストデータとして「ESC-50」[9] を使用した。ESC-50 は AudioSet ほど大規模ではないものの、1 対 1 のラベリングが施されたクリアな音声データを提供する環境音データセットであり、環境音の分類性能を評価するための標準的なベンチマークデータセットとして広く用いられている。

ESC-50 は、50 種類の環境音を均等に含むデータセットであり、各カテゴリごとに 40 サンプル、合計 2,000 の音声クリップが収録されている。これらのデータは、人手によって厳密にラベリングされており、データ品質が高い。

- **データ規模**: 50 クラス、合計 2,000 音声クリップ
- **クリップ長**: 5 秒
- **ラベリング**: 人手による正確なアノテーション
- **カテゴリ**: 5 つの主要グループに分類
  - **動物の鳴き声**（犬、猫、鳥など）
  - **自然音**（雷、雨、水の流れ、風など）
  - **都市環境音**（交通音、電車、車のエンジン音など）
  - **ヒューマンボイス**（話し声、笑い声、叫び声など）
  - **家庭内の音**（掃除機、ドアの開閉音、食器の音など）
- **サンプリングレート**: 44.1kHz / 16bit

AudioSet は YouTube 動画から自動抽出された大規模データセットであるため、一部のデータにはノイズや誤ラベリングが含まれることがある。一方、ESC-50 はすべてのデータが人手でラベル付けされており、分類タスクの基準とし

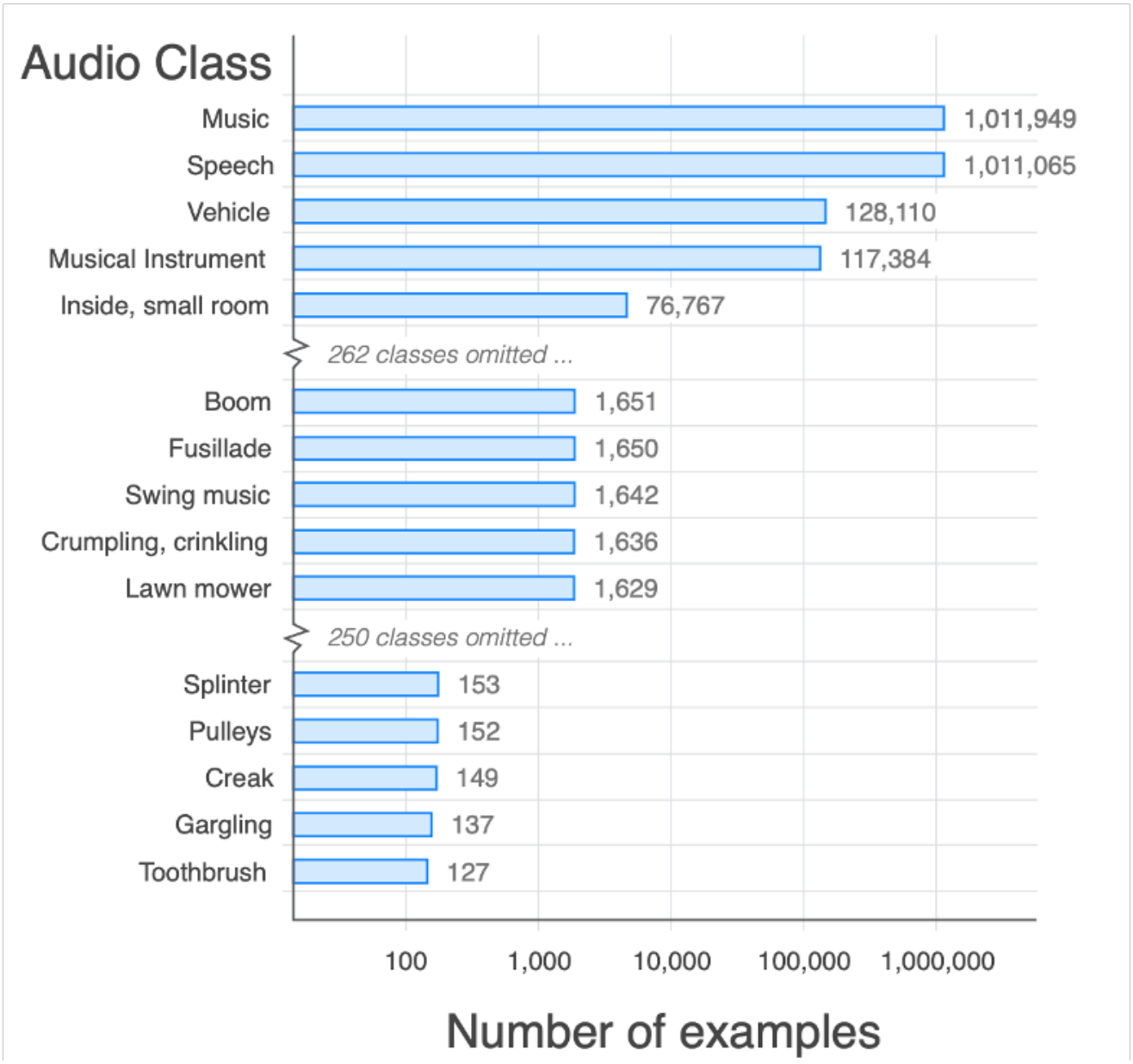

図 5. システムの全体

て非常に信頼性が高い。さらに、AudioSet にはカテゴリ間のオーバーラップやデータの不均衡があるため、分類の基準として使用するのが難しい場合がある。ESC-50 は均等に分布した 50 クラスに分類されており、環境音の識別タスクに適している。

本研究では、ESC-50 をテストデータとして使用し、提案手法の分類性能を評価する。具体的には、VGGish および Transformer ベースの特徴抽出モデル（AST[6]）を用いて、ESC-50 の音響特徴を解析し、既存の楽器音データとの類似度を比較する。さらに、ESC-50 を利用することで、AudioSet に依存しない小規模な実験環境での性能検証を行い、提案手法の汎用性を検証する。

ESC-50 は AudioSet と同様に自然音データセットであるが、データ量が AudioSet に比べて小さいため、学習データとして使用する際にはデータ拡張が必要となる。本研究では、ESC-50 に対して以下の 2 種類のデータ拡張手法を採用した。

AST モデルを ESC-50 で訓練する際、各エポックごとに異なるデータ拡張を適用する方法を採用した。具体的には、以下のデータ拡張をランダムに適用することで、モデルの汎用性を向上させることを目的とする。

- 同カテゴリデータの組み合わせ- 同じクラスに属する別のデータとランダムで結合し、新たなデータを生成。
- 異なるレベルのノイズの付加- ランダムなホワイトノイズを加える。
- スペクトログラムの時間・周波数方向のランダムマスキング- スペクトログラムや MFCC に基づき、ランダムに時間軸または周波数軸の一部をマスキング。

この手法により、異なるデータ分布に対応できるモデルの学習が可能となる。

ESC-50 を本研究の自然音評価システムのテストデータとして使用する際には、別のデータ拡張手法を適用した。この手法は、同じテスト対象でも個々のデータの差異による影響を低減することを目的としている。具体的には、以下の手順でデータを処理する。

1. 1 秒の短いクリップへの分割- 50% のオーバーラップ率で、元の音声を 1 秒間の短いクリップに分割。
2. 異なるレベルのノイズの付加- 各クリップに対して、以下の 3 種類の最大振幅比でノイズを追加：
   - **0.01 倍のノイズ強度**
   - **0.05 倍のノイズ強度**
   - **0.1 倍のノイズ強度**

このデータ拡張手法により、同じテスト対象に対する異なるデータによる誤差の影響を抑え、より安定した評価を実現することができる。

3.1.3 自然音データの追加収集

本研究では、自然音の種類を増加させるために、新たに自然音データを収録し、AudioSet ではカバーしきれない音響情報を補完する。このデータ収集の目的は、楽器と自然音の関係性をより詳細に解析し、自然音の楽器的な特性を評価することである。

1. **対象音の選定**
   - **グラス音**: 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、鉄の棒で叩いた音を収録。
   - **水流音**: 蛇口やシャワーヘッドの水量・水圧を変化させ、異なる流れの音を収録。
   - **サンプリングレート**: 44.1 kHz / 16-bit
2. **データ分割と前処理**
   - 各自然音の録音時間は約 4 分間。
   - 各音源を 1 秒間のクリップに分割し、75% のオーバーラップを適用。
   - 結果として、各音源データは約 1000 サンプルを生成。

これらのデータは、AudioSet の自然音データと組み合わせることで、より多様な自然音の特徴を学習可能なデータセットを構築する。また、VGGish を用いた特徴抽出によって、楽器音との比較が容易になり、自然音を楽器として分類・評価するための基盤データとして活用できる。

3.1.4 RWC2003

本研究では、自然音データセットに加えて、「RWC2003」[8] を使用する。RWC2003 は、人間の歌声を含む 50 種類の楽器にラベル付けされた 90,000 以上のモノフォニック（単音）オーディオクリップを収録したデータセットである。

RWC2003 は、既存の音楽データセットとは異なり、純粋な楽器データセットである。これにより、音楽的な要素（旋律、和声、リズム）を含まず、楽器の音色特性をより正確に捉えることができる。RWC2003 に含まれるデータは各楽器における複数の演奏方法（アタック、サステイン、ビブラートなど）がある。また、演奏楽器のグループごとに統合する。例えば、ドラムとシンバルを同一カテゴリとして扱う。それは、従来の楽器データセットと比べ、より詳細な楽器ごとの特徴を記録しているため、楽器の音色を分析し、楽器分類モデルの学習に適したデータセットである。

本研究では、RWC2003 のラベル付けを精細化し、楽器の音色分類に適したデータを構築する。RWC2003 は、細分システムと楽器分類モデルの学習データとして使う。

RWC2003 のデータ量は楽器によって異なり、一部の楽器（特にドラムや管楽器）では音域が狭く、単音のサンプル数が 100 音未満である場合が多い。このデータ不足を補うために、ランダムコンビネーション法を用いたデータ拡張を実施した。その手法について、以下の手順で行う。

まずは 3 種類の音量変換 (変化なし、増加、減少) と 3 種類の位相変換 (変化なし、左シフト、右シフト) で 9 種類のデータを処理する。そして、同じ型に含まれるすべてのデータから、いくつかの 1 対 1 の組み合わせがある。その後、組み合わせを目標数にランダムで削減する。

以上のことを繰り返し、最終的に、RWC データセットの各機器につき 5,000 データ、合計 460,000 データ（92 データタイプ）に拡張された。

3.2 単音抽出手法

RWC2003 は生のオーディオデータセットであるため、学習データとして使用する前にいくつかの前処理を行う必要がある。特に、長いオーディオデータの中から完全な単音を抽出するために、「エネルギー窓法」を用いた音声セグメンテーションを適用する。

エネルギー窓法では、音声データ上をスライドする窓を設定し、窓内のエネルギーを計算することで音の存在を検出する。エネルギー値 $E(n)$ は、以下の式（式 1）によって求められる。

$$E(n) = \sum_{m=0}^{M-1} x^2(n-m) \tag{1}$$

ここで、

- $x(n)$ は入力音声信号
- M は窓の長さ（フレームサイズ）
- $E(n)$ は時刻 n におけるエネルギー値

この方法では、開始閾値 E_{start} と終了閾値 E_{end} を設定し、以下のルールで単音の開始時刻と終了時刻を決定する。

1. エネルギー値 $E(n)$ が開始閾値 E_{start} を超えた最初のフレームを 開始時刻とする。
2. エネルギー値 $E(n)$ が終了閾値 E_{end} を下回った最初のフレームを終了時刻とする。

この処理により、長いオーディオデータ内の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る。これにより、各単音を正確にカットし、楽器音の特性を保持したまま処理を進めることが可能となる。

データ損失率を最小限に抑えるために、複数のパラメータを調整し、最適な値を設定した。最終的に使用したパラメータは以下の通りである。

- **最大振幅正規化**: 1.0

- **開始閾値** (E_{start}): 0.1
- **終了閾値** (E_{end}): 0.005
- **単音の持続時間**: 最小 0.3 秒、最大 3 秒

このパラメータ設定により、不要なノイズや背景音を排除しながら、楽器ごとの特徴を最大限に保持した単音データの抽出が可能となった。

3.3 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。具体的には、まず楽器単音データを拡張する。拡張したデータセットが AST で出力された楽器の自然音特徴データセットを作成する。同じ HS 分類に基づく種類を「PCA（Principal Component Analysis)」すれば、同じ楽器の各点は母集団として見える。図 6 のように、同じ HS 分類に基づく種類を一つグループとして表示できる。この例は RWC データセットの弦鳴楽器である。図 6 は分かりやすいため、2 次元の PCA を使うが、実験の場合は 4 次元となる。

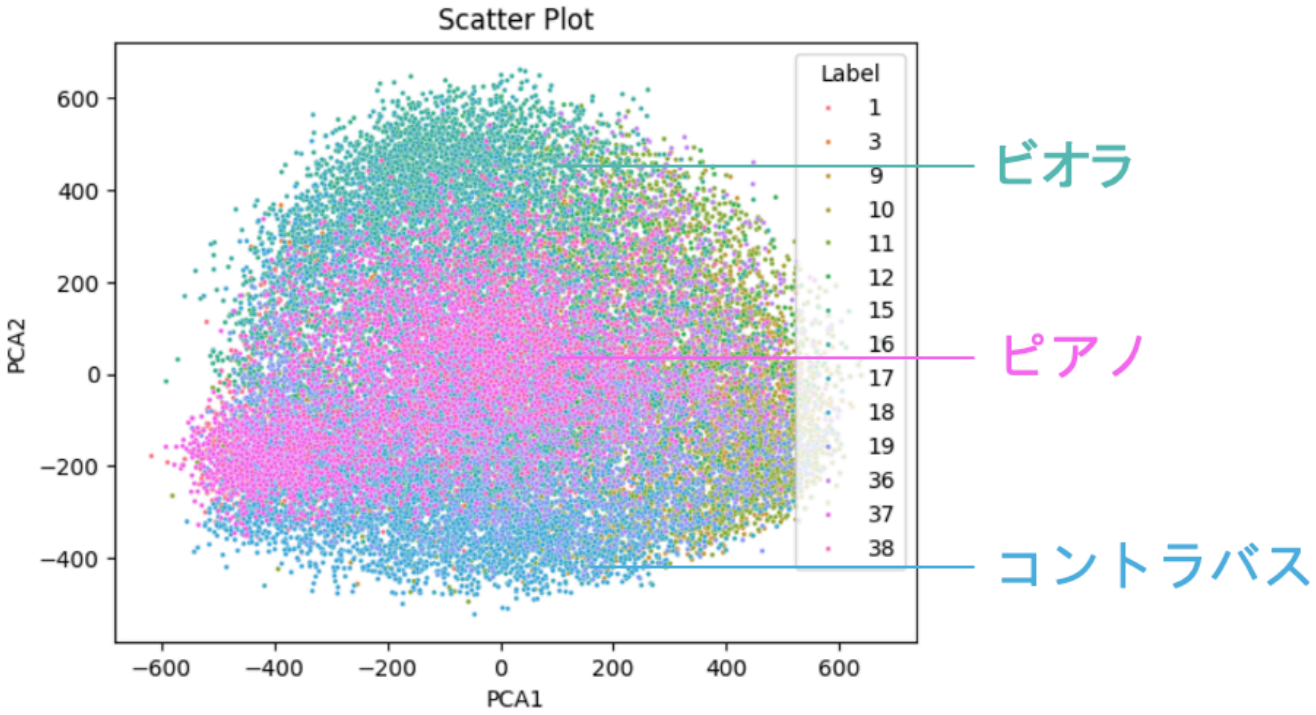

図 6. 弦鳴楽器の母集団例

そして、母集団の中心点を計算する、中心点と原点の距離は楽器の音色を表せる。また、各点と中心点の距離の分散は楽器の音域を表せる。図 7 に示すように、中心から原点までの距離の分散と、各点から中心までの距離の分散は、各楽器セットについて求めることができる。

同じ HS 種類の中心点距離と各点距離分散を標準化して、X 値と Y 値が座標系で見える。その後は「K-means」などでクラスタリングして、HS 種類を細かく分類できる。例えば、弦鳴楽器の場合は図 8 のように分かる。

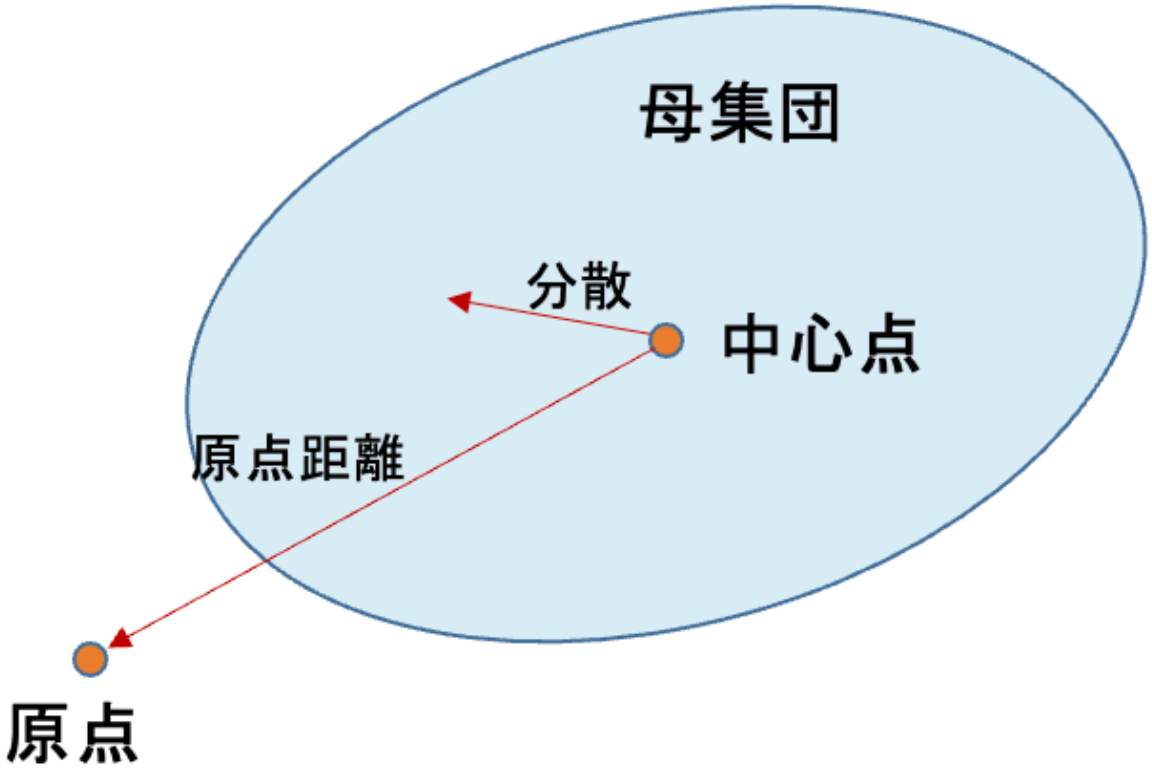

図 7. 細分手法

3.4　評価手法

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。例えば、主観的な評価により、対象が既存の楽器の一つと高い類似度を持つ場合、その楽器と類似しているとみなし、類似した演奏が可能であると判断できる。対象が複数の既存楽器と高い類似度を持つ場合、その楽器は複数の既存楽器の複合体と考えられ、新たな楽器として認識される可能性がある。一方、既存のすべての楽器と類似度が低い場合、その楽器はメロディを奏でる楽器としての価値はないと判断できる (ただし、リズム楽器として使用されたり、音楽制作における背景音として利用される可能性は残る)。

4　実験

4.1　細分手法の検証

提案手法の細分手法の実現可能性を検証するため、RWC2003 データセットに含まれるすべての楽器について実験を行った。図 9 に示すように、中心から原点までの距離の分散と、各点から中心までの距離の分散は、各楽器の母集団に

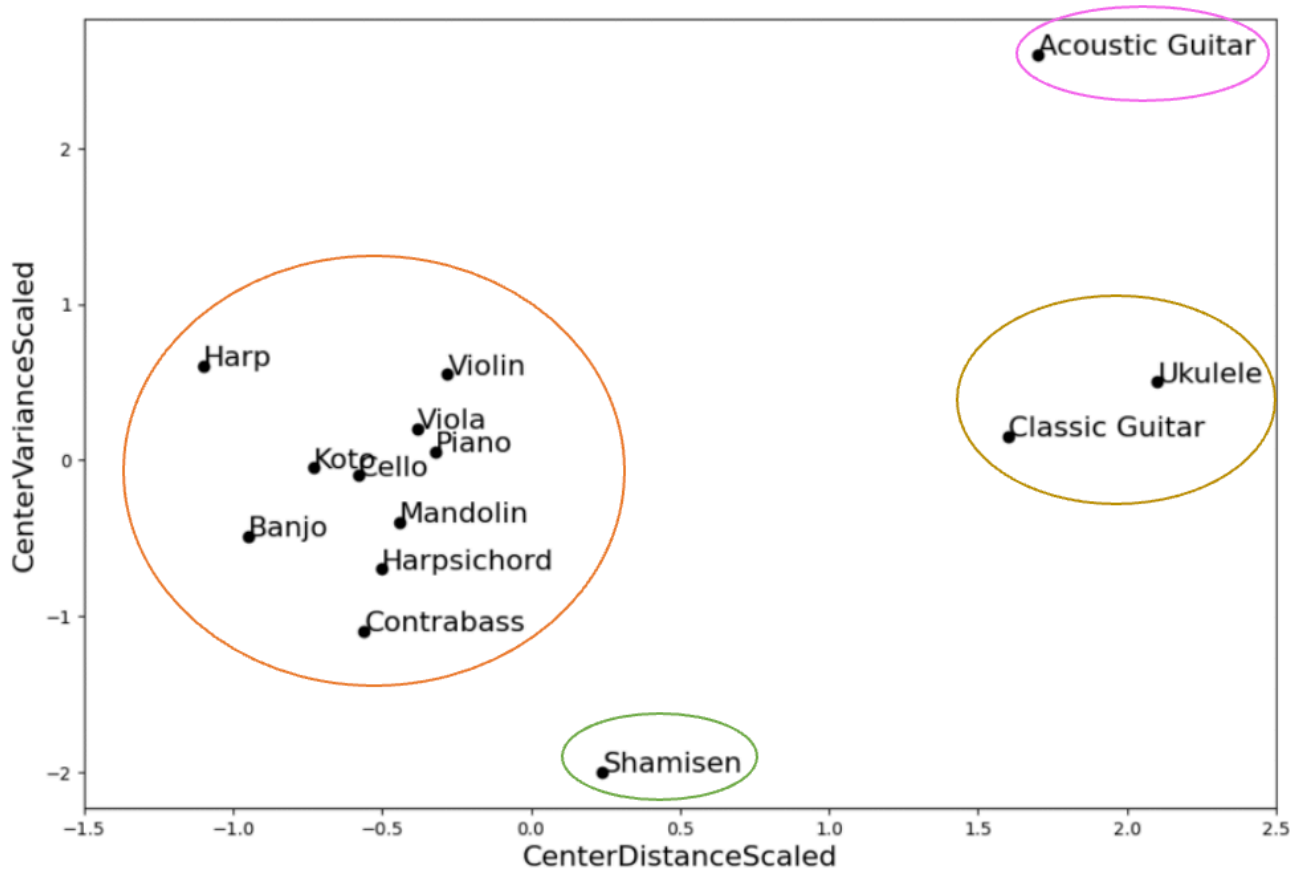

図 8. 弦鳴楽器のクラスタリング結果

ついて求めることができる。

ハイハットのデータは他と大きく異なる。エレクトリック・ピアノとギターは似ており、ベースとは大きく異なる。このような詳細な観察は、提案手法の実現可能性を検証するものである。しかし、すべての楽器を考えると、かなりの混合が生じる。したがって、提案手法は既存のカテゴリーを細分化するのに適しているだけであり、巨大で複雑な、完全に未知のデータを直接分類するのには適していない。

そこで、まずピアノ、バイオリン、ビオラを含む 14 種類の弦鳴楽器を選択した。これらの楽器に提案手法のステップを適用し、各楽器点セットについて中心点から原点までの距離と各点から中心点までの距離の分散を求めた。これによると、音色が非常に似ているピアノと電子ピアノ は、横軸の同じ位置にある。図 8 は、K-means クラスタリ ングの結果を 4 種類に分類したものである。

次に、同様の方法で、「気鳴楽器」、「体鳴楽器」、「膜鳴楽器」、「電鳴楽器」、「人間の歌声」を細分化した。

上記の方法により、最終的に RWC データセットの 92 種類の楽器を HS 分類に基づいて細分化し、自然音の楽器性を評価するための 17 のカテゴリー、すなわち「体鳴楽器 1 5」、「膜 鳴楽器 High」、「膜鳴楽器 Low」、「弦鳴楽器 1 4」、「気鳴 楽器 1 4」、「電鳴楽器」、「ボーカル」を作成した。そこで膜鳴楽器が二つになる原因は、膜鳴楽器はほぼ音域はないので、音色だけで考えると、音色を表せる X 軸だけで 2 種類に細分した。

本手法には、良い点があるが、良くない点もある。例えば、「Acoustic Guitar」を聴くと、「Piano」と「バイオリン」と細分されるべきだが、1 つの種類に細分されている。これは、メルスペクトログラムを用いて直接トレーニングした自然音の特徴量のピッチにおける性能不足を表している。そのため、本研究では、MFCC を用いた AST モデルを試した。

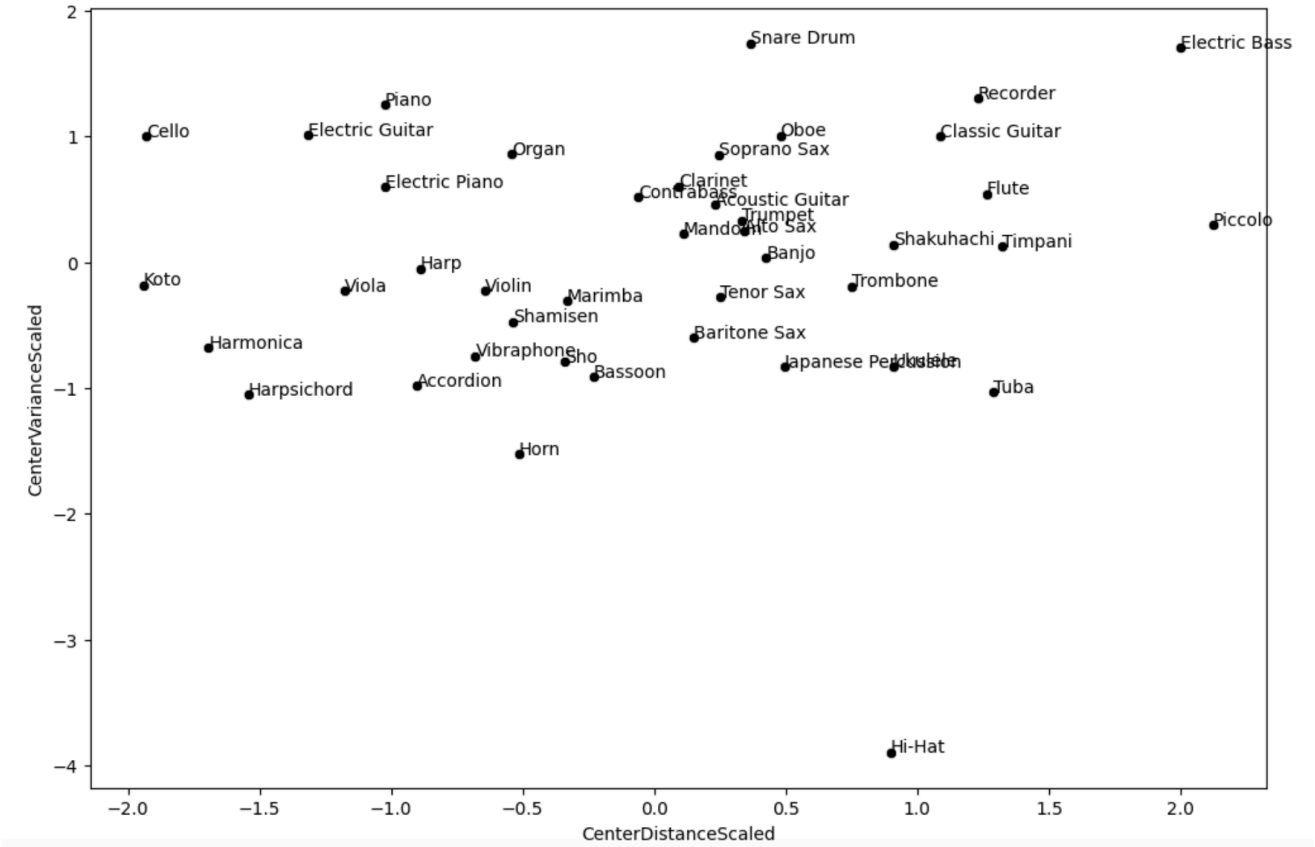

図 9. 全楽器の中心点-原点距離と各点距離分散

4.2 MFCC-AST

本研究では、AST の構造を継承しつつ、従来の 128 次元のメルスペクトログラムを入力としていたモデルを 16 次元の MFCC に変更し、実験を行った。メルスペクトログラムと比較して、MFCC はデータ量が少ないものの、情報がより集中しているという特徴を持つ。特に、大規模データセットである AudioSet を使用する場合、メルスペクトログラムでは音高情報が損失しやすくなり、前述の細分実験において問題が生じる可能性がある。そのため、本研究では AST ベースのモデルに MFCC を適用し、その有効性を検証した。

本研究で 16 次元の MFCC を選択した理由は、AST の基盤となる ViT（Vision Transformer の構造に適応させるためである。AST は、事前学習済みの ViT（ImageNet で学習済み）を AudioSet で転移学習したモデルであり、ViT は画像を 16 × 16 のパッチに分割し、各パッチを Transformer の入力トークンとして処理する。[2] このとき、順序トークンを用いてパッチの位置情報をモデルに学習させている。この構造を音響分野に適用するにあたり、入力特徴量として 16 次元の MFCC を使用することで、縦方向のパッチ移動を行わずに ViT のモデルパラメータをそのまま転移学習に利用できる。この方法により、モデルの収束速度が向上し、汎用性も強化される。

本研究では、モデルの学習データとして ESC-50 を使用し、手法の可行性を検証した。実際の使用環境では、AudioSet を用いた学習が望ましい。AST は Transformer をベースとしたモデルであり、CNN モデルと比較してパラメータ数が多いため、学習に必要なデータ量も増加する。しかし、ESC-50 は 2000 個の 5 秒間の音声データしか含まれていないため、前述の訓練時のランダムデータ拡張手法を適用した。複数回の実験の結果、3000 エポック以内でモデルが完全に収束することが確認されたため、最終的な訓練回数を 3000 エポックとした。また、学習率は初期値を 0.001 に設定し、

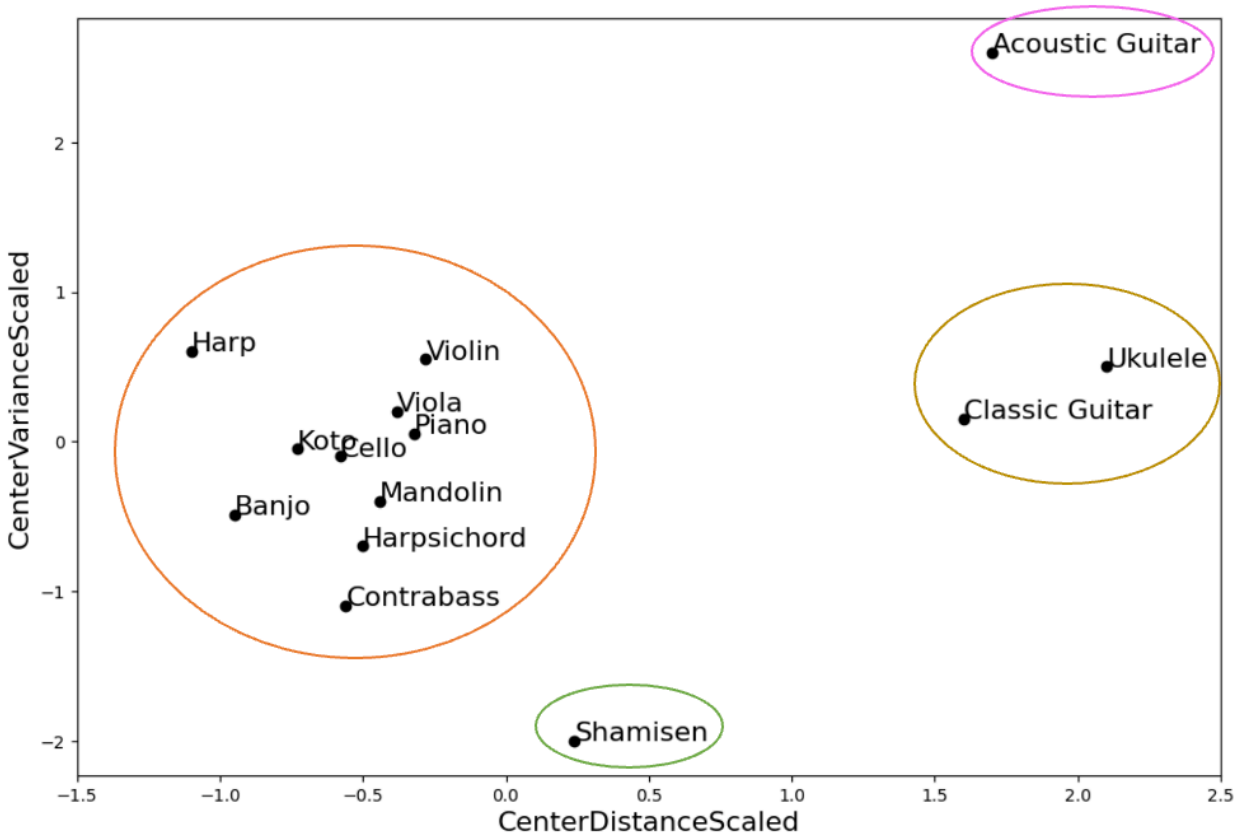

図 10. 弦鳴楽器のクラスタリング例

ウォームアップを有効にした後、学習率スケジューリングを適用し、モデルの精度向上が見られなくなった時点で 0.8 倍に調整した。結果については図 11 に示す。

メルスペクトログラムと MFCC を入力とした AST モデルの訓練結果を比較すると、初期段階ではメルスペクトログラムの方が高い精度を示し、収束速度も速いことが確認された。しかし、最終的には MFCC もメルスペクトログラムと同程度の精度に到達し、場合によってはそれを上回る傾向も見られた。この結果は、ランダムデータ拡張を適用したことによって、学習の安定性が一時的に低下したためと考えられる。メルスペクトログラムの性能が MFCC より優れているのは、メルスペクトログラムが自然音に含まれる微細な情報をより多く保持しているためであり、一方の MFCC は音色や音高といった特徴に焦点を当てるため、自然音の分類タスクではメルスペクトログラムの方が優れた結果を示すことは予測可能であった。ただし、MFCC を用いた場合の精度低下は許容範囲内であり、実用上の問題はないと判断できる。

通常の AST モデルの主要な問題点は、自然音の音高情報を適切に表現できない点にある。そこで、本研究では、MFCC-AST モデルが音高分類において AST よりも優れた性能を発揮するかどうかを検証するため、音高分類モデルを構築し、AST の出力特徴ベクトルと MFCC-AST の出力特徴ベクトルをそれぞれ用いて訓練を行い、その比較を通じて評価を行った。

具体的には、RWC データセットのピアノ音を対象とした音高分類モデルとして、単純な MLP モデルを構築した。RWC データセットから和音データを除外し、23 組のピアノ音の 88 音を用いたデータセットを作成した。各音高データには、訓練時のランダムデータ拡張を適用した。実験の結果、モデルの完全な収束には 10000 エポックが必要であることが判明し、最終的な訓練回数を 10000 エポックとした。学習率は 0.001 に設定し、学習率スケジューリングは適用しなかった。

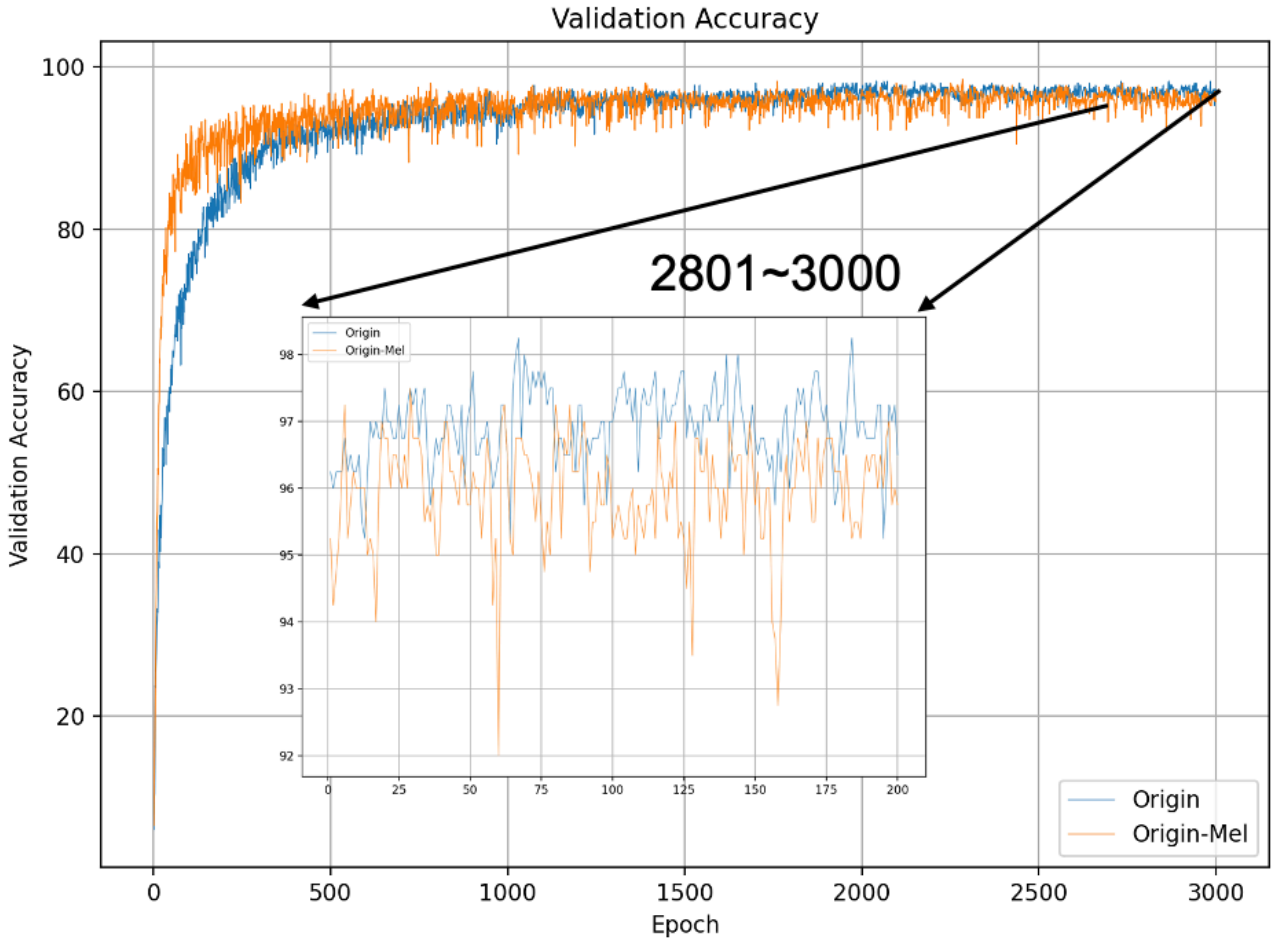

図 11. MFCC-AST と AST のトレーニング結果

実験結果は図 12 に示す。訓練データと検証データの精度を比較すると、両者の間に大きな差異が見られた。訓練データおよび検証データの両方において、MFCC-AST 特徴を用いた場合の学習速度は AST 特徴よりも速いが、早期に上限に達する傾向が見られた。一方、検証データに関しては、AST 特徴が途中で MFCC-AST 特徴を上回る場面も確認された。これは、検証データには学習データには含まれていない情報が多く含まれているため、メルスペクトログラムを用いた AST は時間の経過とともに MFCC が保持しない情報を学習できるためである。この結果は予測可能なものであり、本実験では MFCC-AST が AST よりも音高情報の学習に優れていることを証明できれば十分である。学習時間が経過すると、メルスペクトログラムを用いた AST モデルも MFCC-AST モデルと同等、あるいはそれ以上の精度に達することができるが、自然音データセットには音高情報に関するラベルが含まれていないため、細かく音高特徴を学習できる能力があったとしても、それを適切に活用することが難しい。一方、MFCC は初期段階から音高情報に焦点を当てているため、自然音分類モデルの最終的な精度が低くなったとしても、音高情報を失うことなく学習できる。

音高に基づく詳細な分類手法を適用して評価を行った結果、MFCC-AST の優位性が明確に示された。図 13 に示すように、MFCC-AST はアコースティックギターを正しく分類するだけでなく、ピアノのような最も広い音域を持つ楽器を縦軸の最上部に配置することができた。この結果は、従来のメルスペクトログラムを使用した AST モデルと比較して、MFCC-AST の方がより適切に音高特徴を捉えられることを示唆している。

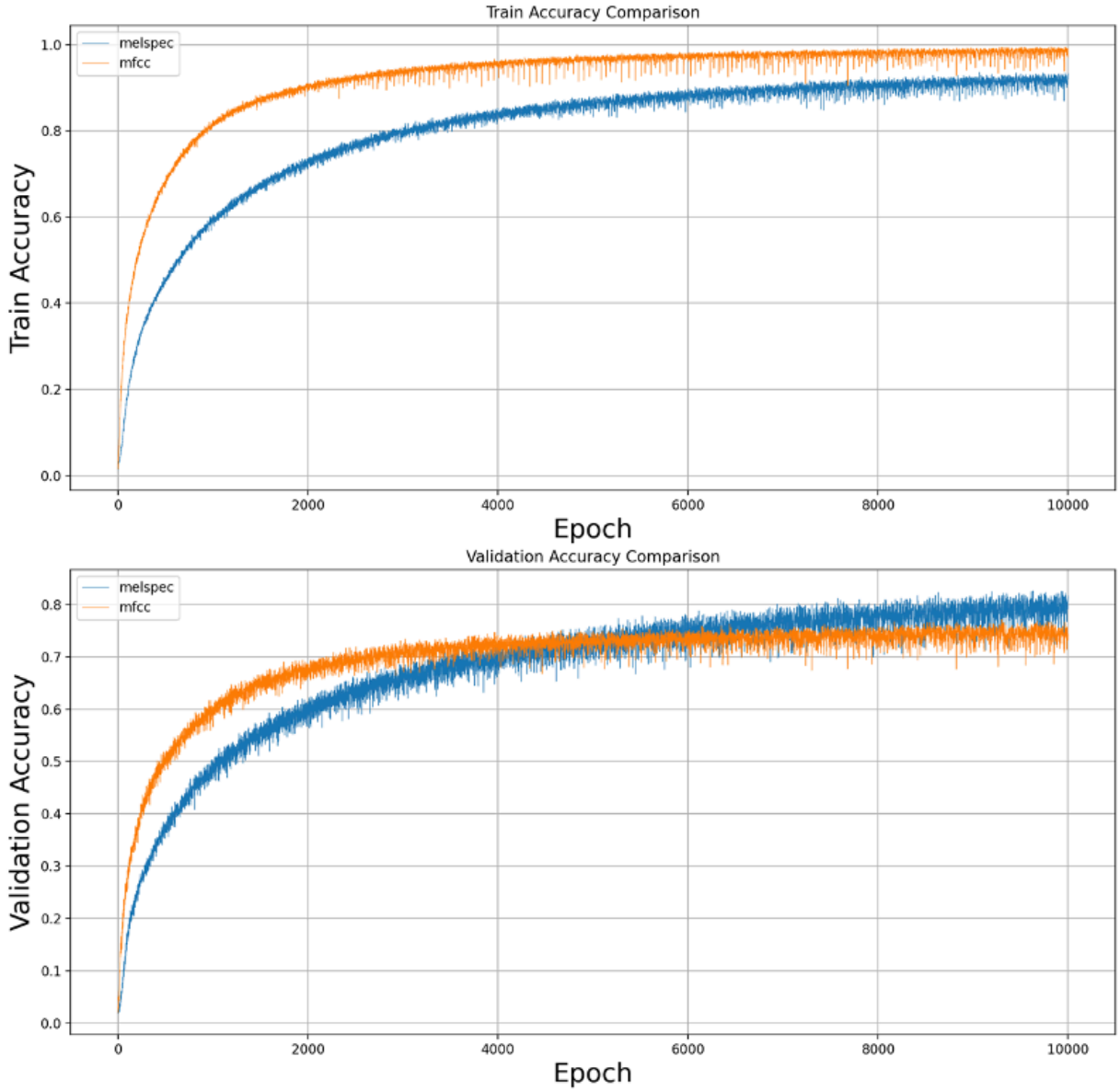

図 12. 音高分類モデルのトレーニング結果

4.3 モデルのテスト

本研究では、まず VGGish と AST の 2 種類の自然音モデルを評価した。評価の内容は、同一のデータセット（RWC2003）から自然音特徴を抽出し、それを用いて単純な全結合ニューラルネットワークを学習させ、楽器音を RWC2003 の 92 種類の原クラスに分類するものである。実験の結果は表 1 に示す。

表 1. VGGish と AST の分類精度

Model	**Test Accuracy**
VGGish	0.7595
AST	0.9726

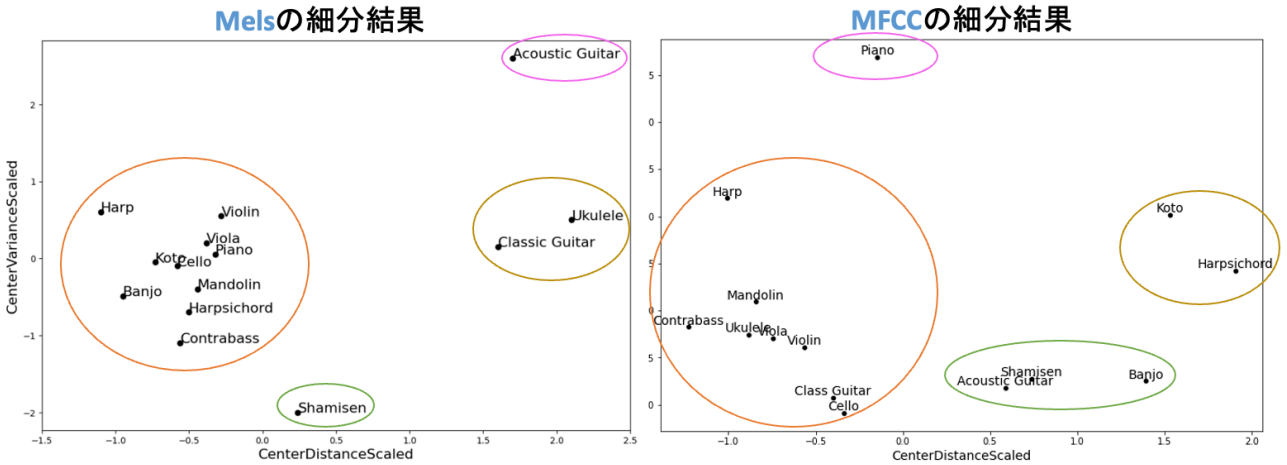

図 13. AST と MFCC-AST 特徴の細分の比較

結果を分析すると、AST の分類精度は VGGish を大きく上回ることが確認された。そのため、本システムでは AST を自然音モデルとして採用することとした。

次に、楽器分類モデルの構造の最適化を行った。AST を自然音モデルとして固定した上で、1D-CNN、全結合ニューラルネットワーク（FCN）、および残差接続ニューラルネットワーク（ResNet）の 3 つのモデルを比較した。その結果は表 2 に示す。

表 2. 3 種類の楽器分類モデルの精度

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

実験結果によると、1D-CNN は 98.1% の分類精度を達成し、全結合ネットワーク（97.63%）をわずかに上回る結果となった。しかし、1D-CNN のパラメータ数は全結合ネットワークの 3 倍に達しており、精度の向上はわずか 0.47% にとどまった。また、過学習は確認されなかったため、残差接続を追加することによるメリットは小さいと判断された。以上の結果から、本システムでは計算コストと精度のバランスを考慮し、最終的に全結合ニューラルネットワーク（FCN）を楽器分類モデルとして採用することとした。

最終的に、表 3 に示す全結合ニューラルネットワークを構築し、訓練を行った。学習結果について、17 種類の楽器の細分カテゴリにおいて 98% 以上の高い分類精度を達成した。この結果から、本システムが楽器分類において高い信頼性を持つことが確認された。

4.4　自然音の評価

4.4.1　一段階実験

本研究のシステムの可行性を検証するため、まず五種類の自然音をテスト対象として選定した。具体的には、鳥の鳴き声、車のエンジン音、カウベル、グラス、流水音を選定した。これらの自然音を選定した理由は以下の通りである。

鳥の鳴き声とグラスの音は、すでに広く認識されている自然音楽器であり、期待通りの結果が得られる可能性が高い。カウベルは、自然音楽器として一般的ではないものの、音色が多くの打楽器、特に金属製の打楽器と類似している。音域の広さが十分ではなく、単独で旋律を演奏する楽器としては適さないものの、多くの野外音楽や民族音楽ではリズム

表 3. 楽器分類モデルの構造

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

楽器として使用されている。したがって、カウベルは体鳴楽器との高い類似度が期待される。

一方で、車のエンジン音と流水音は、一般的な環境音であり、楽器としての利用はほとんど見られない。環境音は、伝統的な音楽制作において楽器として扱われることは少なく、一部の革新的な音楽ジャンルではサンプリングされたり、特定の音楽セクションに直接挿入されたりすることがある。しかし、これらの音は、聴覚的にもスペクトル的にもランダムノイズに近いため（図 14 のように、水流のスペクトログラムはノイスに近い）、良好な結果が得られる可能性は低いと予想される。

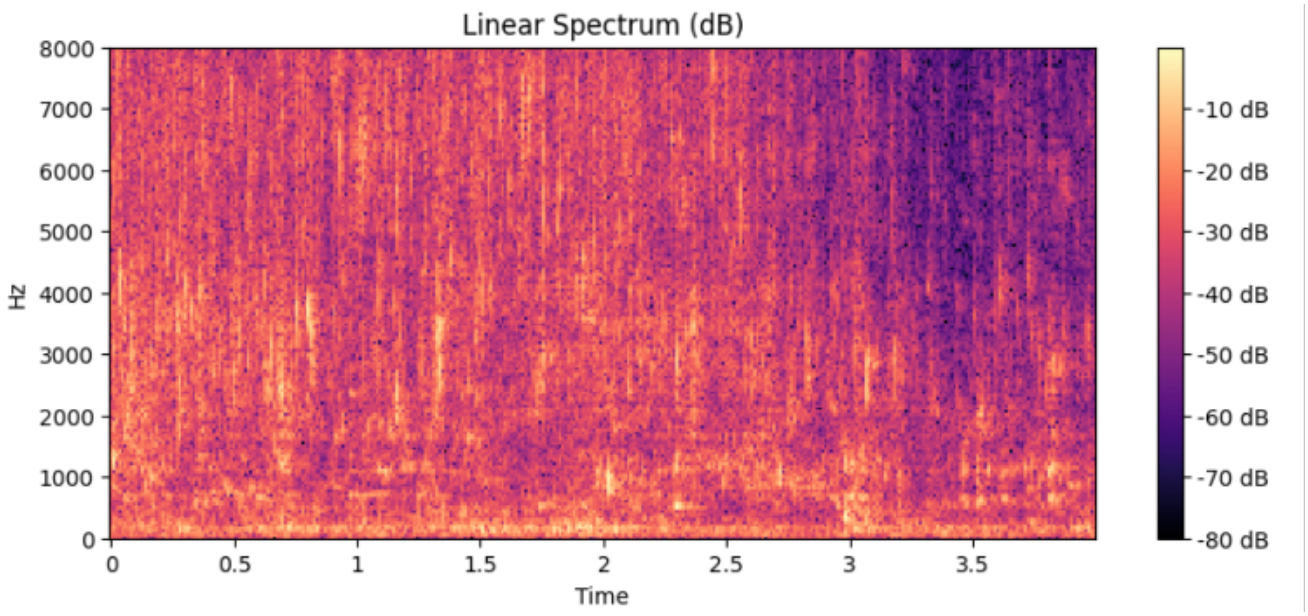

図 14. 水流のスペクトログラム

これら五種類の自然音のうち、鳥の鳴き声、カウベル、車のエンジン音は AudioSet データセットから取得し、グラスと流水音は前章で説明した自ら収録したデータを使用した。

最初に、表 4 に示すように、各自然音対象に最も近い細分した楽器種類を抽出した。しかしながら、この評価方法では意味がほとんどない。それは、最も近い楽器種別を得たとしても、それは単に既存の楽器と同一の自然音を検出しているに過ぎず、本研究が目指す「新たな自然音楽器」の発見には至らないためである。

そこで、各自然音オブジェクトが本研究の評価システム下でどのような結果を示すかをより詳細に把握するため、図 15 に示すヒートマップを得た。このヒートマップでは、各自然音が異なる楽器カテゴリに分類される割合を示しており、全体の合計は 1 となる。

表 4. 各自然音の結果

自然音	一番近い種類	確率値
Bird	Aero2	0.284971
Car	vocal	0.230435
Cowbell	Chordo1	0.226562
Glass	Idio5	0.298551
Water	Low-Membrano	0.205572

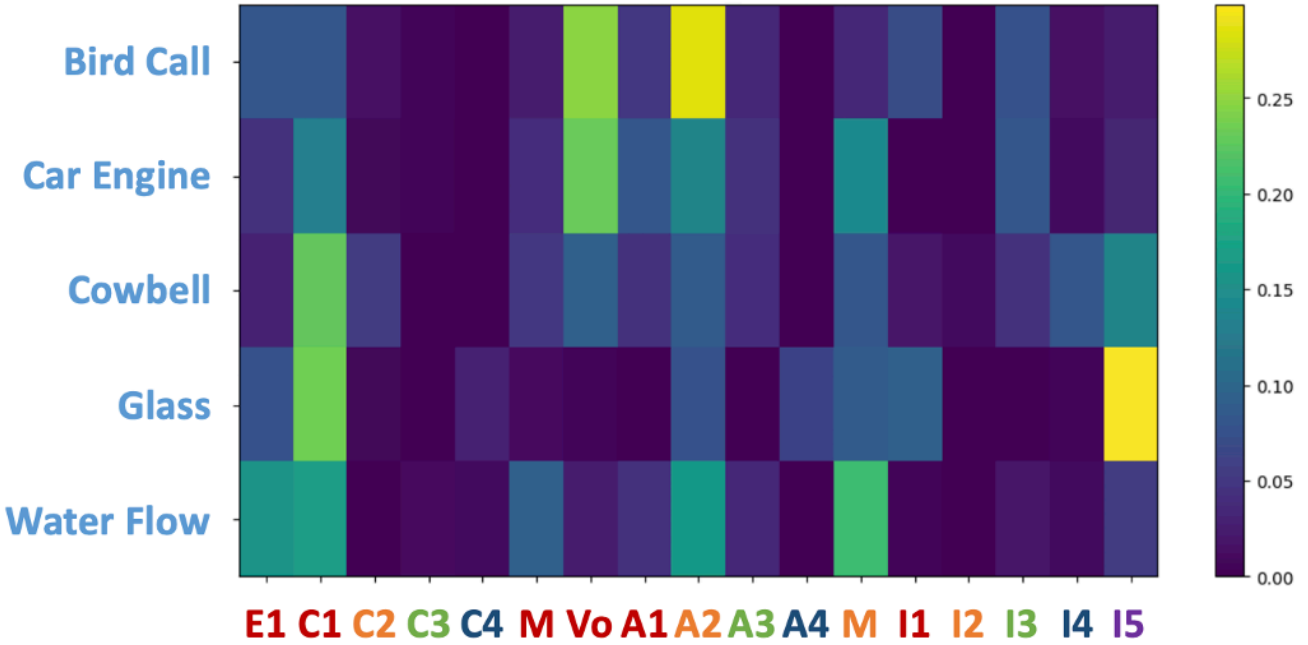

図 15. 5 種類自然音の結果

ヒートマップの結果を分析すると、鳥の鳴き声は気鳴楽器 2 と人の歌声に高い類似度を示していた。このように、二つの異なるカテゴリとの高い相似性が見られる場合、既存の単一の楽器には分類されず、新たな楽器としての可能性が示唆される。この結果は予想通りである。同様に、グラスの音も体鳴楽器 5 と弦鳴楽器 1 の両方に高い類似度を示し、複合的な特性を持つ楽器として認識された。

一方、車のエンジン音と流水音は、すべての楽器カテゴリとの類似度が低く、楽器としての価値はほとんどないと判断された。これらの音は、音色、音高、音域のいずれにおいても楽器として適さない。しかし、音楽制作の際に編曲や特殊効果として使用される可能性がある点には注意が必要である。例えば、多くの楽曲で車のエンジン音を挿入することで、都市の雰囲気を演出し、物語性を強調することがある。このような創作手法は重要であるが、本研究ではあくまで自然音を主要な楽器として使用することを前提としており、特定の演出効果を目的とした利用は評価対象外とした。

予想外の結果として、カウベルのデータは期待とは異なる傾向を示した。本来、カウベルは鳥の鳴き声やグラスと同様に複合的な楽器として分類されるか、体鳴楽器との高い類似度を示すと予想されていた。しかし、実際には弦鳴楽器との類似度が最も高く、全体的な分類結果も車のエンジン音や流水音と同様に、すべてのカテゴリとの類似度が低かった。この結果は予想に反するものである。

この原因を探るため、カウベルのデータを分析した結果、AudioSet の収録方法に起因する問題が見られた。AudioSet は YouTube の動画から音声を収録しているため、カウベルの音が含まれる動画は主に野外で録音されており、背景ノ

イズが多いことが判明した。このノイズがカウベルの特徴を不明瞭にし、分類結果に影響を与えたと考えられる。また、AudioSet のデータには複数のラベルが付与されるため、多くのカウベルのデータには人の話し声や他の環境音が含まれており、特徴の正確性が損なわれた可能性がある。

この影響は、車のエンジン音のデータにも見られた。車のエンジン音の類似度が最も高いカテゴリは人の歌声であった。これは、運転中の YouTube 動画には話し声が頻繁に含まれており、人の話し声と歌声はスペクトル的に類似しているため、このような誤分類が発生したと考えられる。

これらの問題を考慮し、より正確な分析を行うために、追加の実験を実施し、評価対象の自然音の種類を増やして再検証を行った。結果については次章で詳述する。

4.4.2 追加実験

本研究では、前回の実験で使用したカウベルのデータを除外し、新たに目覚まし時計の音、マウスクリック音、雷鳴、ピアノ音、教会の鐘の音の五種類のデータを追加した。選定理由は第一段階と同様である。目覚まし時計の音と教会の鐘の音は、カウベルと同じく明確な金属音を持つ音源であり、カウベルのように民族音楽で使用されることはないものの、発音原理が類似しており、音色も一定の類似性を持つ。そのため、これらの音は体鳴楽器との高い相似性を示すことが期待される。

雷鳴は、打楽器の中でも特に太鼓に近い音響特性を持つため、選定された。雷鳴と太鼓の音は聴感上の類似性が高く、雷鳴が膜鳴楽器と高い相似性を示すことが予想される。マウスクリック音については、特定の楽器に類似しているとは言い難いが、すでに音楽の分野では異なる種類のマウスクリック音をリズム楽器として用いる試みが行われている。そのため、何らかの楽器カテゴリに類似することが期待されるが、全く類似しない（すなわちノイズとみなされる）可能性も考えられる。

最後に、ピアノ音は本実験において特別な役割を持つ。モデルが学習していない既存の楽器データに対する識別精度を確認するため、楽器分類モデルの訓練時にはピアノのデータを除外した。ただし、訓練データには電子ピアノが含まれているため、モデルがピアノに関する基本的な音響特性を完全に失うことはないと考えられる。また、ピアノは弦鳴楽器の中でも特に広い音域を持つため、HS 分類を基にした細分化において独立したクラスに分類されやすい。さらに、ピアノは最も一般的で聴き馴染みのある楽器であるため、テストデータとして使用することでシステムの有効性をより説得力のある形で示すことができる。

これらのデータのうち、ピアノ音を除く四種類（目覚まし時計の音、マウスクリック音、雷鳴、教会の鐘の音）は ESC-50 データセットから取得した。前述の通り、ESC-50 は AudioSet のように動画から収集されたデータではなく、データセット作成者が意図的に収録した音源であり、ラベルの混在やノイズの影響が少ない。その反面、データ量は少なく、各カテゴリの合計音声時間は 200 秒程度に限られる。そのため、本研究では、これらのデータを訓練時のランダムデータ拡張ではなく、RWC データセットで採用した九種類のデータ拡張手法を用いて増強した。一方、ピアノ音については RWC2003 のピアノデータを使用した。RWC には 2000 以上の単音データが含まれており、十分なデータ量が確保されているため、データ拡張は行わずそのまま使用した。データ量のバランスを取るため、前述の四種類の自然音データのデータ拡張後の目標サンプル数をピアノデータの総数に合わせた。

新たに追加した五種類のデータと、前回使用したカウベルを除いた四種類のデータを組み合わせた結果、同じ形式のヒートマップを得ることができた（図 16 に示す）。

ヒートマップの結果を分析すると、目覚まし時計の音は弦鳴楽器 1 と最も高い相似性を示した。この二つの音は聴感上も類似性があるため、妥当な結果であると考えられる。マウスクリック音は、体鳴楽器 3 および 5 と高い相似性を示し、既存の体鳴楽器とは異なる新しい体鳴楽器として分類できる可能性がある。雷鳴については、膜鳴楽器の二つのカテゴリとの相似性が高く、「雷鳴と太鼓の音が類似している」という仮説を支持する結果が得られた。

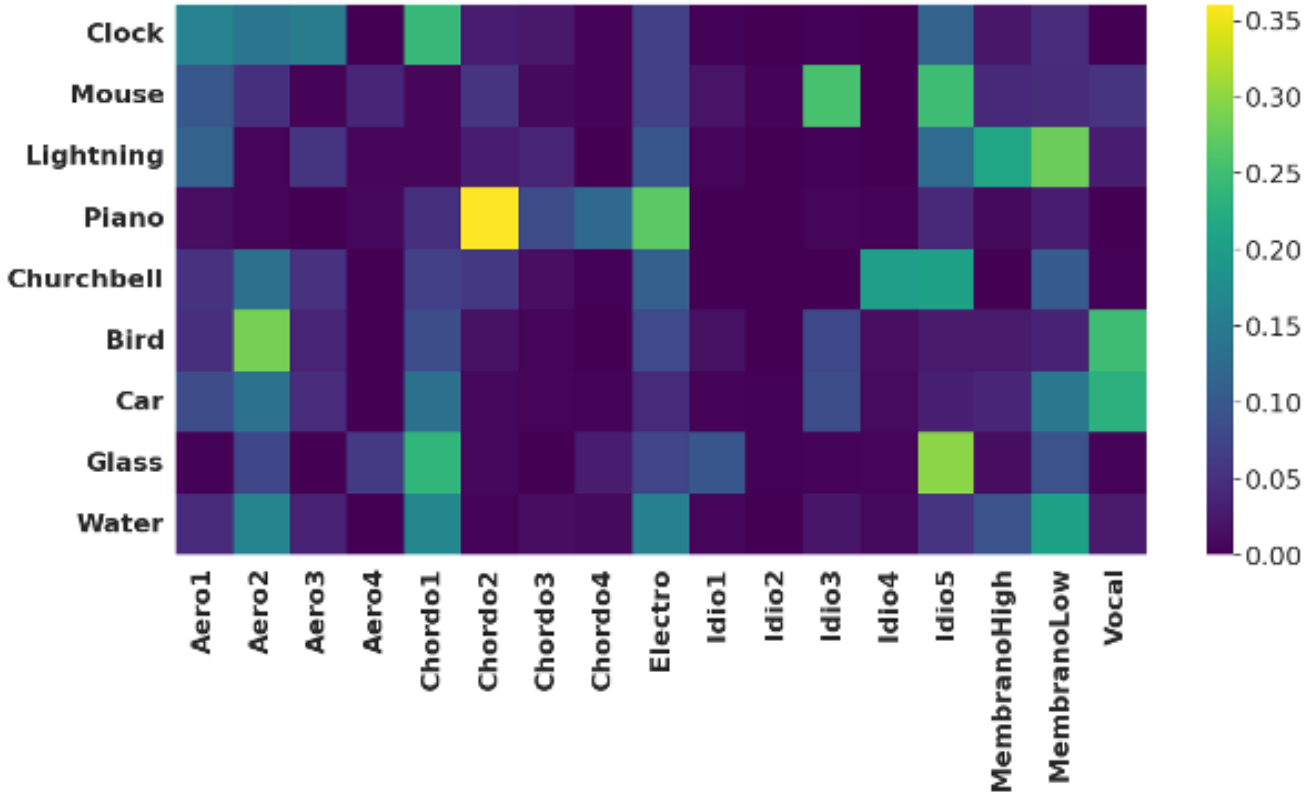

図 16. 8 種類自然音とピアノの結果

教会の鐘の音については、体鳴楽器 4 および 5 と比較的高い相似性を示したが、期待ほどの高い相似性には至らなかった。この原因については、今後の研究で詳しく検討する必要がある。ピアノ音については、完全に予想通りの結果が得られ、相似度が最も高いカテゴリは弦鳴楽器 2 であり、また電鳴楽器とも高い相似度を示した。これは、電子ピアノのデータが訓練データに含まれていたことの影響を示している。

しかし、本研究の結果からいくつかの課題が浮かび上がる。第一に、現在の主観的な評価方法には、目標とする自然音が楽器としての価値を持つかどうかを定量的に判断する明確な基準が存在しない点が挙げられる。例えば、今回の実験では、教会の鐘の音が体鳴楽器 4 および 5 と 20% 程度の相似性を示したが、エンジン音も人の歌声との相似度が 20% 近くに達していた。これは、前述のように AudioSet のデータに含まれる人の話し声の影響によるものであるが、相似度の数値だけでは明確な判断が難しいことを示している。今後の研究では、計算式を用いて各自然音の評価スコアを算出し、一定の閾値を超えた場合にのみ楽器としての価値を認める手法の開発が必要である。また、今回の結果から「相似度が 20% を超えれば楽器としての可能性がある」という仮説が得られたが、この閾値の適切な設定についても今後の検討が求められる。

もう一つの課題は、細分化分類システムに関するものである。本研究で提案した細分化分類システムは HS 分類を基に構築されているが、HS 分類では異なるカテゴリ間で音色が類似する楽器が存在する。例えば、ピアノと電子ピアノは異なる発音原理を持つが、音色は非常に類似している。本研究ではこの問題を完全には解決していない。解決策として、電子ピアノとピアノを同じ弦鳴楽器の細分化カテゴリに統合する方法が考えられるが、この場合、他の複合的な楽器分類、例えば気鳴楽器と人の歌声、体鳴楽器と弦鳴楽器など、さらなる分類課題が生じることになる。これらの問題の解決は、今後の研究において重要な課題となる。

5　結論と今後の課題

本研究では、自然音が楽器としての性能を持つかを評価するため、楽器分類モデルを構築し、既存の楽器との類似性を測定する手法を提案した。具体的には、VGGish および AST を基盤としたモデルを用い、特徴ベクトルを抽出した上で、自然音と楽器の類似度を定量的に評価した。さらに、HS 分類を基にした細分化分類を適用し、従来の楽器分類手法と比較した際の適合性についても検証を行った。

実験の結果、鳥の鳴き声やグラスの音など、従来から自然音楽器として認識されている音は、特定の楽器カテゴリとの高い類似性を示し、新たな楽器としての可能性を持つことが確認された。一方で、車のエンジン音や流水音のような一般的な環境音は、どの楽器とも高い類似性を示さず、楽器としての価値が低いと判断された。また、カウベルに関しては、AudioSet のデータに含まれるノイズや複数ラベルの影響により、期待された体鳴楽器との高い相似性を示さず、データ品質が結果に与える影響の大きさが明らかとなった。

さらに、AST の入力特徴量として、従来の 128 次元メルスペクトログラムではなく、16 次元の MFCC を用いた MFCC-AST モデルを構築し、その有効性を検証した。MFCC-AST は、音高特徴の学習において AST よりも優れた性能を示し、音高を重視する楽器分類タスクにおいて有用であることが確認された。ただし、メルスペクトログラムと比較して情報量が限定されるため、自然音の微細な特徴の識別能力は若干低下することも示された。

また、追加実験として、目覚まし時計の音、マウスクリック音、雷鳴、教会の鐘の音、ピアノ音をテストデータに追加し、細分化分類の精度を向上させる試みを行った。その結果、目覚まし時計の音は弦鳴楽器、雷鳴は膜鳴楽器、教会の鐘は体鳴楽器と一定の類似性を示し、期待通りの結果が得られた。一方で、細分化分類の課題として、HS 分類に基づくカテゴリの相互関係や、音色の類似性を考慮した分類手法の最適化が必要であることが明らかになった。

本研究の成果により、自然音の楽器としての可能性を定量的に評価する手法の有効性が示された。しかし、いくつかの課題も残されている。

第一に、楽器としての価値を定量的に判断する明確な基準の確立が必要である。現在の評価方法では、相似度が一定以上であれば楽器として認めるという主観的な判断に依存しており、より客観的な評価基準の策定が求められる。例えば、特定の相似度閾値を設定し、それを超えた場合にのみ楽器として認定する手法の開発が考えられる。

第二に、HS 分類を基盤とした細分化分類の課題が挙げられる。本研究では、HS 分類に基づく楽器のカテゴリ分けを採用したが、実際の音色の類似性とは異なる場合があることが確認された。例えば、ピアノと電子ピアノは発音原理が異なるが、音色は非常に類似している。今後の研究では、楽器の発音原理だけでなく、音響的特徴の類似性を考慮した分類手法を導入する必要がある。

第三に、データセットの品質の問題がある。特に AudioSet のデータにはノイズや複数のラベルが含まれており、分類精度に影響を与えることが明らかとなった。ESC-50 のような高品質なデータセットを活用することで、より正確な評価が可能になると考えられるが、データ量が少ないという制約もあるため、大規模かつ高品質なデータセットの構築が今後の課題となる。

以上の課題を解決することで、自然音の楽器としての利用可能性をより正確に評価し、新たな楽器の創出や音楽制作への応用が期待される。今後の研究では、細分化分類手法の改良、評価基準の確立、データセットの最適化を進め、より実用的なシステムの構築を目指す。

参考文献

[1] B. S. Dominika Szeliga, Paweł Tarasiuk and P. S. Szczepaniak. Musical instrument recognition with a convolutional neural network and staged training. *Procedia Computer Science*, 207:2493–2502, 2022.

[2] 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.

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

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

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

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

[7] 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.

[8] T. N. Masataka Goto, Hiroki Hashiguchi and R. Oka. Rwc music database: Music genre database and musical instrument sound database. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, pages 229–230, October 2003.

[9] K. J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

[10] K. Tanaka, Y.-J. Luo, K. W. Cheuk, K. Yoshii, S. Morishima, and S. Dixon. On the use of synthesized datasets and transformer adaptors for musical instrument recognition. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[11] S. Umemoto, Y. Naka, N. Yoshida, and T. Yonezawa. Proposal of a sound transformation technique by converting rhythms and scales hidden in environmental sounds into musical elements - pseudo-expression of absolute tone sense. Technical Report 17, Kansai University, feb 2015.

[12] E. M. von Hornbostel and C. Sachs. Classification of musical instruments: Translated from the original german by anthony baines and klaus p. wachsmann. *The Galpin Society Journal*, 14:3–29, 1961.

[13] 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.

[14] 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.